Instituto
Politécnico de Setúbal
Escola Superior de
Tecnologia de Setúbal

Métodos
Estatísctícos

Licenciatura em Engenharia Informática
Ano letivo 2021-2022

Unidade Curricular Métodos Estatísticos

Docente José Palma

Ema Barão | 201400238
Nuno Reis | 202000753
Bernardo Teixeira | 201801954

Conteúdo

Índice de gráficos

Ema Barão | **201400238**
Nuno Reis | **202000753**
Bernardo Teixeira | **201801954**

Índice de tabelas

Índice de figuras

Resumo

No âmbito da disciplina de métodos estatísticos iremos tratar um conjunto de dados relativo a um estudo conduzido pela Coreia do Sul, de 2005 a 2018, que recolheu várias informações sobre os seus cidadãos, particularmente sobre o rendimento das famílias.

Nesta fase iremos apenas caracterizar os dados constantes na base de dados disponibilizada e que é apenas uma amostra dos dados completos e que podem ser consultados no site da KOWEPS[1] (Korea Welfare Panel Study).

A analise dos dados deverá consistir na caracterização dos dados em qualitativos e quantitativos dependendo da variável em estudo. A organização dos dados irá permitir encontrar as medidas de localização e de dispersão. Durante a análise vão ser produzidos os auxiliares gráficos apropriados para cada uma das variáveis.

Palavras-chave: Analise dos dados; Dados qualitativos; Dados quantitativos; Medidas de localização; Medidas de dispersão;

[1] https://www.kaggle.com/datasets/hongsean/korea-income-and-welfare

Ema Barão | **201400238**
Nuno Reis | **202000753**
Bernardo Teixeira | **201801954**

Introdução

Este trabalho foi nos solicitado no âmbito da Unidade Curricular de Métodos Estatísticos e tem como principal objetivo caracterizar aspetos relevantes do conjunto de dados que nos foi fornecido.

Neste processo selecionamos 8 variáveis aleatórias que caracterizamos mediante os seguintes critérios:

Caso a variável seja qualitativa ordinal ou qualitativa nominal:

- A sua representação gráfica será feita através de gráficos de barras e gráficos circulares;
- O único indicador de localização utilizado será a moda;
- Indicação das medidas de assimetria e de curtose.

Caso a variável seja quantitativa discreta ou quantitativa contínua:

- A sua representação gráfica será feita através de gráficos de barras, (histogramas para variáveis aleatórias continuas);
- Os seus indicadores de localização serão a média, a moda, mediana, quartis e decis;
- Os seus indicadores de dispersão serão a amplitude total, a amplitude interquartil, a variância e o desvio padrão;
- Indicação das medidas de assimetria e de curtose.

Na conclusão deste processo será possível conhecer e caracterizar cada uma das variáveis ao detalhe.

Dados fornecidos e tratamento dos dados

Foram fornecidos 2 conjuntos de dados:

1. Korea Income and Welfare
2. job_code_translated

O conjunto de dados Korea Income and Welfare apresentava as seguintes características:

1. Representa os dados que caracterizam o rendimento das famílias em determinada área geográfica e em determinado período de tempo;
2. Têm a dimensão de 92857 linhas (observações) e 14 colunas (variáveis aleatórias);
3. É composto pelas seguintes variáveis aleatórias:
    a. id;
    b. year;
    c. wave;
    d. region;
    e. income;
    f. family_member;
    g. gender;
    h. year_born;
    i. education_level;
    j. marriage;
    k. religion;
    l. occupation;
    m. company_size;
    n. reason_none_worker

O conjunto de dados job_code_translated apresentava as seguintes características:

3. Apresenta os valores da variável occupation do conjunto de dados Korea Income and Welfare;
1. Têm a dimensão de 149 linhas (observações) e 5 colunas (variáveis aleatórias);
2. É composto pelas seguintes variáveis aleatórias:
    a. Id;
    b. Job_Category_Code;
    c. Job_Category_Title;
    d. job_code;
    e. job_title;

Numa analise à priori da limpeza dos dados com que iremos trabalhar, encontramos os valores constantes da tabela 1.

Variável	Range	Null	Classificação da variável aleatória	Outliers	Selecionada para análise	Observações
year	[2005;2018]	-	Quantitativa discreta	-	Sim	
region	[1;7]	-	Qualitativa nominal	-	Sim	Apresenta dados em código[2]
income	[-100 9998]	-	Quantitativa continua	Sim	Sim	
family_member	[1;9]	-	Quantitativa discreta	-	Sim	
gender	[1;2]	-	Qualitativa nominal	-	Sim	Apresenta dados em código[2]
year_born	[1910;2002]	-	Quantitativa discreta	-	Sim	
education_level	[2;9]	-	Qualitativa ordinal	-	Sim	Apresenta dados em código[2]
company_size	[NA,NA]	33642	Quantitativa discreta	Sim	Sim	

Tabela 1- Valores do conjunto de dados antes do tratamento dos dados

No conjunto de dados Korea Income and Welfare, procedemos à remoção dos valores a Null, tal como, removemos os outliers conforme se pode observar no script de R constante na entrega do projeto.

Variável year

É uma variável quantitativa discreta, representa o ano em que o estudo foi efetuado. E organiza se na seguinte tabela de frequências.

i	xi	ni	fi	Ni	Fi
1	2005	7072	0.07616012	7072	0.07616012
2	2006	6580	0.07086165	13652	0.14702176
3	2007	6314	0.06799703	19966	0.21501879
4	2008	6207	0.06684472	26173	0.28186351
5	2009	6034	0.06498164	32207	0.34684515
6	2010	5735	0.06176163	37942	0.40860678
7	2011	7532	0.08111397	45474	0.48972075
8	2012	7312	0.07874474	52786	0.56846549
9	2013	7048	0.07590166	59834	0.64436715
10	2014	6914	0.07445858	66748	0.71882572
11	2015	6723	0.07240165	73471	0.79122737
12	2016	6581	0.07087242	80052	0.86209979
13	2017	6474	0.06972011	86526	0.93181990
14	2018	6331	0.06818010	92857	1.00000000

Tabela 2- Tabela de frequências da variável year

Podemos visualizar as frequências quer relativas quer absolutas da variável nos gráficos 1 e 2.

