Lab Assignment MA661E, VT25

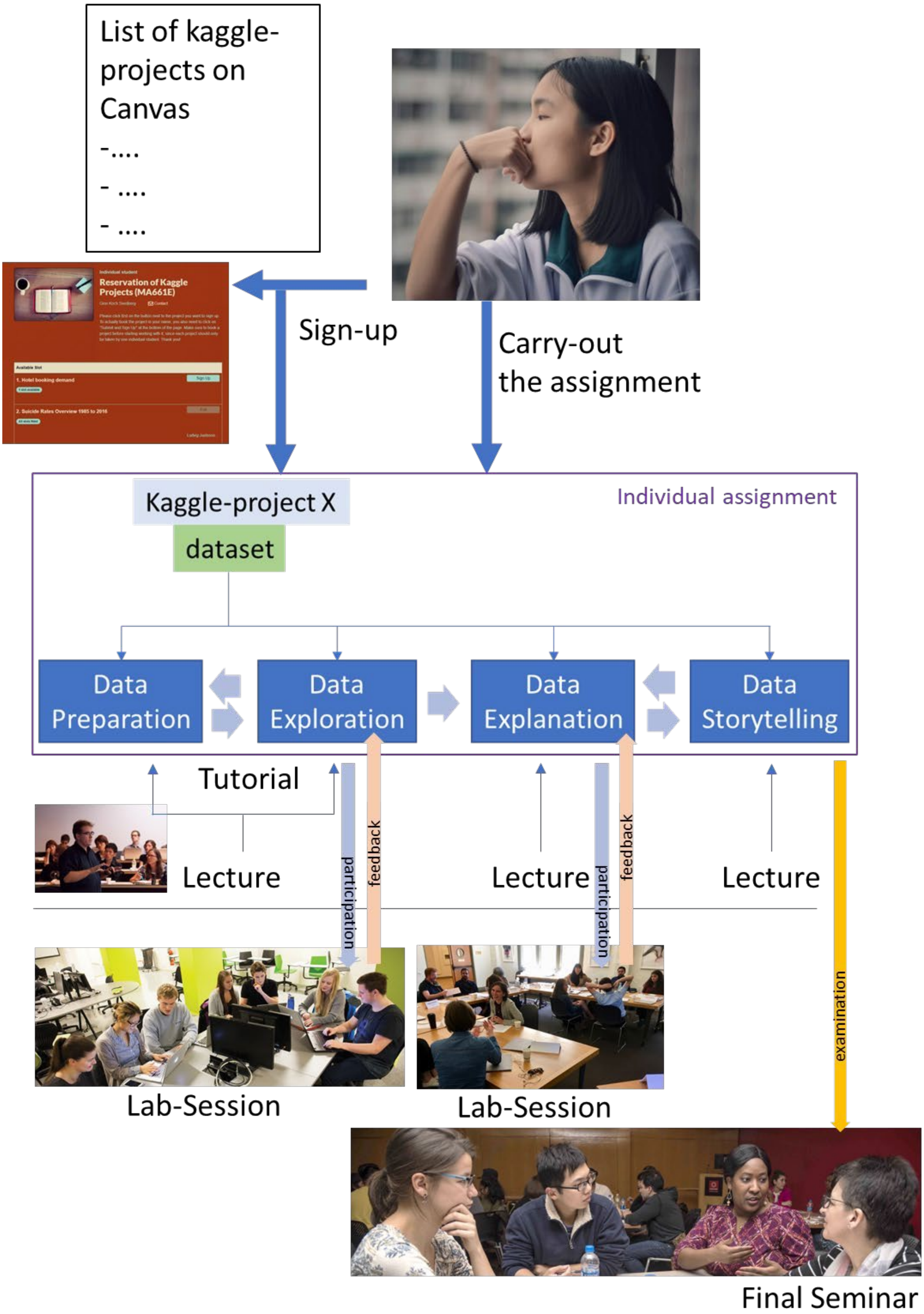
The purpose of this lab-assignment is to practice **data analysis, data visualization and storytelling** with data on an individually selected dataset.

