INSTITUTO UNIVERSITÁRIO DE LISBOA

Curso de Licenciatura em Ciência de Dados

1º ano/ 2º semestre - 2023/2024

Unidade Curricular: Análise Exploratória de Dados

Docentes: Maria do Carmo Botelho

Relatório Final

Caso Prático Nº7

TURMA: CDA2

António Santos, nº 123434

Francisco Rosa, nº 123418

Gonçalo Henriques, nº 123422

José Alberto, nº 121959

Pedro Silva, nº 123404

Lisboa, março 2024

Sumário

INTRODUÇÃO

No âmbito da unidade curricular Análise Exploratória de Dados, durante o segundo trimestre 2023/2024, foi-nos atribuída a elaboração do relatório técnico, com o propósito de retratar o percurso adotado no tratamento dos dados recolhidos através do inquérito realizado em Portugal e Espanha, como parte do Estudo sobre os Valores Europeus. Foi nos atribuído o Caso Prático Nº7 que engloba um vasto número de variáveis tais como, dados demográficos, opiniões, emoções e atividades dos inquiridos, com o objetivo de compreender as diferenças culturais e sociais que influenciam os valores contemporâneos nestes países vizinhos.

A metodologia aplicada consistiu em várias etapas de tratamento e análise de dados, inicialmente utilizando o Excel para a preparação e organização dos dados brutos, seguido pela aplicação do Jamovi para análises estatísticas mais profundas. O processo e as ferramentas escolhidas permitiram-nos uma manipulação eficiente dos dados, facilitando uma interpretação detalhada das variáveis em estudo.

Embora a análise específica da informação extraída seja apresentada no relatório complementar em R Markdown, as imagens ilustradas neste documento mostram as etapas metodológicas e os insights obtidos ao longo do projeto. O objetivo deste relatório é, portanto, fornecer uma descrição transparente e sistemática das tarefas realizadas, desde a preparação dos dados até a sua análise exploratória, ilustrando a integração das várias ferramentas e técnicas utilizadas.

Primeira Parte – Excel

Opção de Filtragem

Para facilitar a observação dos dados, a primeira linha onde se encontram os nomes das 15 variáveis foi congelada para permanecer visível enquanto se percorre a tabela. Foram também inseridas opções de filtragem em todas as colunas para se poder escolher e ordenar os dados conforme necessário.

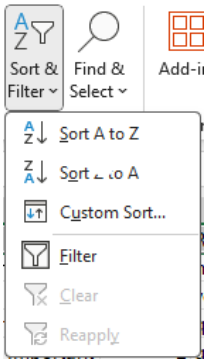


Fig.1 – Opção de filtragem utilizada.

Atribuição de código a uma variável com valores omissos

Seguidamente, para que a variável "v234" não ficasse com 10 espaços em branco, atribuiu-se o código de 99, para uma melhor compreensão imediata destes espaços correspondentes. Para tal, foi selecionada a respetiva variável, e no separador HOME escolheu-se o comando FIND & SELECT e a funcionalidade FIND AND REPLACE, a barra "Find What" deixou-se em branco, de modo que fosse selecionado todos os espaços omissos e com o código usado na barra "Replace with" fosse substituído por "99".

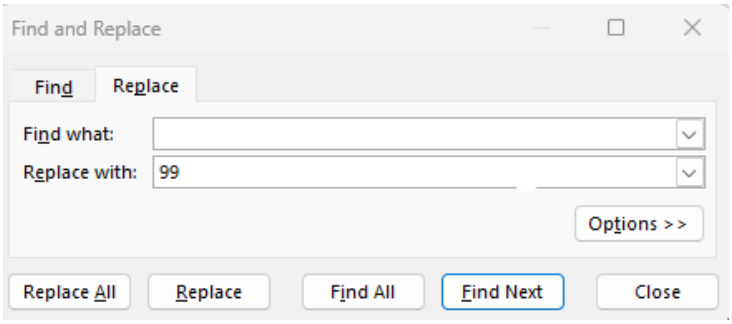


Fig.2 – Utilização do comando Find and Replace para valores omissos

É importante salientar que as restantes variáveis que, mesmo após o tratamento em Excel, apresentavam espaços omissos foram trabalhadas posteriormente em R.

