Analýza dat ve vybraných statistických softwarech

Tereza Mokrenová

Seminární práce
4ST310

Abstract. Práce je zaměřena na porovnání tří statistických softwarů. Postupně budou porovnány programy Excel, SPSS a Statgraphics při různých typech úkolů. První z porovnávaných softwarů vyniká svou přístupností a do jisté míry i uživatelskou přívětivostí, avšak velkou nevýhodou Excelu jsou časté chyby vnizklé při překladu do češtiny. SPSS je intuitivní, výstupy se dají snadno editovat do žádané podoby. Statgraphics však přidává k výstupům i jednoduchou interpretaci výsledků, čímž se stává velmi vhodným nástrojem i pro statistické začátečníky.

Keywords: Excel, SPSS, Statgraphics
JEL Classification: C-10, C-40, C-87

1 Úvod

Pro analýzu dat je použit soubor Cars, který obsahuje záznamy o vozidlech inzerovaných k prodeji na různých serverech. Ukázka dat je k dispozici na obrázku 1. Soubor obsahuje jak numerické, tak i kategoriální proměnné a celkem 358 pozorování. Některé sloupce obsahují chybějící pozorování, a tak se počty pozorování v některých výstupech mohou lišit v závislosti na tom, zda daný software sám vynechá chybějící pozorování a nebo je potřeba je předem ze souboru vyloučit. Data byla získána při úvodním kurzu ekonometrie od pana docenta Zouhara. [2]

Druhým zpracovávaným souborem jsou data o cenách akcií firmy Apple pozorované za 504 měsíčních snapshotů. Ukázka dat je k dispozici na obrázku 2. Zobrazená úkázka demonstruje zároveň problém s proměnnou Date, kdy bylo při načtení dat ze souboru csv následně nutné opravit formát před dalším zpracováním. Data byla získána ze stránky Kaggle [1] vyfiltrováním vybraných období pro zvolenou firmu.

Práce je rozdělena do logických celků podle typu prováděné analýzy. Prvním typem úkolu v semestrální práci jsou popisné statistiky doplněné vhodnými grafy. Následně jsou v jednotlivých softwarech provedeny různé typy analýzy závislostí v datech. Nejprve jsou sestaveny kontingenční tabulky nad nimiž je následně provedena analýza závislosti. Dále je na datech provedena analýza rozptylu (ANOVA) a regresní analýza. Poslední disciplínou pro porovnání statistických prostředí je práce s časovou řadou. Výstupy ze softwarů nebyly nijak editovány.

model	engine volume	price (CZK)	kilometres	transmission	fuel	year	contact	Stáří
Škoda Octavia Combi	0	285000,00	-	-	-	2000	SMS INZEF	Nové
Škoda Octavia Combi	0	330000,00	-	-	-	2000	SMS INZEF	Nové
Škoda Octavia Combi	0	330000,00	-	-	-	2000	SMS INZEF	Nové
Škoda Octavia Combi	0	330000,00	-	-	-	2000	SMS INZEF	Nové
Škoda Octavia Combi	4	289000,00	106000,00	5	diesel	2000	AUTOPRO	Nové
Škoda Octavia Combi	4	349000,00	89000,00	A	diesel	2001	Autodíly K	Nové
Škoda Octavia Combi	4	298000,00	77000,00	5	diesel	1999	Autoland	Nové
Škoda Octavia Combi	2	224000,00	106200,00	5	petrol	1999	AUTOBAZ	Nové
Škoda Octavia Combi	3	271000,00	179000,00	5	petrol	1999	Mechanik	Nové
Škoda Octavia Combi	3	245000,00	179000,00	5	petrol	2002	AUTO - KŘ	Nové
Škoda Octavia Combi	3	455000,00	73622,00	5	petrol	2000	Auto Jarov	Nové
Škoda Octavia Combi	3	597000,00	32736,00	5	petrol	2002	Auto Jarov	Nové
Škoda Octavia Combi	3	289000,00	122000,00	-	petrol	1999	Přerost a S	Nové
Škoda Octavia Combi	4	339000,00	99000,00	5	diesel	2000	Autobazar	Nové
Škoda Octavia Combi	4	439000,00	79832,00	5	diesel	2001	Auto Jarov	Nové
Škoda Octavia Combi	4	639000,00	9174,00	6	diesel	2003	Auto Jarov	Nové
Škoda Octavia Combi	4	549000,00	19153,00	6	diesel	2002	Auto Jarov	Nové
Škoda Octavia Combi	4	579000,00	9368,00	5	diesel	2002	Auto Jarov	Nové
Škoda Octavia Combi	4	539000,00	19081,00	5	diesel	2003	Auto Jarov	Nové
Škoda Octavia Combi	4	579000,00	17380,00	5	diesel	2003	Auto Jarov	Nové
Škoda Octavia Combi	4	880000,00	43330,00	5	diesel	2000	Auto Jarov	Nové
Škoda Octavia Combi	4	315000,00	178077,00	-	diesel	2000	Přerost a S	Nové
Škoda Octavia Combi	4	649000,00	26937,00	5	diesel	2003	Auto Jarov	Nové
Škoda Octavia Combi	2	349000,00	20000,00	-	petrol	2002	Autobazar	Nové
Škoda Octavia Combi	2	289000,00	76600,00	-	petrol	1998	Autotrenc	Staré
Škoda Octavia Combi	2	239000,00	102000,00	-	petrol	1998	Autobazar	Staré