1. Overview

Collections of data are unique. They are usually collected for certain purposes and vary in quality and quantity. To analyze and visualize data is a creative process with the aim of understanding as much as possible about the data. With the collection of vast amounts of data in a digitalized world, the extensive possibilities given by data analysis for data-driven products and services has made this topic very popular. New books, methods, software-libraries, and software tools are released constantly, while companies and organizations desperately try to find experts in this field.

The concept for this year is illustrated in figure 1:

- From the current list of datasets in Canvas, please select one that you find interesting and that no one other in class already has reserved. Put your name in the excel form with different datasets.
- You start working with your dataset by first preparing the data (handling missing or unfeasible values).
- Using your knowledge about statistical characteristics you will carry out data exploration by means of a univariate and a bivariate data analysis.
- A lecture is given to get you started with the Python-based scikit-learn framework and with Jupyter notebooks, and to answer questions you might have.
- In the common lab-sessions you get the chance to share your work with other students in your group, as well as to help each other and to get valuable feedback.
- For the data explanation part, we will test available analytical tools that allow us to visualize relations between data features in an interactive way.
- The lecture will present the tools very roughly and you are expected to select one of the tools and try to use it for the data explanation part of your data set. At the common lab session students will have the opportunity to discuss the use of the tool with other students and to exchange their experiences.
- Once you gained a better understanding for your data, we will introduce you to the concept of ‘Storytelling with data’. In the corresponding lecture we touch upon the main concepts, aware that this will not be enough to master this subject. But we want you to try and to present your story at the final seminar.



Individual student

Reservation of Kaggle Projects (MA661E)

[illegible] Contact

Please click first on the button next to the project you want to sign up. To actually book the project to your name, you also need to click on "Submit and Sign Up" at the bottom of the page. Make sure to book a project before starting working with it, since each project should only be taken by one individual student. Thank you!

Available Slot

1. Hotel booking demand — Sign Up — 1 slot available

2. Suicide Rates Overview 1985 to 2016 — Full — All slots filled

Figure 1: Overview over the lab-assignment

2. Data Preparation & Exploration

Use Google's Colab or Anacondas scikit-learn installation with Jupyter Notebook to carry out the data preparation, cleaning and exploration of your selected dataset. Follow the Tutorial-DataExploration on Canvas to learn about the different steps that are usually needed and apply them to your data. Generate an own copy of the Tutorial on Colab and adapt it to your dataset. Using your own words, you should describe how you:

1. read the csv-file and store it as a DataFrame variable
2. explain the meaning of the column names (features) of the dataset
3. find and handle missing values
4. convert categorical data
5. aggregate the data if necessary
6. analyze the feasibility of the values
7. generate subplots with the histograms and density estimations of the single features and discuss their interpretations
8. make use of boxplots to find out about outliers, and how you handle them
9. remove outliers if you do, and how much data is left, discuss if it still is enough
10. create a heatmap and how you interpret its result. Analyze the results from your heatmap carefully. Decide based on the heatmap:
    - which values to combine in scatterplots and joint distribution plots with regression fit. Explain your reasoning.
    - Find interesting category dependencies and plot the most interesting ones
11. run a t-test of at least one of the category dependencies that you found interesting. Conclude if there is a statistically relevant difference or not.
12. Save your refined data in an excel-file and pickle all of your final version of the data

At the end of this data exploration phase, you should have gained insights about the statistical distributions of the individual features and the feasibility of their values, as well as about the correlations between the different features and interesting category dependencies, and whether or not these differences are statistically relevant or not.

13. ***Do not proceed to the next phase before you have summarized these insights at the end of your Jupyter notebook!*** The insights about interesting correlations and category dependencies are instrumental for the next phase, which is Data explanation.

Do not forget to pickle your final version of the data as well as to save it in a excel-file!

3. Data Explanation

The insights from the data exploration phase are excellent starting points to work with the tools proposed for data visualization, which are Tableau or Qlik Sense.
Elaborate more on the interesting correlations and category dependencies using one of these tools and build interactive visualizations that you can use to explain patterns and relationships present in the data. **If you want to use other tools than these ones, please tell the teacher and ask for confirm if the tool is okay, or not.**

Some practical tips about what to do with these tools on the data include:

- look for trends (for example changes over time) of the different features and their combination

- combine features that are correlated in an interesting way to find the factors that are most important, e.g. relevant information
- investigate possible clusters in your data
- build dashboards for visual overviews
- compare different cases
- propose hypotheses to test
- test using the tool in many different ways
- do not confuse correlation with causation when looking for patterns

Your understanding of your data and its context will guide you when creating worksheets with this kind of tools. Worksheets will be used for your storytelling with data, hence you will most probably iterate back and forth between data visualization and storytelling.

