STRUČNÝ NÁVOD K OVLÁDÁNÍ

IBM SPSS Statistics 19 a IBM SPSS Modeler 14

Pavel PETR

Ústav systémového inženýrství a informatiky

Fakulta ekonomicko – správní

UNIVERZITA PARDUBICE

2012

Obsah

1. Základní informace o programu IBM SPSS Statistics

IBM SPSS Statistics patří mezi celosvětově rozšířené statistické systémy pro aplikace ve vědě, marketingu, personalistice a výzkumu, pro zpracování laboratorních měření a pro sumarizace dat z velkých i menších databází různého typu. Používá se pro finanční analýzy, tvorbu rozhodovacích modelů a analýzu i predikci časových řad. IBM SPSS Statistics poskytuje také uživatelsky příjemné softwarové prostředí pro vybrané metody využívané v oblasti Data Miningu (DM), manažerských analýz a podporu business intelligence.

Modularita systému IBM SPSS Statistics umožňuje složit systém „na míru" dle potřeb a přání uživatele – jak pro jednoduché rychlé tabelace, tak pro kvalitní průběžné inženýrské a marketingové analýzy, ale také pro nejnáročnější matematicko-statistické aplikace, modelování a DM. Jeho univerzalita zaručuje pokrytí potřeb v různých částech organizace, a tím zaručuje kompatibilitu a zjednodušuje přípravu podkladů pro řízení a management. Uživatelská jednoduchost činí jeho ovládání dostupným i administrativním pracovníkům a asistentům.

1.1. Základní moduly IBM SPSS Statistics 19

IBM SPSS Statistics Base: základ celého systému – možnost načtení dat z mnoha formátů a pomocí ODBC, export dat do jiných formátů, manipulace se soubory, datové manipulace (výběr případů, vážení, agregace, identifikace duplikátních případů), transformace dat, základní statistické přehledy a tabulky, složitější statistické metody a postupy (t-testy, ANOVA, korelační a regresní analýza, vyhlazování křivek, neparametrické testy, faktorová analýza, diskriminační analýza, seskupovací analýza, analýza reliability, mnohorozměrné škálování ALSCAL, mnohonásobné odpovědi a další), statistické grafy, snadná editace výstupů (úpravy tabulek a grafů), export výstupů, ovládání programu pomocí syntaxe včetně maker, skripty.

IBM SPSS Custom Tables: jednoduché a interaktivní vytváření komplexních tabulek na míru zákazníkovi.

IBM SPSS Regression: pokročilé mnohorozměrné modely založené na regresi (binární logistická regrese, mnohorozměrná logistická regrese, nelineární regresní modely s okrajovými podmínkami i bez nich, metoda vážených nejmenších čtverců, dvoustupňová metoda nejmenších čtverců).

IBM SPSS Advanced Statistics: sofistikované metody matematicko-statistického modelování vztahů: mnohorozměrný obecný lineární model, metody pro modelování vztahu mezi kategorizovanými proměnnými a metody analýzy délky života.

IBM SPSS Decision Trees: tvorba, ověřování a aplikace klasifikačních stromů.

IBM SPSS Categories: analýza mnohorozměrných kategorizovaných dat včetně grafického zobrazení vztahů (optimální škálování, percepční mapy, různé techniky redukce dimenzí, kategorická regresní analýza).

IBM SPSS Conjoint: analýza vlastností produktu nebo služby na základě preferencí zákazníků, doporučení vhodné kombinace atributů.

IBM SPSS Complex Samples: nástroj pro práci s komplexními výběry od plánování až po analýzy. Korektní odhady statistik, včetně složitějších modelů (obecný lineární model, logistická regrese).

IBM SPSS Forecasting: široká škála metod pro analýzu časových řad.

IBM SPSS Exact Tests: analýza malých datových souborů nebo řídce zastoupených skupin případů s přesnými hladinami významnosti.

IBM SPSS Missing Values: analýza chybějících hodnot – zjišťování struktury, sumarizace, vzory, odhady chybějících pozorování.

IBM SPSS Neural Networks: odhalení komplexní struktury vztahů v datech pomocí neuronových sítí.

IBM SPSS Data Preparation: zefektivnění procesu validace dat a zjednodušení časově náročných postupů při jejich manuální kontrole. Snadná identifikace podezřelých nebo chybných případů, proměnných a hodnot v datech, detekce extrémních hodnot, sledování struktury chybějících hodnot i další postupy umožňují získat z dat přesnější výsledky.

IBM SPSS Direct Marketing: porozumění zákazníkům a optimalizace marketingových kampaní použitím RFM analýzy (Recency, Frequency, Monetary), segmentace zákazníků a analýzy jejich profilů, porovnáním efektivity kampaní a odhadnutím pravděpodobnosti nákupu – to vše v jednotném a snadno ovladatelném rozhraní.

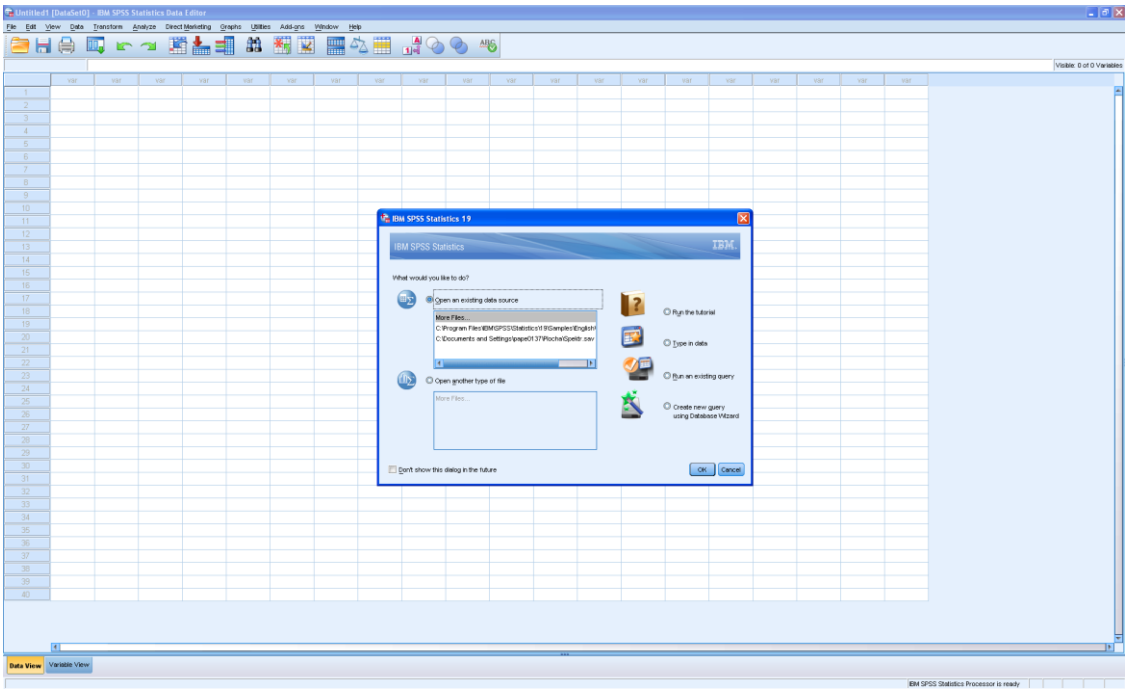
IBM SPSS Bootstrapping: validace modelů metodou bootstrap – robustní odhad směrodatné chyby a intervalů spolehlivosti pro odhadované parametry, např. průměr, medián, percentily, poměř šancí, korelační a regresní koeficienty.

2. Program IBM SPSS Statistics

V této části bude popsáno základní prostředí uvedeného programu a základní operace, které je nutno znát pro efektivní využívání uvedeného programu.

2.1. Prostředí programu IBM SPSS Statistics

Při standardním spuštění programu se objeví následující okno (Obrázek 1).



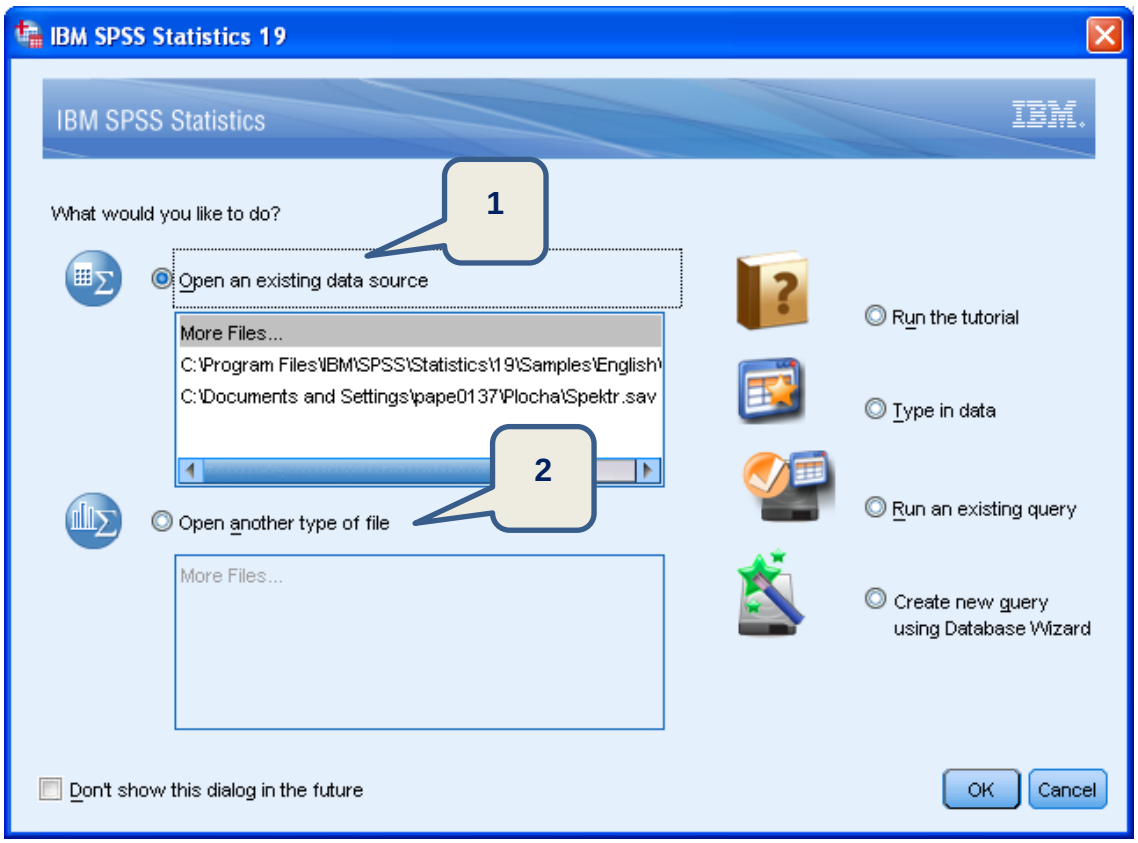
OBRÁZEK 1 VZHLED ÚVODNÍHO OKNA IBM SPSS STATISTICS

V tuto chvíli lze realizovat následující operace (Obrázek 2):

- otevření datového souboru;
- spuštění tutoriálu;
- vytvoření datového souboru
- načtení dat pomocí spuštění existujícího dotazu z databáze;
- vytvoření nového dotazu pro import dat z databáze pomocí průvodce.

Pro naše potřeby zůstaneme u možnosti otevření datového souboru, kterou lze upřesnit pomocí výběru z následující volby:

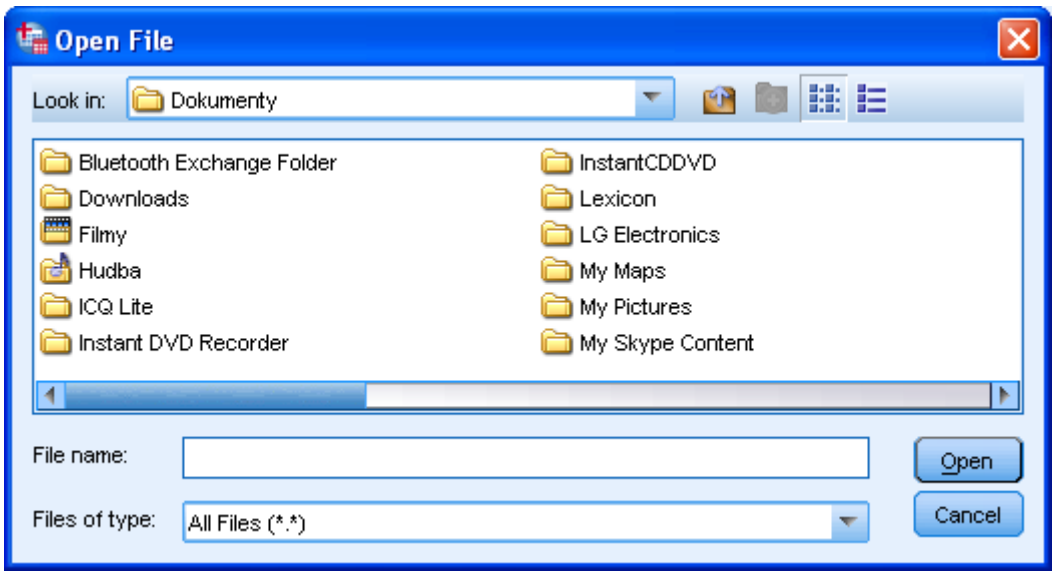
1. „Open an existing data source“;
2. „Open another type of file“.



OBRÁZEK 2 VOLBA ČINNOSTÍ VE VSTUPNÍM OKNĚ IBM SPSS STATISTICS

První volba nám umožní rychlý přístup k datovým souborům ve formátu IBM SPSS Statistics (s příponou *.sav). Můžeme volit ze seznamu naposledy otevřených souborů nebo vybrat jiný soubor ve formátu *.sav.

Druhá volba nám slouží k výběru datového souboru v libovolném formátu (Obrázek 3). Možnosti jsou stejné jako v případě první volby. Je třeba standardním způsobem vyhledat potřebný soubor z jeho úložiště.



OBRÁZEK 3 OKNO VÝBĚRU LIBOVOLNÉHO SOUBORU

Samozřejmé je, že nemusíme využít žádnou z těchto možností a otevře se nám prázdná tabulka v prostředí IBM SPSS Statistics (Obrázek 1). V tuto chvíli můžeme dále využívat všechny možnosti, které jsou určené pro práci s programem IBM SPSS Statistics.

2.2. Okna v programu IBM SPSS Statistics

Program IBM SPSS Statistics využívá čtyři základní typy oken:

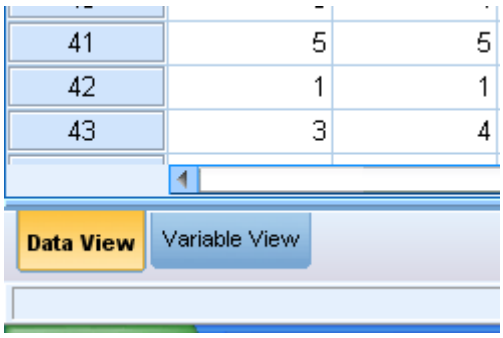
- datové okno – práce se vstupními daty a proměnnými,
- výstupové okno – záznam a zobrazení všech výstupů programu,
- syntaxové okno – je určeno k práci se syntaxí, tj. pomocí speciálního jazyka určeného k zadávání příkazů IBM SPSS,
- skriptové okno – tvorba skriptů (programů) pro automatizaci činností v programu.

Jednotlivá okna mají odlišné funkce a rovněž nabídky, které jsou zde k dispozici, se v každém z nich mírně liší.

Vzhledem k rozsahu a určení tohoto návodu, zde bude objasněna práce pouze v datovém a výstupovém okně.

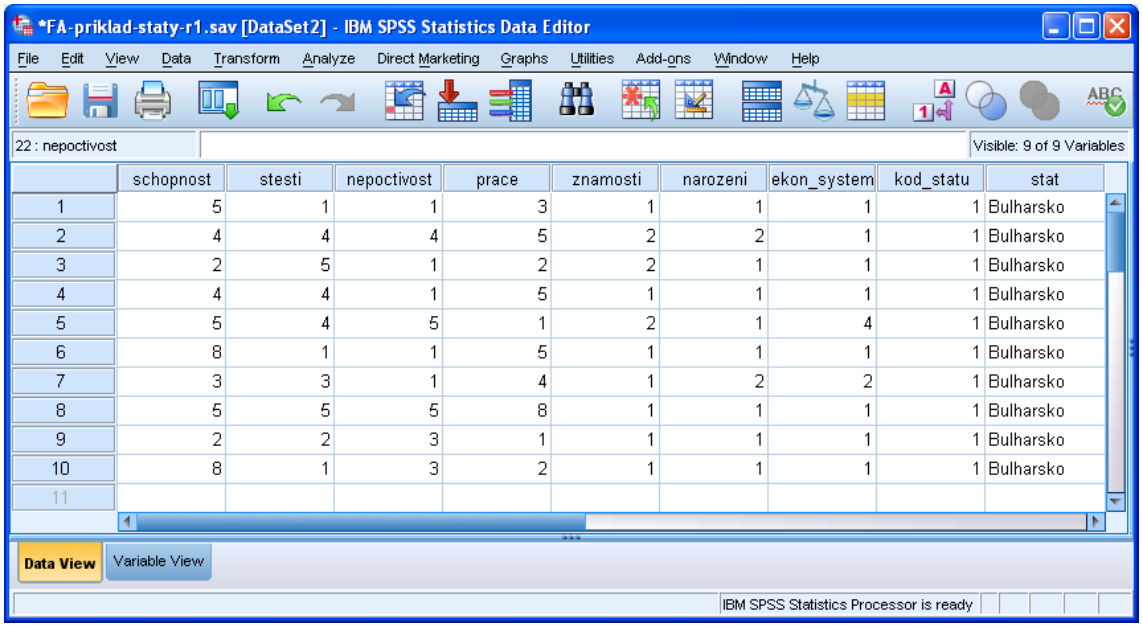
2.2.1. Datové okno

Datové okno je určeno pro práci se vstupními daty a proměnnými. Skládá se ze dvou záložek – pohled na data ***Data View*** a pohled na proměnné ***Variable View*** (Obrázek 4).



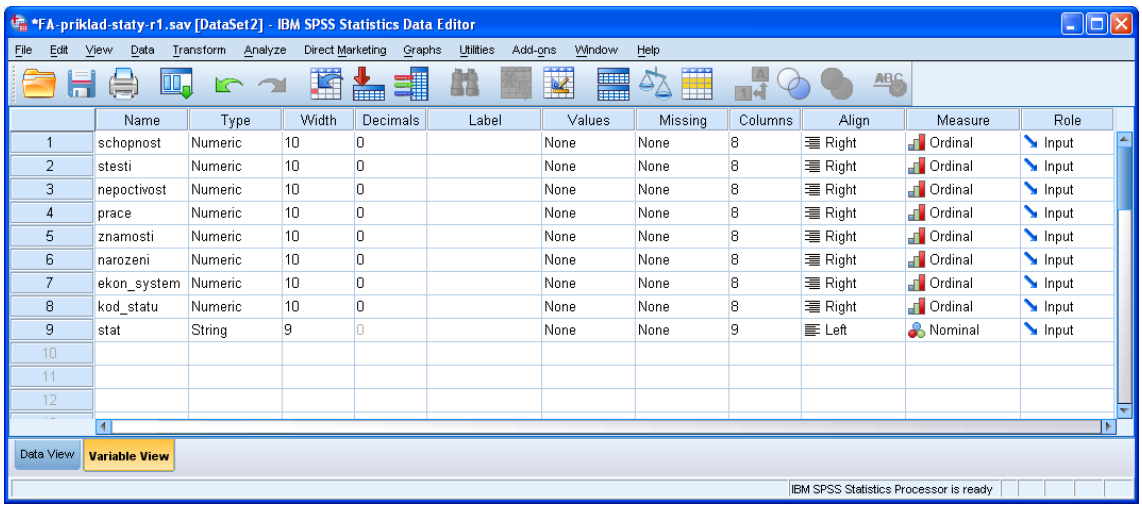
OBRÁZEK 4 ZÁLOŽKY V DATOVÉM OKNĚ

Záložka ***Data View*** zobrazuje pohled na data, kde v prvním řádku jsou uvedeny názvy proměnných a v dalších řádcích jsou jednotlivé záznamy. Ve sloupcích jsou uvedeny hodnoty dané proměnné pro konkrétní záznam (Obrázek 5). Zde lze i jednotlivé hodnoty dat editovat.



OBRÁZEK 5 ZÁLOŽKA DATA VIEW

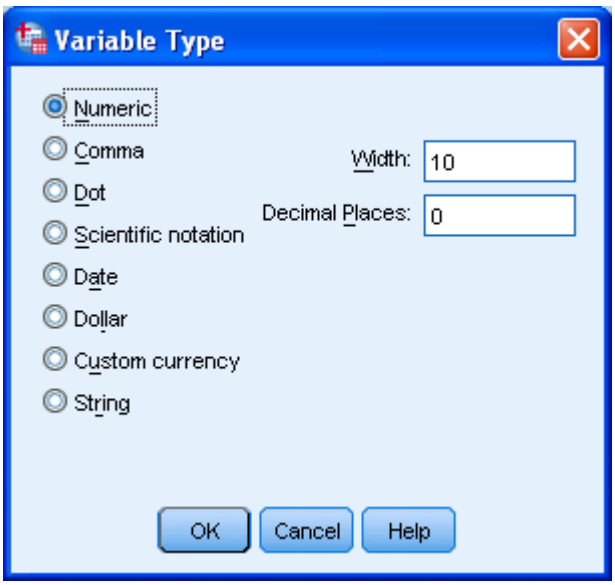
Záložka ***Variable View*** zobrazuje v jednotlivých řádcích názvy a vlastnosti zavedených proměnných (Obrázek 6).