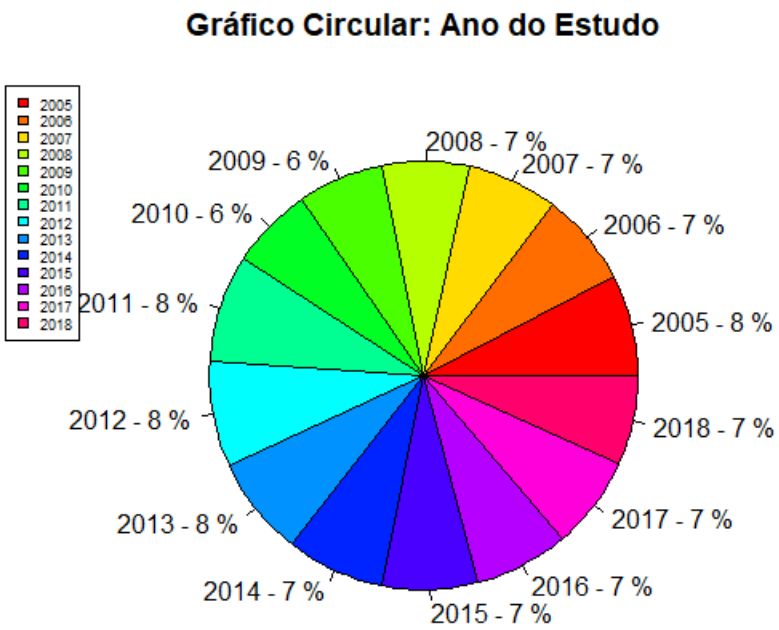

Gráfico 1-Gráfico circular das frequências relativas do ano de estudo

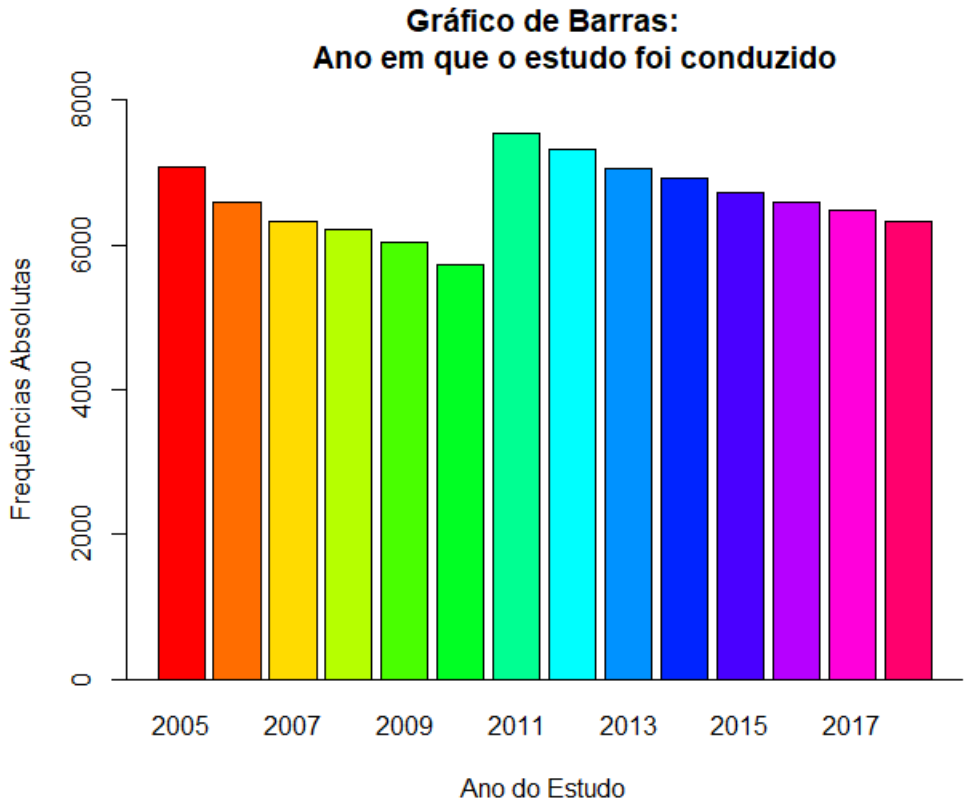

Gráfico 2- Gráfico de barras das frequências absolutas do ano de estudo

Podemos visualizar como se comportam as medidas de dispersão e de localização respetivamente nas tabelas 3 e 4.

Medidas de dispersão	
Variância	16.08991
Desvio Padrão	4.011222
Amplitude Total	13
Amplitude Interquartil	7

Tabela 3 - Medidas de dispersão da variável year

Medidas de Localização					
Moda	2011	Quartis		Decis	
Média	2011.518				
Mediana	2012				
		25%	2008	10%	2006
		50%	2012	20%	2007
		75%	2015	30%	2009
				40%	2010
				50%	2012
				60%	2013
				70%	2014
				80%	2016
				90%	2017

Tabela 4- Medidas de localização da variável year

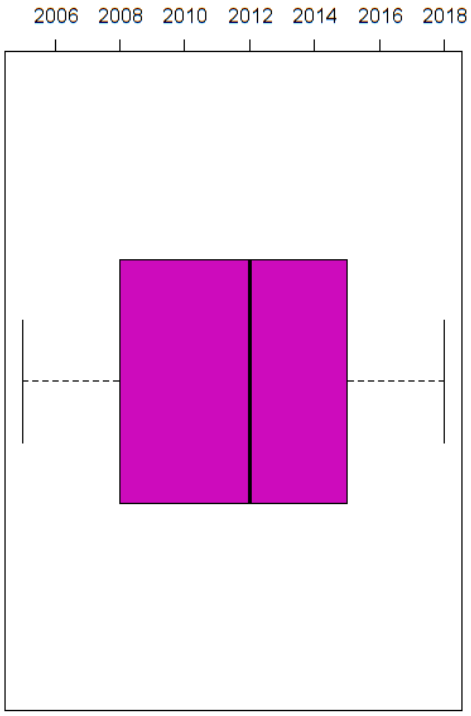

Figura 1- Caixa de bigodes da variável year

Na caraterização da distribuição das frequências verificamos o valor da assimetria é de (b1=-0.03954851). Podemos assim afirmar que a assimetria é negativa pois, b1<0.

Relativamente aos valores de curtose verificamos que a variável apresenta uma curva platicúrtica ou achatada dado que b2<0, (b2=-1.182807).

Na figura 1 podemos visualizar a caixa de bigodes da variável year.

Variável region

É uma variável qualitativa nominal, representa a área geográfica onde o estudo foi efetuado. Apresenta se originalmente de uma forma discreta como forma de codificação dos valores observados e assume a seguinte forma:

1) Seoul 2) Kyeong-gi 3) Kyoung-nam 4) Kyoung-buk 5) Chung-nam 6) Gang-won &. Chung-buk 7) Jeolla & Jeju

Apresenta se na seguinte tabela de frequências.

i	xi	ni	fi	Ni	Fi
1	Seoul	14437	0.15547562	14437	0.1554756
2	Kyeong-gi	19353	0.20841724	33790	0.3638929
3	Kyoung-nam	16154	0.17396642	49944	0.5378593
4	Kyoung-buk	12205	0.13143866	62149	0.6692980
5	Chung-nam	7843	0.08446321	69992	0.7537612
6	Gang-won & Chung-buk	6927	0.07459858	76919	0.8283597
7	Jeolla & Jeju	15938	0.17164026	92857	1.0000000

Tabela 5 - Tabela de frequências da variável region

Podemos visualizar as frequências quer relativas quer absolutas da variável nos gráficos 3,4.

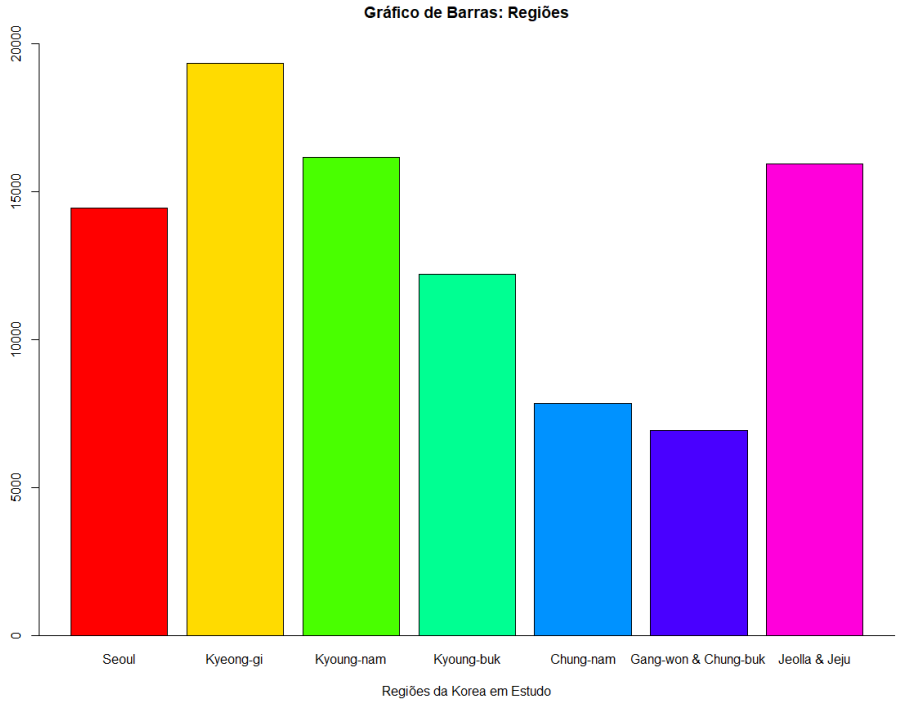

Gráfico 3-Gráfico de barras das frequências absolutas das regiões onde o estudo decorreu