Atribuição de texto a variáveis numéricas

Seguidamente, passou-se à atribuição de texto (valor nominal) a variáveis que continham código (valor numérico), criando-se novas variáveis ao lado das mesmas. Para tal, foi usada a sheet "Variáveis e Códigos" que continha os códigos respetivos e as funções ***HLOOKUP*** e ***VLOOKUP***. O seguinte procedimento foi realizado para a variável v1 e aplicado de modo semelhante (mudando apenas o Lookup_value e a Table_array) para as variáveis v2, v3, v4, v5, v6 e v234 criando-se assim as variáveis: ***Work***, ***Family***, ***Friends***, ***Leisure***, ***Politics***, ***Religion*** e ***Status***, respetivamente.

Este procedimento consiste em criar uma nova coluna à direita da variavél v1, selecionando a opção "Insert" após a seleção da coluna encontrada à direita da mesma (variável v2). Após isto, na primeira célula desta nova coluna (D2) foi introduzida a seguinte função:

Fig.3 – Utilização da função VLOOKUP para atribuição de texto

Tal como na função ***HLOOKUP***, a função ***VLOOKUP***, recebe como parâmetros; a célula com o código que é pretendido substituir; a tabela com a respetiva correspondência do código (fixando as linhas, pois o código foi copiado em linha); a coluna onde se encontra a correspondência do código e opcionalmente foi utilizado para as variáveis de v1 a v6 e v234 o parâmetro "Range_lookup", pois estas apresentavam valores negativos/valores não possíveis de resposta, especificando-se "FALSE" para que a correspondência fosse exata do valor devolvido, evitando assim a indução em erro (que poderia acontecer se não fosse especificado este parâmetro).

Seguindo a mesma ideologia para as variáveis numéricas v21 e v225 foi utilizado o mesmo procedimento, com a diferença de que o código para valor nominal encontrava-se na horizontal e, como tal, foi utilizada a função HLOOKUP do seguinte modo:

Criou-se, assim, as novas variáveis, ***Voluntary*** e ***Gender***, respetivamente.

Correção de Erros

Após a criação destas novas variáveis por atribuição de valor nominal às variáveis com valor numérico, às quais faria sentido a sua correspondência, verificou-se o erro #N/A (como indicação de uma impossibilidade de correspondência), nas variáveis ***Work*** e ***Status*** devido a um "erro" que se encontrava na sua coluna de código correspondente. Na variável v1 (correspondente à variável ***Work***) identificou-se um valor que não se encontrava nas "possíveis respostas" sendo este o atributo "11", que assumimos como um erro, assim apagou-se este valor (desaparecendo o erro #N/A) e mais tarde, como já mencionado, este atributo foi trabalhado em R. De modo semelhante, na variável v234 (correspondente à variável ***Status***) como já era de esperar, identificou-se os 10 valores de código 99 atribuído, pelo que, apagou-se os 10 erros #N/A correspondentes na variável ***Status*** (ficando estes em branco).

Outros erros foram corrigidos nas variáveis v7 e v226 correspondentes ao grau de felicidade com que as pessoas se avaliam e o ano de nascimento das mesmas, respetivamente. Na variável v7, foi aplicado o mesmo método para a variável v1, mas desta vez o atributo encontrado foi "happy" que também não era uma "possível resposta" e, como tal, foi apagado este valor, totalizando 7 valores omissos desta variável.

No que diz respeito ao ano de nascimento dos inquiridos (v226), primeiramente foi criada a variável ***Age*** à direita da mesma, pois para uma análise mais imediata é preferível a utilização das idades, subtraindo a data atual no momento (2017) e o respetivo ano de nascimento. Seguidamente identificou-se um valor de "1800" na variável v226 e uma idade de 217 para a nova coluna criada (***Age***), pelo que considerou-se um erro e então apagou-se estes atributos. Como ainda existiam 2 valores omissos na variável v226 a sua correspondência era 2017 para a nova variável criada, pelo que também foram apagados estes valores, totalizando 3 valores omissos nestas variáveis.

v234	Status
99	#N/A
99	#N/A
99	#N/A
99	#N/A
99	#N/A
99	#N/A
99	#N/A
99	#N/A
99	#N/A
99	#N/A