OBRÁZEK 6 ZÁLOŽKA VARIABLE VIEW

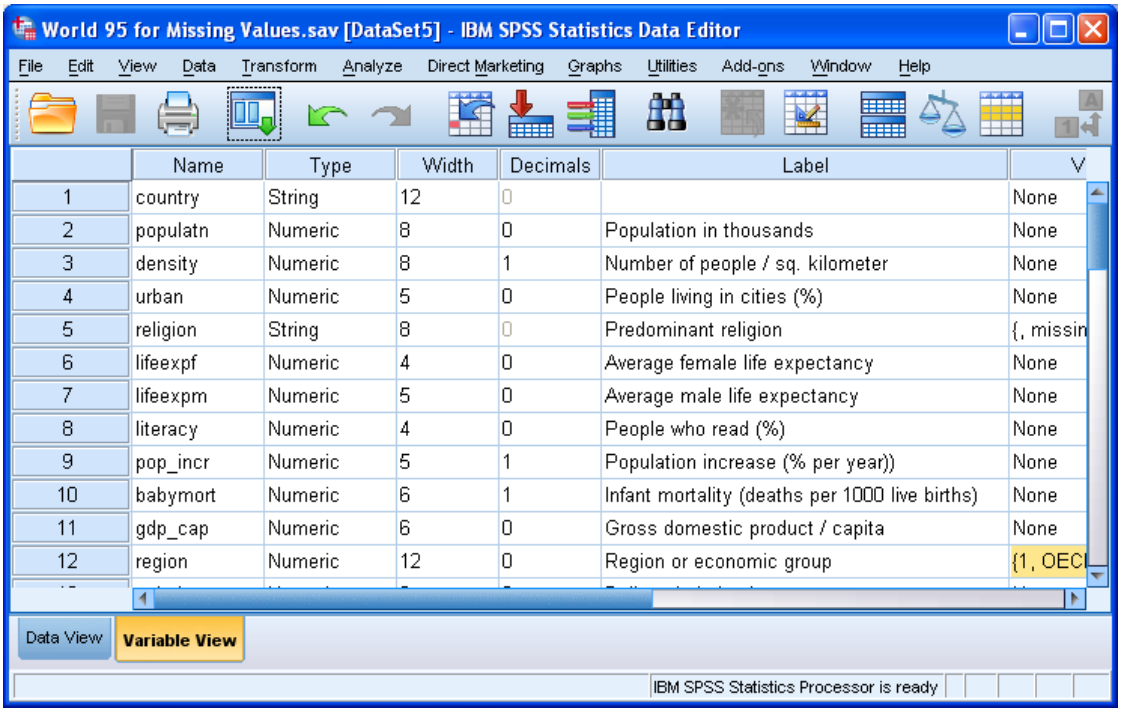
Jedná se o následující vlastnosti:

- Název proměnné (**Name**) – až 64 znaků; nesmí začínat číslicí, obsahovat mezery nebo různé speciální znaky (tečka, dvojtečka, čárka, středník apod.).
- Typ proměnné (**Type**), kde je možné volit z následujících typů (Obrázek 7):
  - číselná (**Numeric** – desetinný oddělovač čárka; **Comma** – desetinný oddělovač tečka, každé tři pozice oddělovač čárka; **Dot** - desetinný oddělovač čárka, každé tři pozice oddělovač tečka; **Scientific Notation**),
  - datum (**Date**),
  - číselná obsahující měnu nebo jednotku (**Dollar**, **Custom currency**),
  - textová (**String)**,
  - počet míst (**Width**),
  - počet desetinných míst (**Decimal Places**),



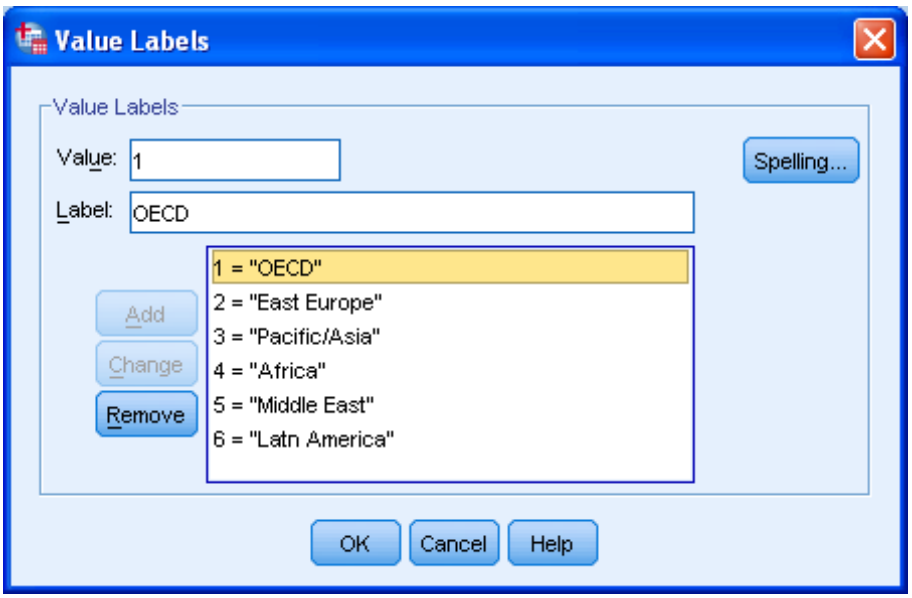
OBRÁZEK 7 VOLBA TYPU PROMĚNNÉ

- popis proměnné – až 256 znaků pro rozšířený popis proměnné oproti jejímu názvu (může se objevovat ve všech tabulkách a grafech místo názvu nebo společně s ním), (Obrázek 8),



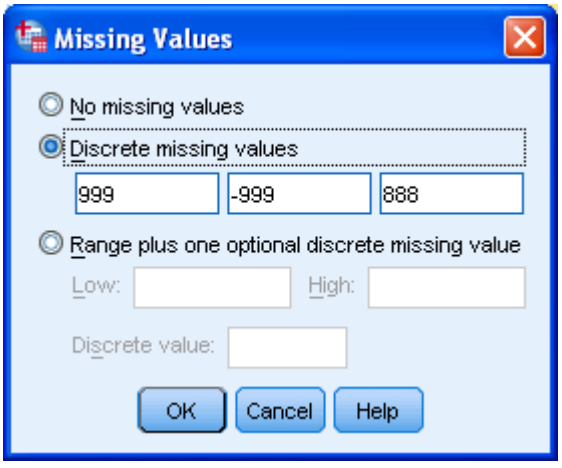
OBRÁZEK 8 POPIS PROMĚNNÝCH

- popis hodnot proměnné – v případech, kdy je vhodnější pracovat s číselnými kódy, je zde uveden význam jednotlivých kódů ve formě textu (může se objevovat ve všech tabulkách a grafech místo názvu nebo společně s ním), (Obrázek 9),



OBRÁZEK 9 POPIS HODNOT PROMĚNNÉ

- definice chybějících hodnot – v tomto případě se jedná o uživatelem definované chybějící hodnoty a jsou zde možné tři možnosti (Obrázek 10):
  - bez chybějících hodnot (**No missing values**),
  - definování maximálně tří diskrétních hodnot (**Discrete missing values**) reprezentujících různé důvody vzniku chybějící hodnoty (např. kód 999, -999, 888),
  - definování rozsahu hodnot nebo jedna hodnota (**Range plus one optional discrete missing value**),



OBRÁZEK 10 DEFINICE CHYBĚJÍCÍCH HODNOT

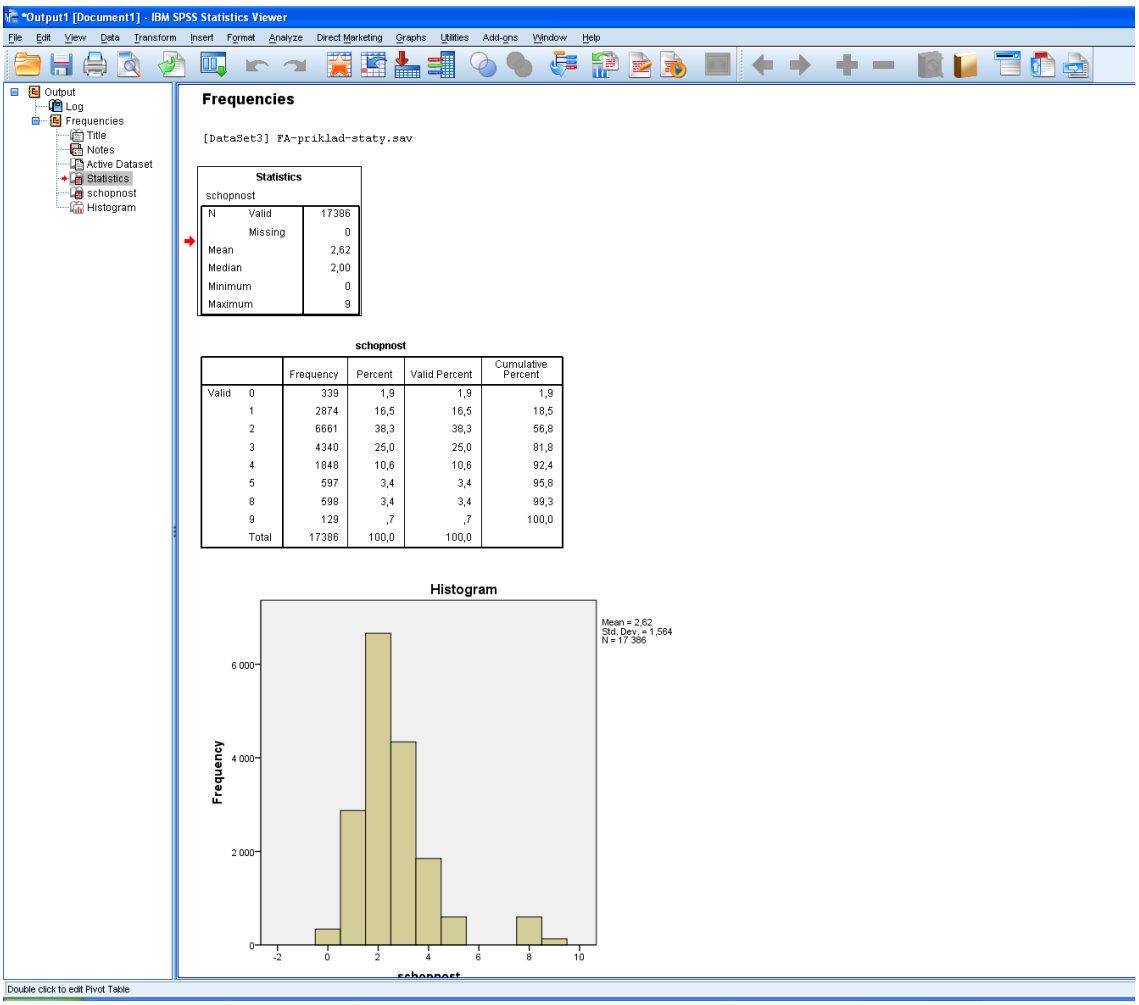
- šířka sloupce v datové matici (**Columns**) – šířku sloupce můžeme upravit přetažením myší na záložce ***Data View*** nebo nastavením tohoto parametru na záložce ***Variable View***,

- zarovnání hodnot proměnných v datové matici (**Align**) – vlevo, na střed nebo vpravo,
- způsob měření (**Measure**) – rozlišujeme, zda se jedná o proměnnou číselnou (**scale**), nominální (**nominal**) nebo ordinální (**ordinal**),
- role proměnné (**Role**) – některé dialogy při nastavování parametrů umožňují využít předdefinované role proměnných (proměnné, které splňují požadavky, se zobrazují v cílovém seznamu); jsou zde následující možnosti:
  - Vstupní (**Input**) – může být použita jako vstupní proměnná (prediktor, nezávislá proměnná).
  - Cílová (**Target**) – bude použita jako výstupní nebo cílová proměnná (závislá proměnná).
  - Obojí (**Both**) – může být použita jako vstupní nebo výstupní proměnná.
  - Nemá roli (**None**) – nemá přidělenou žádnou roli.
  - Rozdělení (**Partition**) – proměnná bude použita pro rozdělení dat do oddělených vzorků (trénovací, testovací, validační).
  - Štěpení (**Split**) – zajišťuje kompatibilitu s IBM SPSS Modelerem.

  Standardně je všem proměnným přiřazena vstupní role.

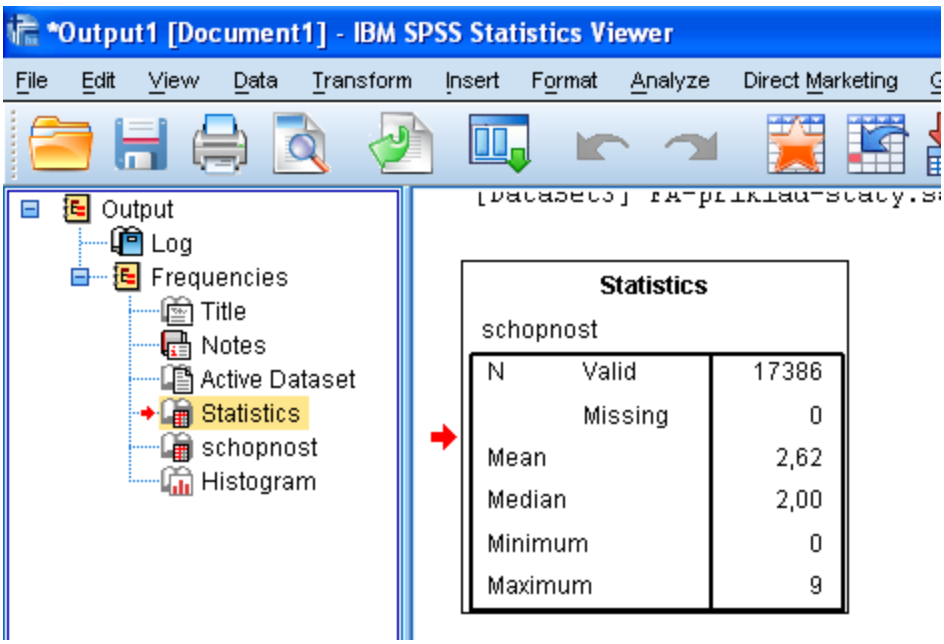
2.2.2. Výstupové okno

Do tohoto okna se zaznamenávají veškeré výstupy programu (tabulky, grafy, texty, hlášení apod.). Jednotlivé výstupy lze do určité míry dále editovat nebo graficky upravovat. V horní části okna je nástrojová lišta, panel nabídek a dolní část je rozdělena na dvě části (Obrázek 11). V levé části jsou ve formě stromové struktury zobrazeny všechny objekty v tomto okně (činnosti a výstupy v programu). Dílčí části v této struktuře lze dále skrývat/zobrazovat a editovat. To zlepšuje přehlednost jednotlivých výstupů a práci s nimi. V pravé části se potom nacházejí jednotlivé objekty ze stromové struktury.



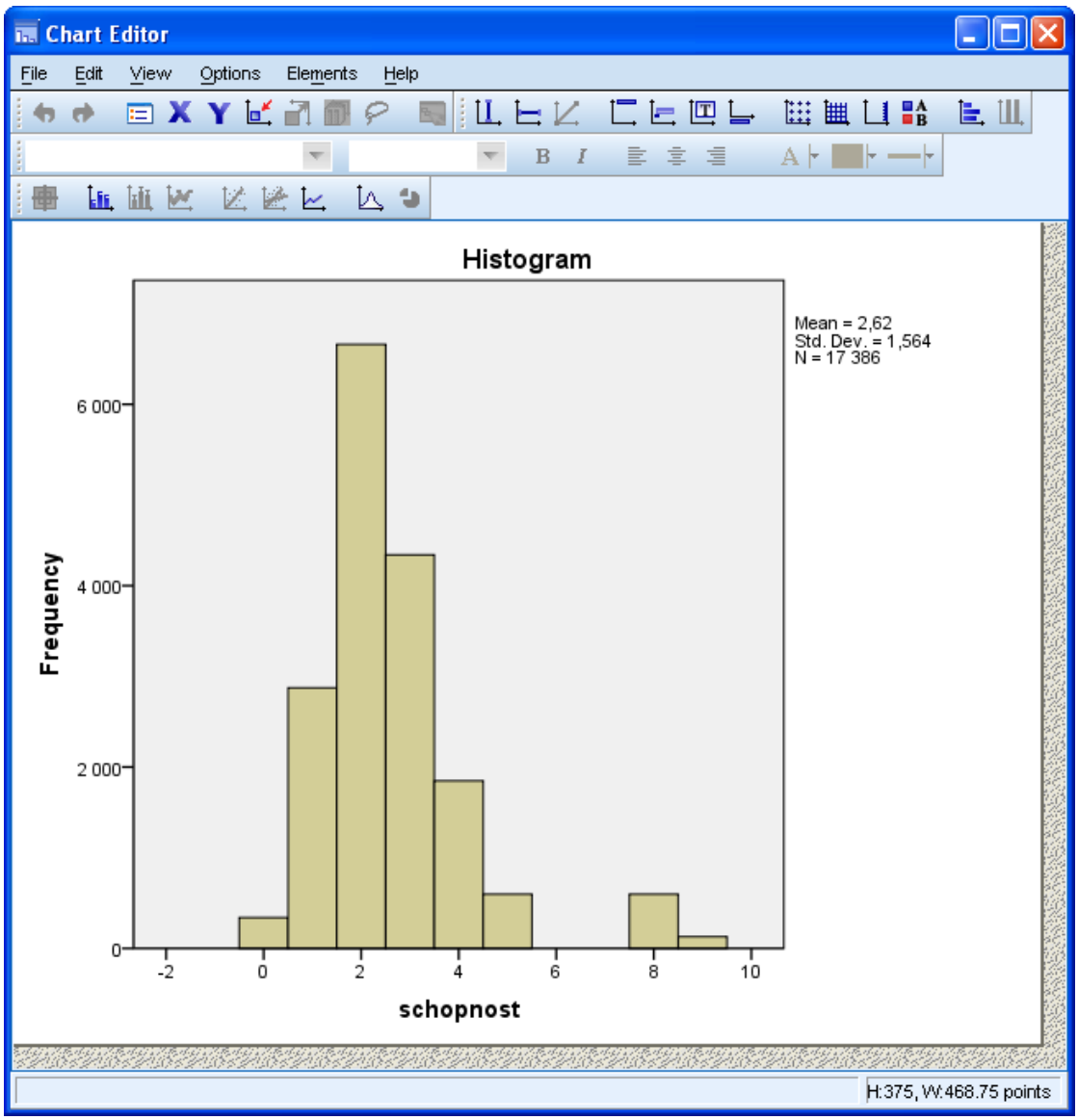
OBRÁZEK 11 STRUKTURA VÝSTUPOVÉHO OKNA

Jednotlivé položky lze upravovat tak, že na ně dvakrát poklepeme myší. Tímto způsobem lze pomocí myši nebo s využitím nabídek a ikon měnit uspořádání položek nebo jejich hierarchii v stromové struktuře výstupového okna. Signalizace editovatelné položky je pomocí červené šipky (Obrázek 12).



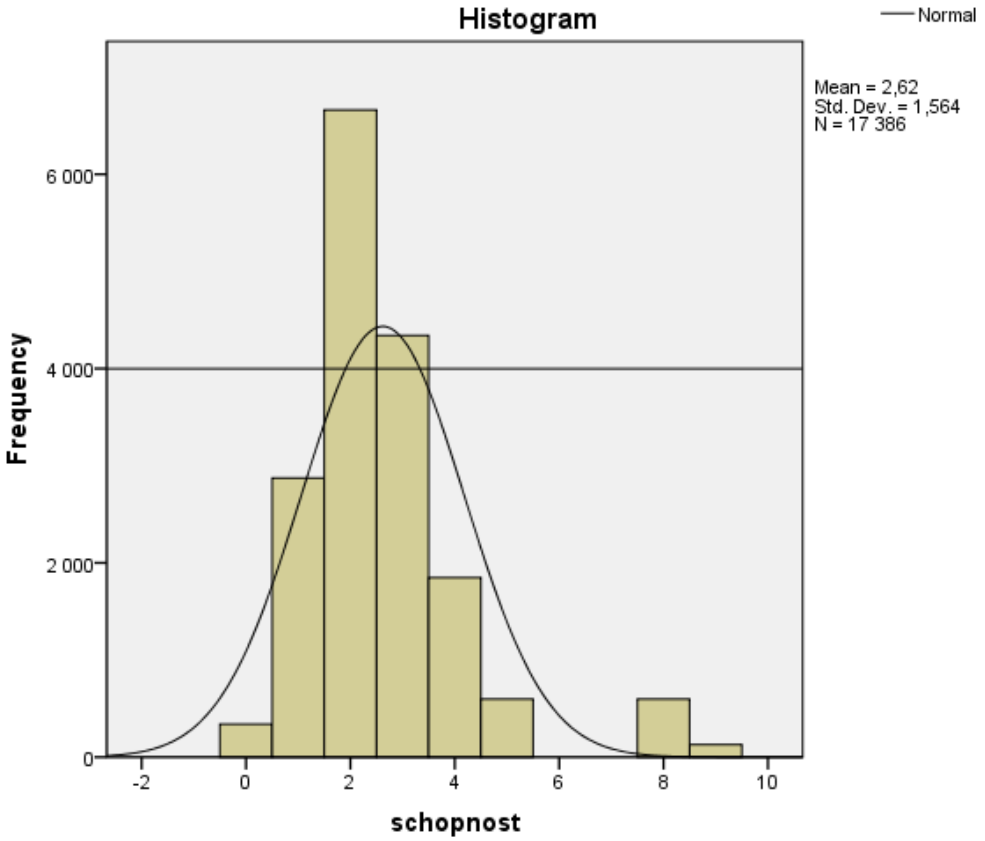
OBRÁZEK 12 OZNAČENÍ EDITOVATELNÉ POLOŽKY

Stejným způsobem lze editovat i grafy (Obrázek 13). Nástrojová lišta i odpovídající menu umožňují editovat veškeré prvky grafu (formát, měřítko a orientaci jednotlivých os, formát a zobrazení pracovní oblasti grafu, kopírovat jednotlivé části grafu, přidávat popisky do grafu, vkládat různé typy např. distribučních křivek, doplňovat referenční úrovně apod.). Všechny prvky jsou zde definovány jako objekty a následně jim lze přidělovat požadované vlastnosti.



OBRÁZEK 13 EDITACE VE VÝSTUPOVÉM OKNĚ

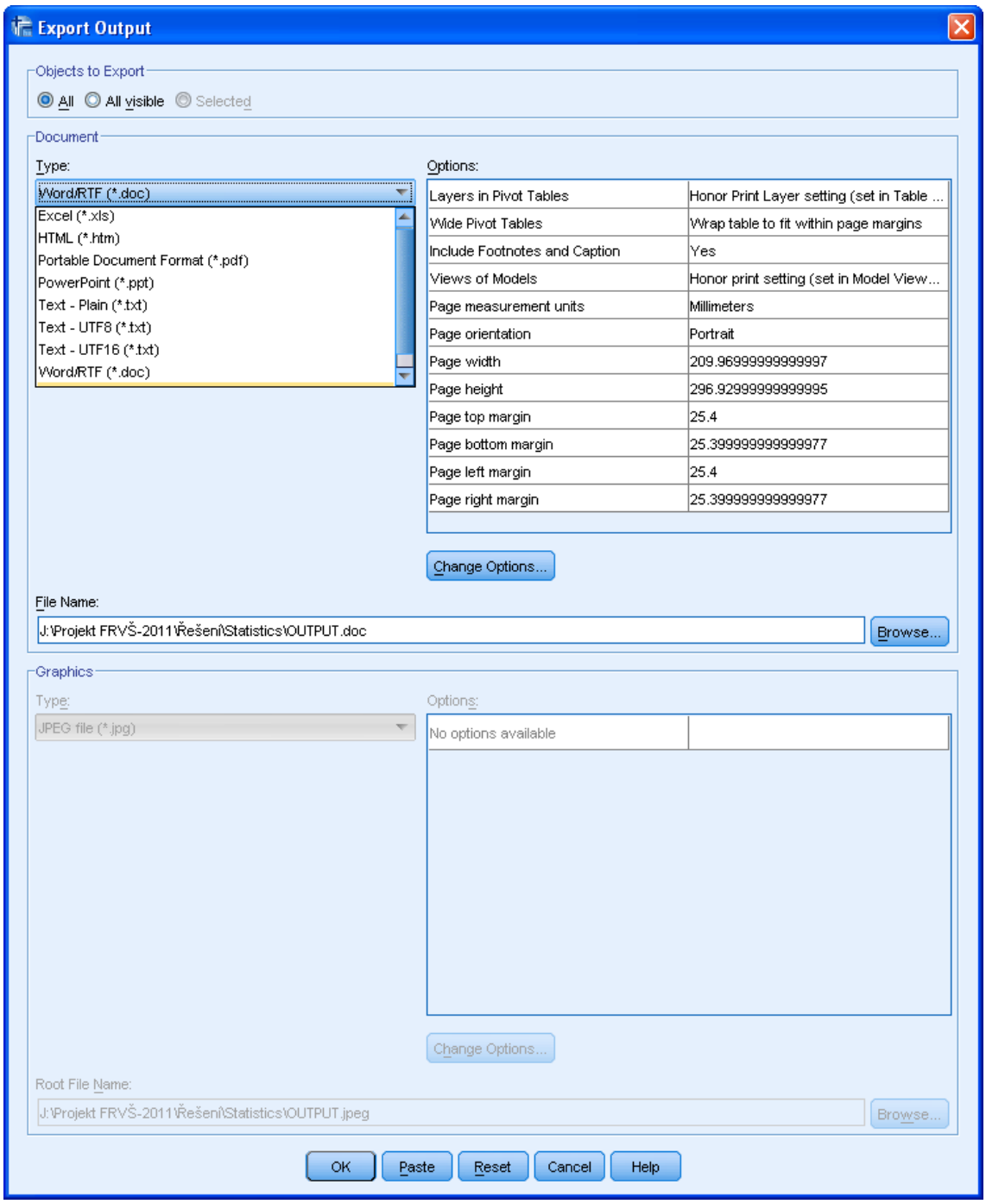
Na následujícím příkladu (Obrázek 14) je ukázáno jedno z možných editování výstupního grafu. Je zde doplněna distribuční křivka a referenční hodnota na úrovni 4000.