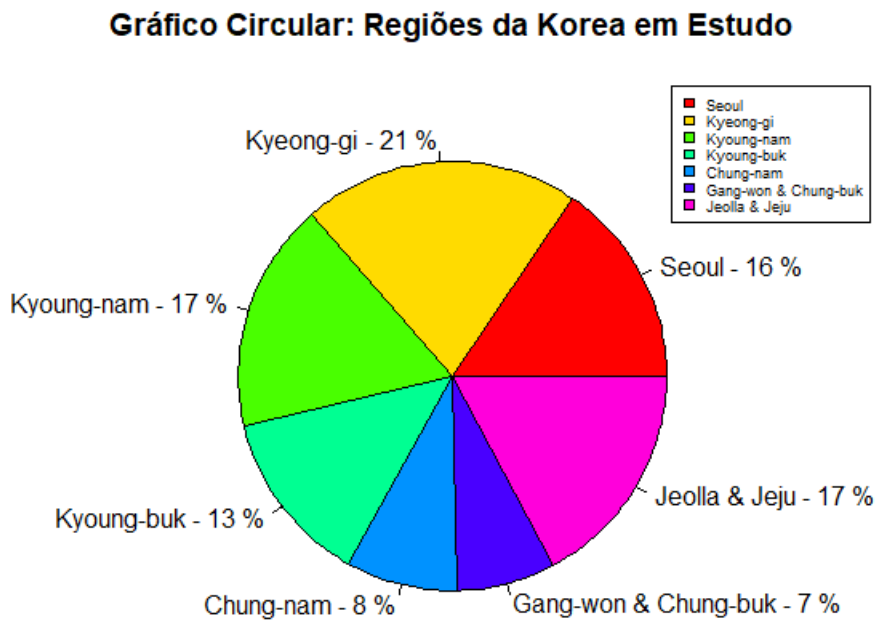

Gráfico 4 - Gráfico circular das frequências relativas das regiões onde decorreu o estudo

Podemos visualizar como se comportam as medidas de localização e de dispersão respetivamente nas tabelas 6 e 7.

Medidas de Localização					
Moda	2 ⇔ "Kyeong-gi"	Quartis		Decis	
Média	Não aplicável[2]				
Mediana	Não aplicável[2]				
		25%	Não aplicável[2]		
		50%	Não aplicável[2]		
		75%	Não aplicável[2]		
				10%	Não aplicável[2]
				20%	Não aplicável[2]
				30%	Não aplicável[2]
				40%	Não aplicável[2]
				50%	Não aplicável[2]
				60%	Não aplicável[2]
				70%	Não aplicável[2]
				80%	Não aplicável[2]
				90%	Não aplicável[2]

Tabela 6 - Medidas de localização da variável region

[2] Não aplicável quando a variável é qualitativa.

Medidas de dispersão	
Variância	Não aplicável[3]
Desvio Padrão	Não aplicável[3]
Amplitude Total	Não aplicável[3]
Amplitude Interquartil	Não aplicável[3]

Tabela 7 - Medidas de dispersão da variável region

Na caraterização da distribuição das frequências verificamos o valor da assimetria é de (b1= 0.3701752). Podemos assim afirmar que a assimetria é positiva pois, b1>0.

Relativamente aos valores de curtose, podemos afirmar que a variável apresenta uma curva platicúrtica ou achatada dado que b2<0, (b2= -1.172695).

[3] Não aplicável quando a variável é qualitativa.

Variável income

É uma variável quantitativa continua, representa o rendimento anual em M KRW(*Million Korean Won. 1100 KRW = 1 USD*). Apresenta se originalmente de uma forma textual, tipo char como forma de codificação dos valores observados, contudo foi convertida para a sua forma numérica homologa.

Apresentava outliers severos superiores e inferiores, foram removidos da amostra dos dados, conforme podemos verificar na figura 3.

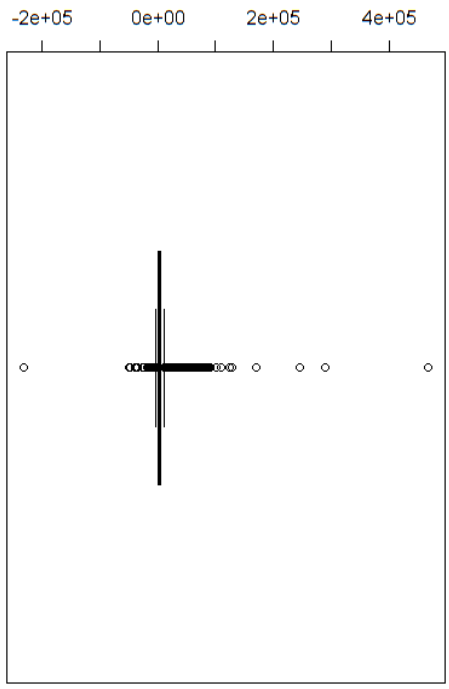

Figura 2-Caixa de bigodes da variável income antes da remoção dos outliers

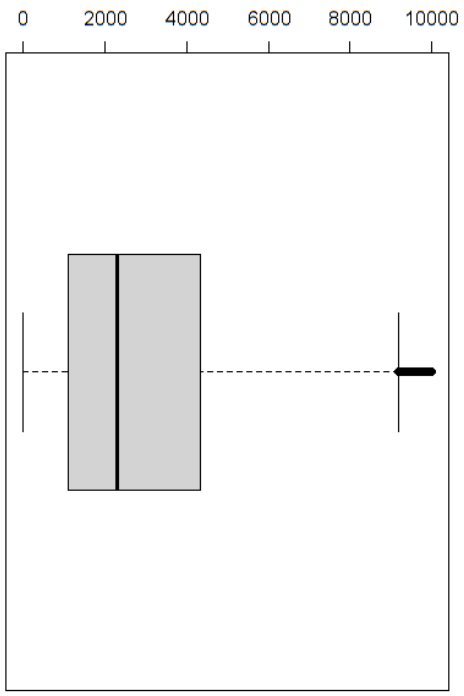

Figura 3-Caixa de bigodes da variável income depois da remoção dos outliers

Aplicando se a regra de Sturges consegui-o apurar 17 classes (K=17), as classes são fechadas à direita e a amplitude de cada classe é de 588.2353, pois h=588.2353 . Apresenta a seguinte tabela de frequências.

	classes	ni	fi	Ni	Fi
1	[0,588]	6356	0.0716	6356	0.0716
2	(588,1.18e+03]	17560	0.1977	23916	0.2693
3	(1.18e+03,1.76e+03]	11956	0.1346	35872	0.4039
4	(1.76e+03,2.35e+03]	9288	0.1046	45160	0.5085
5	(2.35e+03,2.94e+03]	7536	0.0849	52696	0.5933
6	(2.94e+03,3.53e+03]	6318	0.0711	59014	0.6645
7	(3.53e+03,4.12e+03]	5631	0.0634	64645	0.7279
8	(4.12e+03,4.71e+03]	4878	0.0549	69523	0.7828
9	(4.71e+03,5.29e+03]	4135	0.0466	73658	0.8294
10	(5.29e+03,5.88e+03]	3438	0.0387	77096	0.8681
11	(5.88e+03,6.47e+03]	2915	0.0328	80011	0.9009
12	(6.47e+03,7.06e+03]	2320	0.0261	82331	0.9270
13	(7.06e+03,7.65e+03]	1907	0.0215	84238	0.9485
14	(7.65e+03,8.24e+03]	1515	0.0171	85753	0.9656
15	(8.24e+03,8.82e+03]	1245	0.0140	86998	0.9796
16	(8.82e+03,9.41e+03]	1045	0.0118	88043	0.9913
17	(9.41e+03,1e+04]	769	0.0087	88812	1.0000

Tabela 8 - Tabela de frequências da variável income

No histograma do gráfico 5 começamos a vislumbrar uma simetria positiva, mais tarde iremos cruzar com os valores da simetria e curtose para validar este dado.