Fig.5 – "Erro" na variável Status

v1	Work
11	#N/A

Fig.4 – Identificação do erro em v1

v226	Age
	2017
1800	217
	2017

Fig.6 – Erro na variável v226

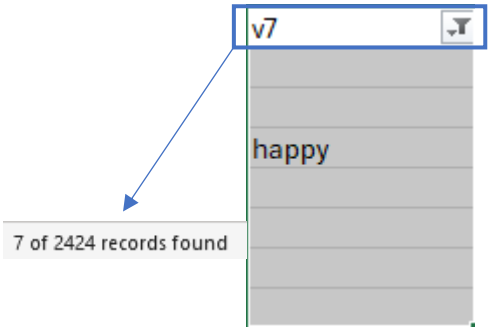


Fig.7 – "Erro" na variável v7

Criação de Tabelas Descritivas

Após a criação da variável ***Age***, foram criadas tabelas descritivas para uma melhor compreensão das idades dos inquiridos. Para tal foram criadas 3 tabelas descritivas (uma para as idades em geral dos inquiridos, outra para Portugal e uma última para Espanha) e 1 PivotTable (tabela dinâmica) com o agrupamento de idades, na sheet "Descrição_Idades".

Para a criação das tabelas descritivas, foi utilizado o add-in "Data Analysis" e o tool "Descriptive Statistics" encontrado no separador DATA após a seleção da variável ***Age***, filtrando pelo país desejado. Para a tabela dinâmica, selecionou-se a mesma variável; inseriu-se uma PivotTable; agrupou-se as idades (através da ferramenta "Group") em intervalos de 10 em 10 começando na idade mínima (15) até aos 75 anos; adicionou-se o filtro para a variável ***Country***, e foram usados os valores de contagem (coluna n) e de percentagem acumulativa (coluna %acum), como mostra a figura 9.

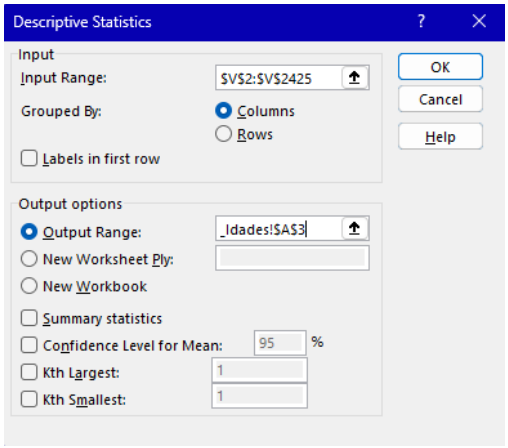


Fig.8 – Criação das Tabelas Descritivas

Country	(All)	
Row Labels	**n**	**%acum**
15-24	209	8,6%
25-34	303	21,1%
35-44	404	37,8%
45-54	397	54,2%
55-64	438	72,3%
65-75	409	89,2%
>75	261	100,0%
Grand Total	**2421**	

Fig.9 – Tabela Dinâmica de frequências

Regra de Validação para Variáveis

No que diz respeito às regras de validação, foram criadas para as variáveis ***v1*** (qualitativa ordinal) e ***v239_r*** (quantitativa discreta), de modo que se fosse inserido qualquer valor indesejado aparecesse uma mensagem de erro.

Assim, para a variável ***v1*** aplicou-se o critério que o valor inserido teria de estar na lista de valores possíveis para esta variável, encontrada no range de C5:C10 na sheet 'Variáveis e códigos'.

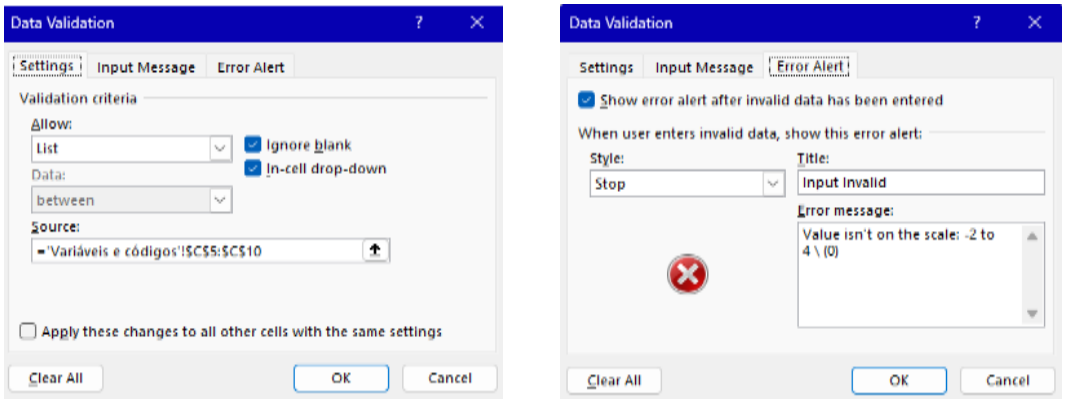


Fig.10 – Validação para a variável v1

Para a variável ***v239_r***¸que diz respeito ao número de filhos dos inquiridos, aplicou-se o critério que o valor inserido não poderia ser negativo para esta variável, logo teria de ser maior ou igual a 0.