OBRÁZEK 14 PŘÍKLAD EDITACE VÝSTUPNÍHO GRAFU

Všechny výstupy z výstupového okna můžeme ukládat. Soubor výstupového okna má příponu *.spv. Způsob ukládání je buď nabídkou Uložit (**Save**) nebo Uložit jako (**Save as…**). Soubor z výstupového okna můžeme později kdykoliv v IBM SPSS Statistics otevřít a dále s ním pracovat (včetně jeho editace).

Pro práci s výstupy v jiném prostředí než je IBM SPSS Statistics, je třeba převést výstupy do jiného formátu přes nabídku **File – Export** (Obrázek 15).

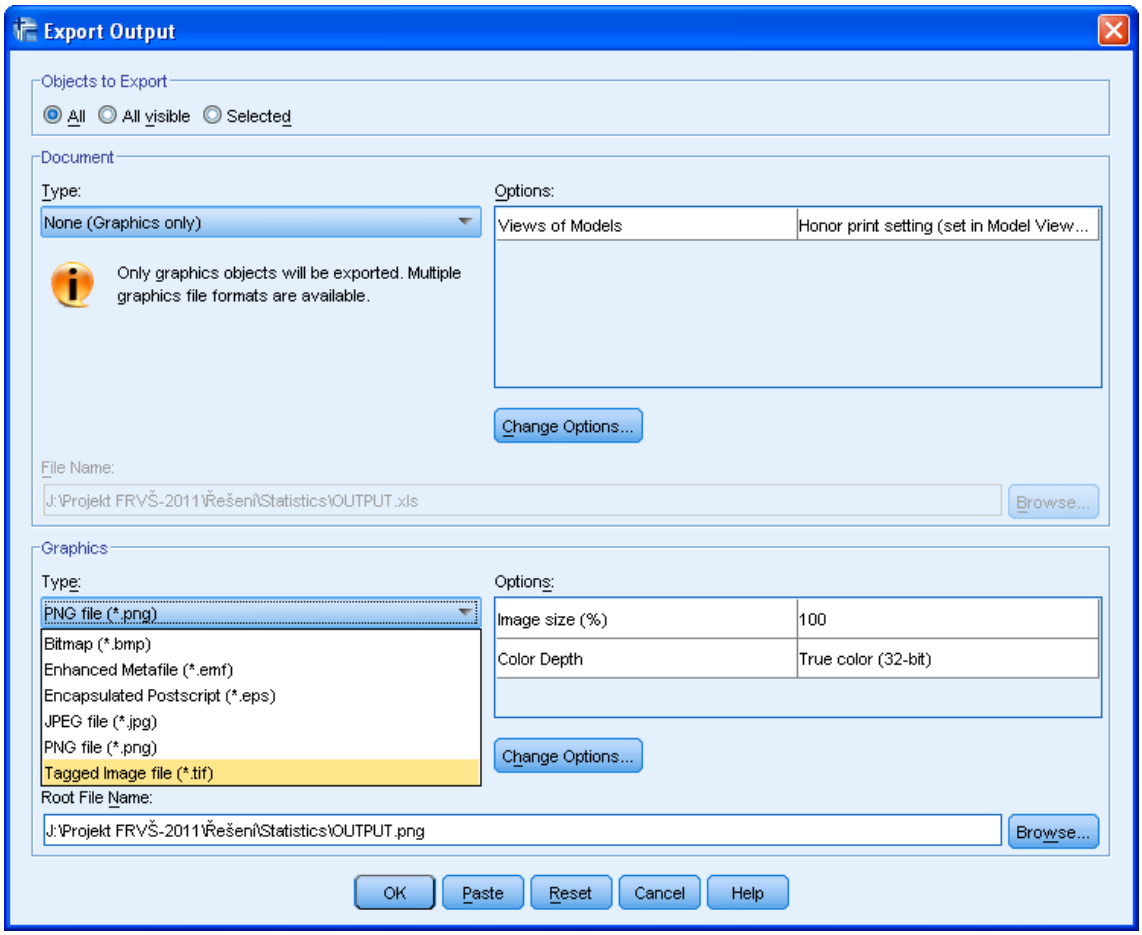


OBRÁZEK 15 DIALOGOVÉ OKNO PRO EXPORT Z VÝSTUPOVÉHO OKNA – MIMO FORMÁT GRAFIKA

V tomto dialogu v části **Object to Export** volíme, zda chceme exportovat všechny objekty výstupového okna (**All**), zobrazené objekty (**All Visible**) nebo jen vybrané objekty (**Selected**). Následně v části **Document** lze vybrat z jednotlivých nabízených formátů pro export (Obrázek 15).

V části **Options** lze upřesnit, co a jakým způsobem bude exportováno u tabulek. Změna nastavení je možná přes volbu **Change Options**. V části **File Name** zadáváme cestu a název ukládaného exportovaného souboru.

V případě, že při exportu zvolíme volbu ukládání v grafickém formátu, zpřístupní se volba pro výběr formátu grafického souboru (Obrázek 16). V tomto případě jsou z výstupového okna exportovány pouze grafické objekty (grafy). Další doplňující nastavení výstupu lze opět ovlivnit pomocí části **Options**.

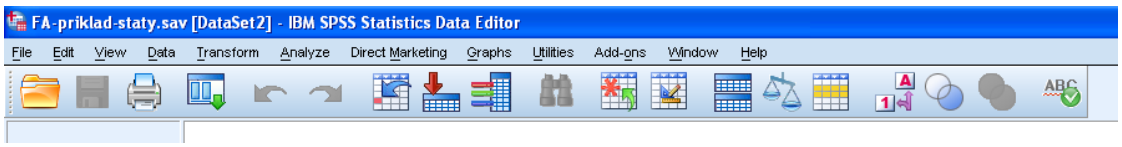


OBRÁZEK 16 DIALOGOVÉ OKNO PRO EXPORT Z VÝSTUPOVÉHO OKNA V GRAFICKÉM FORMÁTU

Tímto končí část potřebná k seznámení se základní obsluhou výstupového okna programu IBM SPSS Statistics.

2.3. Základní ovládání programu IBM SPSS Statistics

Při ovládání programu IBM SPSS Statistics máme k dispozici panel nástrojů v podobě ikon a panel nabídek (Obrázek 17). Kombinací těchto dvou panelů se lze dostat k většině standardních možností tohoto programu. Panel nástrojů je určen pro rychlý přístup k dané funkci bez nutnosti procházet jednotlivé nabídky. Je samozřejmé, že panel nástrojů lze dále editovat a tím přizpůsobovat potřebám jednotlivých uživatelů s cílem zrychlení práce.



OBRÁZEK 17 ZÁKLADNÍ PANEL V IBM SPSS STATISTICS

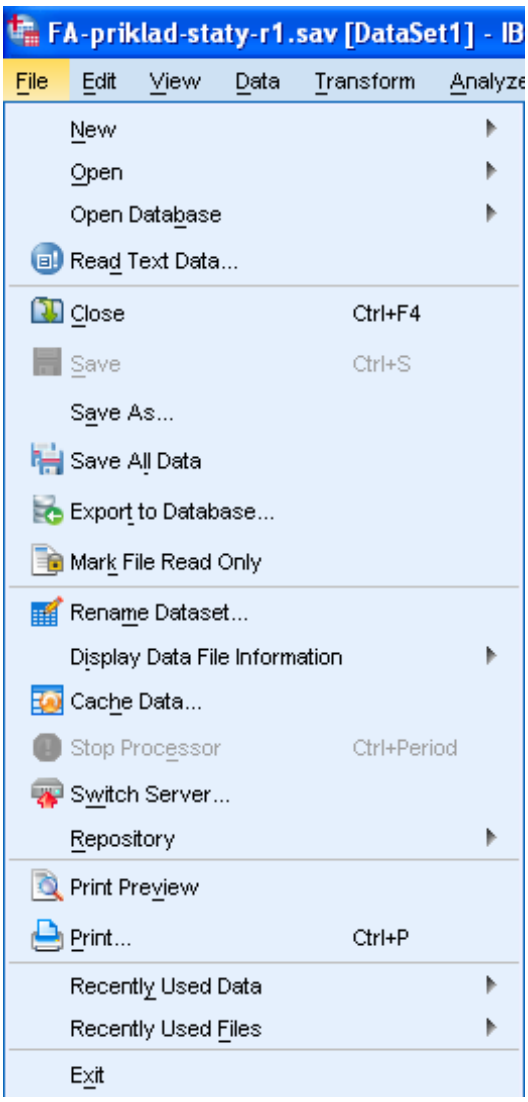
V následující části bude uveden přehled jednotlivých položek z panelu nabídek.

2.3.1. Soubor (File)

Tato nabídka je určena k práci se souborem. Jsou zde k dispozici následující možnosti (Obrázek 18):

- vytvořit nový soubor (**New**) - datový, syntaxový, výstupový, skriptový,
- otevřít existující soubor (**Open**) - datový, syntaxový, výstupový, skriptový,
- načíst data z databáze (**Open Database**) - vytvořit nový dotaz, editovat hotový dotaz, spustit dotaz,
- načíst data z textového souboru (**Read Text Data**),
- zavřít okno (**Close**),
- uložit (**Save**) – uložení pod dříve zadaným jménem nebo zadání jména souboru,
- uložit jako **(Save as**) – uložení souboru pod jiným jménem,
- uložení všech datových souborů (**Save All Data**) – uloží všechny otevřené datové soubory,
- export do databáze (**Export to Database…)** – přidání dat do databáze,
- označení souboru pouze pro čtení/pro psaní a čtení (**Mark File Read Only/Mark File Read Write**) – jedná se o přepínač pro určování vlastností souboru z hlediska možnosti jeho změny,
- přejmenování datového souboru (**Rename Dataset…**) – v případě, že je otevřeno více verzí stejného datového souboru jsou odlišeny názvem v hranaté závorce (např. [DataSet1]); údaj v této závorce lze v této volbě změnit,
- zobrazení informací o souboru (**Display Data File Information**) – doplnění informací o datovém souboru do výstupového okna,

- pracovní záloha souboru (**Cache Data**) – vytvoří pracovní zálohu datového souboru a při následujích operacích s ním pracuje; umožňuje to urychlit práci,
- zastavení výpočtu (**Stop Processor**) – pro zastavení výpočtu nebo spuštěné procedury,
- připojení k serveru (**Switch Server**) – při zpracování většího objemu dat může být výhodné se připojit k aplikaci na výkonnějším počítači,
- datové úložiště (**Repository**) – připojení k datovému úložišti,
- náhled před tiskem (**Print Preview**) – zobrazení v tiskovém formátu,
- tisk (**Print**) – tisk zvolených objektů z aktuálního okna,
- naposledy používané datové soubory (**Recently Used Data**) – zobrazení maximálně devíti naposledy používaných datových souborů,
- naposledy používané soubory (**Recently Used Files**) – zobrazí naposledy používané soubory kromě datových (výstupy, syntax apod.),
- ukončení programu IBM SPSS Statistics (**Exit**).



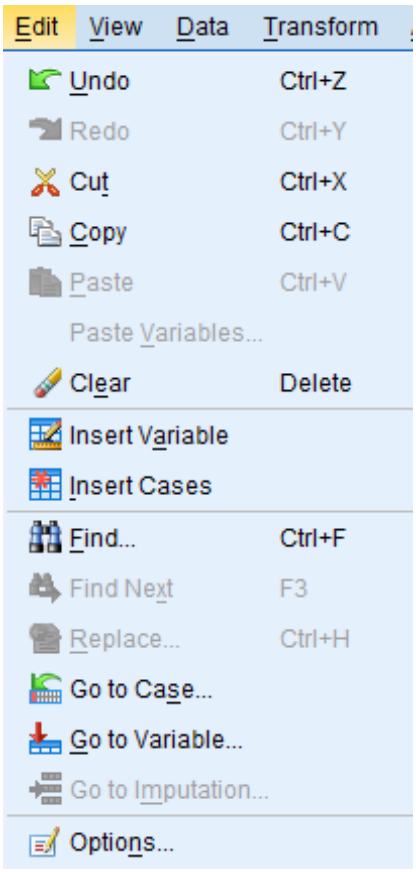
OBRÁZEK 18 NABÍDKA SOUBOR

2.3.2. Úpravy (Edit)

Tato nabídka dovoluje, kromě práce se schránkou v systému Windows, také vracet jednotlivé kroky úprav, editovat, hledat a pohybovat se v datovém okně. Jsou zde následující možnosti (Obrázek 19):

- zpět (**Undo**) – krok zpět při úpravě datové matice,
- vpřed (**Redo**) – krok dopředu při úpravě datové matice,
- vyjmout (**Cut**) – přesune vybrané položky do schránky,
- kopírovat (**Copy**) – kopíruje vybrané položky do schránky,
- vložit (**Paste**) – vloží položku ze schránky na vybrané místo,
- vložit proměnnou (**Paste Variable**) – na základě upřesnění v dialogovém okně vloží požadovaný počet proměnných (na základě dat ve schránce) se zadanými jmény (je aktivní pouze na záložce ***Variable View***),

- smazat (**Clear**) – smaže vybrané údaje,
- vložit novou proměnnou (**Insert Variable**) – vloží novou prázdnou proměnnou před místo označené kurzorem,
- vložit nové případy (**Insert Cases**) – vloží nové prázdné řádky do datové matice (jejich počet odpovídá počtu označených řádků),
- najít (**Find**) – hledá požadovaný řetězec (lze upřesnit možnosti prohledávání),
- najít další (**Find Next**) – najde další záznam podle údajů z položky Find,
- nahradit (**Replace**) – nahradí zadaný řetězec novým řetězcem,
- přejít na případ (**Go to Case…**) – přechod na zadané číslo případu,
- přejít na proměnnou (**Go to Variable**) – přechod na zadanou proměnnou,
- nastavení (**Options…**) – nastavení prostředí IBM SPSS Statistics pro potřeby uživatele.

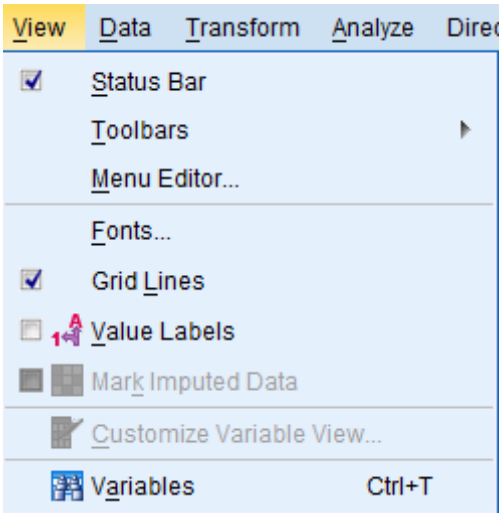


OBRÁZEK 19 NABÍDKA EDIT

2.3.3. Pohled (View)

Tato nabídka definuje zobrazení/skrytí a vzhled prvků datového okna. Jsou zde následující možnosti (Obrázek 20):

- stavový řádek (**Status Bar**) – zobrazení nebo skrytí stavového řádku v IBM SPSS Statistics (jsou zde zobrazovány doplňující informace),
- panely nástrojů (**Toolbars**) – zobrazení/skrytí nebo editace panelu nástrojů v IBM SPSS Statistics,
- úprava nabídek (**Menu Editor…**) – pomocí této volby lze měnit obsah jednotlivých nabídek v oknech IBM SPSS Statistics, lze ji i doplňovat vlastními nabídkami,
- písmo (**Fonts…**) – nastavení písma pro datovou matici,
- zobrazení mřížky (**Grid Lines**) – zapnutí/vypnutí zobrazení mřížky v datové matici,
- popisy hodnot (**Value Labels**) – přepínač pro zobrazení popisu hodnot v datové matici (pokud jsou tyto hodnoty popsány v záložce ***Variable View***),
- zobrazování vlastností proměnných (**Customize Variable View…**) – řídí zobrazování vlastností proměnných na záložce ***Variable View***, lze si vybírat, které proměnné budou zobrazeny a které skryty,
- přepínání mezi záložkami datového okna (**Variables/Data**).



OBRÁZEK 20 NABÍDKA VIEW

Toto byly základní volby z nabídky sloužící k ovládání programu IBM SPSS Statistics.

Další nabídky jako jsou **Data**, **Transform**, **Analyze**, **Direct Marketing**, **Graphs**, **Utilities**, **Add-ons** slouží již k realizaci vlastních analýz a dalších výpočetních operací v tomto programu. Popis těchto nabídek bude součástí dalšího dílu tohoto textu.

Zbývající volby **Windows** a **Help** mají již standardní funkce související se zobrazením jednotlivých oken a využíváním nápovědy.

3. Základní informace o programu IBM SPSS Modeler

Software IBM SPSS Modeler podporuje v každém kroku (od převzetí dat až po předání skórovacích výsledků) mezinárodně uznávanou Data Miningovou metodiku **CRoss-Industry Standard Process for Data Mining (CRISP-DM)**, která je sama o sobě hodnotným rámcem Data Miningových (DM) postupů.

IBM SPSS Modeler podporuje klíčové aktivity, mezi které patří především:

- tvorba zákaznických profilů a určení jejich hodnoty,
- detekce a predikce podvodů,
- detekce a predikce vazeb v datech z webu,
- predikce budoucích prodejních a růstových trendů,
- odhad účinnosti marketingových akcí, kreditní riziko,
- odhad rizik v monitorování procesů,
- predikce odcházejících klientů (churn), klasifikace, segmentace zákazníků,
- analýza velmi rozsáhlých dat, objevování skrytých vazeb a struktur.

Díky vizuálnímu rozhraní a uživatelsky zaměřenému ovládání IBM SPSS Modeler lze veškeré analýzy realizovat pomocí tvorby datového proudu. Toto se uskutečňuje v prostředí IBM SPSS Modeler snadno, přirozeně a intuitivně - pouhým tahem myši. IBM SPSS Modeler vytváří proudy datového procesu, v nichž je obsažen každý krok toku DM úloh. S IBM SPSS Modeler se lze zaměřit na užití metod a postupů bez zbytečné ztráty času plynoucí z nutnosti naučit se složité ovládání softwaru.

IBM SPSS Modeler obsahuje širší rozsah procedur strojového učení i statistických procedur než jiné DM nástroje. Můžete volit mezi algoritmy pro seskupování, klasifikaci, asociaci a predikci.

Modelovací algoritmy v IBM SPSS Modeler jsou rozděleny následovně:

- Základní modul IBM SPSS Modeler
- Modul IBM SPSS Classification
    - Podporuje návrh klasifikačních a regresních modelů.
    - Zahrnuje tyto algoritmy: Neuronové sítě k odhalování slabých nebo skrytých vztahů v datech. C5.0, pokročilý rozhodovací strom a algoritmus pro tvorbu pravidel. Decision List – nový algoritmus založený na pravidlech, který zakomponuje interní

obchodní znalost do pravidel, které odhalil algoritmus. V kombinaci s automatizací modul nabídne navíc i optimální model. Využitím obchodních znalostí lze získat lepší aplikační výsledky v úlohách databázového marketingu, výzkumu trhu, skórování kreditního rizika, cílení programů ve státní správě, výzkumu v organizacích a medicínském výzkumu. Do tohoto modulu patří regresní metody, neuronové sítě, faktorová analýza, diskriminační analýza, rozhodovací stromy.

- Modul IBM SPSS Segmentation
  - Pomůže seskupit data do skupin podobných profilů, rozvrstvit zákazníky do homogenních a podobně reagujících segmentů a detekovat neobvyklé a anomální případy.
  - Nabízí tři výkonné seskupovací algoritmy pro řešení problémů a pro pochopení zákonitostí v datech: algoritmus TwoStep je založený na hierarchických metodách, které se používají k získání nejlepšího počtu klastrů pro daný problém. Kohonenova mapa je seskupovací neuronová síť nazývaná také "self-organization map" (SOM) určená k odhalení vztahů, které mohou být skryté v mnohorozměrných datech. Detekce anomálií - odhalení a vysvětlení neobvyklých vztahů pomocí algoritmu založeného na seskupování. Obsahuje metodu K-means, Kohonenovy mapy, metodu Two Step a detekci anomálií.
- Modul IBM SPSS Association
  - Obsahuje tyto algoritmy: algoritmus Apriori pro odhalování asociací s pokročilými evaluačními funkcemi. Asociační algoritmus Carma, který podporuje pravidla s více závěry. Uzel Sequence pro odhalování asociací závislých na pořadí událostí. Obsahuje metodu Apriori, Carma a Sequence.