Fig. 1. Ukázka dat ze souboru Cars

							řetězový index	bazický index
Date	AAPL	absolutní přírustek	průměrný abs přírustek	index (koeficient rustu)	průměrný koef. Růstu	relativní přírustek	it/t-1	it/t1
12.01.2015	117,34		0,106779324		1,000749546	0,07%		1
12.02.2015	116,28	-1,06		0,990966422		-0,90%	0,990966422	0,990966422
12.03.2015	115,2	-1,08		0,990712074		-0,93%	0,990712074	0,9817624
12.04.2015	119,03	3,83		1,033246528		3,32%	1,033246528	1,014402591
12.07.2015	118,28	-0,75		0,993699067		-0,63%	0,993699067	1,008010908
12.08.2015	118,23	-0,05		0,999577274		-0,04%	0,999577274	1,007584796
12.09.2015	115,62	-2,61		0,977924385		-2,21%	0,977924385	0,985341742
12.10.2015	116,17	0,55		1,004756962		0,48%	1,004756962	0,990028976
12.11.2015	113,18	-2,99		0,974261858		-2,57%	0,974261858	0,964547469
12/14/2015	112,48	-0,7		0,993815162		-0,62%	0,993815162	0,958581899
12/15/2015	110,49	-1,99		0,982307966		-1,77%	0,982307966	0,941622635
12/16/2015	111,34	0,85		1,007693004		0,77%	1,007693004	0,948866542
12/17/2015	108,98	-2,36		0,978803664		-2,12%	0,978803664	0,928754048
12/18/2015	106,03	-2,95		0,972930813		-2,71%	0,972930813	0,903613431
12/21/2015	107,33	1,3		1,012260681		1,23%	1,012260681	0,914692347
12/22/2015	107,23	-0,1		0,999068294		-0,09%	0,999068294	0,913840123
12/23/2015	108,61	1,38		1,012869533		1,29%	1,012869533	0,925600818
12/24/2015	108,03	-0,58		0,994659792		-0,53%	0,994659792	0,920657917
12/28/2015	106,82	-1,21		0,988799408		-1,12%	0,988799408	0,910346003
12/29/2015	108,74	1,92		1,017974162		1,80%	1,017974162	0,92670871
12/30/2015	107,32	-1,42		0,986941328		-1,31%	0,986941328	0,914607125
12/31/2015	105,26	-2,06		0,980805069		-1,92%	0,980805069	0,897051304
01.04.2016	105,35	0,09		1,000855026		0,09%	1,000855026	0,897818306
01.05.2016	102,71	-2,64		0,974940674		-2,51%	0,974940674	0,875319584
01.06.2016	100,7	-2,01		0,980430338		-1,96%	0,980430338	0,858189876

Fig. 2. Ukázka dat ze souboru Apple

2 Popisné statistiky

Sekce obsahuje ukázku zpracování popisných statistik v jednotlivých statistických prostředí, jejich interpretaci a komentář.

2.1 Tabulky

Na první pohled je nejpřehlednější výstup ze Statgraphics 5. Tabulky nesou stejné informace o vybraných proměnných. V případě softwaru Excel 3 a Statgraphic 5 popisné statistiky pro proměnnou *engine volume* a v softwaru SPSS 4 pro proměnné *price(CZK)* a *kilometres*. Díky jednotlivým charakteristikám je možné usuzovat o poměrně vysoké kvalitě dat v souboru. Nepozorujeme neplatné hodnoty typu extrémně vysokých nebo naopak nízkých hodnot. Pro proměnnou *engine volume* je minimum rovno 0, maximum rovno 7, medián roven 2 a modus, tedy nejčastěji se vyskytující hodnota v souboru, rovný 1. Jde o diskrétní proměnnou.

V prostředí SPSS je sestavena i frekvenční tabulka pro nově vytvořenou proměnnou *price_inter*, která je vytvořena transformací proměnné *price(CZK)* do 4 intervalů. Z tabulky 6 je možné přečíst jak absolutní, tak i relativní zastoupení jednotlivých intervalů v souboru. Nejvíce zastoupenou kategorií jsou levná vozidla, která představují téměš 53 % souboru.

engine volume	
Stř. hodnota	2,114525
Chyba stř. hodnoty	0,082045
Medián	2
Modus	1
Směr. odchylka	1,552358
Rozptyl výběru	2,409816
Špičatost	-0,32032
Šikmost	0,589502
Rozdíl max-min	7
Minimum	0
Maximum	7
Součet	757
Počet	358

Fig. 3. Popisné statistiky - Excel

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
price(CZK)	354	929000	66000	995000	241424,82	9997,607	188103,855	3,538E10	1,803	,130	2,913	,259
kilometres	328	229890	110	230000	92234,41	2697,331	48850,749	2,386E9	,099	,135	-,449	,268
Valid N (listwise)	328											

Fig. 4. Popisné statistiky - SPSS

Summary Statistics for engine volume	
Count	354
Average	2,13842
Standard deviation	1,54463
Coeff. of variation	72,2324%
Minimum	0
Maximum	7,0
Range	7,0
Stnd. skewness	4,51479
Stnd. kurtosis	-1,20205

Fig. 5. Popisné statistiky - Statgraphics

Statistics

price_inter

N	Valid	354
	Missing	5

price_inter

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	nejlevnější	57	15,9	16,1	16,1
	levné	190	52,9	53,7	69,8
	středně drahé	67	18,7	18,9	88,7
	drahé	40	11,1	11,3	100,0
	Total	354	98,6	100,0	
Missing	System	5	1,4		
Total		359	100,0		

Fig. 6. Popisné statistiky - Statgraphics

2.2 Grafy

V každém programu je vytvořen frekvenční graf pro proměnnou *engine volume*. Grafy 7, 8 a 9 jsou vzhledově srovnatelné. U všech tří softwarů je uživatel limitován možnostmi dodatečného upravení, ve velké výhodě jsou tak programy využívající programovací jazyky typu R nebo Python. Žádný z grafů není interaktivní, zde jsou vhodnější softwary pro vizualizaci Tableau nebo PowerBI. Boxploty v programech Excel 10 a Statgraphics 11 se liší orientací. Oba nesou stejnou informaci, opět jako u popisných statistik.

