42184 Data Science for Mobility
42577 Introduction to Business Analytics course
Project Statement

Welcome to this year's challenge!

This topic focuses on the company Airbnb, which operates an online marketplace for short-term rentals (complete apartments, private rooms, shared rooms, etc). It acts as a broker and charges a commission from each booking. Analyzing data from Airbnb rentals, and in particular, rental prices, can be useful from several perspectives. First, it can assist prospective renters in making decisions about rentals, for example, determining an appropriate pricing level, minimum nights, location etc. In fact, several companies provide exclusive 'analytics' services to businesses and individuals seeking to enter the short-term rental sphere. They do this by analyzing Airbnb data from around the world (for example: https://www.airdna.co/). Second, Airbnb has also been controversial and attracted criticism, in part due to its effects on housing affordability and long-term rental prices. *Inside Airbnb* is an initiative that provides data and advocacy about Airbnb's impact on residential communities. For this project, you will utilize data from *Inside Airbnb.*

You have access to data from Airbnb rentals across the world. We have downloaded data from a single city, Copenhagen, which is provided in the file 'listings_CPH.csv'. It contains information from 13815 rentals currently listed in Copenhagen. This dataset will be the focus of the prediction challenge (first component of the project).

You can find this data along with csv files of listings in a number of other cities worldwide at: http://insideairbnb.com/get-the-data.

You can also explore the data for Copenhagen more at:
http://insideairbnb.com/copenhagen/

We expect that through this Data Sciences project, you not only address the mandatory questions (below) but also seek for yourself new questions, new data, new insights.

Project

The project has three components:

- *Prediction challenge* (30%): All groups need to address the same problem (30%).
- *Exploratory component* (40%): Each group is invited to choose their own research question and explore the data accordingly.
- *Report (30%) -* Each group should deliver one or more jupyter-notebooks, that should be self-explanatory in each step (or block). This will function as a report, so it should have introduction and conclusions, besides comments and reflections. However, there are some rules about the structure of the report, which should follow the 4-part outline shown below:

Section 1: Introduction + Data analysis and visualization

Section 2: Prediction Challenge

Section 3: Exploratory Component

Section 4: Conclusions

At the end of this document you will find a list of practical information, which will include details on what is expected in each task, and how these aspects contribute to the final grade.

Introduction to the data

Figure 1 shows the variables you will have in this dataset. The data is provided as a CSV file. Notice that the variables require some treatment in order to be usable (e.g. Categorical, strings, different scales, etc).

	0	1	2	3	4
id	6983	26057	26473	29118	31094
name	Copenhagen 'N Livin'	Lovely house - most attractive area	City Centre Townhouse Sleeps 1-10 persons	Best Location in Cool Istedgade	Beautiful, spacious, central, renovated Penthouse
host_id	16774	109777	112210	125230	129976
host_name	Simon	Kari	Julia	Nana	Ebbe
neighbourhood_group	NaN	NaN	NaN	NaN	NaN
neighbourhood	Nrrebro	Indre By	Indre By	Vesterbro-Kongens Enghave	Vesterbro-Kongens Enghave
latitude	55.68641	55.69307	55.67602	55.67023	55.666602
longitude	12.54741	12.57649	12.5754	12.55504	12.555283
room_type	Entire home/apt	Entire home/apt	Entire home/apt	Entire home/apt	Entire home/apt
price	898	2600	3250	725	1954
minimum_nights	3	4	3	7	3
number_of_reviews	172	59	300	24	19
last_review	2022-06-21	2022-08-09	2022-09-10	2022-08-04	2022-08-22
reviews_per_month	1.08	0.55	2.06	0.16	0.13
calculated_host_listings_count	1	1	3	1	1
availability_365	0	303	56	59	0
number_of_reviews_ltm	4	8	7	2	2
license	NaN	NaN	NaN	NaN	NaN

Figure 1. Dataframe view

For the *prediction challenge*, you have two tasks (both are classification problems):

1. Binary Classification: predict whether the price of a rental is either ‘low’ or ‘high’. We leave it up to you to decide how you want to define ‘low’ and ‘high’, for example, ‘low’ could mean that the price of the rental is lower than the median rental price in the city, or it could mean that the price of the rental is lower than the average rental price in the city. You may also use any other threshold to define the classes low and high.
   As a benchmark, we expect that you build a classifier with an f1 score of at least 0.6 on test data.
2. Multi Class Classification: predict whether the price of a rental is either ‘low’, ‘med’ or ‘high’. We again leave it up to you to decide how you want to define

'low', 'med' and 'high'. For example, 'low' could mean that the price of the rental is lower than the 33th percentile of prices in the city, 'medium' is a price between the 33th and 66th percentile, and 'high' is a price higher than the 66th percentile. You may also use any other threshold to define the classes low, medium and high.
Here, there is no benchmark, try to build the best possible classifier you can (considering all the metrics we looked at f1 score, confusion matrix, etc.).

