Mini-Project 1 - Exploratory Data Analysis (Guideline)

Timo Meiendresch

WS 20/21

1 Introduction

Welcome to the first mini project of this course. Here, you'll focus on the process called Exploratory Data Analysis (EDA) based on the concepts from our sessions and beyond. EDA is a crucial first step in the analytic process whenever you encounter a new dataset. The three main steps are:

1. Looking at the raw data
2. Calculating (conditional) descriptive statistics of single variables as well as interactions between them
3. Creating data visualizations.

These steps are generally repeated in an iterative circle with the goal to generate questions about your data which you'll then approach to answer by visualizations or models. In each circle, you may learn something new about your data which can then be used to refine your questions, visualizations, and models.

Unfortunately, EDA is not a strict recipe that automatically leads to the desired outcome as it is a creative process. Feel free to investigate whatever idea comes to your mind and come up with methods that are useful in light of your questions. For every data visualization or model, keep in mind the question that you want to answer.

As your EDA progresses, investigate questions that seem particularly interesting to you and your audience (the course and instructors) and worth to pursue further. EDA is part of any data analysis, whether it is done implicitly or explicitly. In the project, some guidelines are given which may help you, but these should not be understood as strict rules. Focus on telling an interesting story based on your questions and answer them using some neat data visualizations. In addition, practice how to communicate your insights and how to tell a compelling story with data.

Important points for your project:

- Keep it interesting, tell a compelling story with data
- Work question-oriented
- Focus on clean and informative visualizations

It is not always necessary to be able to answer your questions with the given data. Maybe you detect an interesting relationship that you are not able to explain. These are often the most interesting questions and we are particularly interested in hearing about those as well as possible explanation attempts.

Keep in mind that your time to conduct the analysis (one week) as well as to communicate your findings in class (5 minutes per person) is severely limited! We are well aware of this fact and will keep that in mind. Time management is of essence and part of the challenge. We don't expect a perfect analysis to give you a very good grading. However, we expect you to tell an interesting story with data and show interesting aspects of the dataset in your brief presentation. We suggest to limit yourself to 5 to 7 slides of content. A presentation may contain

- 1-2 introductory slides, including the questions you want to answer with your presentation and your storyline,
- 1-2 slides of interesting statistics,

- 2-3 data visualization slides and
- 1 slide to summarize what we have learned.

In the next section, we'll briefly outline how you may conduct and structure your EDA.

2 Exploratory Data Analysis (EDA)

This is the start of the iterative process of the EDA. The goal is to strengthen your understanding of the data in each circle. Part of this is to get to know the variables, identify issues (missing values, outliers) as well as interesting relationships that help inform on possible model construction. You may use this structure in your presentation.

2.1 Look at the data!

- First step as usual: Check out the structure of the data, i.e. variables and observations
  - What variables are in your dataset? Is there a main variable of interest? You may decide on which variable you want to focus in your analysis!
  - Which variables are numerical, categorical?
  - What about their levels, central tendency, variability?
  - How do you think are these variables related to each other?
- Use the `summarize()` and `group_by()` function and report on interesting summary statistics

Note that it is generally helpful to center your analysis around one or two main variables. This becomes particularly useful once you want to check multivariate relations or start modeling. In case you focus on a main variable, clarify what your main variable of interest is and why.

2.2 Develop an understanding of your data

One efficient way to improve your understanding of the data is to explicitly write down questions that guide you through your analysis. These questions are expected to change and improve throughout your analysis.

