MV011 Statistika I – cvičení 4

1) [R] cv. 8 úkol 1,2,3,4,5,6

1. Z údajů v tabulce **Sales**, pro které název pozice (***job_title***) obsahuje řetězec „Rep“, vytvořte html/pdf/rtf obsahující kontingenční tabulku sloupců pohlaví (***gender***) a stát (***country***). Nastavte vhodný nadpis a potlačte výpis datumu. (PROC FREQ)

Sales Rep Frequency Report

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Gender by Country		
	Country		
Gender	AU	US	Total
F	27 16.98 40.30 44.26	40 25.16 59.70 40.82	67 42.14
M	34 21.38 36.96 55.74	58 36.48 63.04 59.18	92 57.86
Total	61 38.36	98 61.64	159 100.00

2. Vytvořte tabulku **Sales1** z tabulky **Sales**, ve které vznikne nový sloupec ***hire_age*** představující věk zaměstnance v okamžiku nástupu do zaměstnání. Vytvořte formát ***HireAge***, který agreguje zadaný sloupec do kategorií low-<20, 20-<25 a 25-high. Následně vytvořte frekvenční tabulku pro sloupec hire_age formátovaný pomocí HireAge. (PROC FREQ)

Hire_age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1. pod 20	43	26.06	43	26.06
2. 20 - 25	68	41.21	111	67.27
3. nad 25	54	32.73	165	100.00

3. Z tabulky **Sales1** z úkolu 2

   a. vypište průměr (mean) a rozsah (range) příjmu (***salary***) pro všechny trojice hodnot sloupců pohlaví (***gender***), stát (***country***) a ***hire_age*** formátovaného pomocí HireAge z úkolu 2. (PROC MEANS)

Analysis Variable : Salary					
Gender	Country	Hire_age	N Obs	Mean	Range
F	AU	1. pod 20	10	27498.00	5600.00
		2. 20 - 25	10	27849.00	5615.00
		3. nad 25	7	27785.00	4695.00
	US	1. pod 20	7	28853.57	7055.00
		2. 20 - 25	18	28285.83	6475.00
		3. nad 25	16	31048.75	57825.00
M	AU	1. pod 20	9	35837.22	82510.00
		2. 20 - 25	19	31463.95	61990.00
		3. nad 25	8	28962.50	5975.00
	US	1. pod 20	17	40757.94	217905.00
		2. 20 - 25	21	27750.95	7600.00
		3. nad 25	23	32950.00	72380.00

   b. uložte výstup procedury (bez specifikace ukládaných údajů) do tabulky a porovnejte výstup bodu a) a b).

Obs	Gender	Country	Hire_age	_TYPE_	_FREQ_	_STAT_	Salary
1			.	0	165	N	165.00
2			.	0	165	MIN	22710.00
3			.	0	165	MAX	243190.00
4			.	0	165	MEAN	31160.12
5			.	0	165	STD	20082.67
6			1. pod 20	1	43	N	43.00
⋮							
175	M	US	2. 20 - 25	7	21	STD	2250.29
176	M	US	3. nad 25	7	23	N	23.00
177	M	US	3. nad 25	7	23	MIN	22710.00
178	M	US	3. nad 25	7	23	MAX	95090.00
179	M	US	3. nad 25	7	23	MEAN	32950.00
180	M	US	3. nad 25	7	23	STD	18155.18

4. Z tabulky **Sales1** z úkolu 2 vytvořte kontingenční tabulku s absolutními četnostmi a řádkově a sloupcově podmíněnými relativními četnostmi. Řádková dimenze bude tvořena kartézským součinem hodnot sloupce ***hire_age*** formátovaného pomocí HireAge (včetně souhrnu (all)) a hodnot sloupce ***country***. Sloupcová dimenze bude tvořena hodnotami sloupce ***gender***. (PROC TABULATE)

		Gender					
		F			M		
		N	RowPctN	ColPctN	N	RowPctN	ColPctN
Hire_age	Country						
1. pod 20	AU	10	52.63	14.71	9	47.37	9.28
	US	7	29.17	10.29	17	70.83	17.53
2. 20 - 25	AU	10	34.48	14.71	19	65.52	19.59
	US	18	46.15	26.47	21	53.85	21.65
3. nad 25	AU	7	46.67	10.29	8	53.33	8.25
	US	16	41.03	23.53	23	58.97	23.71
All	Country						
	AU	27	42.86	39.71	36	57.14	37.11
	US	41	40.20	60.29	61	59.80	62.89