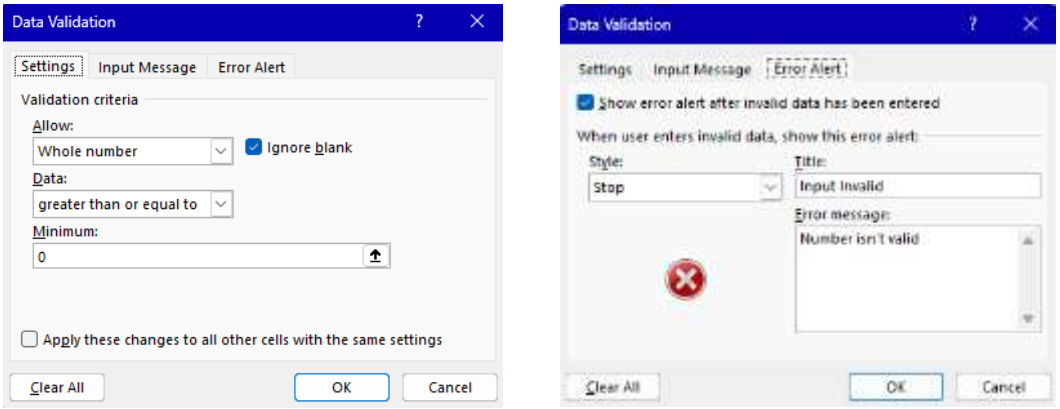


Fig.11 – Validação para a variável v239_r

Tabelas de Frequências Absolutas

Posteriormente para a construção de tabelas de frequências foi utilizado como recurso as funções de contagem (***COUNTIF*** E ***COUNIFS***) para a criação das colunas "n" e para a percentagem selecionou-se a célula à esquerda correspondente e dividiu-se pelo total, por fim, formatou-se a célula como percentagem. Após as formatações manuais obteve-se as tabelas encontradas na sheet "Tabela_Freq_Abs".

Tabelas de Frequências Dinâmicas

Mais uma vez, para a criação da tabela dinâmica encontrada na sheet "Status Civil" foi selecionada a tabela por total e inseriu-se uma PivotTable, colocando em linha a variável ***Status*** e nos valores utilizou-se a contagem (n) e a percentagem normal (%).

Country	(All)	
Status	**n**	**%**
divorced	209	8,7%
married	1156	47,9%
never married and never registered partnership	614	25,4%
registered partnership	57	2,4%
separated	80	3,3%
widowed	298	12,3%
Grand Total	**2414**	**100,0%**

Fig.12 – Distribuições dos Estados Civis

Tabelas de Cruzamento Dinâmica

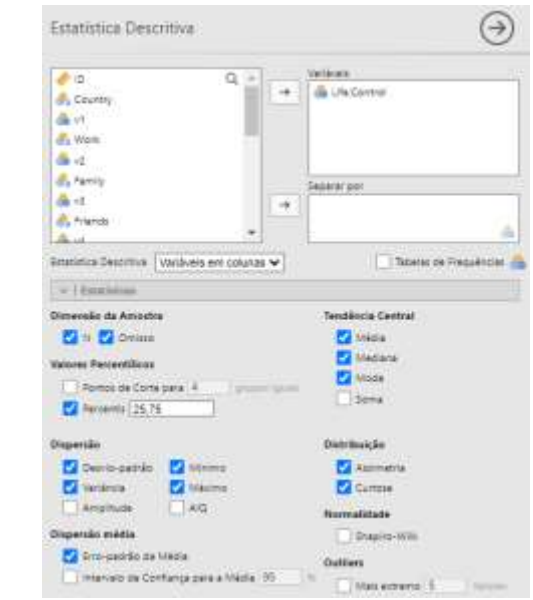
Para terminar o tratamento de dados em Excel, foram elaboradas 5 tabelas de cruzamento dinâmicas entre a variável ***v239_r*** (número de filhos) e algumas das que são classificadas como valorativas, nomeadamente as variáveis ***Work***, ***Friends***, ***Leisure***, ***Politics*** e ***Religion***. Como tal foi selecionado a tabela por total, utilizou-se o comando PivotTable, inseriu-se a variável ***v239_r*** em linha em todas as 5 respetivas tabelas e em cada uma colocou-se em coluna a variável correspondente. Foi utilizado também, para uma melhor compreensão, o valor de contagem (coluna n) e a percentagem em linha (%Linha). É importante salientar que foi ordenado (através da ferramenta "Move") os diversos atributos das diferentes variáveis valorativas colocadas em coluna, seguindo a ordem encontrada na sheet "Variáveis e códigos". Assim ficaram criadas 5 tabelas dinâmicas na sheet "Tabela_Cruzamentos".