- Write down questions about the dataset.
- Questions serve as base to organize your next steps
- The better the questions, the better and more precise your analysis will be
- Return to your initial questions repeatedly, come up with follow up questions and refinements to your initial set of questions as well as to check if they remain unanswered

A useful classification of overarching questions (and maybe to structure your presentation):

- **Variation**: What type of variation occurs within my variables?
- **Covariation**: What relationships occur across variables (covariation)?
  - Simple measures of association (joint variability): Covariance and (Pearson's) correlation coefficient
  - What do they measure and mean? Make sure to understand these concepts if you do not already

2.2.1 Variation

Variables generally vary across measurements. For two observations, the values are generally not always the same. For example, you measure the height of person A and the height of person B. These may be equal but - more often than not - vary between these two observations. This is the case for repeated measurements as well as measurements across individuals, subjects or households. Your dataset consists of observations and variables. Key idea here is to investigate the variation of a variable across observations. Note that this is closely linked to the concept of the distribution of a variable, i.e. the distribution of a variable shows the variation!

- Visualize the variation (distribution) of your main variable(s) of interest. What do you see?
  - How to visualize the distribution of a variable depends on whether it is categorical or continuous.
- Possible questions:

  - Which values are common? Why?
  - Do you see rare values or outliers? Why?
  - Unusual patterns? Possible explanation?
  - What about clusters (accumulations) of similar values? Why?
- Show us an interesting pattern based on these questions and present possible explanations
- Explore the distribution of interesting variables and what did you learn?
- Do you discover anything interesting, surprising, unusual?
  - What do you think is the probable cause here?

In case you encounter unusual or missing values: You may drop or replace these values. Motivate your choice between these two options!

2.2.2 Covariation

So far, variation covers the behavior of (or within) one variable, whereas covariation describes the behavior in relation between variables (two or more). A good way to describe this is to visualize it. Which visualization to choose depends on the variable type as you have seen in previous sessions.

- Visualize covariation or association between your variables of interest.
  - Visualize by subgroups (for example `facetting` or side-by-side boxplots)
  - Motivate the choice for your visualization. Which question do you want to answer with the visualization?
- What other variable is most important for your main variable of interest?

This is your time to shine and show your creative and powerful data visualization skills. However, please don't overdo it. There is a fine line between clean, informative, well motivated visualizations (titles, correct labeling of axis, etc.) and cluttered, confusing ones.

2.2.3 Models

Right now, you may be aware of patterns and systematic relationships in your data which you would like to test more rigorously. The way to proceed is to model these relationships. Present possible relationships that you would like to investigate further using models. Questions to guide you here:

- What is the key variable of interest?
  - What is the relation of this variable of interest to other variables?
- What are possible explanations for this relationship?
- How to describe this relationship?
- How strong is the relationship?
- What other variables might affect the relationship?
- Does the relationship change across subgroups?

Models are tools to explain and investigate patterns as well as to quantify and test relationships. However, you are not expected to do this here as this will be part of the next projects.

3 Groups and Datasets

Group	Group members	Dataset
#1	Hantao Hui, Muhammad Atique Nazar, Antonia Pütz	Christmas Music (link)
#2	Henry Lin, Hammad Ahmed	Coffee ratings (link)
#3	Mehtanin Kabir Rashikh, Tauseef Ahmad	Global Crop Yields (link)
#4	Hatim Ali Ashgar, Sibora Domi	Friends (link)
#5	Yongzhao Li (Eric Li), Steljana Lleshi	European Energy (link)
#6	Pavel Raschetnov, Rishab Bhattacharayya	Spotify Songs (link)
#7	Pruthvi Hegde, Md Razaul Haque	Extinct Plants (link)

4 Outcomes and Grading

The outcome is a short presentation (5 minutes per person) in which you'll communicate your results coded in R. Please take turns in your group. It is probably a good idea to limit yourself to roughly 5 to 7 content slides. One slide should be reserved to summarize your presentation, in particular on "What have we learned?".

After each presentation, there will be time for feedback and questions. All groups are encouraged to ask fair questions and give honest feedback. Use this feedback to improve your communication skills in the upcoming mini projects as well as the final project.

We expect to see nice, informative visualizations and a coherent and interesting short story told with your dataset. Creativity and interesting presentations will be greatly appreciated! Make it fun and interesting for everyone. After each presentation, we should have learned something that we did not know before on the topic that you have covered ("What did we learn that we did not know before?").

We are looking forward to your short presentations!