Project Big Data – Final project

Goal

The goal of this final project is to provide you with practice and experience in exploring a real-life dataset using Python. Your task is to perform a diverse set of analyses to extract knowledge and insights from the dataset and to help the audience learn about the topics represented in the dataset. For this final project, we expect you to ask interesting questions about the data, perform the necessary data processing, visualize trends and relationships, discover correlations (if any), and perform statistical tests of hypotheses.

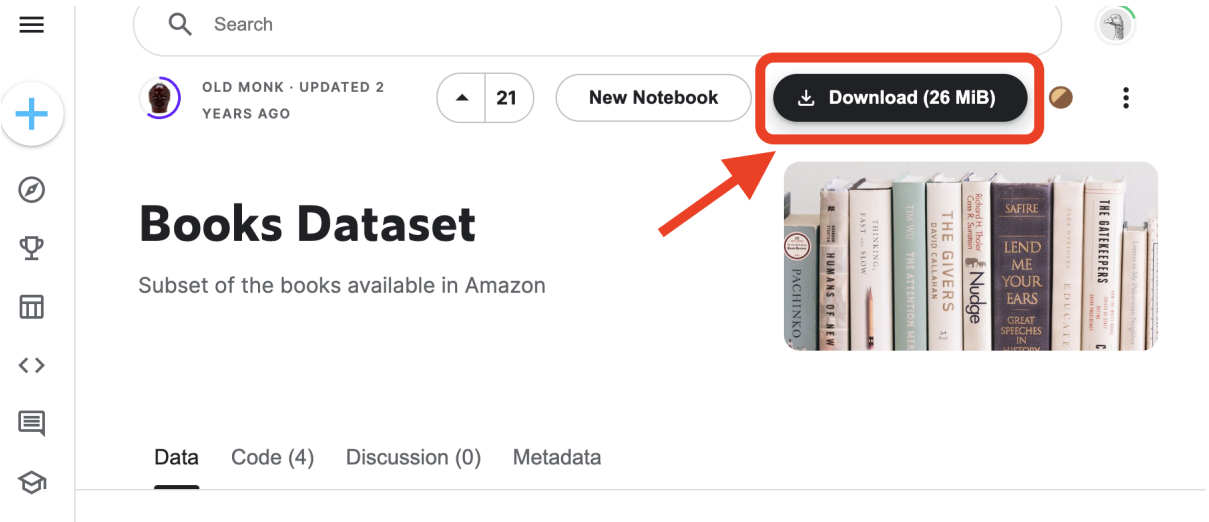
Details about the final project

As specified in the course syllabus, this project consists of a report with an accompanying notebook (35% of your course grade).

Data analysis

All provided datasets are quite interesting to "explore" and rich in potential questions to investigate. In general, we do not specify which analyses you must perform but instead we expect you to ask yourselves (and answer) meaningful and interesting questions. Make sure that the analyses you perform are methodologically valid and that the insights you discover are non-trivial.

Each dataset is provided in the Chapter Data Sets at the bottom of this document. Here you can find a short description and a link to the page where you can find the datasets.
Most datasets are coming from Kaggle. Here you can download a zip file with all the data related to the dataset, by clicking on the button at the top corner (unless mentioned differently in the description):



NOTE: You will need to create an account and log in to download the data.

If you are dealing with a large dataset and the computing time starts to become an issue, feel free to work with a subset of the data and/or to reduce it according to some meaningful criteria. Please comment extensively about this choice in your notebook and in your report.

A few example/starting questions for your data set (which you are free to use or ignore). You can also get a lot of inspiration by looking at case studies on Kaggle or other sources. We encourage you to do this. However, make sure to write your own code and report and try to come up with some original questions. Reference both in the notebook and in the report all the blog posts, forums, or web pages that you consulted/copied code from.
We suggest that you document all questions, answers, and analyses that you perform (assuming they are valid), even in the case of negative/inconclusive results. These are part of any thorough data analysis and this type of findings may still be insightful/useful.

What to submit?

You are required to hand in two files: a final report and a companion Jupyter notebook.