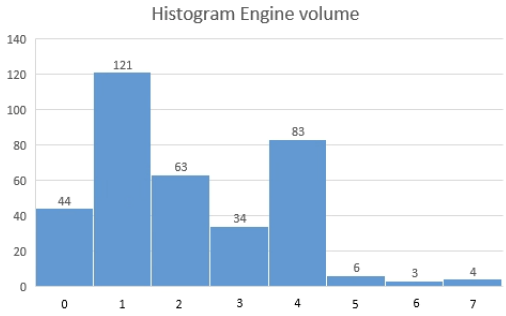


Fig. 7. Frekvenční graf pro proměnnou *engine volume* - Excel

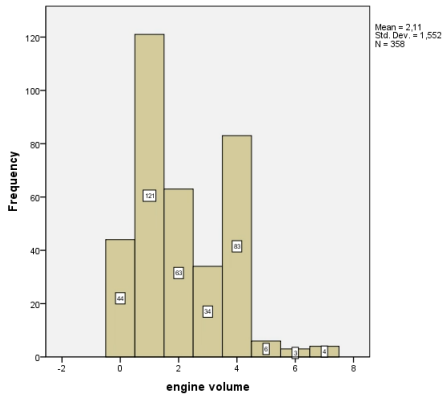


Fig. 8. Frekvenční graf pro proměnnou *engine volume* - SPSS

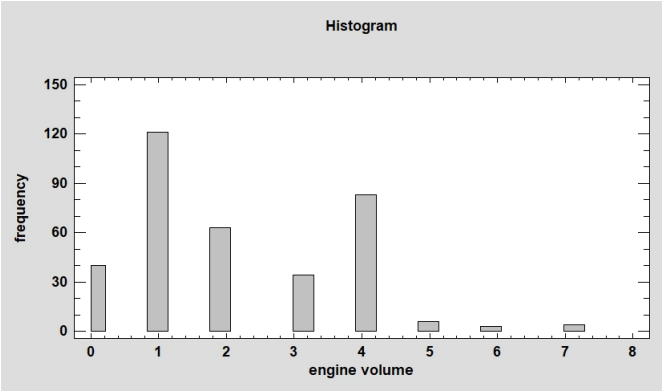


Fig. 9. Frekvenční graf pro proměnnou *engine volume* - Statgraphics

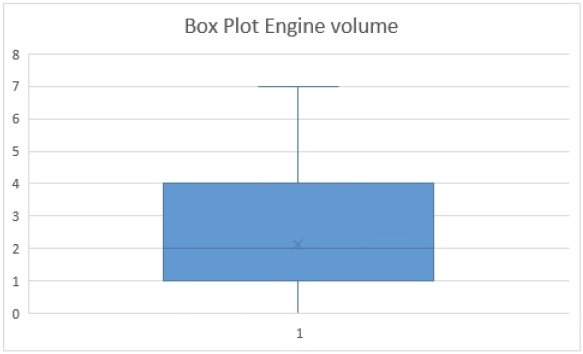


Fig. 10. Boxplot pro proměnnou *engine volume* - Excel

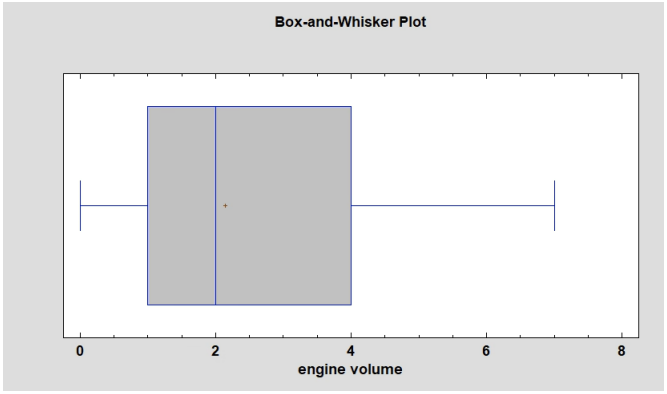


Fig. 11. Boxplot pro proměnnou *engine volume* - Statgraphics

3 Analýza závislosti

Sekce obsahuje analýzu závislosti pro různé typy proměnných. Pro dvě kategoriální proměnné jde o kontingenční tabulku. Pro jednu kategoriální proměnnou a jednu numerickou jde o analýzu rozptylu (ANOVA). A nakonec pro dvě kategoriální proměnné je provedena regresní analýza. Analýza rozptylu a regresní analýza předpokládá normálně rozdělené proměnné. V práci je přijat tento předpoklad jako asymptoticky splněný vzhledem k počtu pozorování, ačkoliv pozdější testy ukázaly, že tomu tak nemusí být. Vhodným nástrojem by byla logaritmická transformace proměnných, avšak s ohledem na zjednodušené modely typu jednoduchá regrese je kladen větší důraz na interpretaci výsledků, než jejich reálnou informační hodnotu. Zobrazené výsledky jsou tedy pro původní proměnné bez transformace. Dalšími předpoklady pro ANOVU jsou nezávislost pozorovaných dat, homoskedasticita a stejná velikost vzorků. Homoskedasticita není testována a taktéž v kontextu nevyrovnanosti zastoupení jednotlivých skupin ve vzorku je

třeba k vytváření závěrů o souboru dat po provedené analýze rozptylu přistupovat opatrně. Předpoklady pro regresní analýzu jsou: linearita vztahů mezi proměnnými, nulová střední hodnota náhodné složky, homoskedasticita, konstantní rozptyl náhodné složky, nulová kovariance náhodné složky a její normalita. Předpoklady nejsou testovány, výsledky odhadů i testů tedy mohou být zkreslené a nekonzistentní.