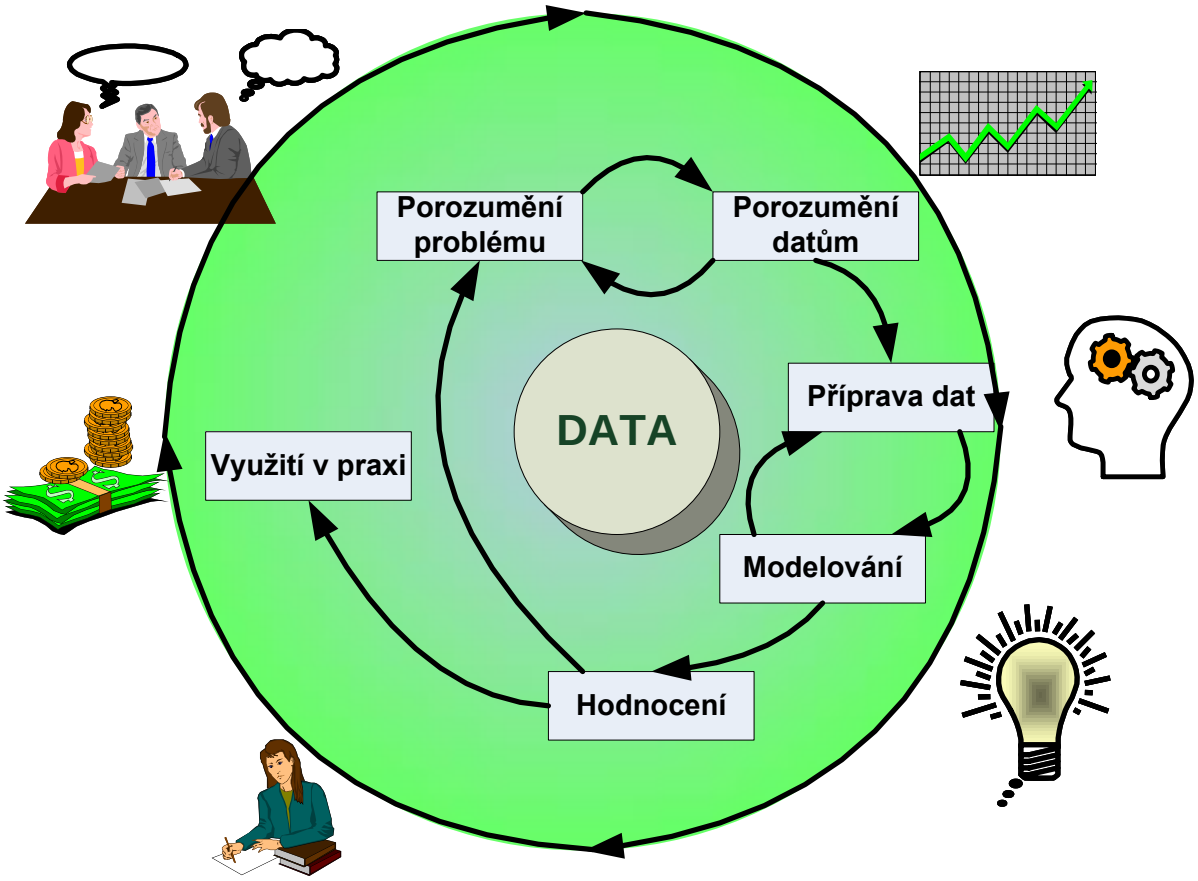
3.1. Fáze CRISP-DM

Životní cyklus projektu DM je podle metodiky CRISP-DM tvořen následujícími šesti fázemi (Obrázek 21):

1. porozumění problému,
2. porozumění datům,

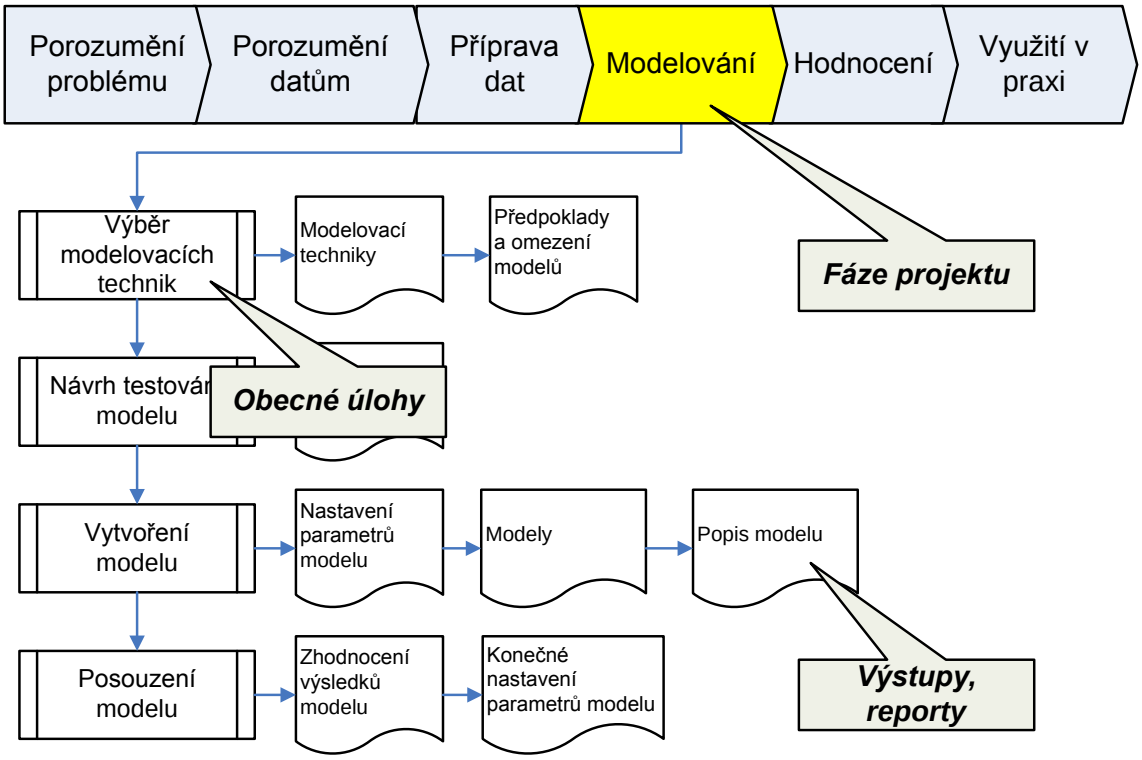
3. příprava dat,
4. modelování,
5. hodnocení,
6. využití v praxi.

Pořadí jednotlivých fází není pevně dáno. Výsledek dosažený v jedné fázi ovlivňuje volbu kroků následujících, často je třeba se k některým krokům a fázím vracet. Vnější kruh na obrázku symbolizuje cyklickou povahu procesu dobývání znalostí z databází jako takovou.



OBRÁZEK 21METODIKA CRISP-DM

Pro následující objasňování jednotlivých etap a úkolu řešených podle metodiky CRISP-DM budeme používat následující strukturní schéma (Obrázek 22).



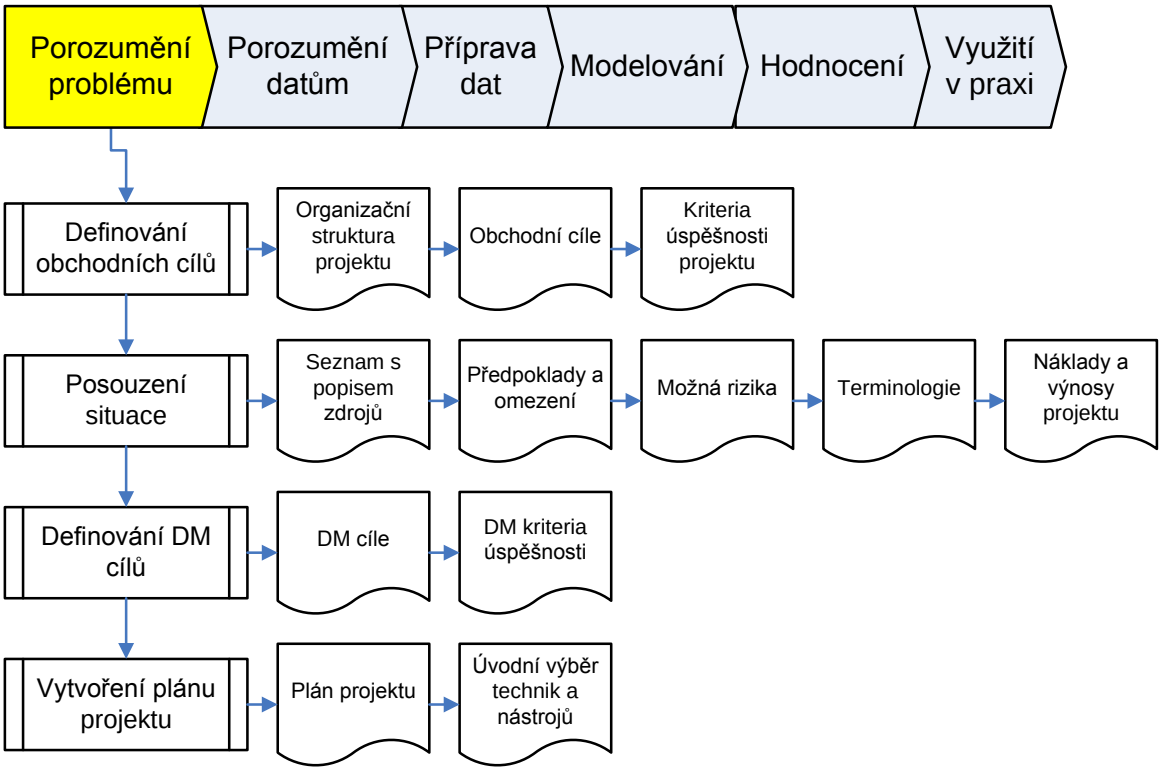
OBRÁZEK 22 PŘÍKLAD STRUKTURY FÁZE PODLE CRISP-DM

3.1.1. Porozumění problému (Business Understanding)

Tato úvodní fáze je zaměřena na pochopení cílů úlohy a požadavků na řešení formulovaných z manažerského hlediska (Obrázek 23). Manažerská formulace musí být následně převedena do zadání úlohy pro dobývání znalostí z databází.

Manažerský problém, ke kterému pomocí metod DM hledáme informace, může být formulován (téměř) bez vazby na informace získávané pomoci metod DM z dostupných dat. Příkladem může být snaha nabídnout uložení části peněz na zvláštní účet s delší výpovědní lhůtou pomoci reklamy vhodně zacílené na nadějnou skupinu klientů (i potenciálních) banky. Pro DM to znamená nalézt takovou charakteristiku klientů, která zajišťuje, že ve skupině klientů s touto charakteristikou bude velká část klientů mít stále dostatečně vysoký zůstatek na účtu. V tomto případě je zadání pro DM formulováno relativně přesně. Přesto je však třeba počítat s možností přeformulovat nebo upřesnit manažerský problém na základě provedených analýz. Jinou možnou úlohou je otázka včasného rozpoznáni klientů, kteří představují rizikovou skupinu z hlediska spláceni poskytnutého úvěru.

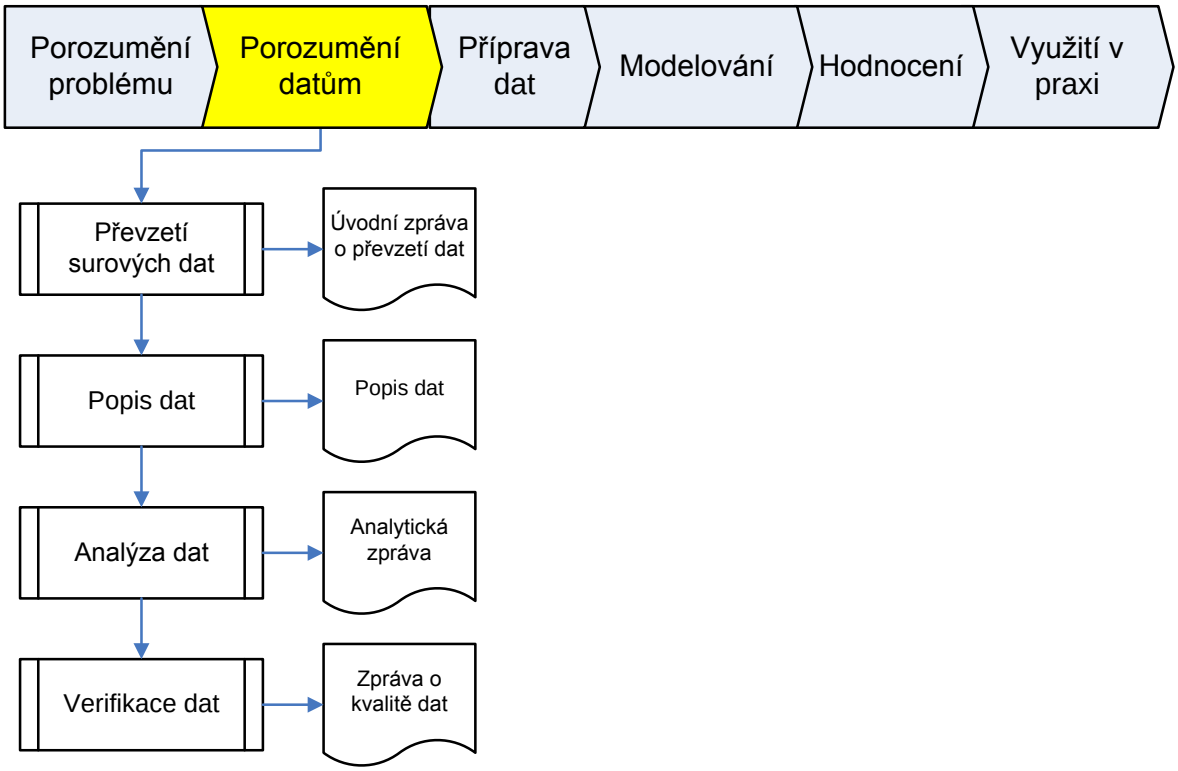
V této fázi se rovněž provádí inventura zdrojů (datových výpočetních i lidských), hodnotí se možná rizika, náklady a přínos použití metod KDD a stanovuje se předběžný plán prací.



OBRÁZEK 23 SCHÉMA ČINNOSTÍ A VÝSTUPŮ ETAPY POROZUMĚNÍ PROBLÉMU

3.1.2. Porozumění datům (Data Understanding)

Fáze porozumění datům začíná prvotním sběrem dat (Obrázek 24). Následují činnosti, které umožní získat základní představu o datech, která jsou k dispozici (posouzení kvality dat, první „náhled" do dat, vytipování zajímavých podmnožin záznamů v databázi…). Obvykle se zjišťují různé deskriptivní charakteristiky dat (četnosti hodnot různých atributů, průměrné hodnoty, minima, maxima apod.). S výhodou se využívají i různé vizualizační techniky.



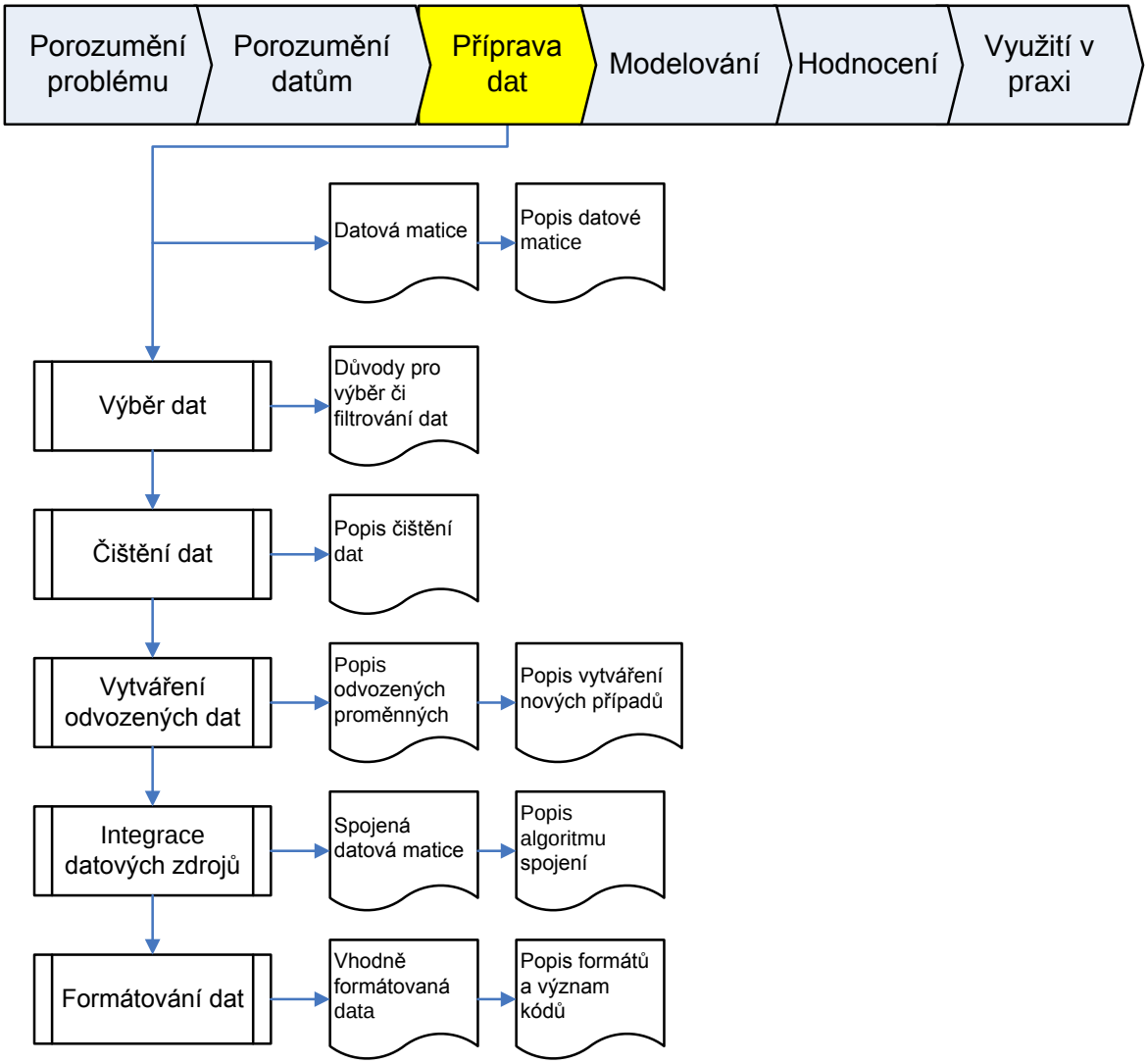
OBRÁZEK 24 SCHÉMA ČINNOSTÍ A VÝSTUPŮ ETAPY POROZUMĚNÍ DATŮM

3.1.3. Příprava dat (Data Preparation)

Příprava dat zahrnuje činnosti, které vedou k vytvoření datového souboru, který bude zpracováván jednotlivými analytickými metodami (Obrázek 25). Tato data by měla:

- obsahovat údaje význačné pro danou úlohu,
- mít podobu, která je vyžadována vlastními analytickými algoritmy.

Příprava dat tedy zahrnuje selekci dat, čištění dat, transformaci dat, vytváření dat, integrování dat a formátování dat. Tato fáze je obvykle nejpracnější částí řešení celé úlohy. Jednotlivé úkony jsou obvykle prováděny opakovaně, v nejrůznějším pořadí.



OBRÁZEK 25 SCHÉMA ČINNOSTÍ A VÝSTUPŮ ETAPY PŘÍPRAVY DAT

3.1.4. Modelování (Modeling)

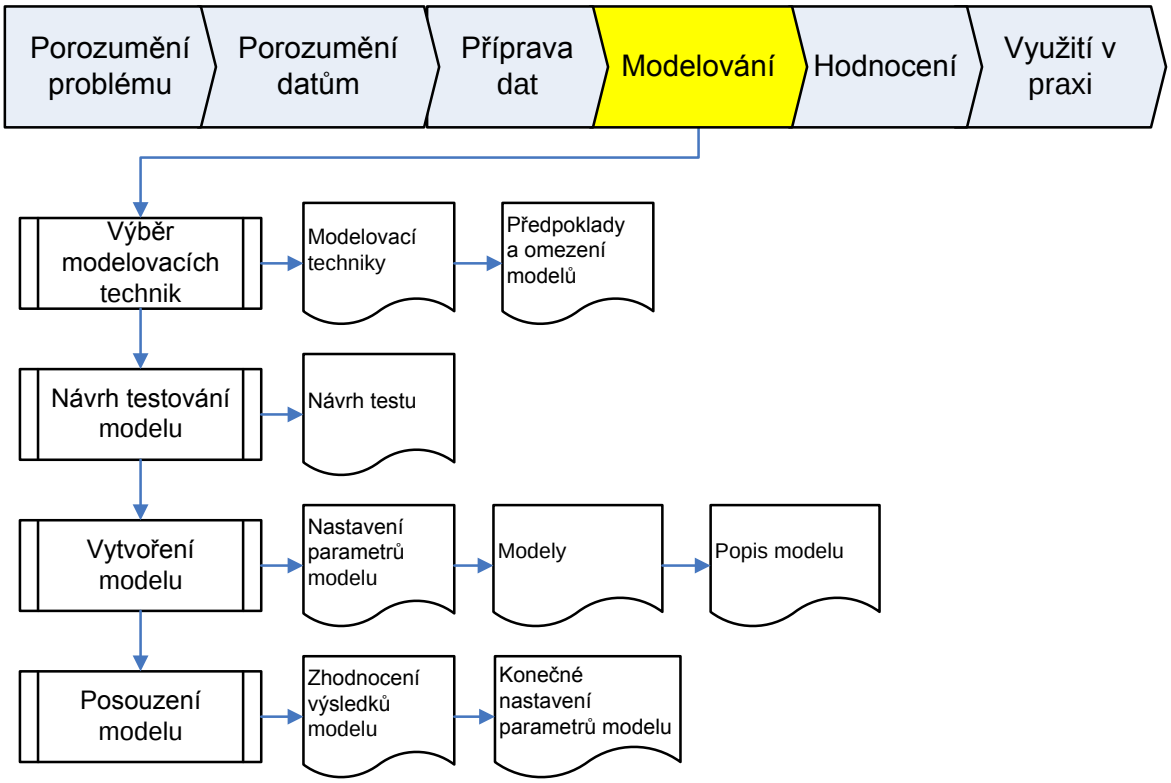
V této fázi jsou nasazeny analytické metody - algoritmy pro dobývání znalosti (Obrázek 26). Obvykle existuje řada různých metod pro řešení dané úlohy, je tedy třeba vybrat ty nejvhodnější (doporučuje se použít více různých metod a jejich výsledky kombinovat) a vhodně nastavit jejich parametry. Jde tedy opět o iterační činnost (opakovaná aplikace algoritmů s různými parametry), použití analytických algoritmů může navíc vést k potřebě modifikovat data, a tedy k návratu k datovým transformacím z předcházející fáze.

Pro hledání „zajímavých“ skupin klientů je možné použít metody shlukování nebo asociační pravidla. Pro rozpoznání rizikových klientů z hlediska půjček jsou (vzhledem k tomu, že data o klientech obsahují informace o průběhu splácení) vhodné

například algoritmy pro tvorbu rozhodovacích stromů nebo rozhodovacích pravidel. Tyto metody je vhodné kombinovat. Například shluková analýza může rozdělit klienty do skupin a rozhodovací strom potom umožní jednotlivé skupiny charakterizovat dostatečné srozumitelným způsobem.

Součástí této fáze je rovněž ověřování nalezených znalostí z pohledu metod dobývání znalostí. To může představovat např. testování klasifikačních znalostí na nezávislých datech.

Znalosti „deskriptivní" (charakteristiky skupiny klientů „zajímavých" z hlediska připravovaného produktu) jsou předkládány expertům z banky. Klasifikační znalosti (umožňující „rozpoznat“ klienty, kteří nesplácejí úvěr) jsou testovány například na novém vzorku dat.