SEGUNDA PARTE – JAMOVI

No que diz respeito ao Software Jamovi, que permite uma análise mais imediata dos dados, foi dada a importação dos dados já tratados e limpos através do R de modo a puderem serem criadas tabelas descritivas, gráficos de cruzamento entre variáveis e gráficos de barras, permitindo retirar insights valiosos e concretos.

No que toca a uma preparação de dados neste Software, apenas foi ordenado os diferentes atributos das variáveis valorativas (as variáveis em código e em texto) criadas em Excel.

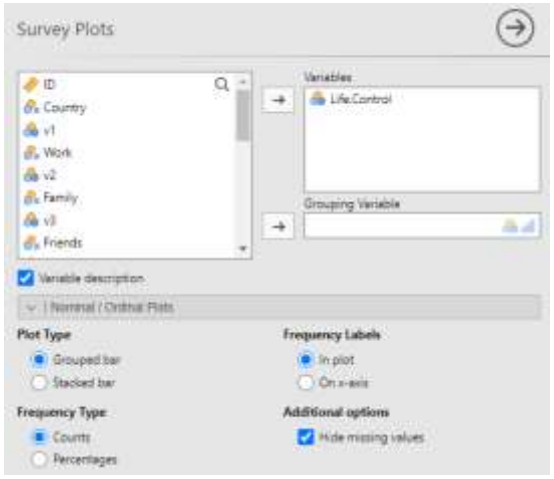
Começamos por criar a tabela descritiva (univariada) referente ao número de filhos e para uma melhor análise criou-se um Survey Plot da tabela correspondente.



Estatística Descritiva

Estatística Descritiva

	Life.Control
N	2424
Omisso	0
Média	7.24
Erro-padrão da média	0.0374
Mediana	7.00
Moda	8.00
Desvio-padrão	1.84
Variância	3.39
Mínimo	1
Máximo	10
Assimetria	−0.587
Erro-padrão da Assimetria	0.0497
Curtose	0.449
Erro-padrão da Curtose	0.0994
25º percentil	6.00
75º percentil	8.00



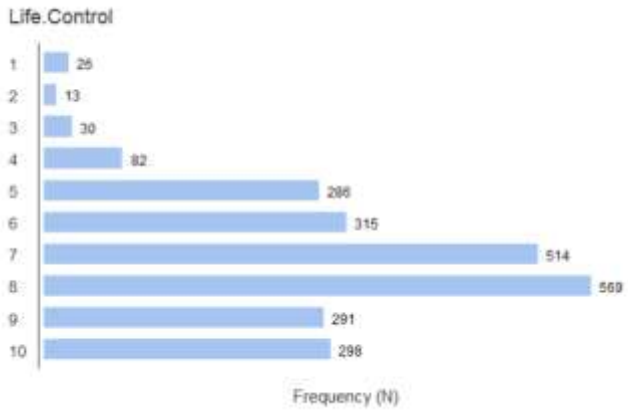
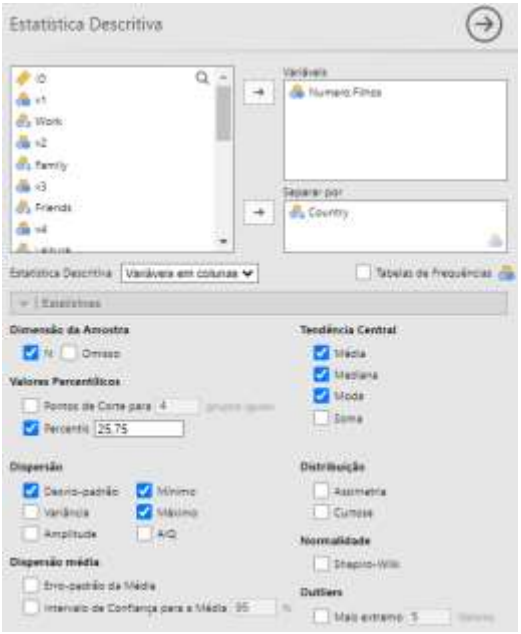