3.1 Kontingenční tabulky

Kontingenční tabulky jsou sestrojeny pro proměnné *Stáří* a *Model* v programech Excel 12, SPSS 13 a Statgraphics 14. V programu Excel se uživatel při hledání kontingenční tabulky rychle zorientuje. Zbylé dva programy zprvu pro práci s kontingenčními tabulkami nejsou příliš intuitivní, avšak zobrazují zároveň s výslednou tabulkou i výsledky jednotlivých testů závislosti, viz 13, 15 a 16, kterých je poměrně široký výběr. Pro porovnání je dopočteno Pearsnovo V a Cramerovo C taktéž v programu Excel. Z výsledků usuzujeme o středně silné závislosti mezi proměnnou *Stáří* a *Model*, výsledky jsou vcelku konzistentní napříč programy.

Skutečné četnosti

	Nové	Staré	Celkový součet
Škoda Superb	20		20
Škoda Felicia Combi		42	42
Škoda Felicia	14	133	147
Škoda Octavia Combi	28	10	38
Škoda Octavia	47	64	111
	109	**249**	**358**

Teoretické četnosti

	Nové	Staré	Celkový součet
Škoda Superb	6,08938547	13,91061453	20
Škoda Felicia Combi	12,7877095	29,2122905	42
Škoda Felicia	44,7569832	102,2430168	147
Škoda Octavia Combi	11,5698324	26,4301676	38
Škoda Octavia	33,7960894	77,20391061	111
	109	**249**	**358**

Porovnání

	Nové	Staré	Celkový součet
Škoda Superb	31,7774589	13,91061453	45,68807339
Škoda Felicia Combi	12,7877095	5,597832671	18,38554217
Škoda Felicia	21,1361881	9,252387576	30,38857571
Škoda Octavia Combi	23,332266	10,21372287	33,54598888
Škoda Octavia	5,15868134	2,258217934	7,416899269
	94,1923038	**41,23277558**	**135,4250794**

Pearson	0,52388859
Cramerův	0,61504662

Fig. 12. Kontingenční tabulka pro proměnné *Stáří* a *Model* - Excel

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
model * Stáří	354	100,0%	0	,0%	354	100,0%

model * Stáří Crosstabulation

Count

		Stáří		Total
		Nové	Staré	
model	Škoda Superb	20	0	20
	Škoda Felicia	34	111	145
	Škoda Felicia Combi	5	36	41
	Škoda Octavia	68	43	111
	Škoda Octavia Combi	33	4	37
Total		160	194	354

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	110,442[a]	4	,000
Likelihood Ratio	125,580	4	,000
N of Valid Cases	354		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,04.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,559	,000
	Cramer's V	,559	,000
N of Valid Cases		354	

Fig. 13. Výsledek testování závislosti v kontingenční tabulce - SPSS

Frequency Table for model by Stari

	Nové	Staré	Row Total
Škoda Felicia Combi	5	36	41
	1,41%	10,17%	11,58%
Škoda Felicia	34	111	145
	9,60%	31,36%	40,96%
Škoda Octavia Combi	33	4	37
	9,32%	1,13%	10,45%
Škoda Octavia	68	43	111
	19,21%	12,15%	31,36%
Škoda Superb	20	0	20
	5,65%	0,00%	5,65%
Column Total	160	194	354
	45,20%	54,80%	100,00%

Fig. 14. Kontingenční tabulka pro proměnné *Stáří* a *Model* - Statgraphics

Tests of Independence

Test	Statistic	Df	P-Value
Chi-Square	110,442	4	0,0000

Fig. 15. Výsledek testování závislosti v kontingenční tabulce I - Statgraphics

Statistic	Value	P-Value	Df
Contingency Coeff.	0,4876		
Cramer's V	0,5586		
Conditional Gamma	-0,6831		
Pearson's R	-0,4812	0,0000	352
Kendall's Tau b	-0,4429	0,0000	
Kendall's Tau c	-0,5240		

Fig. 16. Výsledek testování závislosti v kontingenční tabulce II - Statgraphics

3.2 ANOVA

Analýza variance (ANOVA) je statistická metoda používaná k testování hypotézy o rozdílech mezi průměry u více skupin dat. Tento test rozděluje celkovou varianci dat na dvě části: varianci mezi skupinovou a varianci v rámci skupin.

Nulovou hypotézou (H0) je tvrzení, že neexistují žádné rozdíly mezi průměry skupin. Alternativní hypotézou (H1) je tvrzení, že existuje alespoň jeden rozdíl mezi průměry skupin. Testování se obvykle provádí pomocí F-testu, který porovnává variability mezi skupinami a v rámci skupin.

F-test se vypočítá jako poměr variability mezi skupinami a variability v rámci skupin. Pokud je hodnota F-výsledku statisticky významná, zamítáme nulovou hypotézu a přijímáme alternativní hypotézu o existenci rozdílů mezi průměry skupin.

V jednotlivých programech byla provedena analýza rozptylu u proměnné *Fuel* a *price(CZK)*. Závěry jsou shodné napříč softwary a usuzujeme o přítomnosti signifikantní závislosti na 5% hladině významnosti. Výsledky jsou zobrazeny

ve výstupech ze softwaru viz 17, 18 a 19. Softwary SPSS 18 a Statgraphics 20 doplňují výsledky i o grafické znázornění průměrů u jednotlivých skupin. V případě Statgraphics včetně 95% intervalu.