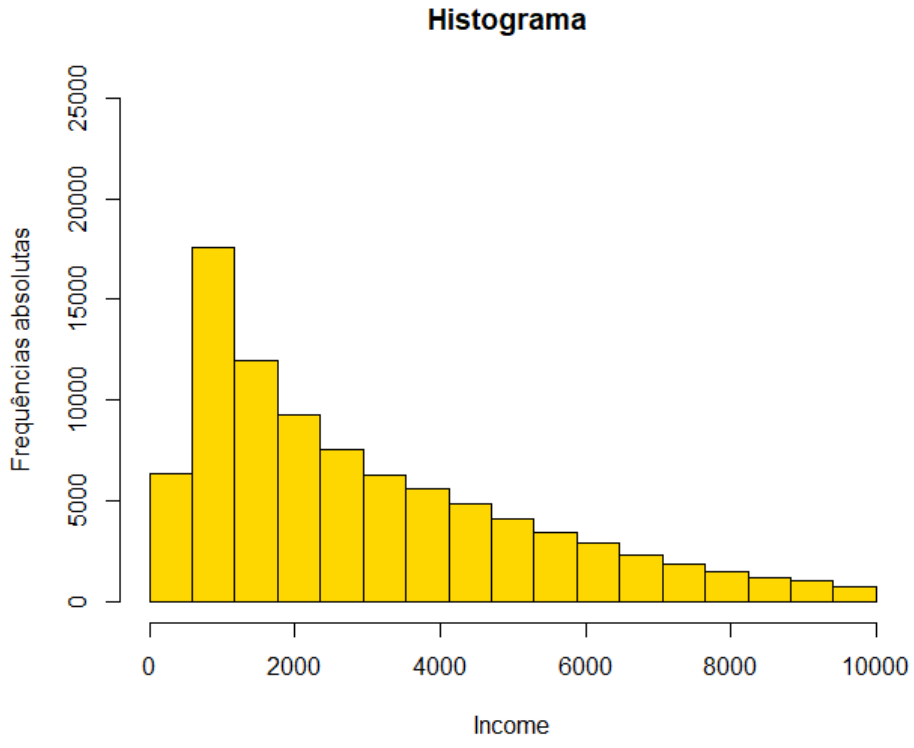

Gráfico 5 -Histograma da variável income

Na caixa de bigodes verificamos que 75% dos rendimentos ficam abaixo dos 4337.25 *KRW*.

Podemos também deduzir que uma pequena porção de indivíduos observados(25%) têm rendimentos superiores a 4337.25 KRW e 10000 KRW.

Podemos deduzir uma grande disparidade de rendimentos entre os dois grupos.

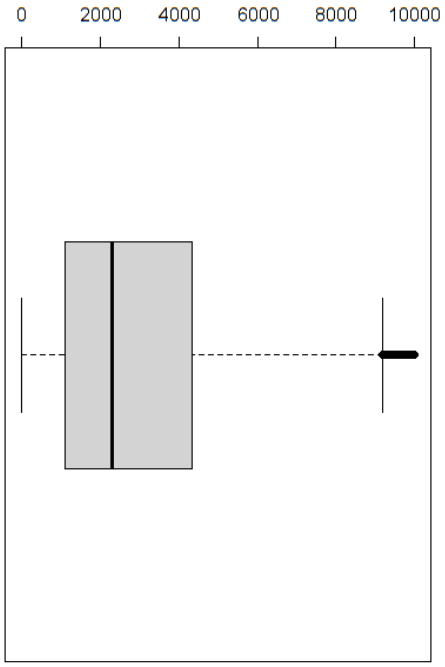

Figura 4-Caixa de bigodes da variável income

Podemos visualizar como se comportam as medidas de localização e de dispersão respetivamente nas tabelas 9 e 10.

Medidas de Localização					
Moda (classe)	(588,1.18000]	Quartis		Decis	
Média	2978.94				
Mediana	2300				
		25%	1106.00		
		50%	2300.00		
		75%	4337.25		
				10%	665.0
				20%	947.0
				30%	1294.0
				40%	1745.0
				50%	2300.0
				60%	2994.0
				70%	3845.0
				80%	4920.0
				90%	6450.8

Tabela 9-Medidas de localização da variável income

Medidas de dispersão	
Variância	5189756
Desvio Padrão	2278.104
Amplitude Total	10000
Amplitude Interquartil	3231.25

Tabela 10 - Medidas de dispersão da variável income

Na caraterização da distribuição das frequências verificamos o valor da assimetria é de (b1= 0.986383). Podemos agora afirmar que a assimetria é positiva pois, b1>0.

Relativamente aos valores de curtose, podemos afirmar que a variável apresenta uma curva leptocúrtica, alongada dado que b2>0, (b2= 0.1676495).

Variável family_member

É uma variável quantitativa discreta, representa o número de elementos do agregado familiar a que foi efetuado o estudo. O estudo incidiu em famílias com 1 elemento até famílias de grandes dimensões com 9 elementos.

Apresenta a seguinte tabela de frequências.

i	xi	ni	fi	Ni	Fi
1	1	25086	2.701573e-01	25086	0.2701573
2	2	28668	3.087328e-01	53754	0.5788901
3	3	16030	1.726310e-01	69784	0.7515212
4	4	16857	1.815372e-01	86641	0.9330584
5	5	4845	5.217700e-02	91486	0.9852354
6	6	1123	1.209386e-02	92609	0.9973292
7	7	211	2.272311e-03	92820	0.9996015
8	8	28	3.015389e-04	92848	0.9999031
9	9	9	9.692323e-05	92857	1.0000000

Tabela 11- Tabela de frequências da variável family_member

Podemos visualizar as frequências quer relativas quer absolutas da variável nos gráficos 7,8.

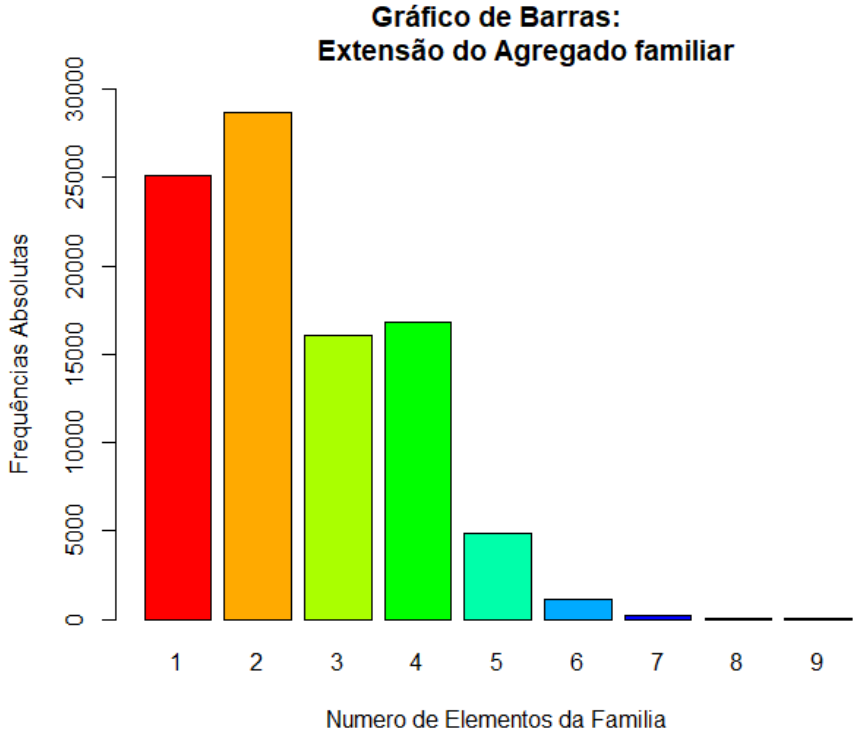

Gráfico 6-Gráfico de barras das frequências absolutas por extensão de agregado familiar