The **report** should be formatted as a PDF document, which must explicitly contain page numbers at the bottom of the page. It should be no more than 15 pages (including all text and figures). The report must have a front page containing the names and student numbers of the authors, the group number on Canvas, the title of the report, the name of the course, and the date of submission. Your final report should be written in English and not contain any Python code, but should include relevant graphs and/or tables for visualization. It is good to mention in the report the packages you have used, especially the new ones (if any). The structure, content, and grading criteria for the report are explained below in this document.

The **Jupyter notebook** should contain all the Python code you used for the final project and show how all data processing, visualizations, and analyses have been performed. Feel free to use any Python packages you want, feel free to explore new ones for the purpose of your analysis. Your Python code will not be graded separately, but will be checked to verify the claims in your report. Make sure that dedicated Markdown cells or comments in your Python code make it easy to navigate between the code and the report. Note that this notebook file must include all the code that you used to obtain the visualizations and results presented in your report.

Both the report and your code must be handed in via Canvas before Sunday July 3rd 23:59. There are two separate assignments on Canvas, one for your report and one for your Jupyter notebook. You must submit one file to each assignment (in .pdf and .ipynb format, respectively). Only one person from your group has to submit and he/she should hand in only one report and one companion Jupyter notebook. If multiple .ipynb or .pdf files are submitted, only the *last submitted ones will be graded*. We suggest that you start looking into the datasets as soon as possible and thinking about what questions you want to investigate and how to split the work within your group.

Grading Criteria for Final Report

1. Formal requirements and written text (10%)
    - The report should be handed in as a PDF document before the deadline.
    - Both the report and the Jupyter notebook have the correct filenames.
    - The report contains a cover page that includes the names and student numbers of the authors, the group number on Canvas, the name of the course, and the date on which the report was handed in.
    - The report contains page numbers in the bottom right corner.
    - The report is no more than 15 pages (including all text and figures).
    - The text has a clear and logical structure.
    - The text is readable and is mostly free from grammatical errors and spelling mistakes.
    - The report contains sources where appropriate (e.g., references, URLs, etc.).

2. Quality and relevance of the questions investigated (20%)
    - There is a sufficient number of interesting questions investigated.
    - The questions that you investigated are relevant for the given dataset and are valuable/insightful/useful for an interested party.

3. Data analyses and interpretations (45%)
    - Your Python code is correct and coherent with the report (i.e., it is error-free and does what you say it does in the report/comments/markdown cells).
    - The appropriate variables are considered for the intended analysis.
    - The appropriate statistical tests have been applied and the corresponding assumptions have been checked.
    - Your interpretations are adequately motivated and, when pertinent, alternative interpretations are discussed.
    - The insights that you identified are supported by the data.
    - You implemented meaningful and relevant regression/classification models, justified why you chose them and reported the results in a clear way.

4. Visualizations (25%)
    - Each graph has an informative caption/title, descriptive labels on its axes, the values on the axes have units, the intervals of the values on the axes are suitable, the graph is clear and legible, the legend (when applicable) is clear and legible. The use of color in the graphs is helpful for understanding the graphs.
    - The graphs are properly sized and properly placed within the report.
    - The graphs are relevant for the analyses performed. Each graph is specifically referred to in the text where the corresponding analysis/conclusions are presented.

Data Sets

There are eleven different data sets available to choose from, listed below:

1) Airbnb
2) Books
3) CardiovascularDisease
4) BikeShare
5) Etsy
6) Fitbit
7) GooglePlay
8) IncidentLog
9) MoviesandSeries
10) News
11) UsedCarPrices

Your group will be assigned to one dataset of your top 5.

Important: File naming convention for submission

The report should be formatted as a PDF document, using the following **naming convention: group*_topic.pdf** where * is your group number for the final project and topic is the keyword of your dataset as mentioned in the list above. For the Jupyter Notebook, please use the similar naming convention: **group*_topic.ipynb**.

Refer to the section Data analysis for more information about analysing the data and what is expected from you.

1) Airbnbs Dutch cities: Amsterdam, Rotterdam, The Hague

The webpage in the provided link contains datasets with the quarterly data for Airbnbs in different cities. For your project, you will investigate the data achieved for the Dutch cities Amsterdam, Rotterdam and The Hague.