Anova: jeden faktor

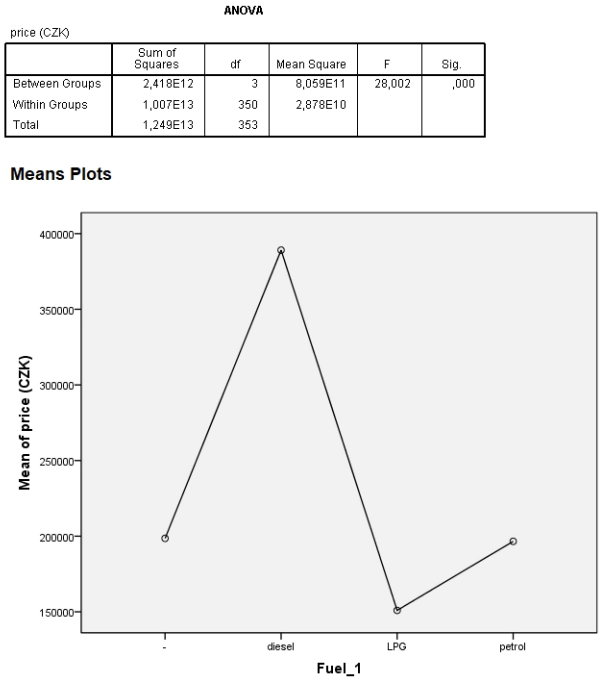
Faktor

Výběr	*Počet*	*Součet*	*Průměr*	*Rozptyl*
diesel	84	32682750	389080,3571	50121108676
lpg	8	1207600	150950	1511545714
petrol	227	44623940	196581,2335	23824540330

ANOVA

Zdroj variabil	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
Mezi výbě	2,35E+12	2	1,17315E+12	38,79814977	8,56E-16	3,024313
Všechny v	9,55E+12	316	30237275173			
Celkem	1,19E+13	318				

Fig. 17. Výsledky analýzy rozptylu pro proměnné *Fuel* a *price(CZK)* - Excel



ANOVA

price (CZK)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2,418E12	3	8,059E11	28,002	,000
Within Groups	1,007E13	350	2,878E10		
Total	1,249E13	353			

Fig. 18. Výsledky analýzy rozptylu pro proměnné *Fuel* a *price(CZK)* - SPSS

ANOVA Table for price (CZK) by fuel

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Between groups	2,41762E12	3	8,05873E11	28,00	0,0000
Within groups	1,00726E13	350	2,87789E10		
Total (Corr.)	1,24902E13	353			

Fig. 19. Výsledky analýzy rozptylu pro proměnné *Fuel* a *price(CZK)* - Statgraphics

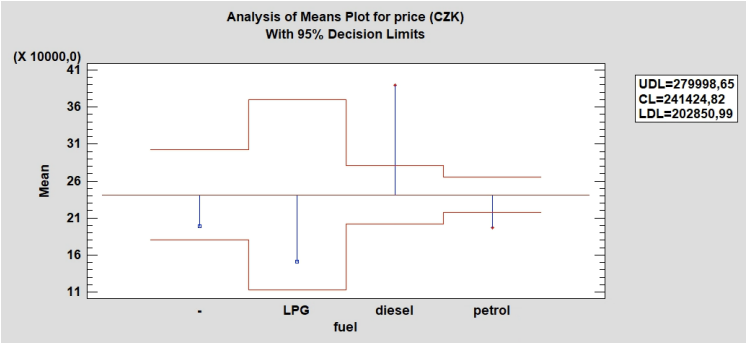


Fig. 20. Grafické zobrazení výsledku analýzy rozptylu pro proměnné *Fuel* a *price(CZK)* - Statgraphics

3.3 Regresní analýza

Jednoduchá lineární regresní analýza je statistická metoda používaná k modelování vztahu mezi dvěma proměnnými a k predikování hodnoty závisle proměnné na základě hodnoty vysvětlující proměnné. Vztah je možno zapsat lineární rovnicí ve tvaru:

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{1}$$

kde y je závisle proměnná, x je vysvětlující proměnná a β_0 je intercept a β_1 odhadovaný koeficient.

Statistická významnost závislosti se testuje na odhadnutém koeficientu β_1 pomocí t-testu s nulovou hypotézou (H0) o nulové hodnotě daného koeficientu. Alternativní hypotéza (H1) je pak o nenulové hodnotě daného koeficientu, tedy o přítomnosti statisticky významné závislosti mezi proměnnou x a y.

Dalším prováděným testem je koeficient determinace R^2, který měří, jak velkou část variability závisle proměnné y se podařilo díky vysvětlující proměnné x vysvětlit.

U zkoumaného souboru byla provedena jednoduchá lineární regrese pro závisle proměnnou *price(CZK)* a vysvětlující proměnnou *kilometres*. Výsledky zobrazené z jednotlivých programů včetně grafického znázornění jsou vcelku konzistentní. Výsledný předpis odhadnuté lineární rovnice:

$$y = 473655 + 2,46 kilometres + \epsilon, \tag{2}$$

je, při zaokrouhlení na dvě desetinná místa a drobném posunu interceptu, shodný pro všechny softwary 23, 24 a 22.

