DSL SPSS

Part II: Data Exploration

Welcome to the second SPSS workshop from the Digital Skills Lab at LSE.

In the first workshop, we have learned how to open SPSS datasets and work with Excel and CSV files in SPSS. We have also explored the SPSS interface and got familiar with recoding variables and computing new ones. In this session, we will start building your data analysis toolbox and learn the basics of Exploratory Data Analysis (EDA, for short). By the end of this session, you will learn how to:

- explore categorical variables
- visualise categorical variables
- explore continuous variables
- visualise continuous variables
- explore joint distributions of variables

We hope you will enjoy this workshop!

Table of Contents

- Intro to EDA
- Categorical variables
- Continuous variables
- Working with two variables
- Course survey

Getting started

We will be using the same dataset we have used in previous sessions, the ESS (European Social Survey) data that is called **ESS5GB.sav**.

Open this file now, either by double clicking on the file in finder (Mac) or Windows Explorer, or using *file -> open -> data* in SPSS.

Follow the example steps provided, then have a go at the challenges for each section.

Intro to EDA

EDA or Exploratory Data Analysis is a key step in any data analytics or research task. Without a proper understanding of your data, you cannot conduct profound research. Hence, it is very important to explore and interrogate your data before building models and moving to something more advanced. In this workshop, we will learn how to explore different types of variables and their combinations.

You might remember levels of measurement from the previous session, such as scale, ordinal and nominal. These are SPSS names. Usually, in social sciences and data science, we call them continuous and categorical variables.

Categorical variables are the ones that imply *separate* categories as values. Hence, the value of this variable cannot be in-between two values. For instance, a city a person was born in is a categorical variable. A person cannot be born a bit in London and a bit in Manchester. They are born in either one or another. In SPSS there are two types of categorical variables:

1. **Ordinal** - those that imply an order of categories. Example: highest achieved level of education (High School, Bachelors's Degree, Master's Degree, etc.). You can order these categories from lowest to highest.
2. **Nominal** - those that imply the absence of the order of categories. Example: favourite colour (Green, Blue, Yellow, etc.). There is no order of these colours, none of them is "higher" or "lower" than any other.

Continuous (or interval) variables are the variables that imply the existence of an interval. A value of the variable can be anywhere in this interval. For instance, your age is a continuous variable. You can be 18 or 34 years old. But also you can be 18.5 or 34.6 years old. The level of granularity differs from variable to variable, but the key logic remains.

In SPSS continuous or interval variables are called **Scale** variables.

💡 Take a minute and think of examples of continuous variables. Name at least two and discuss them in the breakout room or with the person next to you.

Let's now shift to particular variable types and understand how to explore them.

Categorical variables

You might understand that as the categorical variables imply that their categories are separate, we **cannot** use such descriptives as an average or a standard deviation. Imagine if

you are working with a variable with people's favourite colours, where "Green" is encoded as 1, "Red" as 2, and "Blue" as 3. Let's say you calculate an average and you get 2,3. What does it mean? It is meaningless as (a) these categories are separate and act as independent entities, and (b) there is no order of these categories. To explore categorical variables we can use several methods.

Frequencies

The first way to explore a categorical variable is to merely calculate the frequencies of categories. To do that go to `SPSS -> Analyze -> Descriptive Statistics -> Frequencies`. Put the needed variable into the **Variable(s)** box and click **OK**. This will give you 2 tables. The first one will show you how many valid and missing values the variable has. The second one shows category frequencies, percent of each category in the dataset, valid percent that takes missing values into account, and a cumulative percent.

> What you can also do while exploring a categorical variable is to calculate its mode. Mode is the value of a categorical variable that is encountered the most. You can include a mode in your exploratory analysis by going to `SPSS -> Analyze -> Descriptive Statistics -> Frequencies -> Statistics...` and ticking the **Mode** box there.

Challenge 1: Calculating frequencies

Find frequencies and shares (percentages) of categories in a `aesfdrk` variable - *Feeling of safety of walking alone in local area after dark*.

Bar charts

Frequencies are a good way to show frequencies and/or percentages in your categorical variables. However, sometimes you wish to make it more visual. In this case, you might consider using bar charts. To build a bar chart go to `SPSS -> Graphs -> Legacy Dialogs -> Bar...`. In the appeared window select **Simple** and **Summaries for groups of cases** (!do pay attention to both!). In the new window, you will see an area called **Bars represent**. You can choose what you want these bars to show using this area. It can be the number of cases in a category, a share of a particular category and many more. For example, if you choose *N of cases* each bar in the bar chart will show how many observations of a particular category there are in this dataset. Once you decided what exactly you want to show, select a variable that you wish to explore and put it in the box called **Category Axis**.

Challenge 2: Building a bar chart

For the same variable that you used for Challenge 1, build a bar chart, where each bar represents a share of a category in the dataset.

Continuous variables

Now when we have learned how to work with categorical variables we can move to continuous ones. Continuous variables open more possibilities for exploration such as calculating means, variances, ranges, etc. The quickest and easiest way to access a variety of descriptive statistics for a continuous variable is to go to `SPSS -> Analyze —> Descriptive Statistics -> Explore...`. Put your variable(s) in the box called **Dependent list**. Then, we would recommend selecting **Statistics** in the box called *Display* to avoid bulky visualisations. Once you did that click **OK**. In the resulting tables, you will get a lot of important information about your variable ranging from its mean, median, variance to skewness and kurtosis. This is an easy and useful tool to explore your continuous variables.