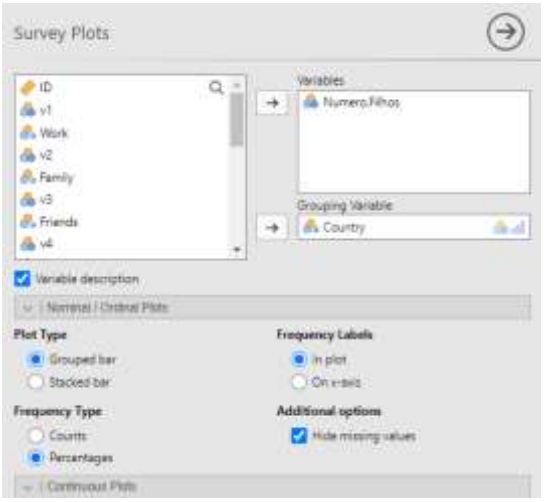
Fig.13 - Análise descritiva e gráfica da variável ***Life.Control***

De seguida criou-se uma tabela descritiva (bivariada) referente ao número de filhos em Portugal e Espanha e para uma melhor análise criou-se um Survey Plot da tabela correspondente.



Estatística Descritiva

	Country	Numero.Filhos
N	Portugal	1215
	Spain	1209
Média	Portugal	1.58
	Spain	1.29
Mediana	Portugal	2
	Spain	1
Moda	Portugal	2.00
	Spain	0.00
Desvio-padrão	Portugal	1.30
	Spain	1.21
Mínimo	Portugal	0
	Spain	0
Máximo	Portugal	5
	Spain	5
25º percentil	Portugal	0.00
	Spain	0.00
75º percentil	Portugal	2.00
	Spain	2.00



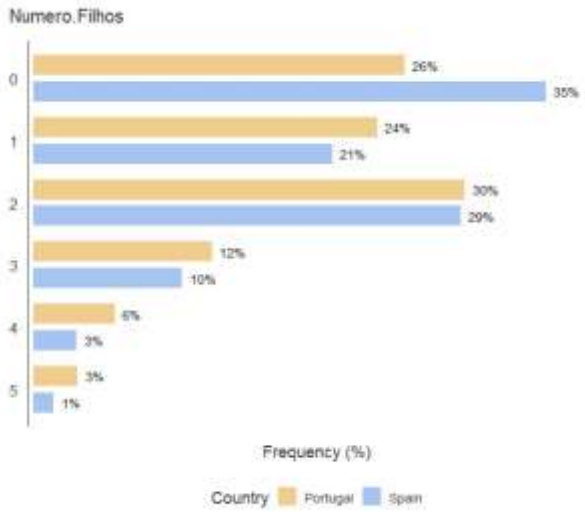
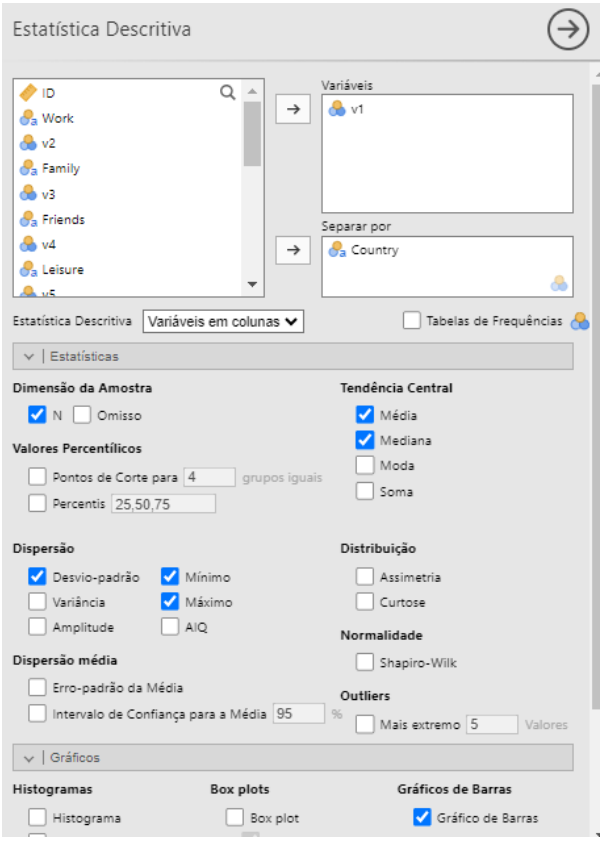