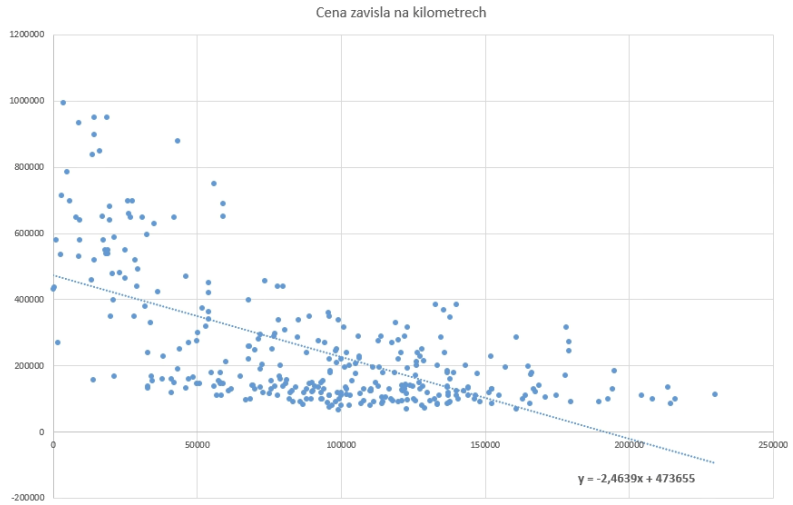


Fig. 21. Grafické znázornění výsledné regresesní přímky pro závisle proměnnou *price(CZK)* a vysvětlující proměnnou *kilometres* pro pozorovaná data - Excel

Taktéž se přibližně rovnají vypočtené koeficienty determinace, které říkají, že se podařilo vysvětlit zhruba 39 % variability proměnné *price(CZK)* pomocí proměnné *kilometres*. Ve všech případech zamítáme nulovou hypotézu t-testu. Na základě 5 % hladině významnosti se nám podařilo nalézt oporu v datech pro usuzování o signifikantní závislosti mezi proměnnou *price(CZK)* a *kilometres*. Graf 21 znázorňuje výslednou regresní přímku proloženou pozorovanými daty. Grafy 25 a 26 pak doplňují informace o výsledných rezidua. Ačkoliv lehké vychýlení reziduí je dle zmíněných grafů pravděpodobně přítomné, zdá se být v toleranci o usuzování o normalitě reziduí. Pro exaktní závěry by bylo třeba provést příslušné testy. Výhodou programů SPSS a Statgraphic je intuitivní procházení výsledků. Oproti tomu Excel působí neuspořádaně, především při zobrazování grafů, které je vždy nutné manuálně posunout nebo umístit na nový list.

Simple Regression - price (CZK) vs. kilometres
Dependent variable: price (CZK)
Independent variable: kilometres
Linear model: Y = a + b*X
Number of observations: 328

Coefficients

Parameter	Least Squares Estimate	Standard Error	T Statistic	P-Value
Intercept	473655,	17851,0	26,5338	0,0000
Slope	-2,46394	0,171089	-14,4015	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	4,73752E12	1	4,73752E12	207,40	0,0000
Residual	7,4465E12	326	2,2842E10		
Total (Corr.)	1,2184E13	327			

Correlation Coefficient = -0,623563
R-squared = 38,883 percent
R-squared (adjusted for d.f.) = 38,6956 percent
Standard Error of Est. = 151136,
Mean absolute error = 116931,
Durbin-Watson statistic = 1,41292 (P=0,0000)
Lag 1 residual autocorrelation = 0,275913

Fig. 22. Výsledky regresesní analýzy pro závisle proměnnou *price(CZK)* a vysvětlující proměnnou *kilometres* - Statgraphics

VÝSLEDEK

Regresní statistika	
Násobné R	0,623562656
Hodnota spolehlivosti R	0,388830387
Nastavená hodnota spolehlivosti R	0,386955633
Chyba stř. hodnoty	151135,7831
Pozorování	328

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	4,73752E+12	4,74E+12	207,4034821	9,82714E-37
Rezidua	326	7,4465E+12	2,28E+10		
Celkem	327	1,2184E+13			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*
Hranice	473655,1924	17850,99775	26,53382	1,93203E-83	438537,5043	508772,8805	438537,5043	508772,8805
kilometres	-2,463940537	0,171089045	-14,4015	9,82714E-37	-2,800518458	-2,127362616	-2,800518458	-2,127362616

Fig. 23. Výsledky regresesní analýzy pro závisle proměnnou *price(CZK)* a vysvětlující proměnnou *kilometres* - Excel

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,624[a]	,389	,387	151135,783	,389	207,403	1	326	,000

a. Predictors: (Constant), kilometres

b. Dependent Variable: price(CZK)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4,738E12	1	4,738E12	207,403	,000[a]
	Residual	7,447E12	326	2,284E10		
	Total	1,218E13	327			

a. Predictors: (Constant), kilometres

b. Dependent Variable: price(CZK)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	473655,192	17850,998		26,534	,000	438537,504	508772,881
	kilometres	-2,464	,171	-,624	-14,402	,000	-2,801	-2,127

a. Dependent Variable: price(CZK)

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-93051,13	473384,16	246395,08	120365,340	328
Residual	-281360,031	530481,125	,000	150904,512	328
Std. Predicted Value	-2,820	1,886	,000	1,000	328
Std. Residual	-1,862	3,510	,000	,998	328

a. Dependent Variable: price(CZK)

Fig. 24. Výsledky regresesní analýzy pro závisle proměnnou *price(CZK)* a vysvětlující proměnnou *kilometres* - SPSS

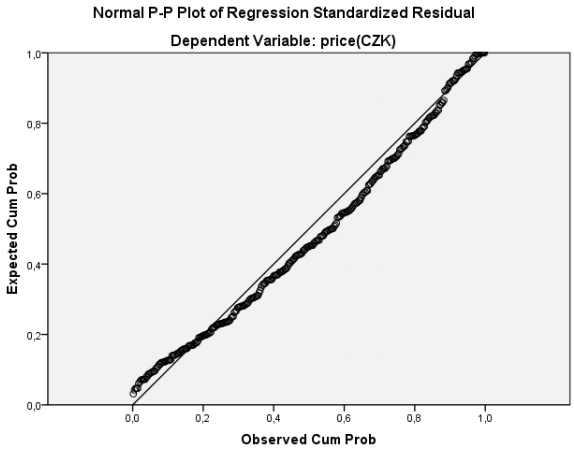


Fig. 25. Grafické znázornění výsledných reziduí při regresní analýze pro závisle proměnnou *price(CZK)* a vysvětlující proměnnou *kilometres* - SPSS