Links to datasets:
http://data.insideairbnb.com/the-netherlands/north-holland/amsterdam/2022-03-08/data/listings.csv.gz
http://data.insideairbnb.com/the-netherlands/south-holland/rotterdam/2022-03-23/data/listings.csv.gz
http://data.insideairbnb.com/the-netherlands/south-holland/the-hague/2022-03-23/data/listings.csv.gz
The page containing all links above is:
http://insideairbnb.com/get-the-data/
The files linked above are listings.csv.gz. If you want, you can also use the other data files, such as reviews and neighbourhoods.

Some ideas for analysis
Below are some ideas that could be interesting to investigate. It is not required for you to investigate them, rather it could be a starting point if you cannot think of your own ideas. Some of these ideas on the list are simple, and others may be complicated.

- Which city/neighbourhood gets the best rating? (review_scores_location)
- What is the most expensive city/neighbourhood?
    - How does the most expensive neighbourhood in one city compare to the others?
- Which city/neighbourhood has the most reviews?
- How do the number of amenities affect the price / rating of a room?
- Which room/property type gives the best ratings on average?
- What is the average price per room type?
    - What is the average price per person per night?
- Can you fit one or more relevant regression models for this dataset?
- Can you implement a relevant and suitable classification/clustering model to this dataset?
- If you implement more than one, report which one is best and argue why their performance differs.

2) Amazon Books

Online data for books from Amazon along with user ratings and users who bought them. This data can be used to create recommending systems.

Link to dataset:
https://www.kaggle.com/datasets/saurabhbagchi/books-dataset

Some ideas for analysis
Below are some ideas that could be interesting to investigate. It is not required for you to investigate them, rather it could be a starting point if you cannot think of your own ideas. Some of these ideas on the list are simple, and others may be complicated.

- What is the average rate for a book? What book is the top 10 best/worst rated?
- Which book is most popular by people who dislike the most popular book?
- After categorizing users based on age:
    - Can you find a relation between age group and the average ratings they give? Is there a group which is more optimistic?
    - Are there publishers that seem to focus more on a certain age group?
    - Which book is most liked by each age group?
    - Which author is most liked by each age group?
- Which author is in general the best rated author?
- Which publisher publishes the best rated books?
- Can you fit one or more relevant regression models for this dataset?
- Can you implement a relevant and suitable classification/clustering model to this dataset?
- If you implement more than one, report which one is best and argue why their performance differs.

3) Cardiovascular Disease

Dataset to examine the effect of different features on the presence or absence of cardiovascular disease.

Link to dataset:
https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

Some ideas for analysis
Below are some ideas that could be interesting to investigate. It is not required for you to investigate them, rather it could be a starting point if you cannot think of your own ideas. Some of these ideas on the list are simple, and others may be complicated.

- What is the average age of the people that have been examined?
- What is the average BMI of men / women?
    - Is this different when you separate per age group?
- Are men / women more likely to have cardiovascular disease?
- How does smoking affect your chances of getting cardiovascular disease?
- What is the average gap between Systolic blood pressure and Diastolic blood pressure?
    - Is this different for men / women?
    - Or per age group?
- Can you fit one or more relevant regression models for this dataset?
- Can you implement a relevant and suitable classification/clustering model to this dataset?
- If you implement more than one, report which one is best and argue why their performance differs.

4) Cyclistic Bike Share

These datasets are used for the case study as the capstone project in Google Data Analytics course on Coursera. This is public data that you can use to explore how different customer types are using Cyclistic bikes.

Link to dataset:
https://www.kaggle.com/datasets/evangower/cyclistic-bike-share
This dataset contains 12 large files. This is a lot of data, which makes this case very interesting, but you need to make sure to use it in a smart way, see also Data analysis.

Some ideas for analysis
Below are some ideas that could be interesting to investigate. It is not required for you to investigate them, rather it could be a starting point if you cannot think of your own ideas. Some of these ideas on the list are simple, and others may be complicated.

- In which period of the year are the most bike rentals?
- Looking at longitude / latitude, what is the average distance between start / end?
- Between which stations are the most trips?
- What type of bike is rented more?
    - Is there a seasonal effect?
    - Is there a relationship between the distances / duration?
    - What about membership?
- During what time of the day are most bikes rented? Seasonal effect?
- Can you fit one or more relevant regression models for this dataset?
- Can you implement a relevant and suitable classification/clustering model to this dataset?
- If you implement more than one, report which one is best and argue why their performance differs.