In the *exploratory component*, each group needs to address at least one new research question. Here, we expect you to formulate your own question. The project will be positively valued with one or more of the following extensions:

- Extension of the dataset with other relevant data (for example, data from other cities, data on 'Points of Interest' such as tourist attractions from Open Street Maps, data on accessibility to public transit stations, etc.)
- Generation and analysis of insightful visualizations;
- Usage of the breadth of techniques from the class beyond classification and data preparation (e.g. clustering)

Some example research questions:

- What are the main factors affecting the price of a rental?
- Are these factors the same across cities? Can data from different cities be pooled to generate better predictions?
- Can a model trained on data from one city perform well on another city?
- Can we better understand/predict the price of a rental using other information not in the dataset, for example accessibility or proximity to bus/transit stations, residential vs commercial area vs business district?

Note: The ordering of tasks we mention is **not** mandatory. In other words, if you prefer to start with the exploratory component, and then go to the prediction challenge, this is perfectly acceptable. You can mention that in the report (or invert Sections 2 and 3). Similarly, data analysis might appear after the introduction (if relevant). However, please be aware that a simple descriptive analysis of the data is not sufficient to complete the task. Make sure to go one step forward and try at least one of the techniques discussed in the course. Note also that in the exploratory part, it is perfectly fine to apply a technique and then find that it is not very insightful (make sure you discuss the possible reasons for it though). For example, you try to do clustering but find that you do not obtain meaningful clusters.

Evaluation

The evaluation of the report will be based on the following criteria:

- Clarity - self-explanatory nature of the notebooks
- Thoroughness - Each research question deserves to be explored to the right amount of depth
- Insightfulness - It's important to go beyond the surface of the conclusions
- Technical aspects: Data have been properly analyzed (data cleaning data preparation, data pre-processing).

        - Which model has been used (only one model, multiple models, only linear models, non-linear models)?
        - Is the model appropriate?
        - Which performance metric (how performance was evaluated)? Were they appropriate?
        - How was the approach benchmarked (how conclusions were drawn)?
- Honesty - While it's fine to use others' code (as a starting point), these shouldn't generally be the actual deliverable **and** the appropriate ethical practice is to **always** reference the source of that code you used.

Rules

- Each group should consist of 3 to 4 students.
- The submission of the project shall be a zip file with all the notebooks. This zip file should contain the surnames of the group members (for example, for Pablo, Anders, Suarez, and Mila, it should be Pablo_Anders_Mila.zip).
- At the end of the report, there must be a section where **individual contributions are clearly clarified**. In case of doubts on individual contributions or authenticity of the report, the teachers will call the group for an oral defence. This section should **not be part of the page counts.**
- Meeting the deadlines for the milestones is important, including for non-evaluated milestones. A penalty of 10% is given for each extra day of delay

PLEASE INDICATE NAME, SURNAME, and STUDENT NUMBER IN THE REPORT

Report length

The report must be in the form of a jupyter notebook. The structure should be the one described in page 1. There is no overall page limit. However, the project (description of the research questions and results) should not exceed 4 pages. This limit does not apply to figures and codes.

To be more precise, the report can include unlimited figures, and there is no limit to the length of the code. The 4 pages limit only applies to markdown cells. As a referece, you can use this code[1] to make the word count of your markdown cells. One document page is about 500 words (3000 characters including space). The project should be approximately 2000-2500 words. Again, this apply only to markdown. **While this is not a strong constraint**, **excessively long reports will be penalized**.

Important dates

- October 18 – Announcement of this challenge statement
- October 31 – Communication of group members and selection of topic (through DTU learn)
- December 2 – Final submission – all materials, including report notebook. Submit through DTU learn

[1] https://stackoverflow.com/questions/71194571/word-count-of-markdown-cells-in-jupyter-notebook