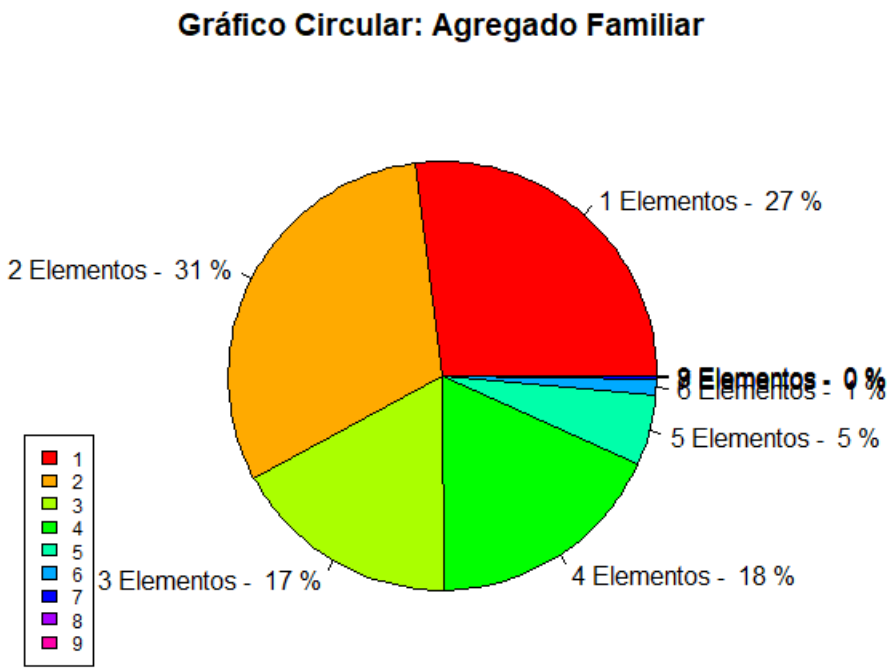

Gráfico 7-Gráfico circular das frequências relativas por extensão de agregado familiar

Podemos visualizar como se comportam as medidas de localização e de dispersão respetivamente nas tabelas 12 e 13.

Medidas de Localização					
Moda	2	Quartis		Decis	
Média	2.484304				
Mediana	2				
		25%	1		
		50%	2		
		75%	3		
				10%	1
				20%	1
				30%	2
				40%	2
				50%	2
				60%	3
				70%	3
				80%	4
				90%	4

Tabela 12-Medidas de localização da variável family_member

Medidas de dispersão	
Variância	1.669912
Desvio Padrão	1.292251
Amplitude Total	8
Amplitude Interquartil	2

Tabela 13 - Medidas de dispersão da variável family_member

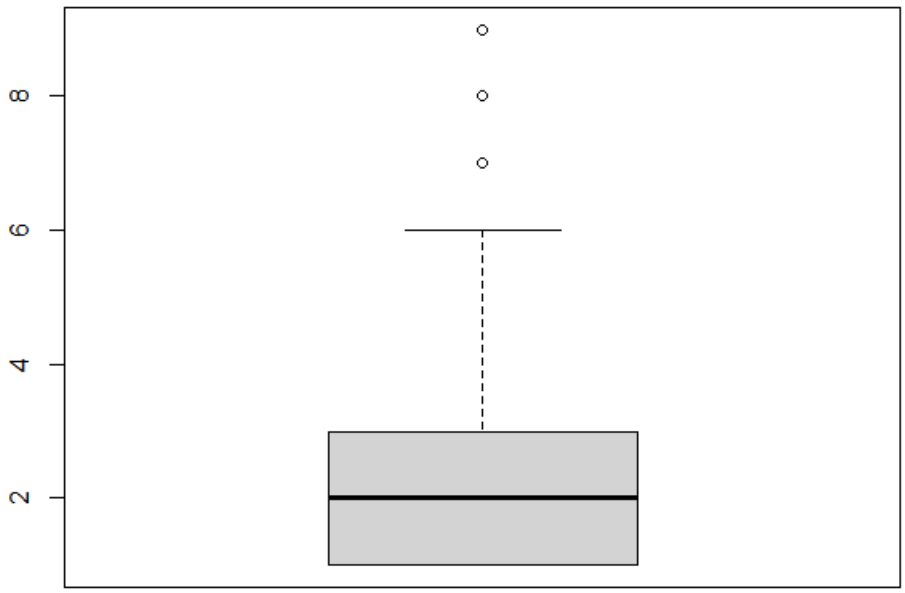

Figura 5 Caixa de bigodes da variável family_member

Na caixa de bigodes podemos verificar que o estudo abrangeu maioritariamente famílias com 1 e 2 elementos, pois 50% dos dados encontram se até ao 2ª Quartil.

Mais esporadicamente foram estudadas famílias com 7,8 e 9 elementos, podemos verificar que são outliers, sem expressão.

Na caraterização da distribuição das frequências verificamos o valor da assimetria é de (b1= 0.6395767). Podemos agora afirmar que a assimetria é positiva pois, b1>0.

Relativamente aos valores de curtose, podemos afirmar que a variável apresenta uma curva platicúrtica ou achatada dado que b2<0, (-0.2621798).

Variável gender

É uma variável qualitativa nominal, representa o género da observação. Apresenta se originalmente de uma forma discreta como forma de codificação dos valores observados e assume a seguinte forma:

1) Masculino 2) Feminino

Apresenta se na seguinte tabela de frequências.

i	xi	ni	fi	Ni	Fi
1	1	65342	0.7036842	65342	0.7036842
2	2	27515	0.2963158	92857	1.0000000

Tabela 14-Tabela de frequências da variável gender

Podemos visualizar as frequências quer relativas quer absolutas da variável nos gráficos 9 e 10.

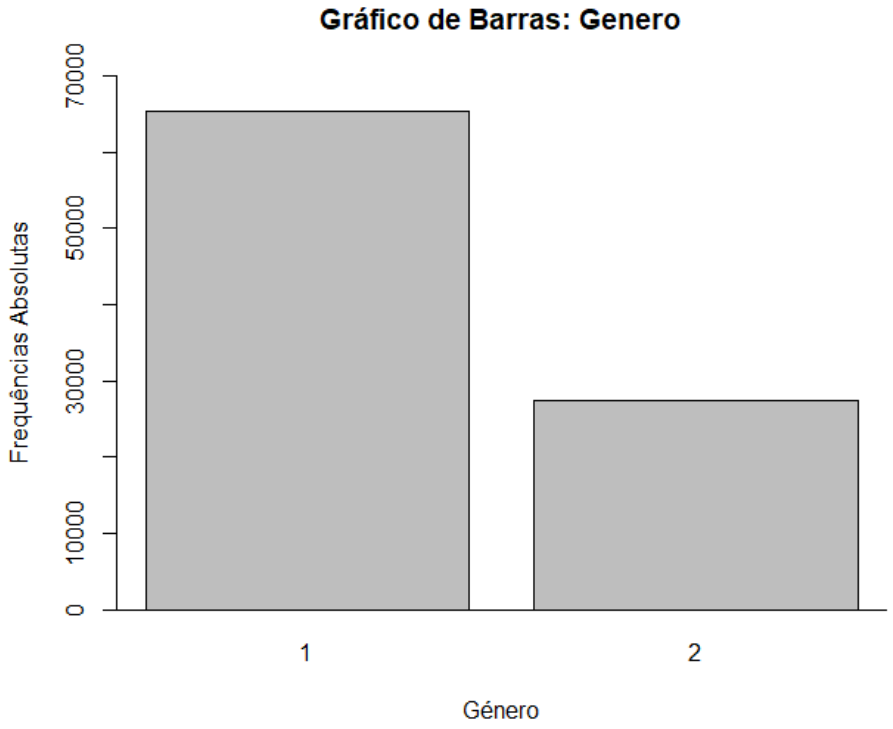

Gráfico 8- Gráfico de barras das frequências absolutas da variável gender

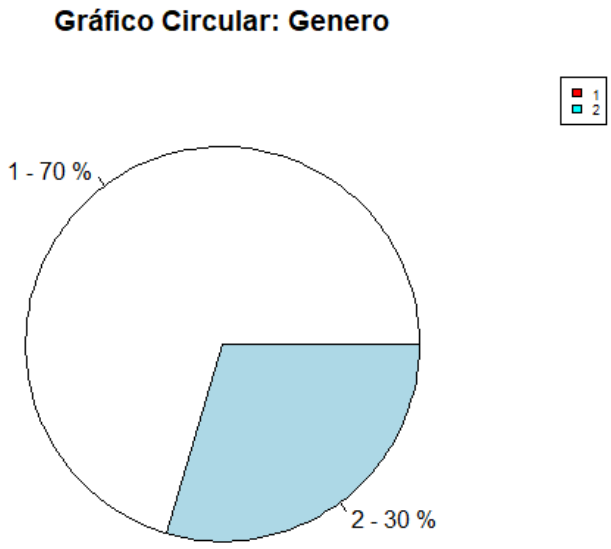

Gráfico 9-Gráfico circular das frequências relativas da variável gender

Podemos visualizar como se comportam as medidas de localização e de dispersão respetivamente nas tabelas 15 e 16.