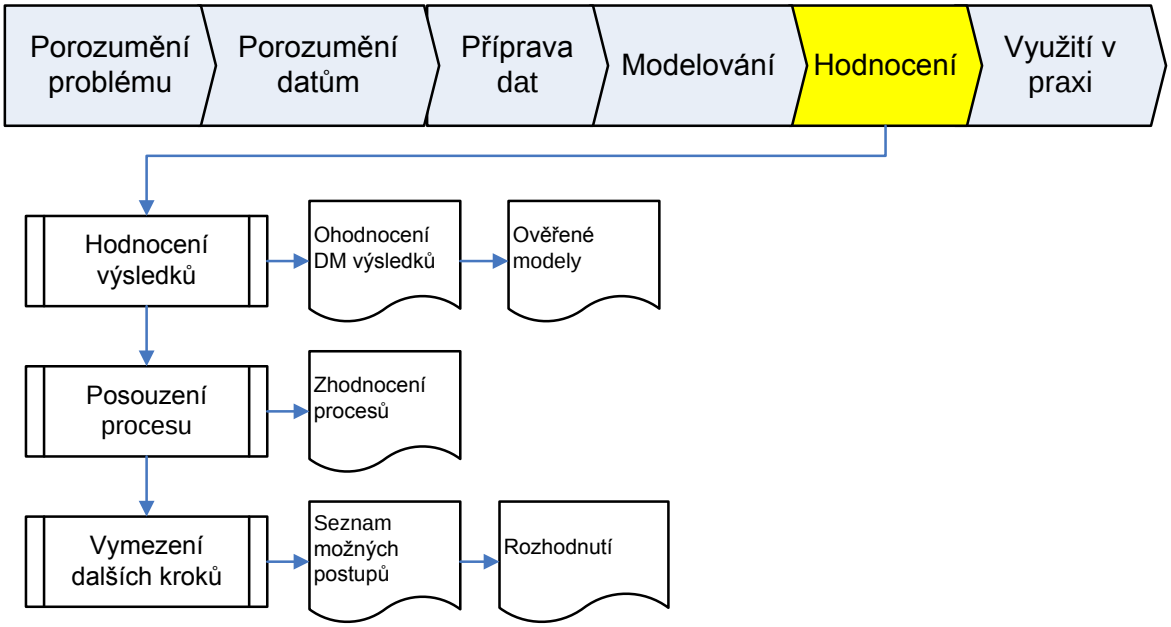
OBRÁZEK 26 SCHÉMA ČINNOSTÍ A VÝSTUPŮ ETAPY MODELOVÁNÍ

3.1.5. Vyhodnocení výsledků (Evaluation)

V této fázi jsme se dopracovali do stavu, kdy jsme nalezli znalosti, které se zdají být v pořádku z hlediska metod dobývání znalostí. Dosažené výsledky je ale ještě třeba vyhodnotit z pohledu manažerů, zda byly splněny cíle formulované při zadání úlohy (Obrázek 27).

Některé nalezené skupiny klientů experty nepřekvapily, vědělo se o nich a banka se připravovala je oslovit dopisem. Jiné (rovněž bonitní skupiny) byly shledány zajímavými, ale budou ještě podrobeny dalšímu zkoumání. Výsledky testování klasifikačních znalostí ukázaly, že systém byl příliš „přísný“, tedy správně rozpoznával klienty rizikové, ale v určitých případech (obzvláště u vyšších půjček) za rizikové označil i klienty bonitní. Bylo tedy rozhodnuto, že ve všech pobočkách banky bude využíván program, který bude rozhodovat o úvěrech do určité částky.

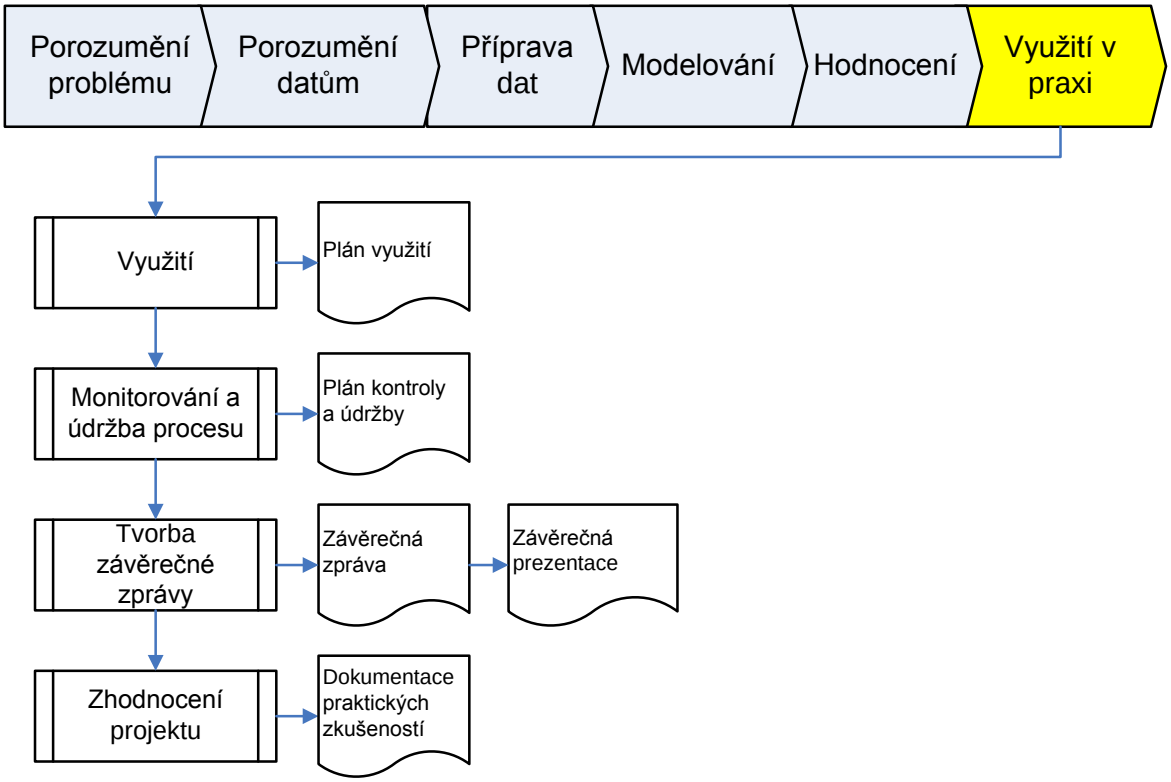
Na závěr této fáze by mělo být přijato rozhodnutí o způsobu využití výsledků.



OBRÁZEK 27 SCHÉMA ČINNOSTÍ A VÝSTUPŮ ETAPY HODNOCENÍ

3.1.6. Využití výsledků (Deployment)

Vytvořením vhodného modelu řešení úlohy obecně nekončí (Obrázek 28). Dokonce i v případě, že řešenou úlohou byl pouze popis dat, je třeba získané znalosti upravit do podoby použitelné pro zákazníka (manažera – zadavatele úlohy).



OBRÁZEK 28 SCHÉMA ČINNOSTÍ A VÝSTUPŮ ETAPY VYUŽITÍ V PRAXI

Podle typu úlohy může využití (nasazení) výsledků znamenat na jedné straně prosté sepsání závěrečné zprávy, na straně druhé pak zavedení (hardwarové, softwarové, organizační) systému pro automatickou klasifikaci nových případů.

Ve většině případů je to zákazník a nikoliv analytik, kdo provádí kroky vedoucí k využívání výsledků analýzy. Proto je důležité, aby pochopil, co je nezbytné učinit pro to, aby mohly být dosažené výsledky využívány efektivně.

3.1.7. Shrnutí

V této kapitole jsme si objasnili jednu z používaných metodik. Jedná se o metodiku CRISP-DM, kterou podporuje i produkt IBM SPSS Modeler. Ukázali jsme si strukturu a obsah jednotlivých etap podle této metodiky.

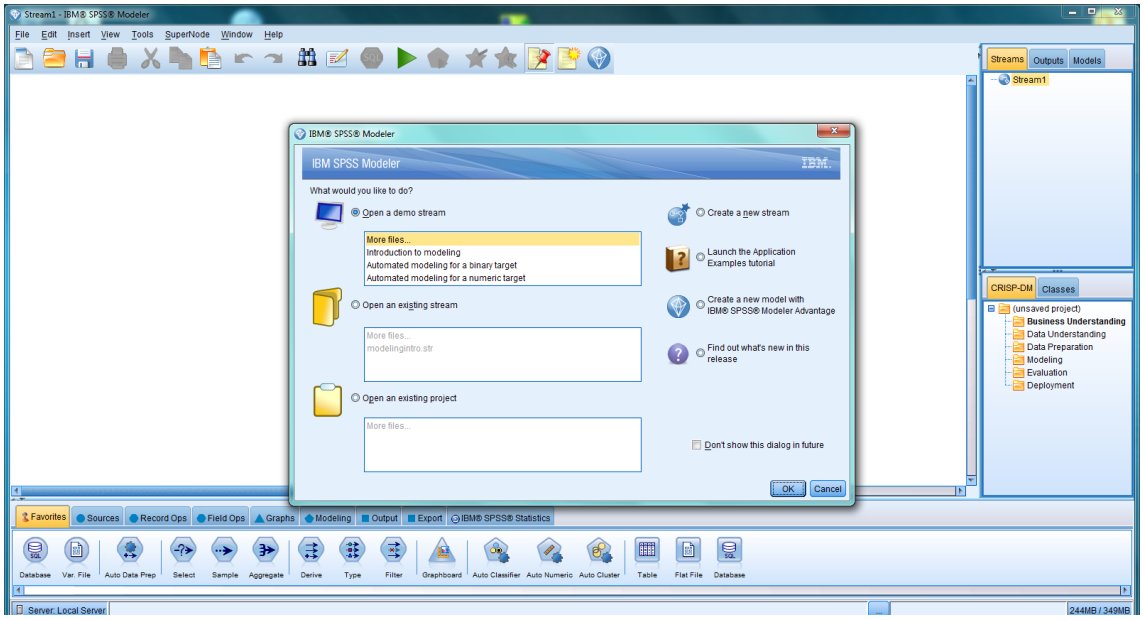
Z uvedeného je zřejmé, že se jedná o vcelku širokou škálu řešených úloh a výstupů. K jejímu úspěšnému řešení je třeba věnovat jednotlivým etapám potřebnou pozornost.

4. Program IBM SPSS Modeler

V této části bude popsáno základní prostředí uvedeného programu a základní operace, které je nutno znát pro efektivní využívání uvedeného programu.

4.1. Prostředí programu IBM SPSS Modeler

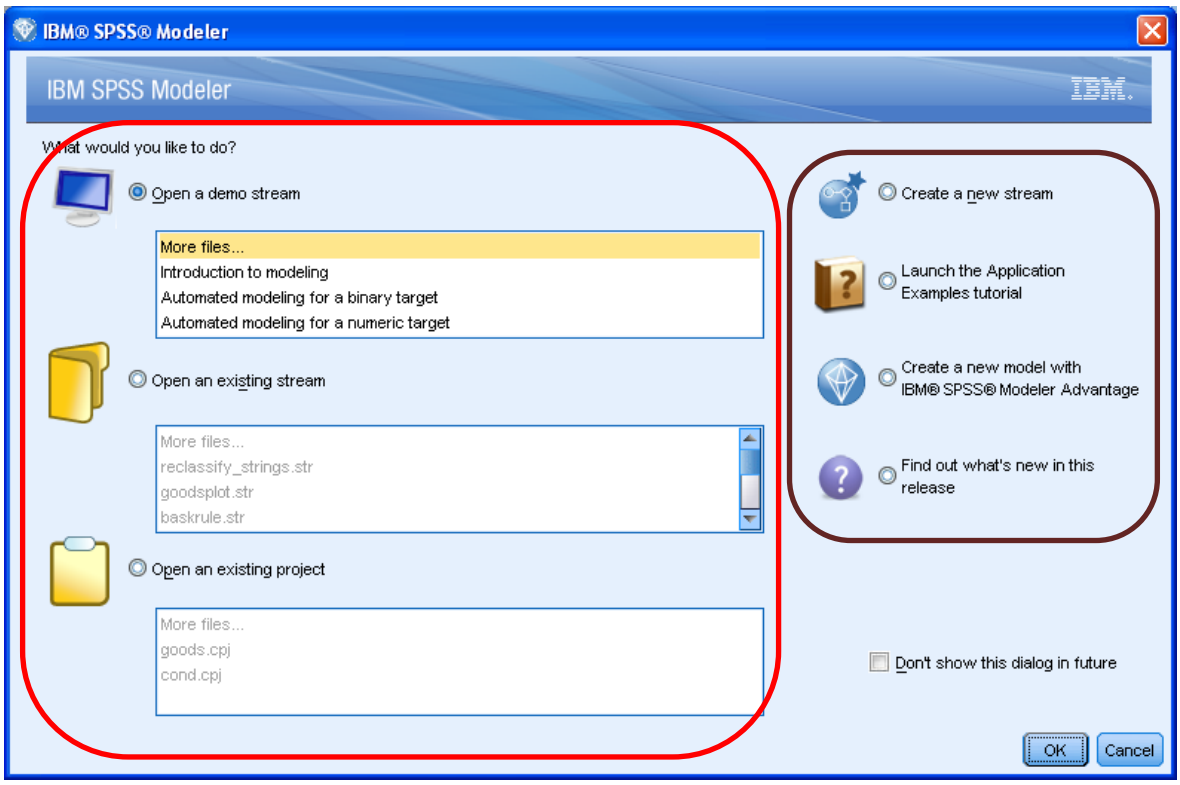
Při standardním spuštění programu je objeví následující okno (Obrázek 1).



OBRÁZEK 29 VZHLED PROSTŘEDÍ IBM SPSS MODELER

V tuto chvíli lze realizovat následující operace – orámovaná pravá část vstupního okna (Obrázek 30):

- vytvoření nového datového streamu (Create a new stream);
- spuštění tutoriálu k ukázkové aplikaci (Launch the Application Examples tutorial;
- vytvoření nového modelu s využitím serveru (Create a new model with IBM SPSS Modeler Advantage);
- zjistit, co je nového v nové verzi programu (Find dout what's new in this release).

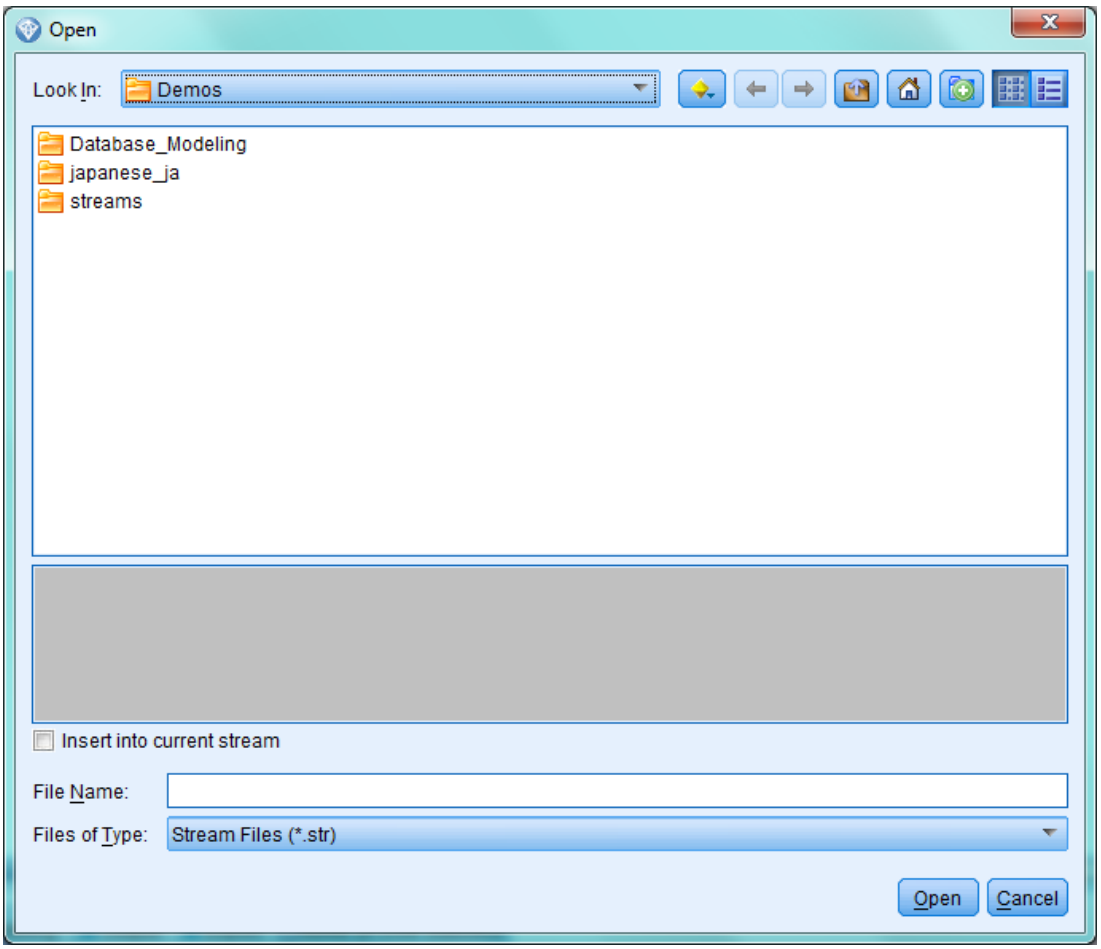


OBRÁZEK 30 ÚVODNÍ OKNO V PROSTŘEDÍ IBM SPSS MODELER

Pro naše potřeby zůstaneme u možnosti otev ření datového streamu, kterou lze upřesnit pomocí výběru z následující volby – orámovaná levá část vstupního okna (Obrázek 30):

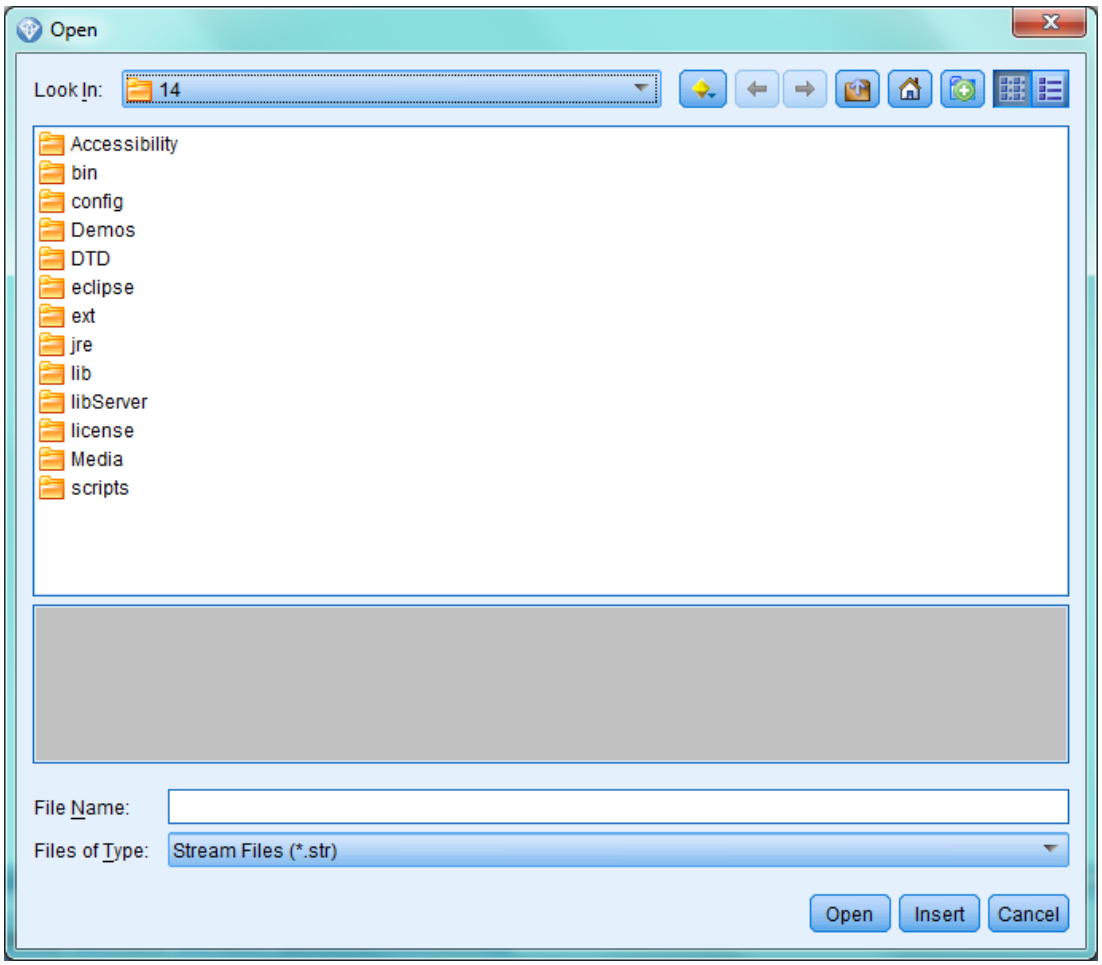
3. Otevřít stream ukázkového příkladu (Open a demo stream);
4. Otevřít existující stream (Open an existing stream);
5. Otevřít existující projekt (Opent an existing project).

První volba nám umožní rychlý přístup k souborům s datovými streamy ve složce „Demos“ programu IBM SPSS Modeler (Obrázek 31). Zároveň je možno vyhledat libovolný jiný soubor se streamem pomocí standardního prohledávání adresářů v prostředí Windows. Tyto soubory mají příponu *.str.