Fig.14 - Análise descritiva e gráfica do Número de Filhos por país

Ainda analisando as diferenças entre países, cruzou-se esta variável ***Country*** com a variável em código que diz respeito à importância no trabalho (v1), onde foi usada a variável numérica para tornar possível este cruzamento. Elaborou-se assim a tabela descritiva e o gráfico de barras .



Estatística Descritiva

Estatística Descritiva

	Country	v1
N	Portugal	1215
	Spain	1209
Média	Portugal	1.56
	Spain	1.35
Mediana	Portugal	1
	Spain	1
Desvio-padrão	Portugal	0.775
	Spain	0.618
Mínimo	Portugal	-2
	Spain	1
Máximo	Portugal	4
	Spain	4

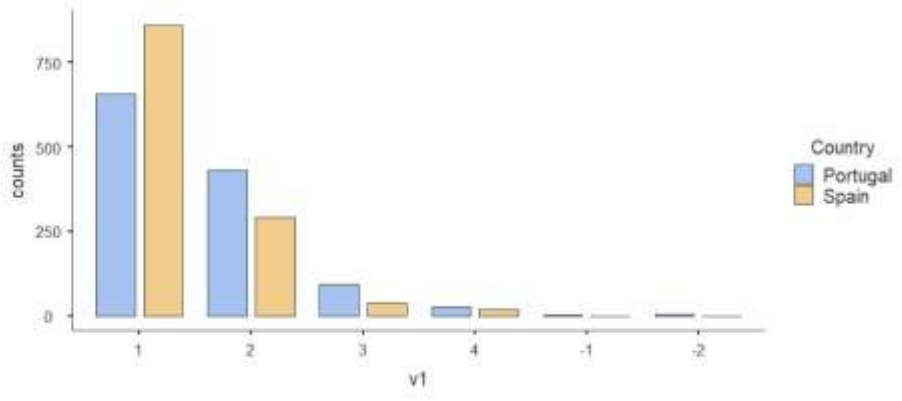
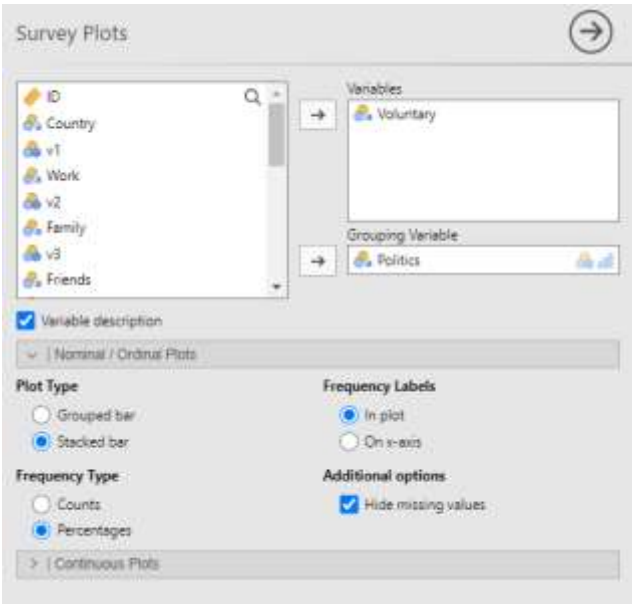


Fig.15 - Análise descritiva e gráfica da variável correspondente à importância do trabalho por país

Foi, também, feita um cruzamento entre as variáveis ***Voluntary*** e ***Politics***, ambas categóricas. Com este cruzamento criou-se, apenas, o Survey Plot com frequência em percentagem.



