Data Science I

Overall Points: / 100

Exercise 5: Data Visualization & Machine Learning Intro

Submission Deadline: December 15 2025, 07:00 UTC
University of Oldenburg
Winter 2025/2026
Instructors: Jannik Schröder, Wolfram "Wolle" Wingerath

Submitted by: <your names here>

Part 1: Exploratory Data Analysis

/ 25

1.) Provide answers to the questions associated with the following data sets, available at http://www.data-manual.com/data.

a) Analyze the movie data set. What is the range of movie gross in the United States? Which type of movies are most likely to succeed in the market? Comedy? PG-13? Drama? Why?

Solution:

< your answer here >

b) Analyze the Manhattan rolling sales data set. Where in Manhattan is themost/least expensive real estate located? What is the relationship between sales price and gross square feet?

Solution:

< your answer here >

c) Analyze the 2012 Olympic data set. What can you say about the relationship between a country's population and the number of medals it wins? What can you say about the relationship between the ratio of female and male counts and the GDP of that country?

Solution:

< your answer here >

d) Analyze the GDP per capita data set. How do countries from Europe, Asia, and Africa compare in the rates of growth in GDP? When have countries faced substantial changes in GDP, and what historical events were likely most responsible for it?

Solution:

< your answer here >

e) Analyze the following data set on electricity demand: https://www.kaggle.com/datasets/albertovidalrod/electricity-consumption-uk-20092022 . Is energy demand different on national holidays compared to normal weekdays or weekends? Can you identify a general trend?

Solution:

< your answer here >

2.) For one data set of your own choosing, answer the following basic questions:

a) Who constructed it, when, and why?

Solution:

< your answer here >

b) How big is it?

Solution:

< your answer here >

c) Identify a few familiar or interpretable records.

Solution:

< your answer here >

d) Find out and describe what Tukey's five number summary is and then provide one for at least 3 different columns.

Solution:

< your answer here >

e) State at least one interesting or noteworthy thing that you Learned from your data set.

Solution:

< your answer here >

Part 2: Interpreting Visualizations / 25

3.) Search your favorite news websites until you find 4 interesting charts/plots, ideally half good and half bad. For each, please critique along the following dimensions:

a) Does it do a good job or a bad job of presenting the data? Why?
b) Does the presentation appear to be biased, either deliberately or accidentally?
c) Is there chartjunk in the figure? Where?
d) Are the axes labeled in a clear and informative way?
e) Is the color used effectively?
f) How can the graphic be improved?

Solution:

< your answer here >

4.) Visit https://viz.wtf and find five laughably bad visualizations. Explain why they are both bad and amusing.

Solution:

< your answer here >

Part 3: Creating Visualizations & Storytelling / 25

5.) Construct a revealing visualization of some aspect of your favorite data set, using:

a) A well-designed table.

Solution:

< your answer here >

b) A dot and/or line plot.

Solution:

< your answer here >

c) A scatter plot.

Solution:

< your answer here >

d) A heatmap.

Solution:

< your answer here >

e) A bar plot and/or a pie chart.

Solution:

< your answer here >

f) A histogram.

Solution:

< your answer here >

6.) Find and tell a Story with data! To this end, first go to https://uol.de/planung-entwicklung/akademisches-controlling/studium-und-lehre and select one of the following data sets:

- **Studienanfänger*innen / Studienanfänger*innen (Fallstatistik) nach Studiengang**
- **Fachstudiendauer / Übersicht über die Fachstudiendauer**

Then explore the data and find a story to tell with it.

a) Define an audience and a goal.
(Example: A data viz that highlights a potential issue to the university council or one that tries to win new students for a certain subject area.)

Solution:

< your answer here >

b) Create a (communicative) data visualization to help your cause.

Solution:

< your answer here >

Part 4: Machine Learning Intro / 25

7.) Give decision trees to represent the following Boolean functions:

a) A and B.

Solution:

< your answer here >

b) A or (B and C).

Solution:

< your answer here >

c) (A and B) or (C and D)

Solution:

< your answer here >

8.) Consider the following titanic dataset:
https://www.kaggle.com/competitions/titanic/data

a) Load the test and training data sets. Briefly describe the dataset.

Solution:

< your answer here >

b) Train a random forest classifier to predict survival chances for Titanic passengers.
(Hint: You can use one of the tutorials/submissions as a starting point.)

Solution:

< your answer here >

c) Evaluate the performance of your model and iterate on it to improve it!

Solution:

< your answer here >

Finally: Submission

Save your notebook and submit it (as both **notebook and PDF file**). And please don't forget to ...

- ... choose a **file name** according to convention (see Exercise Sheet 1, but please **add your group name as a suffix** like `_group01`) and to
- ... include the **execution output** in your submission!