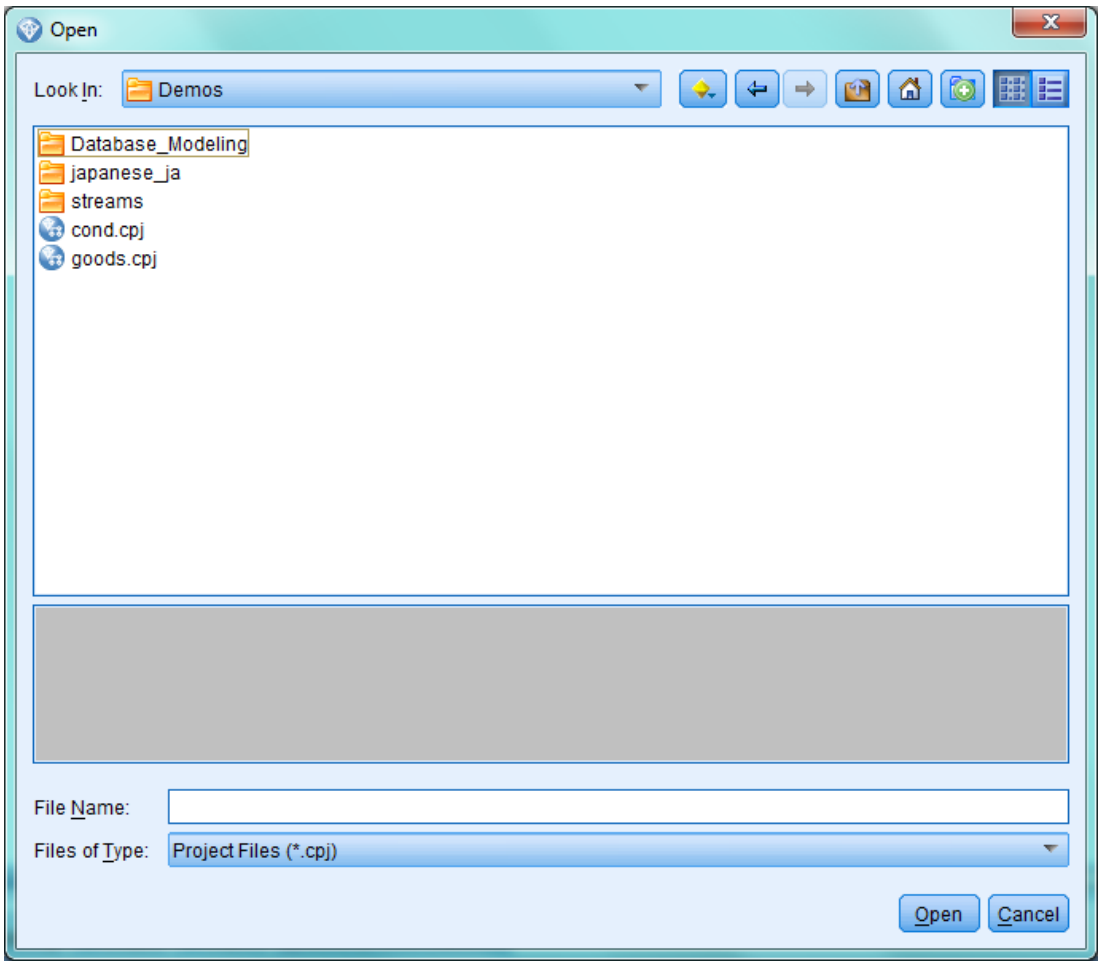
OBRÁZEK 31 OTEVŘENÍ DEMO SOUBORŮ

Druhá volba umožňuje otevření libovolného souboru pomocí vyhledání v adresářích (Obrázek 3) nebo volbou jednoho z devíti posledních otevřených souborů.

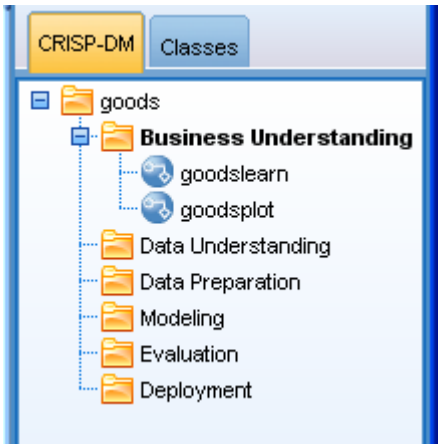


OBRÁZEK 32 OKNO VÝBĚRU LIBOVOLNÉHO SOUBORU

Poslední, třetí volba, slouží k otevření některého z existujících projektů v programu IBM SPSS Modeler. Jedná se o soubory s příponou *.cpj (Obrázek 33). V tomto případě se jedná nejen o načtení odpovídajícího streamu ale i všech souborů, které byly s tímto streamem uloženy v jednotlivých složkách podle metodiky CRISP-DM (Obrázek 34).

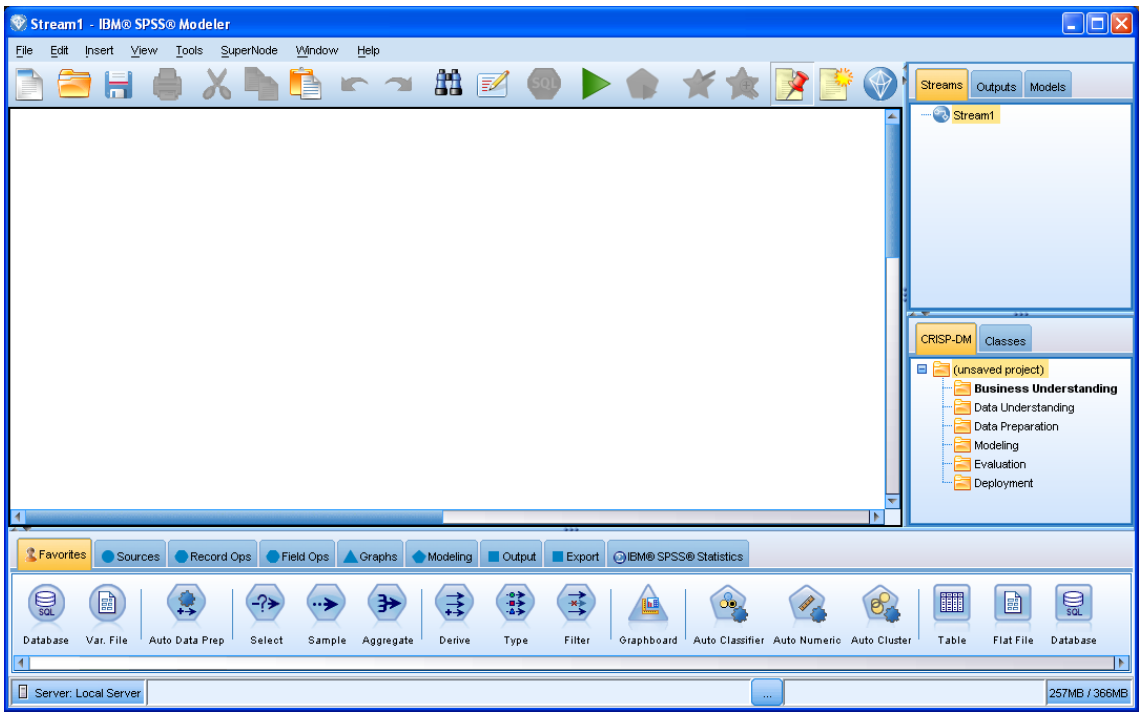


OBRÁZEK 33 OKNO VÝBĚRU EXISTUJÍCÍHO PROJEKTU



OBRÁZEK 34 STRUKTURA SLOŽEK PROJEKTU

Samozřejmou možností je, že nemusíme využít žádnou z těchto možností a otevře se nám prázdná okno v programu IBM SPSS Modeler (Obrázek 35). V tuto chvíli můžeme dále využívat všechny možnosti, které jsou určené pro práci s programem IBM SPSS Modeler.



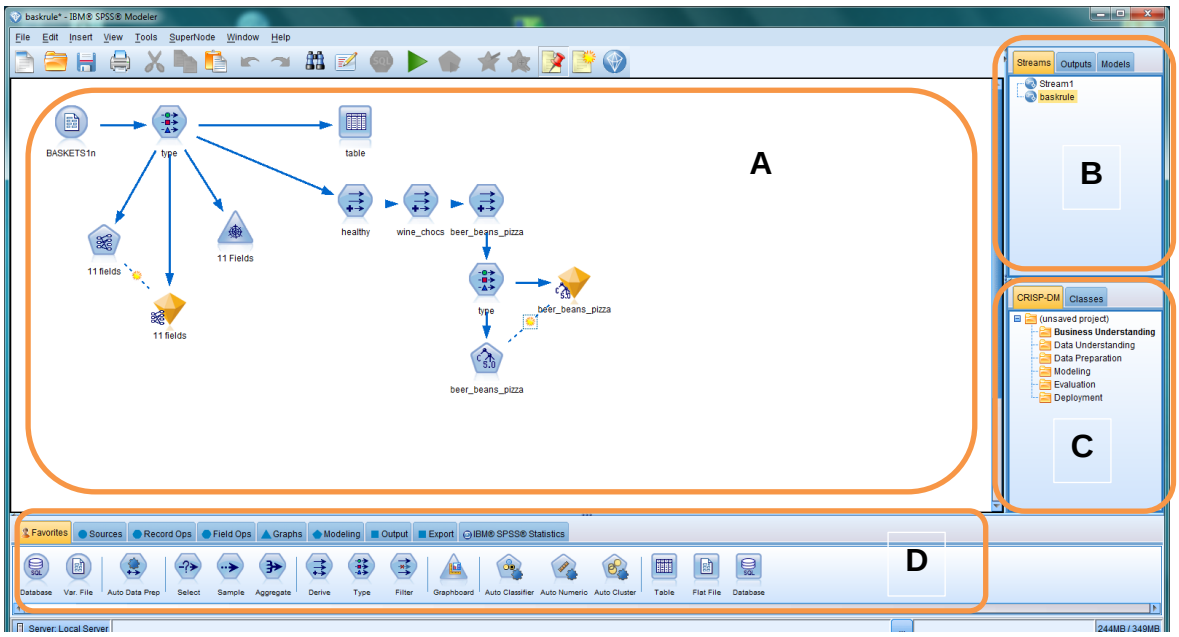
OBRÁZEK 35 VZHLED PRÁZDNÉHO ÚVODNÍHO OKNA IBM SPSS MODELER

4.2. Okna v programu IBM SPSS Modeler

Program IBM SPSS Modeler využívá rozdělení pracovní plochy na čtyři základní typy oken (Obrázek 36):

- okno pro tvorbu datového streamu (datové okno) – část A, slouží pro vlastní konstrukci postupu zpracování ve formě datového streamu;
- okno správce výstupů (výstupové okno) – část B, slouží k ukládání a řízení práce s výsledky jednotlivých datových streamů; zároveň zobrazuje všechny současně spuštěné datové streamy,
- okno správce projektu (projektové okno) – část C,
- okno palety uzlů – část D, zde jsou zobrazeny jednotlivé uzly (reprezentují činnosti a metody), které lze využít při tvorbě datového streamu.

Jednotlivá okna mají odlišné funkce, záložky a rovněž nabídky, které jsou zde k dispozici, se v každém z nich mírně liší.



OBRÁZEK 36 ZÁKLADNÍ TYPY OKEN V IBM SPSS MODELER

Vzhledem k rozsahu a určení tohoto návodu zde bude objasněna práce pouze v datovém a výstupovém okně.

4.2.1. Okno pro tvorbu datového streamu

Při práci s IBM SPSS Modeler se využívají vizuální nástroje ovládání. Pro zjednodušení si lze představit obecný postup řešení DM úlohy jako sled tří kroků:

- načtení dat;
- datové manipulace;
- výstup.

Každý z těchto kroků je složen z jedné a více činností (procesů), které jsou reprezentovány odpovídajícími ikonami – uzly.

Každý uzel tak představuje elementární část procesu v rámci jednotlivých kroků zpracování (např. načtení dat ze souboru nebo databáze, zobrazení dat do tabulky nebo grafu, realizaci matematické operace apod.). Jednotlivé uzly se spojují do proudů – streamů, které představují posloupnost realizovaných operací při zpracování dané úlohy.

Operace s uzly

S jednotlivými uzly lze realizovat následující operace: přidání, odstranění, spojení, odpojení, editace nebo sloučení.

Přidání uzlu do streamu

Přidání uzlu do streamu lze jedním z následujících postupů:

- kliknout myší na uzel ve streamu, ke kterému bude nový uzel připojen a následně dvojklikem na připojovaném uzlu z palety je nový uzel připojen;
- vybraný uzel z palety přetáhnout myší do datového okna;
- kliknutím označit požadovaný uzel na paletě a následně kliknout do datového okna;
- kliknout na požadovaný uzel z palety a z místní nabídky (pravé tlačítko myši) zvolit ***Add To Stream***;
- volbou z nabídky ***Insert.***

Odstranění uzlu z proudu

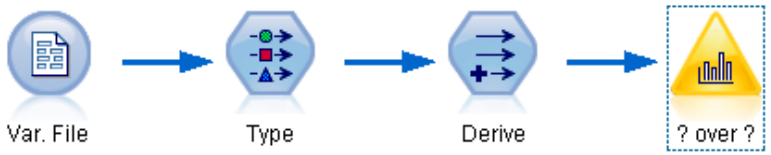
Uzel z datového okna lze odstanit:

- označením uzlu (kliknout na uzel) a stiskem klávesy ***Delete***;
- kliknout na požadovaný uzel z palety a z místní nabídky zvolit ***Delete***;
- označit požadovaný uzel a volbou z nabídky ***Edit → Delete → Selected Canvas Objects***.

Spojení uzlu v streamu

Spojování uzlů do streamů se realizuje z důvodu zabezpečení návazností jednotlivých operací při zpracování DM úlohy (Obrázek 37). Základní způsoby napojování jsou:

- kliknout na uzel, ke kterému chceme připojit další uzel a potom dvakrát kliknout na zvolený uzel z palety;
- umístit připojovaný uzel do datového okna, kliknout na uzel, ke kterému chceme připojovat a z místní nabídky zvolit ***Connect*** a kliknout na připojovaný uzel;
- umístit připojovaný uzel do datového okna, kliknout na uzel, ke kterému chceme připojovat, stlačit klávesu ***F2*** a následně kliknout na připojovaný uzel.



OBRÁZEK 37 SPOJENÍ UZLŮ DO DATOVÉHO STREAMU

POZOR: K uzlům, které jsou na záložce ***Graphs*** a ***Output*** v paletě uzlů, nelze připojovat žádné další uzly.

Odpojení uzlu v streamu

Odpojení jednotlivých uzlů ze streamu lze opět realizovat více způsoby. Například:

- označit odpojovaný uzel a z místní nabídky zvolit ***Disconnect***;
- označit odpojovaný uzel a stlačit klávesu ***F3***;
- umístit kurzor myši na vazbu, kterou chceme zrušit a z místní nabídky zvolit ***Delete Connection***;
- označit všechny odpojované uzly (označit myší oblast nebo s využitím klávesy ***CTRL***) a z místní nabídky zvolit ***Disconnect Nodes*** nebo stlačit klávesu ***F3***.

POZOR: První dvě možnosti odpojí všechny vazby daného uzlu (od předcházejících i od následujících. Třetí možnost zruší pouze vybranou vazbu.

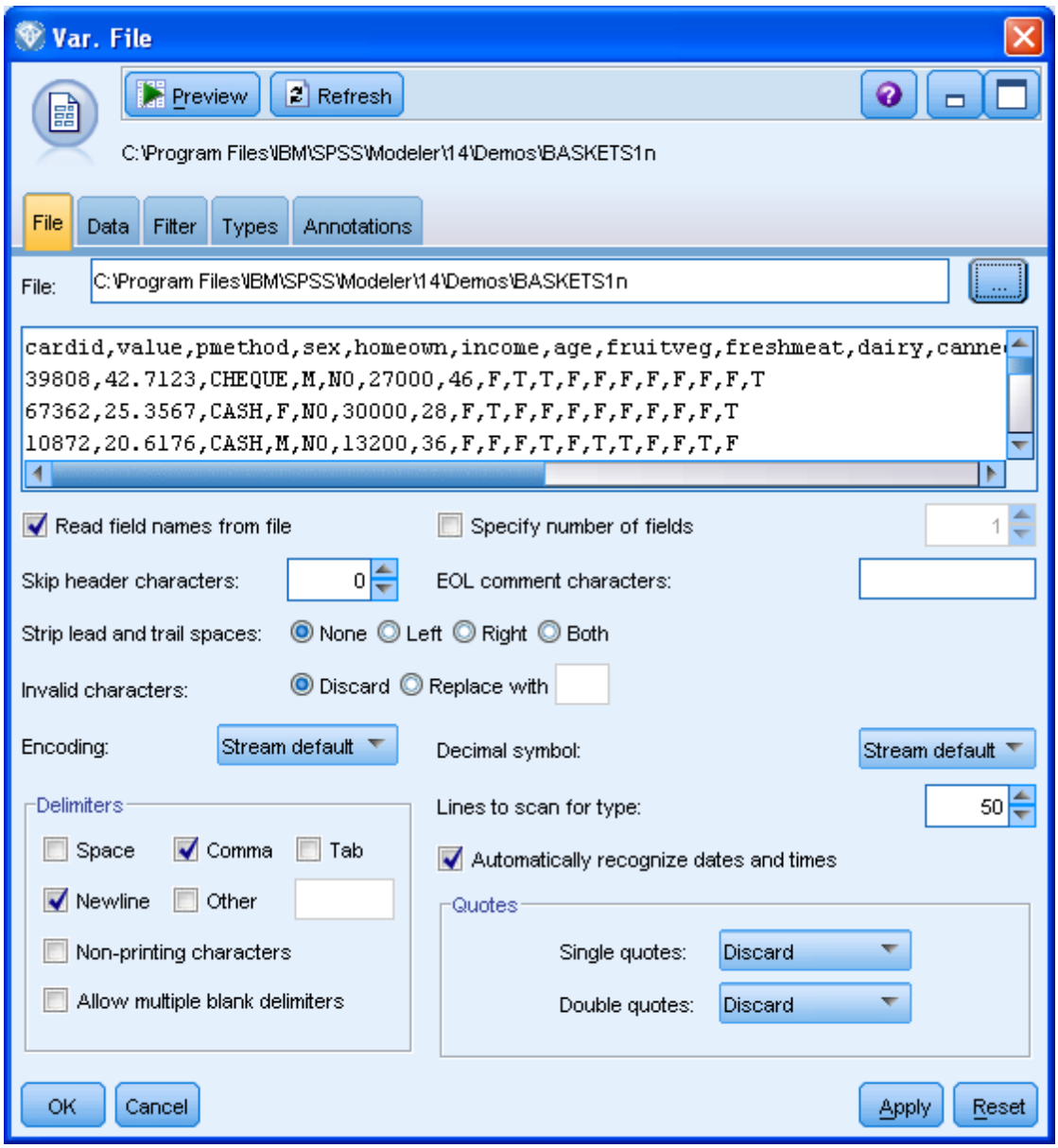
Editace uzlu

Každý uzel lze editovat. To znamená, že lze nastavovat jednotlivé parametry uvnitř jednotlivých uzlů. Například u uzlu ***Var. File*** ze skupiny ***Source***, lze na záložce ***File*** nastavit tyto parametry uzlu (Obrázek 38): cestu k datovému souboru, parametry ovlivňující import datového souboru a celou řadu dalších parametrů. Pro přehlednou práci s jednotlivými uzly a hlavně následně s jejich výstupy, je vhodné editovat i názvy uzlů a připojit k nim i další poznámky. Toto se realizuje u všech uzlů na záložce ***Annotations*** (Obrázek 39).

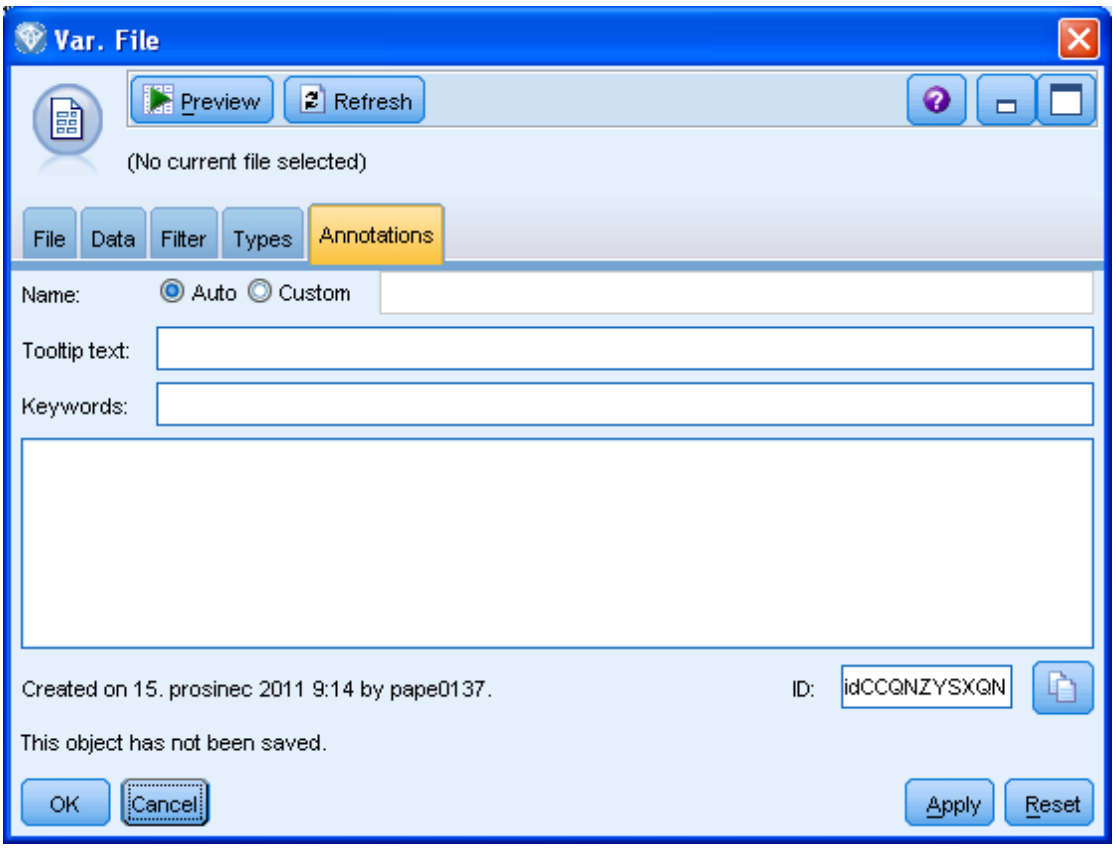
Vlastní editace uzlu se provádí následujícím postupem:

- označit myší uzel, který chceme editovat;
- z místní nabídky zvolit položku ***Edit***;
- zvolit požadovanou záložku pro editaci.

Další možností je po označení uzlu zvolit z nabídky ***Edit → Node → Edit***.



OBRÁZEK 38 OKNO PRO EDITACI UZLU VAR. FILE



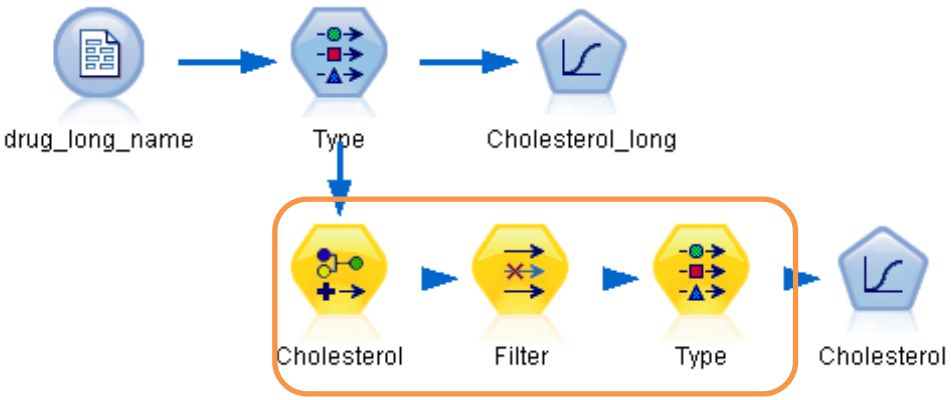
OBRÁZEK 39 OKNO ZÁLOŽKY ANNOTATIONS

Sloučení uzlu

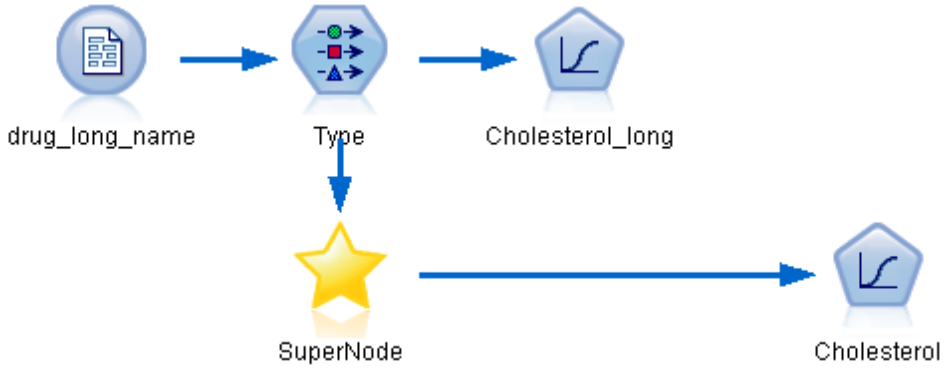
Pro přehledné uspořádání většího počtu uzlů v datovém okně (Obrázek 40) je možnost tyto uzly sloučit do jednoho tzv. Super uzlu (Super Node) (Obrázek 41). Toto je vhodné využít i při sloučení ucelených částí procesu zpracování do logicky navazujících celků. Každý super uzel lze následně opět rozložit na jednotlivé uzly. Další možností je detailní zobrazení pouze zvoleného super uzlu (Obrázek 42).