5. Z tabulky **Sales1** z úkolu 2 vytvořte kontingenční tabulku, která bude obsahovat minimum, medián a maximum příjmu (***salary***). Řádková dimenze bude tvořena kartézským součinem hodnot sloupce ***hire_age*** formátovaného pomocí HireAge a hodnot sloupce ***country***. Sloupcová dimenze bude tvořena hodnotami sloupce ***gender***. U řádkové i sloupcové dimenze včetně všech souhrnů („all“). To vše ve formátu pdf se stylem sasweb. (PROC TABULATE)

		Gender						All		
		F			M					
		Salary			Salary			Salary		
		Min	P50	Max	Min	P50	Max	Min	P50	Max
Hire_age	Country									
1. pod 20	AU	25185.00	27362.50	30785.00	25745.00	26780.00	108255.00	25185.00	26970.00	108255.00
	US	25930.00	28325.00	32985.00	25285.00	27325.00	243190.00	25285.00	27400.00	243190.00
	All	25185.00	27465.00	32985.00	25285.00	27227.50	243190.00	25185.00	27260.00	243190.00
2. 20 - 25	Country									
	AU	25275.00	27445.00	30890.00	25985.00	27115.00	87975.00	25275.00	27440.00	87975.00
	US	25390.00	28132.50	31985.00	25125.00	27100.00	32725.00	25125.00	27425.00	32725.00
	All	25275.00	27742.50	31985.00	25125.00	27107.50	87975.00	25125.00	27432.50	87975.00
3. nad 25	Country									
	AU	25795.00	26850.00	30490.00	26515.00	28495.00	32490.00	25795.00	28480.00	32490.00
	US	25680.00	27510.00	63505.00	22710.00	27410.00	95090.00	22710.00	27460.00	95090.00
	All	25680.00	27460.00	63505.00	22710.00	27485.00	95090.00	22710.00	27472.50	95090.00
All	Country									
	AU	25185.00	27440.00	30890.00	25745.00	27165.00	108255.00	25185.00	27260.00	108255.00
	US	25390.00	28010.00	63505.00	22710.00	27260.00	243190.00	22710.00	27442.50	243190.00
	All	25185.00	27470.00	63505.00	22710.00	27240.00	243190.00	22710.00	27425.00	243190.00

6. Analyzujte (zajímá nás základní sada popisných statistik, test pro charakteristiku polohy, kvantily, odlehlá pozorování) sloupec ***salary*** z tabulky **Sales**. Vytvořte výstup ve formátu rtf se stylem sasweb. (PROC UNIVARIATE)

Moments			
N	165	Sum Weights	165
Mean	31160.1212	Sum Observations	5141420
Std Deviation	20082.6671	Variance	403313519
Skewness	8.16761992	Kurtosis	78.5622611
Uncorrected SS	2.26351E11	Corrected SS	6.61434E10
Coeff Variation	64.4499006	Std Error Mean	1563.43352

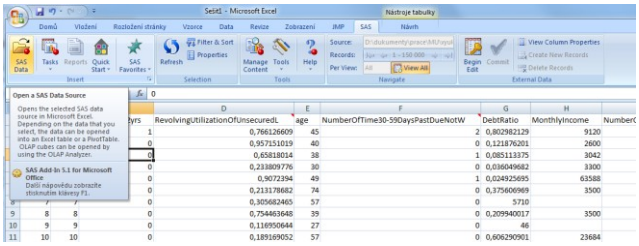
Basic Statistical Measures			
Location		Variability	
Mean	31160.12	Std Deviation	20083
Median	27425.00	Variance	403313519
Mode	26600.00	Range	220480
		Interquartile Range	2825

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	19.93057	Pr > |t|	<.0001
Sign	M	82.5	Pr >= |M|	<.0001
Signed Rank	S	6847.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	243190
99%	108255
95%	32985
90%	31750
75% Q3	29385
50% Median	27425
25% Q1	26560
10%	25965
5%	25680
1%	25110
0% Min	22710