5) Etsy Shops Sales

This data set contains data from Etsy shops. Metrics include shop profiles, sales, products etc.

Link to dataset:
https://www.kaggle.com/datasets/polartech/1m-etsy-shops-sales-data
Use the “Click to download” link to download the data. (instead of the big button on the right top)
Note: The dates are in unix epoch time, see:
https://note.nkmk.me/en/python-unix-time-datetime/#:~:text=Unix%20time%20

Some ideas for analysis
Below are some ideas that could be interesting to investigate. It is not required for you to investigate them, rather it could be a starting point if you cannot think of your own ideas. Some of these ideas on the list are simple, and others may be complicated.

- In what period are most shops created?
- What is the average sales per day, per shop since they have been created?
- Does free shipping increase the number of sales?
- Shops in which country have the most sales?
- Is there a relationship between the number of countries to ship to and the number of sales of a shop?
- What is the relation between the average rating and the number of sales?
- Which country code is included most to ship to?
- Which country code can be best included to ship to?
- Can you fit one or more relevant regression models for this dataset?
- Can you implement a relevant and suitable classification/clustering model to this dataset?
- If you implement more than one, report which one is best and argue why their performance differs.

6) Fitbit

This dataset is generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring.

Link to dataset:
https://www.kaggle.com/datasets/arashnic/fitbit
Note that this dataset contains multiple files, some of which might be more interesting than others. It is up to you to decide what you want to use for your research.

Some ideas for analysis
Below are some ideas that could be interesting to investigate. It is not required for you to investigate them, rather it could be a starting point if you cannot think of your own ideas. Some of these ideas on the list are simple, and others may be complicated.

- What hour of the day / what day of the week are people in general most active?
- When grouping people based on activity:
    - Can you see a relationship between active people and their sleeping pattern (long, short, irregular)?
    - Are active people falling asleep sooner?
    - Can you see a relationship between weight / BMI and active people?
    - Is there a relation between sleeping pattern and BMI?
- When grouping days based on activity:
    - Can you see a relationship between active days and the sleep recorded?
    - Are people falling asleep on active days sooner?
- Can you fit one or more relevant regression models for this dataset?
- Can you implement a relevant and suitable classification/clustering model to this dataset?
- If you implement more than one, report which one is best and argue why their performance differs.

7) Google Play Store

This is a data set about the apps in the Google Play Store. Each app (row) has values for category, rating, size, and more.

Link to dataset:
https://www.kaggle.com/datasets/lava18/google-play-store-apps

Some ideas for analysis
Below are some ideas that could be interesting to investigate. It is not required for you to investigate them, rather it could be a starting point if you cannot think of your own ideas. Some of these ideas on the list are simple, and others may be complicated.

- What is the average size of an app?
- What is the average price of a paid app?
- Which category app…
    - … is rated best on average?
    - … is most often paid/ free?
    - … is installed most often?
    - … has the best ratings?
    - …
- How does the date in the last updated column affect the rating?
- What is the best app per category (or genre)?
    - Is this different when you filter on at least > 500 reviews?
- What is the most popular category / genre apps for teenagers?
- Can you fit one or more relevant regression models for this dataset?
- Can you implement a relevant and suitable classification/clustering model to this dataset?
- If you implement more than one, report which one is best and argue why their performance differs.

8) Incident management process enriched event log

This event log was extracted from data gathered from the audit system of an instance of the ServiceNow platform used by an IT company and enriched with data loaded from a relational database.

Link to dataset:
https://archive-beta.ics.uci.edu/ml/datasets/incident+management+process+enriched+event+log
The following article gives more information on the meaning of the columns:
http://processmining.each.webhostusp.sti.usp.br/wp-content/uploads/2021/02/Incidents_event_log_description.pdf
SLA: service level agreement

Some ideas for analysis
Below are some ideas that could be interesting to investigate. It is not required for you to investigate them, rather it could be a starting point if you cannot think of your own ideas. Some of these ideas on the list are simple, and others may be complicated.