Pro vytvoření Super uzlu existují následující postupy:

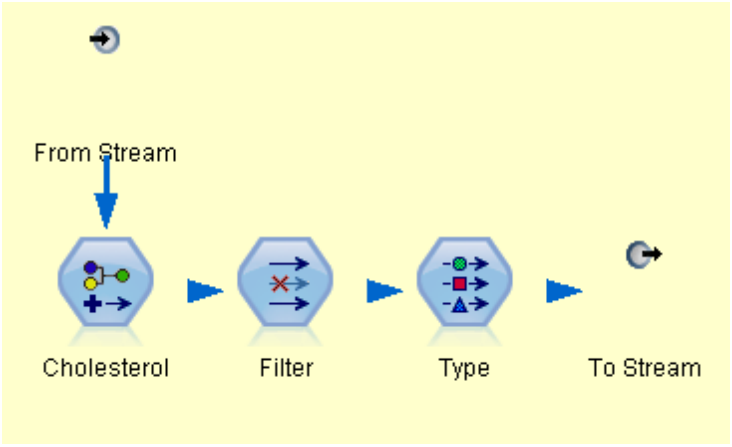
- označit všechny uzly, které mají byt sloučeny do Super uzlu a přes místní nabídku zvolit ***Create SuperNode***;
- označit všechny uzly, které mají byt sloučeny do Super uzlu a v panelu ikon zvolit následující ikonu pro vytvoření Super uzlu  ;
- označit všechny uzly, které mají byt sloučeny do Super uzlu a z nabídky zvolit: ***SuperNode → Create SuperNode → From Selection***.



OBRÁZEK 40 OZNAČENÍ UZLU PRO VYTVOŘENÍ SUPER UZLU



OBRÁZEK 41 STREAM S VYTVOŘENÝM SUPER UZLEM



OBRÁZEK 42 VNITŘNÍ STRUKTURA SUPER UZLU

Pro zrušení Super uzlu existují následující postupy:

- označit rozbalovaný Super uzel a přes místní nabídku zvolit ***Expand***;
- označit rozbalovaný Super uzel a z nabídky zvolit: ***SuperNode*** → ***Expand***.
-

Pro detailní rozbalení obsahu Super uzlu (Obrázek 42) slouží:

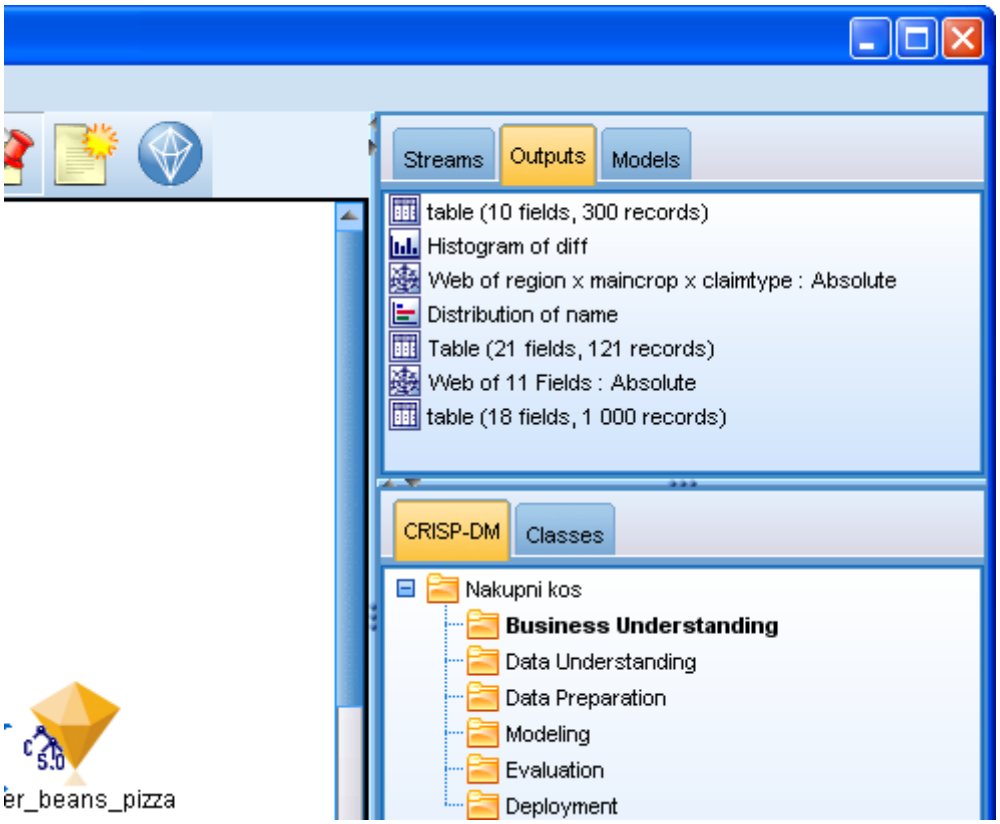
- ikona ***Zoom into SuperNode*** ;
- zvolení volby ***Zoom In*** z místní nabídky nebo z nabídky ***SuperNode***.

Pro opětovné sbalení obsahu Super uzlu do podoby (Obrázek 41) slouží:

- ikona ***Zoom out of SuperNode*** ;
- zvolení volby ***Zoom Out*** z místní nabídky nebo z nabídky ***SuperNode***.

4.2.2. Okno správce výstupů

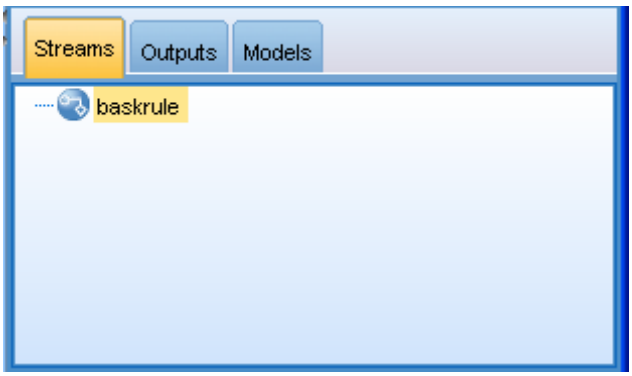
Okno správce výstupů slouží řízení a prohlížení výsledků jednotlivých streamů (Obrázek 44).



OBRÁZEK 43VÝSTUPOVÉ OKNO

Toto okno obsahuje tři záložky:

- ***Streams*** – zobrazuje seznam všech aktuálně otevřených streamů a umožňuje přepínání mezi nimi (Obrázek 44);
- ***Outputs*** – zobrazuje seznam všech výstupů generovaných jednotlivými streamy (Obrázek 45);
- ***Models*** – zobrazuje všechny vytvořené modely (Obrázek 47).

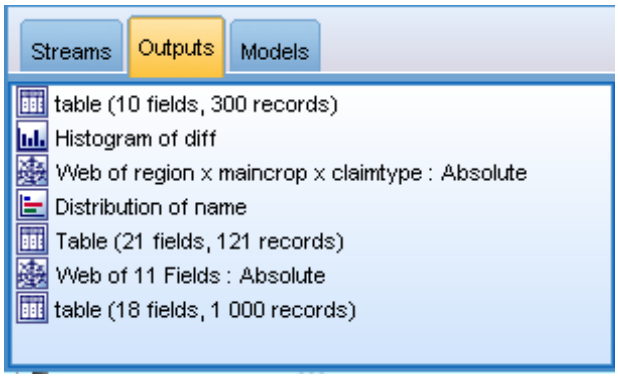


OBRÁZEK 44 OKNO ZÁLOŽKY STREAMS

Z důvodu přehlednosti jednotlivých modelů a výstupů je vhodné, vždy při generování odpovídajícího modelu nebo výstupu, využít možnost pojmenovat daný model nebo výstup v režimu editace uzlu s využitím záložky ***Annotations*** (Obrázek 39).

Záložka Outputs

Zobrazuje všechny výstupy vygenerované po spuštění jednotlivých streamů od spuštění IBM SPSS Modeler (Obrázek 45).

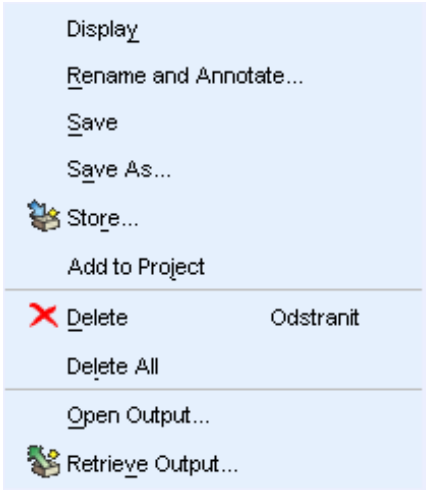


OBRÁZEK 45 OKNO ZÁLOŽKY OUTPUTS

Ze záložky ***Outputs*** (Obrázek 45) lze s jednotlivými výstupy realizovat následující operace (Obrázek 46):

- ***Display*** – zobrazení zvoleného výstupu (lze použít i kvojklik na zvolený výstup);
- ***Rename and Annotate…*** - přejmenování názvu výstupu a úprava poznámky;
- ***Save*** – uložení výstupu do souboru (přípona *.cou);
- ***Save as…*** - uložení výstupu do souboru s možností změnit název souboru;
- ***Store…*** - ukládání do úložiště na serveru;

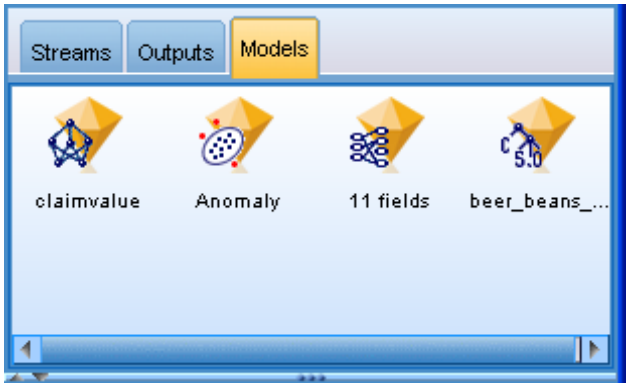
- ***Add to Project*** – doplnění souboru do projektu (zobrazí se v složce v okně správce projektu);
- ***Delete*** – vymazání výstupu (lze i použít pouze klávesu Delete);
- ***Delete All*** – vymazání všech výstupů;
- ***Open Output…*** - otevření již uloženého výstupu;
- ***Retrieve Output…*** - otevření výstupu ze serveru.



OBRÁZEK 46 OPERACE S VÝSTUPY NA ZÁLOŽCE OUTPUTS

Záložka Models

Zobrazuje všechny modely vygenerované po spuštění jednotlivých streamů od spuštění IBM SPSS Modeler (Obrázek 47).

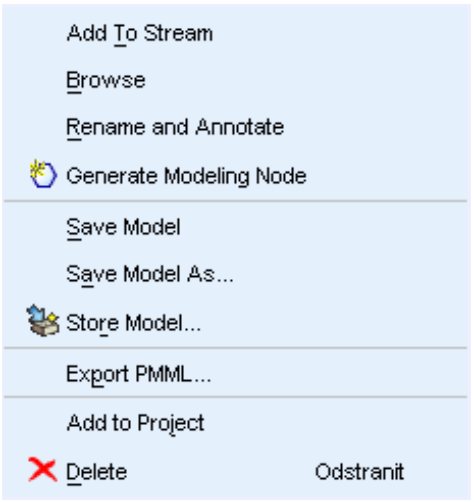


OBRÁZEK 47 OKNO ZÁLOŽKY MODELS

Ze záložky ***Models*** (Obrázek 47) lze s jednotlivými výstupy realizovat následující operace (Obrázek 48):

- ***Add to Stream*** – doplní model do aktuálního streamu (lze použít i kvojklik na zvolený model nebo jej přetáhnout myší);
- ***Browse*** – umožní prohlížet si hotový model a jeho parametry;

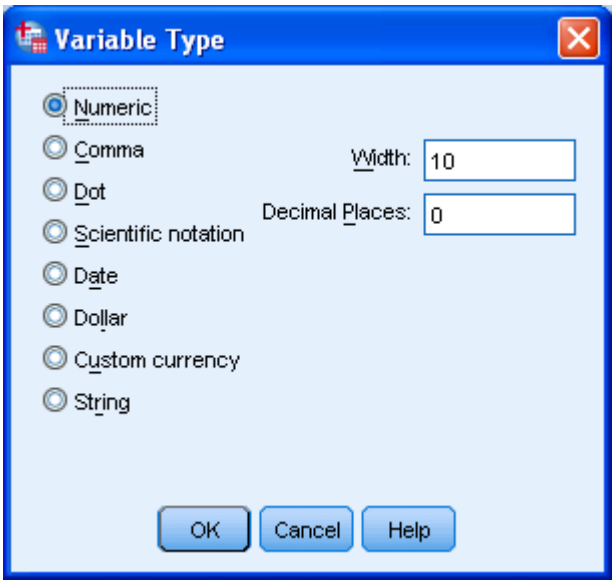
- ***Rename and Annotate…*** - přejmenování názvu modelu a úprava poznámky;
- ***Generate Modeling Node*** – vytvoření uzlu z modelu se zachováním nastavených parametrů;
- ***Save Model*** – uložení modelu do souboru (přípona *.gm);
- ***Save Model as…*** - uložení výstupu do souboru s možností změnit název souboru;
- ***Store Model…*** - ukládání do úložiště na serveru;
- ***Export PMML…*** - export modelu ve specifickém formátu (*.xml);
- ***Add to Project*** – doplnění modelu do projektu (uloží se a zobrazí se v složce v okně správce projektu);
- ***Delete*** – vymazání modelu.



OBRÁZEK 48 OPERACE S VÝSTUPY NA ZÁLOŽCE MODELS

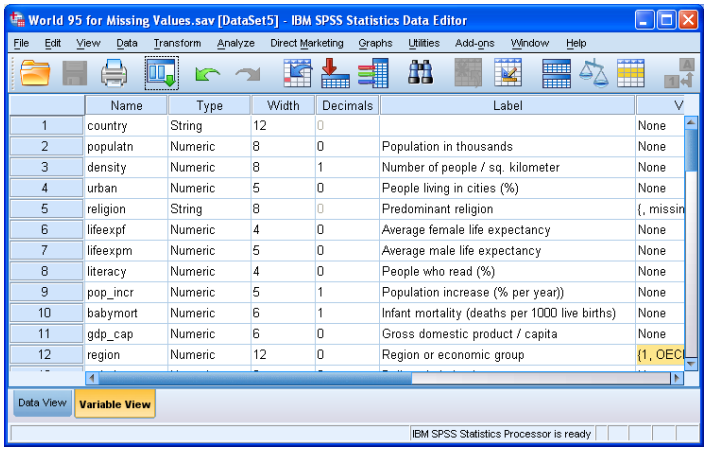
- Název proměnné (**Name**) – až 64 znaků; nesmí začínat číslicí, obsahovat mezery nebo různé speciální znaky (tečka, dvojtečka, čárka, středník apod.).
- Typ proměnné (**Type**), kde je možné volit z následujících typů (Obrázek 7):
  - číselná (**Numeric** – desetinný oddělovač čárka; **Comma** – desetinný oddělovač tečka, každé tři pozice oddělovač čárka; **Dot** - desetinný oddělovač čárka, každé tři pozice oddělovač tečka; **Scientific Notation**),
  - datum (**Date**),

- číselná obsahující měnu nebo jednotku (**Dollar, Custom currency**),
- textová (**String),**
- počet míst (**Width**),
- počet desetinných míst (**Decimal Places**),



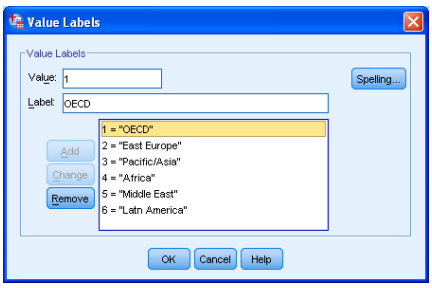
OBRÁZEK 49 VOLBA TYPU PROMĚNNÉ

- popis proměnné – až 256 znaků pro rozšířený popis proměnné oproti jejímu názvu (může se objevovat ve všech tabulkách a grafech místo názvu nebo společně s ním), (Obrázek 8),



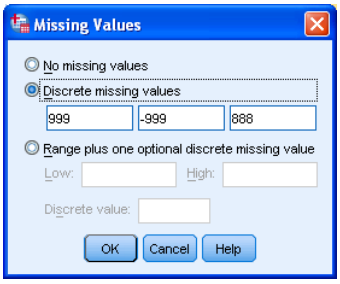
OBRÁZEK 50 POPIS PROMĚNNÝCH

- popis hodnot proměnné – v případech, kdy je vhodnější pracovat s číselnými kódy je zde uveden význam jednotlivých kódů ve formě textu (může se objevovat ve všech tabulkách a grafech místo názvu nebo společně s ním), (Obrázek 9),



OBRÁZEK 51 POPIS HODNOT PROMĚNNÉ

- definice chybějících hodnot – v tomto případě se jedná o uživatelem definované chybějící hodnoty a jsou zde možné tři možnosti (Obrázek 10):
  - bez chybějících hodnot (**No missing values**),
  - definování maximálně tří diskrétních hodnot (**Discrete missing values**) reprezentujících různé důvody vzniku chybějící hodnoty (např. kód 999, -999, 888),
  - definování rozsahu hodnot nebo jedna hodnota (**Range plus one optional discrete missing value**),



OBRÁZEK 52 DEFINICE CHYBĚJÍCÍCH HODNOT

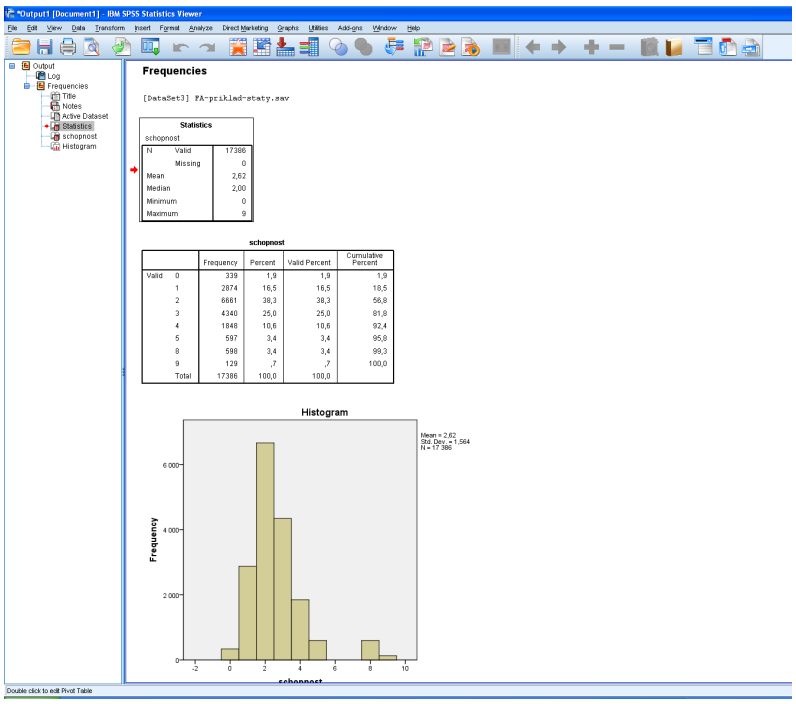
- šířka sloupce v datové matici (**Columns**) – šířku sloupce můžeme upravit přetažením myší na záložce ***Data View*** nebo nastavením tohoto parametru na záložce ***Variable View***,
- zarovnání hodnot proměnných v datové matici (**Align**) – vlevo, na střed nebo vpravo,
- způsob měření (**Measure**) – rozlišujeme, zda se jedná o proměnnou číselnou (scale), nominální (nominal) nebo ordinální (ordinal),
- role proměnné (**Role**) – některé dialogy při nastavování parametrů umožňují využít předdefinované role proměnných (proměnné, které splňují požadavky se zobrazují v cílovém seznamu); jsou zde následující možnosti:

- Vstupní (**Input**) – může být použita jako vstupní proměnná (prediktor, nezávislá proměnná),
- Cílová (**Target**) – bude použita jako výstupní nebo cílová proměnná (závislá proměnná),
- Obojí (**Both**) – může být použita jako vstupní nebo výstupní proměnná,
- Nemá roli (**None**) – nemá přidělenou žádnou roli,
- Rozdělení (**Partition**) – proměnná bude použita pro rozdělení dat do oddělených vzorků (trénovací, testovací, validační),
- Štěpení (**Split**) – zajišťuje kompatibilitu s IBM SPSS Modelerem.

Standardně je všem proměnným přiřazena vstupní role.

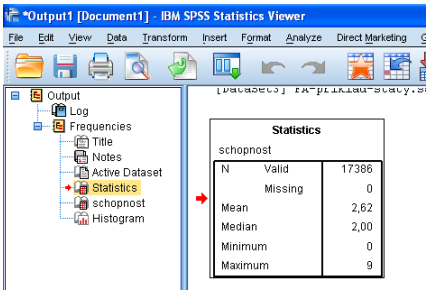
4.2.3. Výstupové okno

Do tohoto okna se zaznamenávají veškeré výstupy programu (tabulky, grafy, texty, hlášení apod.). Jednotlivé výstupy lze do určité míry dále editovat nebo graficky upravovat. V horní části okna je nástrojová lišta a panel nabídek a dolní část je rozdělena na dvě části (Obrázek 11). V levé části je ve formě stromové struktury zobrazen obsah všechny objekty v tomto okně (činnosti a výstupy v programu). Dílčí části v této struktuře lze dále skrývat/zobrazovat a editovat. To zlepšuje přehlednost jednotlivých výstupů a práci s nimi. V pravé části se potom nacházejí jednotlivé objekty z obsahu.