Medidas de Localização					
Moda	1⇔ Masculino	Quartis		Decis	
Média	Não aplicável[4]				
Mediana	Não aplicável[4]				
		25%	Não aplicável[4]		
		50%	Não aplicável[4]		
		75%	Não aplicável[4]		
				10%	Não aplicável[4]
				20%	Não aplicável[4]
				30%	Não aplicável[4]
				40%	Não aplicável[4]
				50%	Não aplicável[4]
				60%	Não aplicável[4]
				70%	Não aplicável[4]
				80%	Não aplicável[4]
				90%	Não aplicável[4]

Tabela 15 - Medidas de localização da variável gender

[4] Não aplicável quando a variável é qualitativa.

Medidas de dispersão	
Variância	Não aplicável[5]
Desvio Padrão	Não aplicável[5]
Amplitude Total	Não aplicável[5]
Amplitude Interquartil	Não aplicável[5]

Tabela 16 - Medidas de dispersão da variável gender

Na caraterização da distribuição das frequências verificamos o valor da assimetria é de (b1= 0.8921008). Podemos assim afirmar que a assimetria é positiva pois, b1>0.

Relativamente aos valores de curtose, podemos afirmar que a variável apresenta uma curva platicúrtica ou achatada dado que b2<0, (b2= -1.204169).

[5] Não aplicável quando a variável é qualitativa.

Variável year_born

É uma variável quantitativa discreta, representa o ano do nascimento do individuo observado. Por apresentar tantos níveis(90), foi agrupada em classes.

Aplicando se a regra de Sturges consegui-o apurar 17 classes (K=17), as classes são fechadas à direita e a amplitude de cada classe é de 5.411765, pois h= 5.411765. Apresenta a seguinte tabela de frequências.

Apresenta se na seguinte tabela de frequências.

	classes	ni	fi	Ni	Fi
1	[1910,1915]	37	0.0004	37	0.0004
2	(1915,1921]	261	0.0028	298	0.0032
3	(1921,1926]	1766	0.0190	2064	0.0222
4	(1926,1932]	4647	0.0500	6711	0.0723
5	(1932,1937]	11895	0.1281	18606	0.2004
6	(1937,1942]	12397	0.1335	31003	0.3339
7	(1942,1948]	9058	0.0975	40061	0.4314
8	(1948,1953]	8541	0.0920	48602	0.5234
9	(1953,1959]	7807	0.0841	56409	0.6075
10	(1959,1964]	10657	0.1148	67066	0.7223
11	(1964,1970]	7972	0.0859	75038	0.8081
12	(1970,1975]	8218	0.0885	83256	0.8966
13	(1975,1980]	6225	0.0670	89481	0.9636
14	(1980,1986]	2236	0.0241	91717	0.9877
15	(1986,1991]	873	0.0094	92590	0.9971
16	(1991,1997]	244	0.0026	92834	0.9998
17	(1997,2002]	23	0.0002	92857	1.0000

Tabela 17- Tabela de frequências da variável year_born

No histograma do gráfico 11 podemos visualizar a distribuição das observações desta variável.

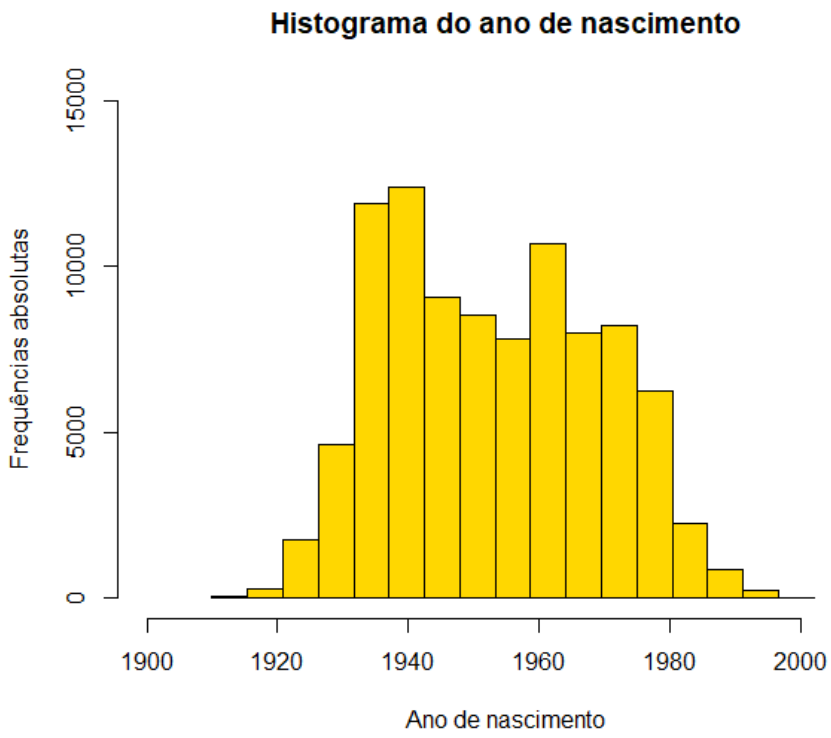

Gráfico 10- Histograma das frequências absolutas da variável year_born

Podemos visualizar como se comportam as medidas de localização e de dispersão respetivamente nas tabelas 18 e 19.

Medidas de Localização					
Moda	1942	Quartis		Decis	
Média	1952.957				
Mediana	1952				
		25%	1939		
		50%	1952		
		75%	1966		
				10%	1933
				20%	1937
				30%	1941
				40%	1946
				50%	1952
				60%	1958
				70%	1963
				80%	1969
				90%	1975

Tabela 18 - Medidas de localização da variável year_born

Medidas de dispersão	
Variância	256.1941
Desvio Padrão	16.00607
Amplitude Total	92
Amplitude Interquartil	27

Tabela 19 - Medidas de dispersão da variável year_born

Na caraterização da distribuição das frequências verificamos o valor da assimetria é de (b1= 0. 0.1768795). Podemos agora afirmar que a assimetria é positiva pois, b1>0.

Relativamente aos valores de curtose, podemos afirmar que a variável apresenta uma curva platicúrtica ou achatada dado que b2<0, (b2= -0.9478836).

Podemos verificar que os quartis da caixa de bigodes tem uma concentração de dados muito uniforme.