4. Storytelling with Data

When communicating your findings and ideas regarding the data, you should go behind just simply showing the data and describing the different digital methods used (=exploratory analysis). Instead, you should take the time and effort to turn the data into information, to find a specific thing you want to explain, a specific story you want to tell. The relationships between the variables within their context should have brought about some insights, hypotheses, questions, reflections, bewilderment, clarity, or anything else worthwhile noting and telling about.

We have composed some excerpts from the book 'storytelling with data. a data visualization guide for business professionals' by Cole Nussbaumer Kaflic, Wiley, 2015. While the whole book and other books on this subject are worthwhile reading, we want you to focus mostly on 'Understanding the Context' and on 'Storytelling' in this assignment.

Referring to the excerpts from the book, you should try to create a story by following the following steps:

1. *Big Idea and 3-minute story*
   - Describe your 'Big Idea' in one to three sentences first and expand it afterwards into a text that would take about three minutes to read.
2. *Storyboarding and crafting the story*
   - Sketch a story board to articulate your story and to plan the desired content and flow. You can write on paper, use post-it or a white board. Include a picture that shows the result.
   - Craft your story with clear beginning (plot), middle (twists), and end (call to action). Leverage conflict and tension to grab and maintain your audience's attention. Consider the order and manner of your narrative.
   - *Creating the story*
     Create your story by combining your narratives, worksheets, pictures from Jupyter or Excel, and other material in Tableau or Qlik Sense. Choose the right kind of visuals (graphics) for your story; and make them as simple as possible.
   - Save your story with the tool you used and as a .pdf-file and upload it on Canvas

5. Ethical principles for communicating digital information

For the preparation of the final seminar, you should review another student's story regarding the ethical principles as described in Nicholas Diakopoulos' chapter on 'Ethics in Data-Driven Visual Storytelling' in the book of Riche et. al. on 'Data-Driven Storytelling'. Focus on the following chapters:

- *Ethical concerns regarding data acquisition*
- *Ethical concerns regarding data transformation*
- *Ethical concerns regarding conveying and connecting insights*

1. Identify at least two ethical questions that are mentioned in the text ***for each of these three ethical concerns***. On presentations day you will ask them to your review-match student.

6. Final Seminar

Turn in this lab-assignment before the deadline. There will be tw deadlines, one before the ordinary final seminar and one for the retakes. If your assignment passes the requirements as stated above and below, you will be invited to participate at the final seminar. You will also be assigned another students assignment for the review of the ethical principles. If your assignment does not pass you can complete it and turn it in again before the next retake of the final assignment. At the final seminar, everybody will present his or her story with help of the tool used. A discussion regarding the presentation will follow as well as the review regarding the ethical concerns.

7. Requirements for passing the assignment

Make sure to document your work in a way **that others can follow your line of reasoning.** Your code needs to be explained and the results discussed with your own wordings. You need to be able to answer questions regarding your work and your code during the final seminar.

Follow the instructions carefully given in the different sections above. When uploading your assignments to Canvas, you need to consider the following:

1. **Combine and upload a pdf with the following structure: NO CODE IS REQUIRED IN THIS PDF, BUT ALL STEPS EXPLAINING YOUR CODE.**
    1) Index of the contents of your report
    2) Storytelling with Data (Big Idea and 3-minute story)
    3) Tableau or Qlik Sense story or Matplotlib plots (if you used these tools provide the dashboards that you created)
    4) Data Preparation & Exploration: list the steps that you did in your code
    5) Data Explanation: write about the steps you did to explore data, and why you did so, and the explanation of the outcome of your exploration
2. For the presentation**, you will not use this pdf (the one mentioned above in point 1), but you will use slides PPT / Canva to tell the story** of your dataset and the key-insights you got from it. **You will submit these slides in the assignment related to the presentation (on the day of the presentation, it is not necessary that you upload them before it).**