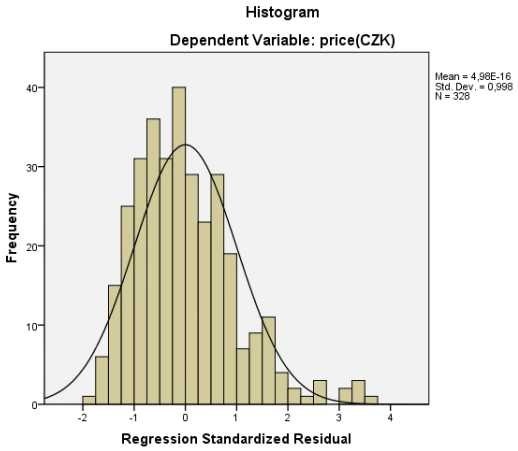


Fig. 26. Grafické znázornění výsledných reziduí při regresní analýze pro závisle proměnnou *price(CZK)* a vysvětlující proměnnou *kilometres* - SPSS

4 Analýza časové řady

Pro analýzu časové řady jsou využita především grafická znázornění dostupná v jednotlivých softwarech. Dále byla taktéž provedena jednoduchá regrese. Opět jde pouze o ilustrativní příklad, kdy nejsou ověřeny jednotlivé předpoklady, nepracuje se se sezónní složkou či autokorelací.

Graf vypracovaný v Excelu 27 zobrazuje časovou řadu ceny akcií firmy Apple v čase. Zároveň zachycuje namodelovaný lineární trend, avšak lze vidět, že lineární trend zdaleka nezachystuje volatilitu finanční časové řady. Ten zachycuje i graf ze Statgraphic 28. Vhodnějším přístupem se jeví klouzavé průměry modelované v prostředí Statgraphics 30. Další graf ze stejného prostředí zobrazuje průběh časové řady a její medián 29.

Statgraphics k výslednému regresnímu modelu 33 zobrazuje i porovnání dalších možných modelů pro trend v přehledné tabulce i s jejich vyhodnocením 34. Zde je vidět, že zvolený lineární trend nebyl nejvhodnější volbou.

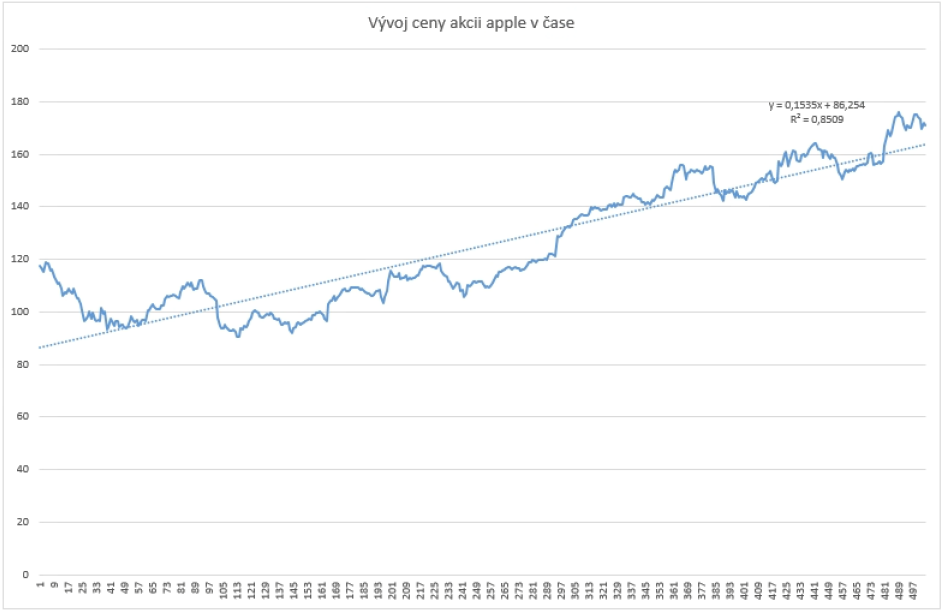


Fig. 27. Zobrazaní vývoje ceny akcií firmy Apple v čase proložené namodelovaným lineárním trendem - Excel

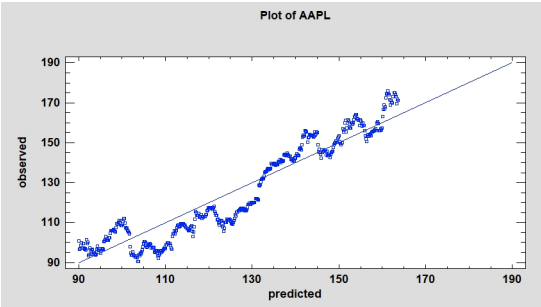


Fig. 28. Zobrazaní vývoje ceny akcií firmy Apple v čase proložené namodelovaným lineárním trendem - Statgraphics

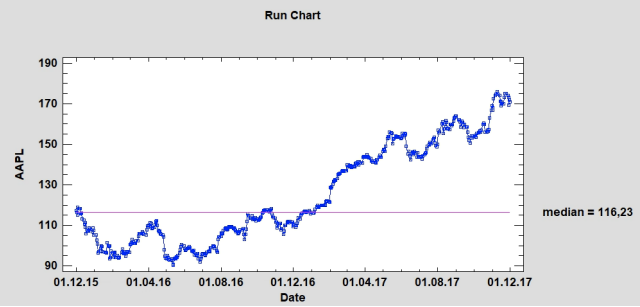


Fig. 29. Zobrazaní vývoje ceny akcií firmy Apple v čase proložené mediánem dané časové řady - Statgraphics