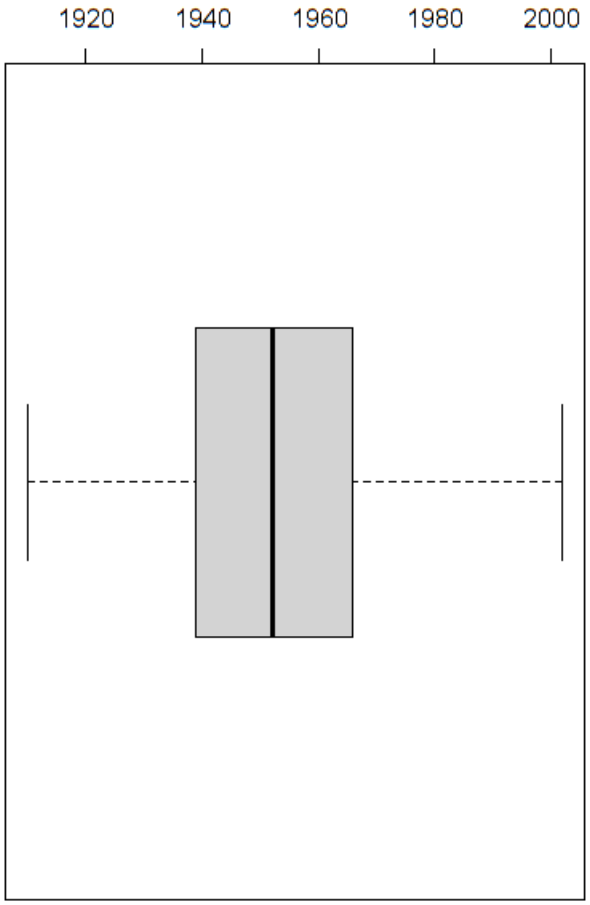

Figura 6- Caixa de bigodes da variável year_born

Variável education_level

É uma variável qualitativa ordinal, representa a escolaridade do individuo observado. Apresenta se originalmente de uma forma discreta como forma de codificação dos valores observados e assume a seguinte forma:

1) no education(under 7 yrs-old) 2) no education(7 & over 7 yrs-old) 3) elementary 4) middle school 5) high school 6) college 7) university degree 8) MA 9) doctoral degree

Apresenta a seguinte tabela de frequências.

i	xi	ni	fi	Ni	Fi
1	2	10858	0.11693249	10858	0.1169325
2	3	21149	0.22775881	32007	0.3446913
3	4	12219	0.13158943	44226	0.4762807
4	5	26181	0.28194966	70407	0.7582304
5	6	5912	0.06366779	76319	0.8218982
6	7	14038	0.15117869	90357	0.9730769
7	8	2221	0.02391850	92578	0.9969954
8	9	279	0.00300462	92857	1.0000000

Tabela 20 - Tabela de frequências da variável education_level

Podemos visualizar as frequências quer relativas quer absolutas da variável nos gráficos 12 e 13.

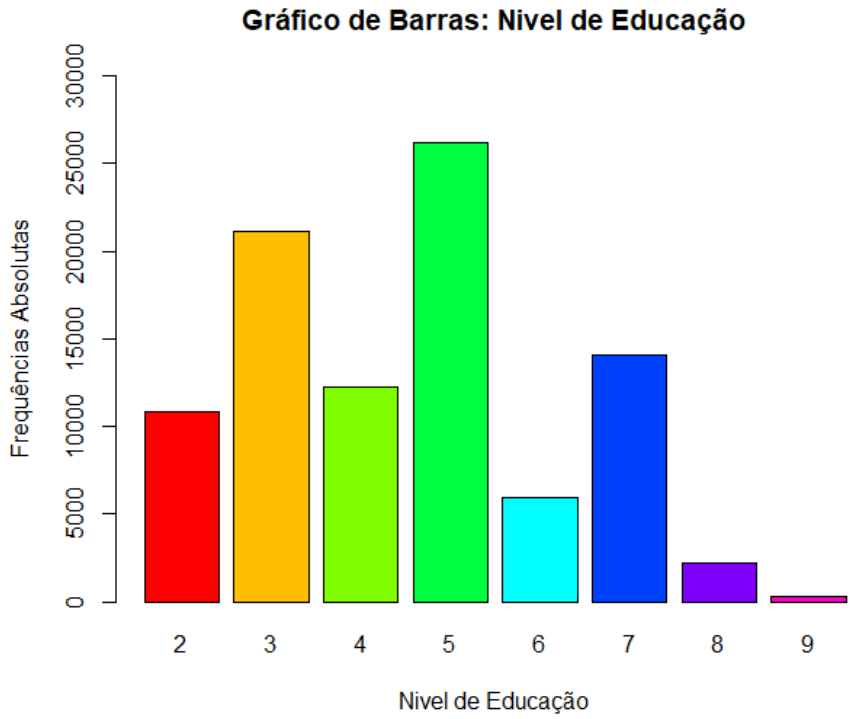

Gráfico 11- Gráfico de barras das frequências absolutas da variável education_level

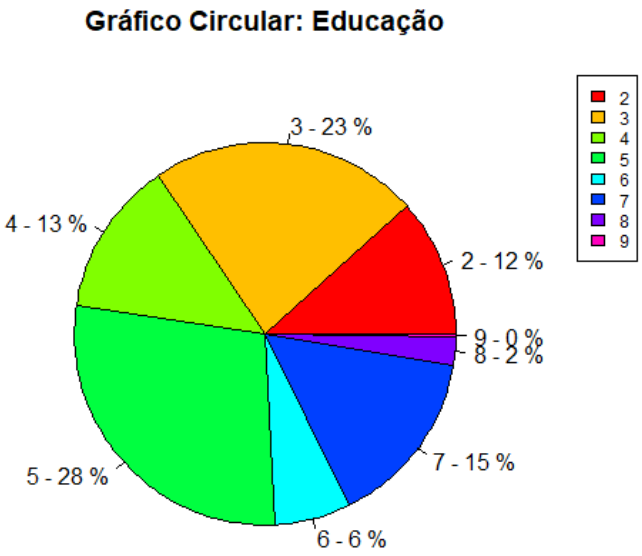

Gráfico 12-Gráfico circular das frequências relativas da variável education_level

Podemos visualizar como se comportam as medidas de localização e de dispersão respetivamente nas tabelas 21 e 22.

Medidas de Localização					
Moda	5⇔ high school	Quartis		Decis	
Média	Não aplicável[6]				
Mediana	Não aplicável[6]				
		25%	Não aplicável[6]		
		50%	Não aplicável[6]		
		75%	Não aplicável[6]		
				10%	Não aplicável[6]
				20%	Não aplicável[6]
				30%	Não aplicável[6]
				40%	Não aplicável[6]
				50%	Não aplicável[6]
				60%	Não aplicável[6]
				70%	Não aplicável[6]
				80%	Não aplicável[6]
				90%	Não aplicável[6]

Tabela 21- Medidas de localização da variável education_level

[6] Não aplicável quando a variável é qualitativa.

Medidas de dispersão	
Variância	Não aplicável[7]
Desvio Padrão	Não aplicável[7]
Amplitude Total	Não aplicável[7]
Amplitude Interquartil	Não aplicável[7]

Tabela 22 - Medidas de dispersão da variável education_level

Na caraterização da distribuição das frequências verificamos o valor da assimetria é de (b1= 0.2576775). Podemos assim afirmar que a assimetria é positiva pois, b1>0.

Relativamente aos valores de curtose, podemos afirmar que a variável apresenta uma curva platicúrtica ou achatada dado que b2<0, (b2= -0.8476749).

[7] Não aplicável quando a variável é qualitativa.

Variável company_size

É uma variável quantitativa discreta, representa o tamanho de uma companhia através do número de funcionários. O estudo incidiu em empresas com um intervalo do número de funcionários de 1 a 99.

Apresenta a seguinte tabela de frequências.

i	xi	ni	fi	Ni	Fi
1	1	28319	0.478304930	28319	0.4783049
2	2	5612	0.094786089	33931	0.5730910
3	3	6497	0.109733646	40428	0.6828247
4	4	2669	0.045079129	43097	0.7279038
5	5	1860	0.031415204	44957	0.7593190
6	6	1346	0.022733798	46303	0.7820528
7	7	3478	0.058743054	49781	0.8407959
8	8	1031	0.017413482	50812	0.8582093
9	9	1097	0.018528215	51909	0.8767375
10	10	6905	0.116624723	58814	0.9933623
11	11	393	0.006637729	59207	1.0000000

Tabela 23- Tabela de frequências da variável company_size

Podemos visualizar as frequências quer relativas quer absolutas da variável nos gráficos 14 e 15.