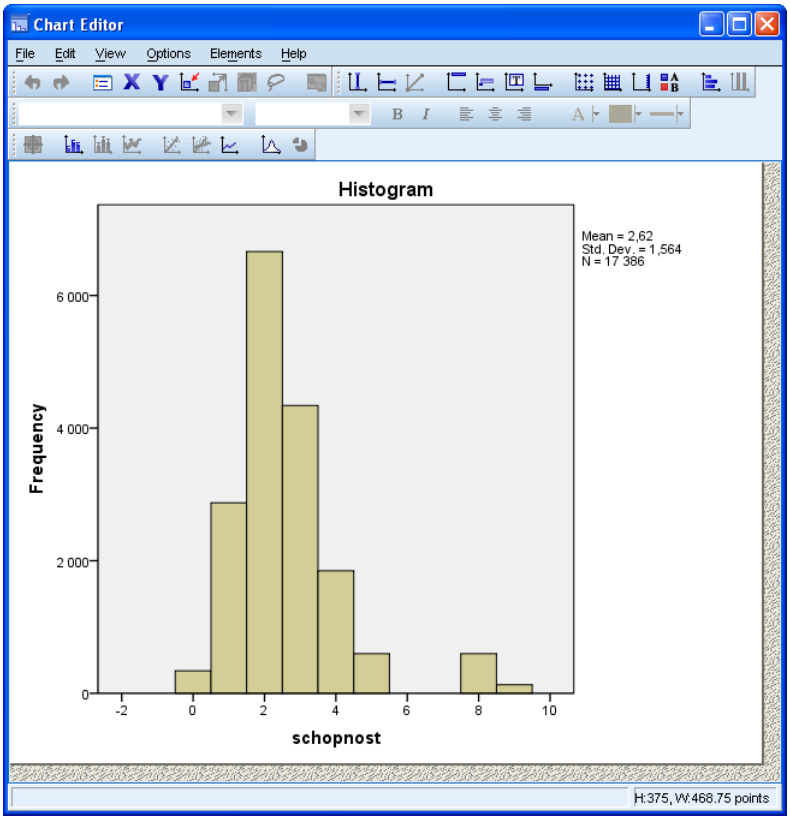
OBRÁZEK 53 STRUKTURA VÝSTUPOVÉHO OKNA

Jednotlivé položky lze upravovat tak, že na ně dvakrát poklepeme myší. Tímto způsobem lze pomocí myši nebo s využitím nabídek a ikon měnit uspořádání položek nebo jejich hierarchii v stromové struktuře výstupového okna. Signalizace editovatelné položky je pomocí červené šipky (Obrázek 12).



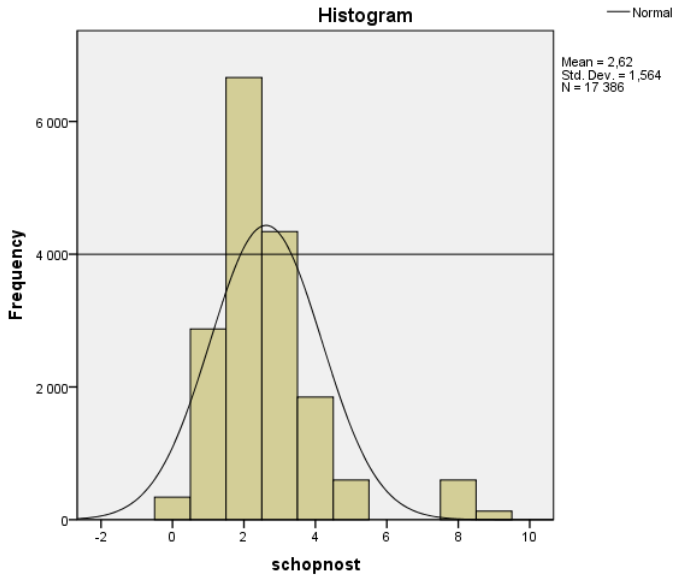
OBRÁZEK 54 OZNAČENÍ EDITOVATELNÉ POLOŽKY

Stejným způsobem lze editovat i grafy (Obrázek 13). Nástrojová lišta i odpovídající menu umožňují editovat veškeré prvky grafu (formát, měřítko a orientaci jednotlivých os, formát a zobrazení pracovní oblasti grafu, kopírovat jednotlivé části grafu, přidávat popisky do grafu, vkládat různé typy např. distribučních křivek, doplňovat referenční úrovně apod.). Všechny prvky jsou zde definovány jako objekty a následně jim lze přidělovat požadované vlastnosti.



OBRÁZEK 55 EDITACE VE VÝSTUPOVÉM OKNĚ

Na následujícím příkladu (Obrázek 14) je ukázáno jedno z možných editování výstupního grafu. Je zde doplněna distribuční křivka a referenční hodnota na úrovni 4000.

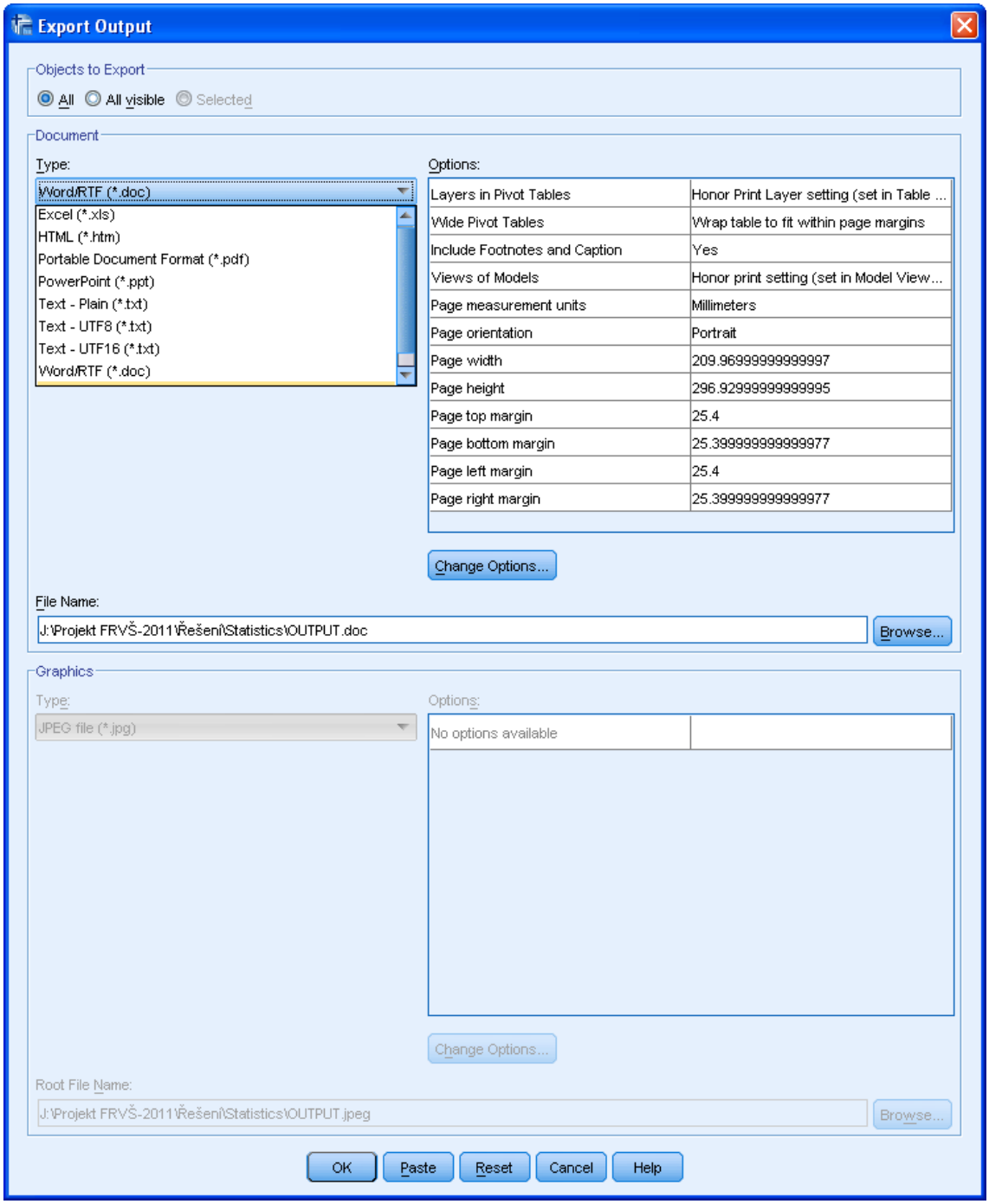


OBRÁZEK 56 PŘÍKLAD EDITACE VÝSTUPNÍHO GRAFU

Všechny výstupy z výstupového okna můžeme ukládat. Soubor výstupového okna má příponu *.spv. Způsob ukládání je buď nabídkou Uložit (**Save**) nebo Uložit

jako (**Save as**). Soubor z výstupového okna můžeme později kdykoliv v IBM SPSS Statistics otevřít a dále s ním pracovat (včetně jeho editace).

Pro práci s výstupy v jiném prostředí než je IBM SPSS Statistics, je třeba převést výstupy do jiného formátu přes nabídku File – Export (Obrázek 15).



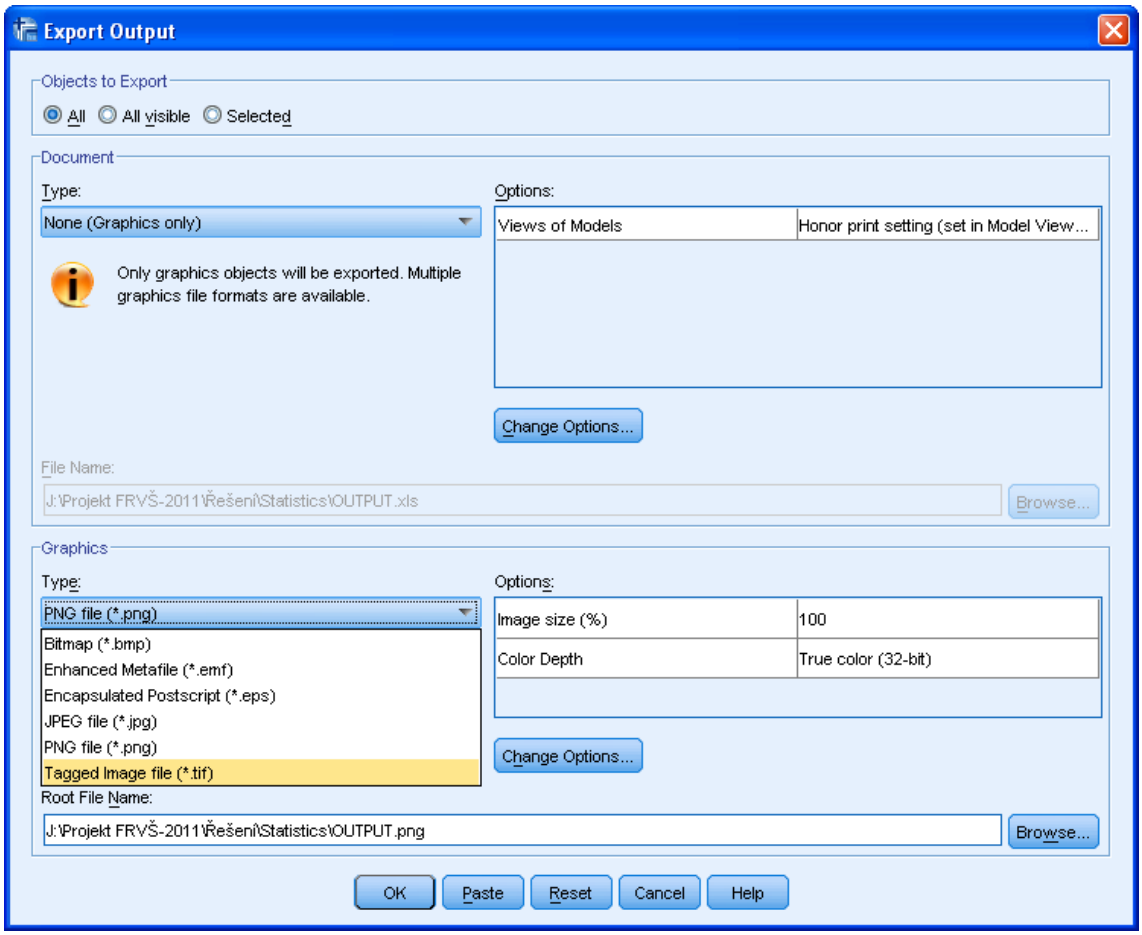
OBRÁZEK 57 DIALOGOVÉ OKNO PRO EXPORT Z VÝSTUPOVÉHO OKNA – MIMO FORMÁT GRAFIKA

V tomto dialogu v části Object to Export volíme, zda chceme exportovat všechny objekty výstupového okna (**All**), zobrazené objekty (**All Visible**) nebo jen

vybrané objekty (**Selected**). Následně v části Document lze vybrat z jednotlivých nabízených formátů pro export (Obrázek 15).

V části Options lze upřesnit, co a jakým způsobem bude exportováno u tabulek. Změna nastavení je možná přes volbu Change Options. V části File Name zadáváme cestu a název ukládaného exportovaného souboru.

V případě, že při exportu zvolíme volbu ukládání v grafickém formátu, zpřístupní se volba pro výběr formátu grafického souboru (Obrázek 16). V tomto případě jsou z výstupového okna exportovány pouze grafické objekty (grafy). Další doplňující k nastavení výstupu lze opět ovlivnit pomocí části Options.



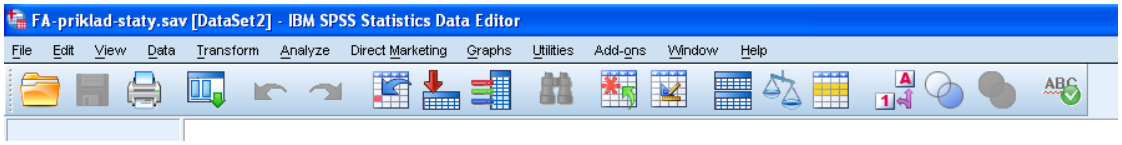
OBRÁZEK 58 DIALOGOVÉ OKNO PRO EXPORT Z VÝSTUPOVÉHO OKNA V GRAFICKÉM FORMÁTU

Tímto končí část potřebná k seznámení se základní obsluhou výstupového okna programu IBM SPSS Statistics.

4.3. Základní ovládání programu IBM SPSS Statistics

Při ovládání programu IBM SPSS Statistics máme k dispozici panel nástrojů v podobě ikon a panel nabídek (Obrázek 17). Kombinací těchto dvou panelů se lze dostat

k většině standardních možností tohoto programu. Panel nástrojů je určen pro rychlý přístup k dané funkci bez nutnosti procházet jednotlivé nabídky. Je samozřejmé, že panel nástrojů lze dále editovat a tím přizpůsobovat potřebám jednotlivých uživatelů s cílem zrychlení práce.



OBRÁZEK 59 ZÁKLADNÍ PANEL V IBM SPSS STATISTICS

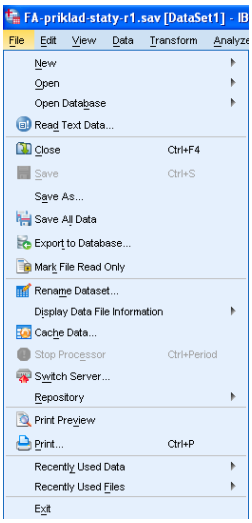
V následující části bude uveden přehled jednotlivých položek z panelu nabídek.

4.3.1. Soubor (File)

Tato nabídka je určena k práci se souborem. Jsou zde k dispozici následující možnosti (Obrázek 18):

- vytvořit nový soubor (**New**) - datový, syntaxový, výstupový, skriptový,
- otevřít existující soubor (**Open**) - datový, syntaxový, výstupový, skriptový,
- načíst data z databáze (**Open Database**) - vytvořit nový dotaz, editovat hotový dotaz, spustit dotaz,
- načíst data z textového souboru (**Read Text Data**),
- zavřít okno (**Close**),
- uložit (**Save**) – uložení pod dříve zadaným jménem nebo zadání jména souboru,
- uložit jako **(Save as**) – uložení souboru pod jiným jménem,
- uložení všech datových souborů (**Save All Data**) – uloží všechny otevřené datové soubory,
- export do databáze (**Export to Database…)** – přidání dat do databáze,
- označení souboru pouze pro čtení/pro psaní a čtení (**Mark File Read Only/Mark File Read Write**) – jedná se o přepínač pro určování vlastností souboru z hlediska možnosti jeho změny,
- přejmenování datového souboru (**Rename Dataset…**) – v případě, že je otevřeno více verzí stejného datového souboru jsou odlišeny názvem v hranaté závorce (např. [DataSet1]); údaj v této závorce lze v této volbě změnit,

- zobrazení informací o souboru (**Display Data File Information**) – doplnění informací o datovém souboru do výstupového okna,
- pracovní záloha souboru (**Cache Data**) – vytvoří pracovní zálohu datového souboru a při následujích operacích s ním pracuje; umožňuje to urychlit práci,
- zastavení výpočtu (**Stop Processor**) – pro zastavení výpočtu nebo spuštěné procedury,
- připojení k serveru (**Switch Server**) – při zpracování většího objemu dat může být výhodné se připojit k aplikaci na výkonnějším počítači,
- datové úložiště (**Repository**) – připojení k datovému úložišti,
- náhled před tiskem (**Print Preview**) – zobrazení v tiskovém formátu,
- tisk (**Print**) – tisk zvolených objektů z aktuálního okna,
- naposledy používané datové soubory (**Recently Used Data**) – zobrazení maximálně devíti naposledy používaných datových souborů,
- naposledy používané soubory (**Recently Used Files**) – zobrazí naposledy používané soubory kromě datových (výstupy, syntax apod.),
- ukončení programu IBM SPSS Statistics (**Exit**).



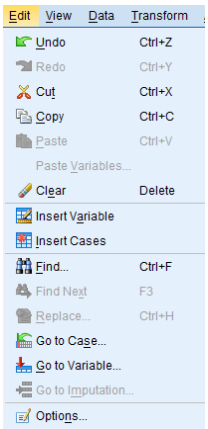
OBRÁZEK 60 NABÍDKA SOUBOR

4.3.2. Úpravy (Edit)

Tato nabídka dovoluje kromě práce se schránkou v systému Windows také vracet jednotlivé kroky úprav, editovat, hledat a pohybovat se v datovém okně. Jsou zde následující možnosti (Obrázek 19):

- zpět (**Undo**) – krok zpět při úpravě datové matice,

- vpřed (**Redo**) – krok dopředu při úpravě datové matice,
- vyjmout (**Cut**) – přesune vybrané položky do schránky,
- kopírovat (**Copy**) – kopíruje vybrané položky do schránky,
- vložit (**Paste**) – vloží položku ze schránky na vybrané místo,
- vložit proměnnou (**Paste Variable**) – na základě upřesnění v dialogovém okně vloží požadovaný počet proměnných (na základě dat ve schránce) se zadanými jmény (je aktivní pouze na záložce *Variable View*),
- smazat (**Clear**) – smaže vybrané údaje,
- vložit novou proměnnou (**Insert Variable**) – vloží novou prázdnou proměnnou před místo označené kurzorem,
- vložit nové případy (**Insert Cases**) – vloží nové prázdné řádky do datové matice (jejich počet odpovídá počtu označených řádků),
- najít (**Find**) – hledá požadovaný řetězec (lze upřesnit možnosti prohledávání),
- najít další (**Find Next**) – najde další záznam podle údajů z položky Find,
- nahradit (**Replace**) – nahradí zadaný řetězec novým řetězcem,
- přejít na případ (**Go to Case…**) – přechod na zadané číslo případu,
- přejít na proměnnou (**Go to Variable**) – přechod na zadanou proměnnou,
- nastavení (**Options…**) – nastavení prostředí IBM SPSS Statistics pro potřeby uživatele.

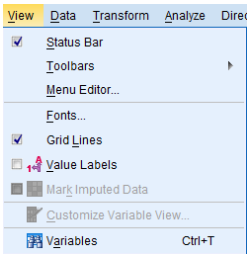


OBRÁZEK 61 NABÍDKA EDIT

4.3.3. Pohled (View)

Tato nabídka definuje zobrazení/skrytí a vzhled prvků datového okna. Jsou zde následující možnosti (Obrázek 20):

- stavový řádek (**Status Bar**) – zobrazení nebo skrytí stavového řádku v IBM SPSS Statistics (jsou zde zobrazovány doplňující informace),
- panely nástrojů (**Toolbars**) – zobrazení/skrytí nebo editace panelu nástrojů v IBM SPSS Statistics,
- úprava nabídek (**Menu Editor…**) – pomocí této volby lze měnit obsah jednotlivých nabídek v oknech IBM SPSS Statistics, lze ji i doplňovat vlastními nabídkami,
- písmo (**Fonts…**) – nastavení písma pro datovou matici,
- zobrazení mřížky (**Grid Lines**) – zapnutí/vypnutí zobrazení mřížky v datové matici,
- popisy hodnot (**Value Labels**) – přepínač pro zobrazení popisu hodnot v datové matici (pokud jsou tyto hodnoty popsány v záložce *Variable View*),
- zobrazování vlastností proměnných (**Customize Variable View…**) – řídí zobrazování vlastností proměnných na záložce *Variable View*, lze si vybírat, které proměnné budou zobrazeny a které skryty,
- přepínámi mezi záložkami datového okna (**Variables/Data**).



OBRÁZEK 62 NABÍDKA VIEW

Toto byl základní volby z nabídky sloužící k ovládání programu IBM SPSS Statistics.

Další nabídky jako jsou Data, Transform, Analyze, Direct Marketing, Graphs, Utilities, Add-ons slouží již k realizaci vlastních analýz a dalších výpočetních operací v tomto programu. Popis těchto nabídek bude součástí další části tohoto textu.

Zbývající volby Windows a Help mají již standardní funkce související se zobrazením jednotlivých oken a využíváním nápovědy.

5. Seznam použité literatury

[1] *CRoss Industry Standard Process* [online]. [cit. 2004-06-25]. <http://www.crisp-dm.org>.

[2] SPSS for Windows [online]. [cit. 2011-04-015]. <http://psych.utoronto.ca/courses/c1/spss/page1.htm>.

[3] Resources to help you learn and use SPSS [online]. [cit. 2011-04-015]. <http://www.ats.ucla.edu/stat/spss/default.htm>.

[4] SPSS for Windows: Getting Started [online]. [cit. 2011-04-015]. <http://ssc.utexas.edu/images/stories/ssc/files/tutorials/SPSS_GettingStarted_Tutorial.pdf>.

6. Seznam obrázků

Název	Stručný návod k ovládání IBM SPSS Statistics a IBM SPSS Modeler
Autor	doc. Ing. Pavel Petr, Ph.D.
Vydavatel	Univerzita Pardubice Fakulta ekonomicko-správní Studentská 84, 532 10 Pardubice
Vydáno	březen 2012
Stran	65
Vydání	první

ISBN 978-80-7395-477-2 (online)