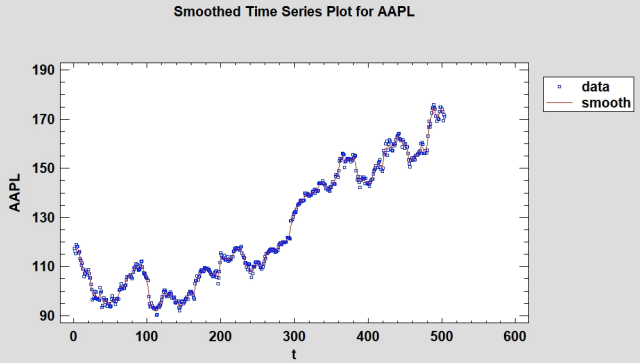


Fig. 30. Zobrazaní vývoje ceny akcií firmy Apple v čase proložené vypočtenými klouzavými průměry - Statgraphics

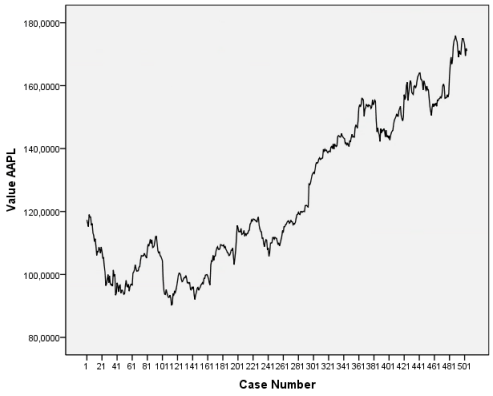


Fig. 31. Zobrazaní vývoje ceny akcií firmy Apple v čase - SPSS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	86,254	,836		103,208	,000	84,612	87,896
	t	,153	,003	,922	53,519	,000	,148	,159

a. Dependent Variable: AAPL

Fig. 32. Výsledky modelování lineárního trendu pomocí regresesní analýzy pro vývoj ceny akcií firmy Apple - SPSS

Coefficients

Parameter	*Least Squares Estimate*	*Standard Error*	*T Statistic*	*P-Value*
Intercept	86,2541	0,835731	103,208	0,0000
Slope	0,153484	0,00286781	53,5195	0,0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	251324,	1	251324,	2864,34	0,0000
Residual	44046,8	502	87,7426		
Total (Corr.)	295371,	503			

Correlation Coefficient = 0,92243
R-squared = 85,0876 percent
R-squared (adjusted for d.f.) = 85,0579 percent
Standard Error of Est. = 9,3671
Mean absolute error = 7,64748
Durbin-Watson statistic = 0,0296918 (P=0,0000)
Lag 1 residual autocorrelation = 0,973664

Fig. 33. Výsledky modelování lineárního trendu pomocí regresesní analýzy pro vývoj ceny akcií firmy Apple - Statgraphics

Comparison of Alternative Models

Model	Correlation	R-Squared
Double squared	0,9625	92,65%
Squared-X	0,9565	91,49%
Square root-Y squared-X	0,9515	90,54%
Logarithmic-Y squared-X	0,9452	89,34%
Reciprocal-Y squared-X	-0,9288	86,27%
Square root-Y	0,9225	85,11%
Linear	0,9224	85,09%
Exponential	0,9213	84,89%
Squared-Y	0,9183	84,33%
Reciprocal-Y	-0,9149	83,70%
Logarithmic-Y square root-X	0,8482	71,94%
Double square root	0,8474	71,80%
Square root-X	0,8453	71,45%
Squared-Y square root-X	0,8374	70,12%
Multiplicative	0,6743	45,46%
Square root-Y logarithmic-X	0,6733	45,33%
Logarithmic-X	0,6712	45,05%
Squared-Y logarithmic-X	0,6640	44,08%
Squared-Y reciprocal-X	-0,1265	1,60%
Reciprocal-X	-0,1216	1,48%
S-curve model	-0,1151	1,32%
Double reciprocal	0,1072	1,15%
Reciprocal-Y square root-X	<no fit>	
Reciprocal-Y logarithmic-X	<no fit>	
Square root-Y reciprocal-X	<no fit>	
Logistic	<no fit>	
Log probit	<no fit>	

Fig. 34. Porovnání výsledků modelování lineárního trendu pomocí regresesní analýzy pro vývoj ceny akcií firmy Apple s dalšími možnými modely - Statgraphics

Závěr

Ve všech vybraných programech byla zpracována analýza představených datových souborů. Byly diskutovány limity práce v kontextu relevantnosti dosažených výsledků. Nejpřehlednější výstupy pocházejí z programu Statgraphic, který je vhodně doplňuje interpretací jednotlivých využitých testů a případně upozorňuje uživatele i na jejich předpoklady. Oproti tomu SPSS produkuje vzhledné tabulky použitelné i pro případnou publikaci. Podobně by tomu bylo i u programu Excel, avšak český překlad kaží celkový dojem. Pro začátečníky lze nejlépe doporučit program Statgraphic. České překlady v Excelu by mohly způsobit nemalé potíže při interpretaci dosažených výsledků.

References

1. ABDULLAH M ALGHAMDI: Big Five Stocks (2019) https://www.kaggle.com/datasets/abdullahmu/big-five-stocks
2. ZOUHAR JAN: Cars [Dataset] https://nb.vse.cz/~zouharj/econCZ.html