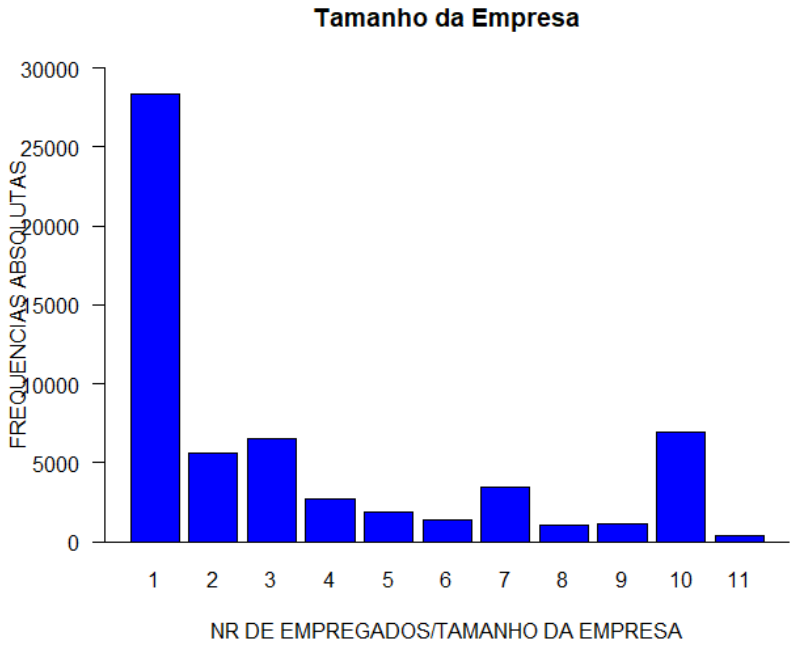

Gráfico 13- Gráfico de barras das frequências absolutas da variável company_size

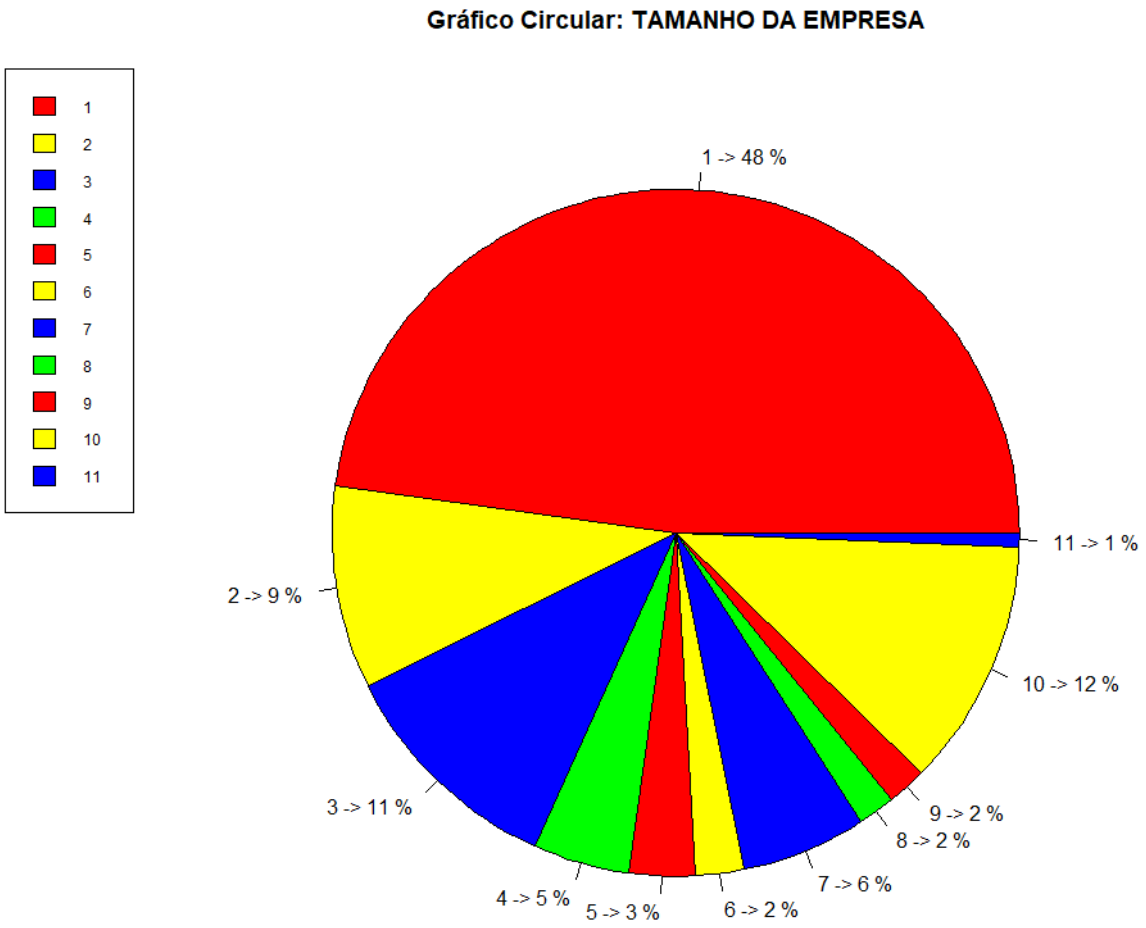

Gráfico 14-Gráfico circular das frequências relativas da variável company_size

Podemos visualizar como se comportam as medidas de dispersão e de localização respetivamente nas tabelas 24 e 25.

Medidas de dispersão	
Variância	10.38252
Desvio Padrão	3.222192
Amplitude Total	10
Amplitude Interquartil	4

Tabela 24- Medidas de dispersão da variável company_size

Medidas de Localização					
Moda	1	Quartis		Decis	
Média	3.427399				
Mediana	2				
		25%	1		
		50%	2		
		75%	5		
				10%	1
				20%	1
				30%	1
				40%	1
				50%	2
				60%	3
				70%	4
				80%	7
				90%	10

Tabela 25 - Medidas de localização da variável company_size

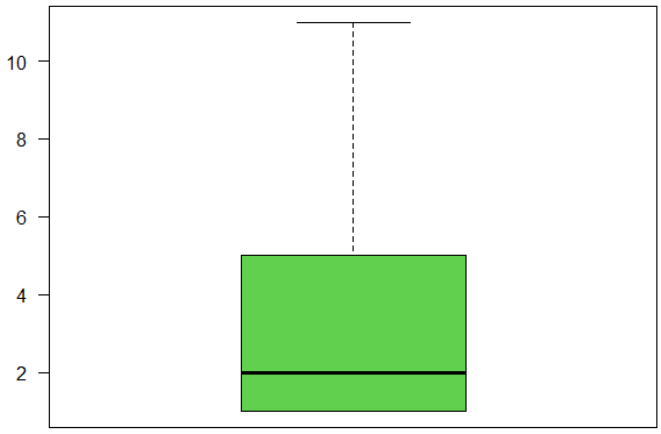

Figura 7-Caixa de bigodes variável company_size

Na caraterização da distribuição das frequências verificamos o valor da assimetria é de (b1= 1.104122). Podemos assim afirmar que a assimetria é positiva pois, b1>0.

Relativamente aos valores de curtose verificamos que a variável apresenta uma curva platicúrtica ou achatada dado que b2<0, (b2= -0.284578).

Na figura 6 podemos visualizar a caixa de bigodes da variável company_size e verificamos o estudo focou se em empresas de pequena dimensão pois a mediana situa se em empresas com 2 funcionários .

Conclusões

Neste projeto abordamos o conjunto de dados intitulado '*Korea Income and Welfare'*, com o objetivo de analisar no âmbito da estatística descritiva 8 das suas variáveis. Com o auxílio do RStudio e através da linguagem de R produzimos um script que nos permitiu caraterizar em detalhe as mesmas.

Cumprimos todos os objetivos a que nos tínhamos proposto nomeadamente a limpeza e o tratamento dos dados, a caracterização das variáveis, tal como a produção de documentos de suporte à análise e pesquisa efetuada.

Este projeto teve uma importância valiosa na aquisição de conhecimentos no âmbito da estatística descritiva e da linguagem de R, pois obrigou todos os elementos deste grupo a pesquisar e analisar e aperfeiçoar técnicas fundamentais nesta área.

Referências bibliográficas

Departamento de Matemática Escola Superior de Tecnologia de Setúbal. Capítulo 1 – Estatística Descritiva. 2021–2022. Materiais de apoio. Disponível em: <https://moodle.ips.pt/2122/mod/resource/view.php?id=3386 >. Acesso em: 10/04/2022