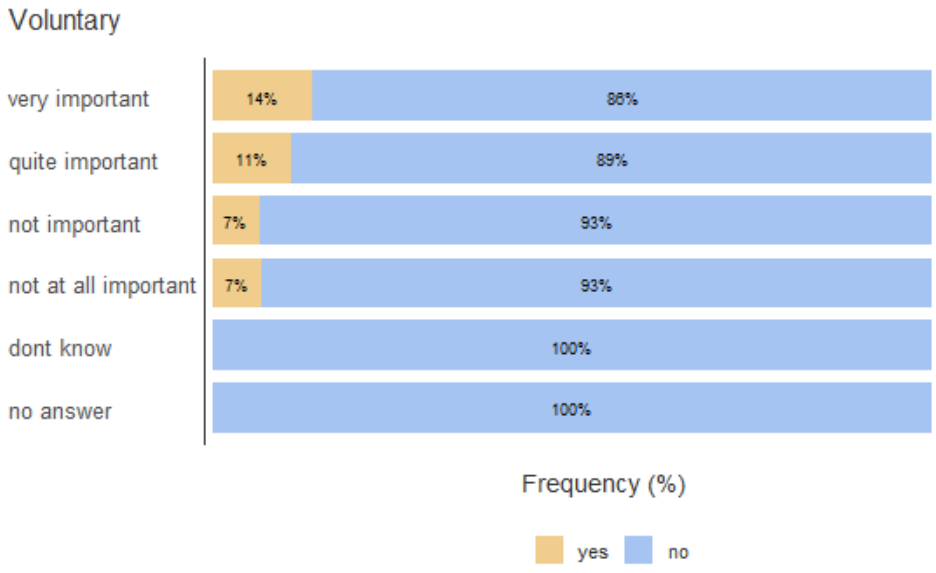


Fig.16 - Análise gráfica entre ***Voluntary*** e ***Politics***

Por fim, criou-se, ainda, um Survey Plot cruzando a variável ***Age*** com a ***Work***.

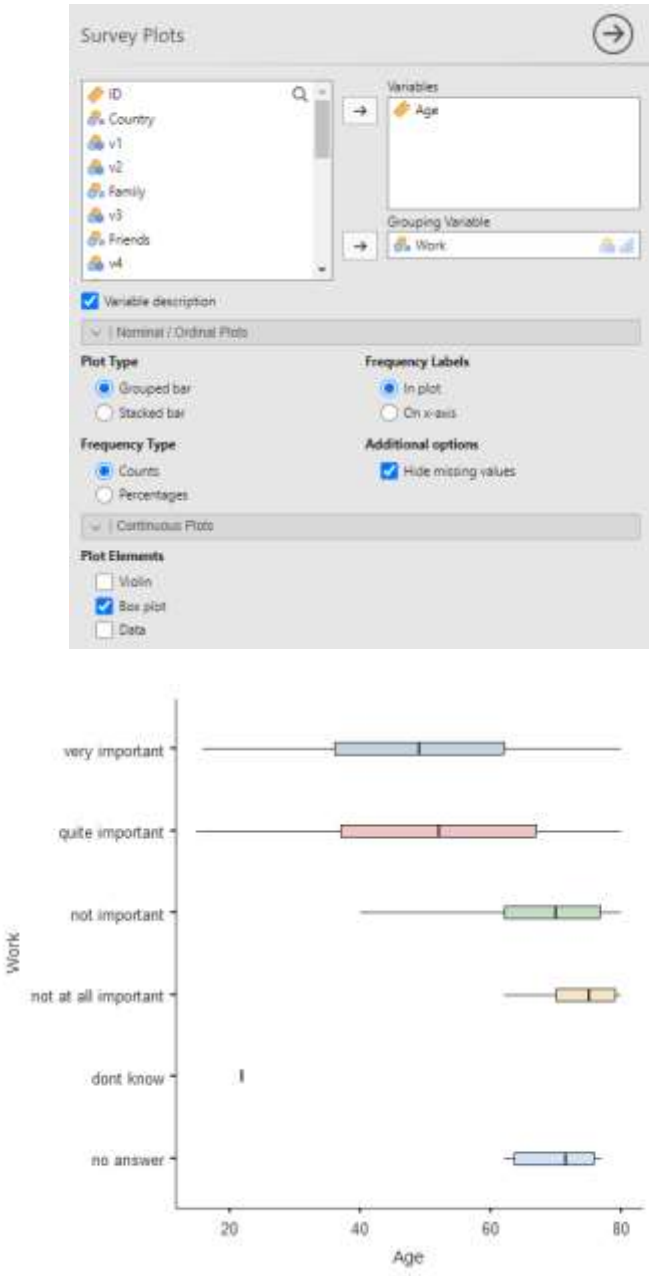


Fig.17 - Análise gráfica entre ***Age*** e ***Work***