- What percentage of the incident is solved within their SLA?
- Which category causes the most incidents?
- Which category usually takes the longest time to solve?
- What is the average number of days between the opening day and closing day for a high priority incident?
- Who solves the most incidents?
- Are incidents of a certain category solved by specific persons?
- Is there a relation between the location and the time an incident is created?
- What is the average number of updates (sys_mod_count) per incident state?
- Can you fit one or more relevant regression models for this dataset?
- Can you implement a relevant and suitable classification/clustering model to this dataset?
- If you implement more than one, report which one is best and argue why their performance differs.

9) Movies and TV series

There are three datasets provided. One about movies and one about tv-shows on Netflix, Prime Video, Hulu and Disney+. Furthermore, there is also a dataset containing the top 1000 movies on IMDB. It is possible to combine the different datasets to answer interesting questions.

Links to datasets:
https://www.kaggle.com/datasets/ruchi798/movies-on-netflix-prime-video-hulu-and-disney
https://www.kaggle.com/datasets/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney
https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows

Some ideas for analysis
Below are some ideas that could be interesting to investigate. It is not required for you to investigate them, rather it could be a starting point if you cannot think of your own ideas. Some of these ideas on the list are simple, and others may be complicated.

- Which platform has the most movies (series)?
    - What about if you filter on movies / series just for kids?
- Which platform has the most recent / old movies or series?
- Which platform has the best content according to the rating in Rotten Tomatoes for someone of 14 years old?
- Which channel provides the most movies from the IMDB top 1000?
    - What if you are only interested in a specific genre?
- In which period were the most top rated movies released?
- How does the IMDB score compare to the Rotten Tomatoes score?
- Can you fit one or more relevant regression models for this dataset?
- Can you implement a relevant and suitable classification/clustering model to this dataset?
- If you implement more than one, report which one is best and argue why their performance differs.

10) Online News Popularity

This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to predict the number of shares in social networks (popularity).

Link to dataset:
https://archive-beta.ics.uci.edu/ml/datasets/online+news+popularity
Useful resource:
A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News

Some ideas for analysis
Below are some ideas that could be interesting to investigate. It is not required for you to investigate them, rather it could be a starting point if you cannot think of your own ideas. Some of these ideas on the list are simple, and others may be complicated.

- What is the best weekday to publish an article looking at the shares?
- Are people more likely to share positive content compared to negative content?
- What is the influence of images and videos on the number of shares?
- On what weekday are the most articles published?
- Are there specific weekdays on which people are more interested in a certain topic (e.g. entertainment)?
- Is there a period in which people have been sharing more articles (looking at time delta)?
- Can you fit one or more relevant regression models for this dataset?
- Can you implement a relevant and suitable classification/clustering model to this dataset?
- If you implement more than one, report which one is best and argue why their performance differs.

11) Used car prices

These datasets include different features of cars, such as year of purchase, manufacturer, model, mileage etc. and show the price paid for the cars. One dataset shows the prices for cars sold in an auction and the other shows the prices sold in a normal way. You can try to combine the sets to answer interesting questions, or treat them separately.

Link to dataset:
https://www.kaggle.com/datasets/tunguz/used-car-auction-prices
https://www.kaggle.com/datasets/harikrishnareddyb/used-car-price-predictions

Some ideas for analysis
Below are some ideas that could be interesting to investigate. It is not required for you to investigate them, rather it could be a starting point if you cannot think of your own ideas. Some of these ideas on the list are simple, and others may be complicated.

- Which manufacturers produce the most expensive cars.
- What has more impact on the price, the mileage or the age of the car?
- Cars from which manufacturer are most sold in each state?
- Assuming the dataset was collected in 2019, what is the average miles / year?
- What is the best place to buy a cheap car?
    - Are the same type of cars really cheaper in certain states? Or are certain (cheaper) models / manufacturers more common in this state?
- What is the average price for a car from your favourite manufacturer (or model)?
- How does the color or the body of a car influence the price?
- Can you better sell your car via an auction or in a normal way?
- Can you fit one or more relevant regression models for this dataset?
- Can you implement a relevant and suitable classification/clustering model to this dataset?
- If you implement more than one, report which one is best and argue why their performance differs.