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
22710	131	84260	165
25110	111	87975	2
25125	104	95090	163
25185	49	108255	1
25275	50	243190	64

2) V MS Excel (nebo ekvivalentní) nad cs-training (nejlépe neimportované do MS Excel ze SASí tabulky pomocí SAS Add-in for MS Office): spočítat průměr, medián, modus, rozptyl, dolní a horní kvartil, šikmost, špičatost a pearsonův koeficient korelace pro věk a DebtRatio; dále vytvořit kontingenční tabulku NumberofDependents a SeriousDlqin2yrs s absolutními četnostmi, řádkově podmíněnými relativními četnostmi a sloupcově podmíněnými relativními četnostmi.



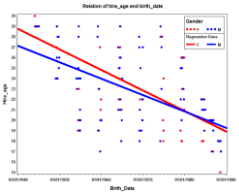
průměr	52,30
medián	52,00
modus	49,00
rozptyl	218,21
Q1	41,00
Q3	63,00
šikmost	0,19
špičatost	-0,49
korelace	0,02

NumberofDependents (row) x SeriousDlqin2yrs (col)

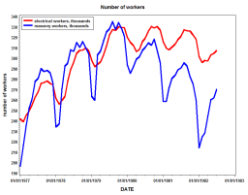
Popisky sloupců	0			1			Celkem N	Celkem RowPctn	Celkem ColPctn
Popisky řádků N		RowPctn	ColPctn	N	RowPctn	ColPctn			
0	81807	94,14%	58,44%	5095	5,86%	50,82%	86902	100,00%	57,93%
1	24381	92,65%	17,42%	1935	7,35%	19,30%	26316	100,00%	17,54%
2	17938	91,89%	12,82%	1584	8,11%	15,80%	19522	100,00%	13,01%
3	8646	91,17%	6,18%	837	8,83%	8,35%	9483	100,00%	6,32%
4	2565	89,62%	1,83%	297	10,38%	2,96%	2862	100,00%	1,91%
5	678	90,88%	0,48%	68	9,12%	0,68%	746	100,00%	0,50%
6	134	84,81%	0,10%	24	15,19%	0,24%	158	100,00%	0,11%
7	46	90,20%	0,03%	5	9,80%	0,05%	51	100,00%	0,03%
8	22	91,67%	0,02%	2	8,33%	0,02%	24	100,00%	0,02%
9	5	100,00%	0,00%		0,00%	0,00%	5	100,00%	0,00%
10	5	100,00%	0,00%		0,00%	0,00%	5	100,00%	0,00%
13	1	100,00%	0,00%		0,00%	0,00%	1	100,00%	0,00%
20	1	100,00%	0,00%		0,00%	0,00%	1	100,00%	0,00%
(Prázdné)	3745	95,44%	2,68%	179	4,56%	1,79%	3924	100,00%	2,62%
Celkový součet	**139974**	**93,32%**	**100,00%**	**10026**	**6,68%**	**100,00%**	**150000**	**100,00%**	**100,00%**

3) [R] cv.9 úkol 1,2

1. Z údajů v tabulce **Sales1** vytvořte bodový graf závislosti ***hire_age*** na ***birth_date*** s rozlišením pohlaví (***gender***). Graf doplňte o regresní přímky a upravte vzhled podle vzoru (PROC GPLOT)… formát x-ové osy mmddyy10., tloušťka reg. přímek = 5, font popisu os i legendy = (arial bold, výška 12 bodů, resp. 10 bodů u „regression lines“), font hodnot na osách a hodnot v legendě= (arial bold, výška 10 bodů), výška nadpisu = 12 bodů.



2. Z údajů v tabulce **Sashelp.workers** vytvořte graf počtu elektrikářů (***electric*)** a počtu zedníků (***masonry*)** v *čase(**date***). Upravte vzhled podle vzoru (PROC GPLOT s overlay)… formát x-ové osy mmddyy10., tloušťka křivek = 5, font popisu os = (arial bold, výška 12 bodů), font hodnot na osách a hodnot v legendě= (arial bold, výška 10 bodů), výška nadpisu = 12 bodů, offset legendy = 1%.



- [R] Martin Řezáč – DM1 cvičení https://www.math.muni.cz/~mrezac/vyuka/DM1_cviceni_v2013.pdf