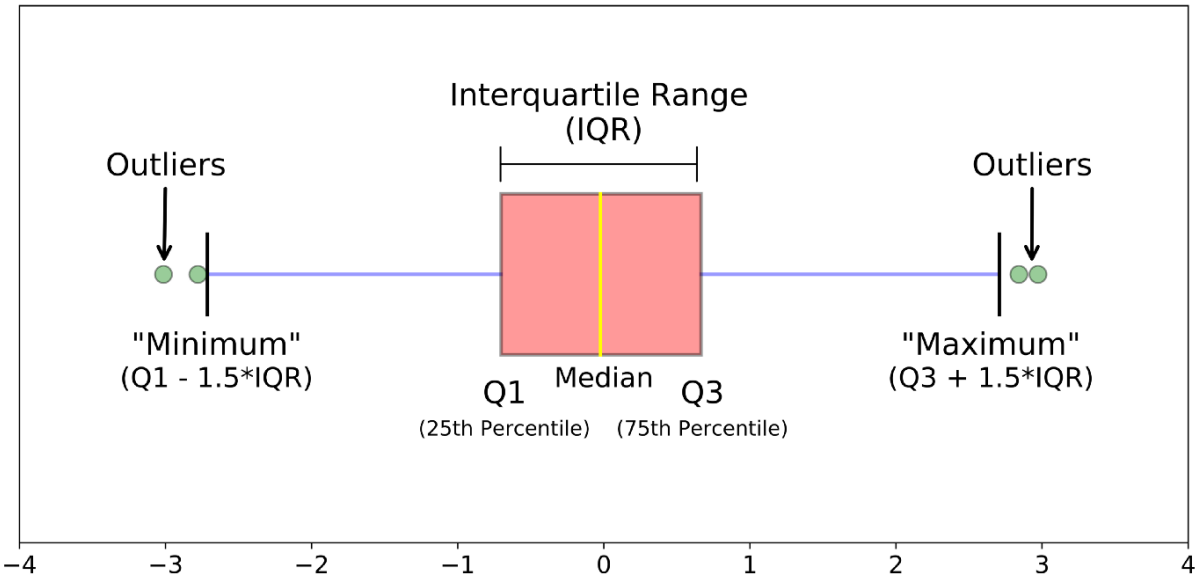
Challenge 3: Exploring trust

Take the variable called `agea` - *Age of respondent, calculated* and explore the descriptive statistics of this variable. What is the mean age in this dataset?

Box plots

As always you might wish to visualise your findings. For continuous variables there are several options to explore them visually:

1. **Box-plots**: `SPSS -> Graphs —> Legacy Dialogs -> Boxplot...`. Next, you select **Simple** and **Summaries of separate variables**. After clicking **Define** you can put your variable(s) in the box called **Boxes Represent** and click **OK**. The elements of the boxplot are explained below:



Source: Understanding Boxplots

1. **Histograms**: `SPSS -> Graphs —> Legacy Dialogs -> Histogram...`. You can put your variable(s) in the box called **Variable** and click **OK**.

These two options will provide you with a nice overview of a variable distribution. Now it's your turn to practice!

Challenge 4: Visualisations of continous variables

Take the variable called `agea` - *Age of respondent, calculated* and visualise it using first a boxplot and then a histogram. Which of them is more informative?

Another useful trick that we wish to show you is exploring a variable distribution across groups. For instance, you wish to see how alcohol consumption is different between men and women. You could build two boxplots — one for women, one for men - and compare them. You can build them side-by-side by going to `SPSS -> Graphs —> Legacy Dialogs -> Boxplot...`. Next, you select **Simple** and **Summaries for groups of cases**. In the **Variable** box you put a variable you wish to explore and in the **Category Axis** box you put a variable with groups that you want to use. Let's try and do it!

Challenge 5: Comparing distributions

Take the variable called `trstplc` - *Trust in the police* and compare distributions of trust between men and women (using the variable called `gndr` - *Gender*). Who trusts the police

more?

Working with two variables

In the last challenge, you have already worked with two variables at the same time. You added a new dimension and explored how men and women are different in their trust in the police. However, there are other ways to explore two variables together. The simplest one is **crosstabulation** that works with two **categirical!** variables. The idea behind crosstabulation is to create a table where rows represent one variable and columns represent another. Let's consider an example. We wish to see whether men and women feel differently walking home at night. We can create a table where columns show gender and rows show how safe a person feels. To do it we can go to `SPSS -> Analyze —> Descriptive Statistics -> Crosstabs...` In the new window you can put the needed variables into columns and rows. Once you put the variables into the corresponding boxes, simply click **OK**. Go ahead and do it as a part of your next challenge.

Challenge 6: Your first table

Create a crosstab where columns use the variable called `gndr` - *Gender* and rows use the variable called `aesfdrk` - *Feeling of safety of walking alone in local area after dark*. Can you spot a relationship between these variables?

Not only can we use simple frequencies in these tables, but also we can use percentages. To do that go to `SPSS -> Analyze —> Descriptive Statistics -> Crosstabs...` and in the appeared window click the **Cells** button. In the new window, you will find an area called *Percentages*. There you can select *row*, *column*, or *total* percentages. This can be done to understand different distributions. For instance, in our case, you can use row percentages to see how men and women are distributed across different levels of safety.

Challenge 7: Spotting a relationship

Create a crosstab where columns use the variable called `gndr` - *Gender* and rows use the variable called `aesfdrk` - *Feeling of safety of walking alone in local area after dark*. Add row percentages to the table. Is there a difference in how men and women feel walking home after dark?

Final challenge - complete the course survey

Click this link to complete the survey