**Assignment 1**

Date on which assignment was provided: 7/8/2024

Marks: 25 Marks

This assignment integrates your understanding of data formats with approaches for exploratory data analysis.

Prepare assignments based on the questions. The utilization of the same dataset and comparable work by multiple groups will result in zero marks. Use Python programming whenever it is necessary. As it is a collaborative activity, all students must participate equally.

**Week 1 Date: 12-16 Aug.: Data Selection**

First, pick a subject that interests you and look for a collection that can help you learn more about it. Follow the links below to find information that will help you get started.   
Once you've chosen a topic and a dataset, but before you start analyzing it, you should make a list of questions you want to find answers to. Clean up the data if you need to in order to move on to the next step.

The questions below will help you look into it. You can add to this list of questions to help you explain your answer.

**Question:** Pick one area (like healthcare or business) and explain how data science is used in that area. Give an example of a data science project or app that works in this area, and then use that example to answer other questions.

**Question:** Give some examples of different ways to get data, such as web scraping, APIs, polls, and databases. Give an example of some code that can be used to get data using one of the ways.

**Week 2: 19-30 Aug.: Exploratory Visual Analysis**

Next, use Python to do an experimental study of your information. You should think about two different stages of exploring.

In the first step, you should try to get a sense of your dataset's style and shape. What kinds of factors does the dataset have? How are they given out? Are there any big problems with the quality of the data? Are there any surprising links between the variables? Also, make sure to do "sanity checks" on trends you think you should see!

You should look into your first questions and any new ones that come up during your research in the second part. For each question, start by drawing something that could help you figure out the answer. Then, improve the display (by adding more factors, changing the sorting or axis sizes, filtering or subsetting data, etc.) to see things from a different angle, look into facts that you didn't expect, or make sure your ideas are still reasonable. You should do this for each of your questions, but you can change them or start looking into other questions if the data prompts you to.

The questions below will help you look into it. You can add to this list of questions to help you explain your answer.

**Question:** Explain some important ideas in statistics and chance, like the mean, median, mode, variance, standard deviation, and probability distributions. Give some examples of how these ideas are used in data research. Find the summary figures (mean, median, mode, range, variance, and standard deviation) for the collection. Draw and explain the spread of the numbers.

**Question:** Explain the three types of data: number, categorical, and order. Give some examples of each type of information.

**Question:** Talk about how to show category data, text data, and time series data. Give some code examples for each type of data.

**Question:** Explain how you can use data science to work with voice, video, and picture data. Give examples of methods that are used in each medium, such as video analysis, picture classification, speech recognition, and so on.

**Question:** Describe why data representation is important. Give instructions on how to make good drawings.

**Question:** Use the information to make different kinds of plots, like bar charts, histograms, and scatter plots. Figure out what each plot's results mean.

**Question:** Explain the difference between graphical and non-graphical EDA methods. Give examples of both one-variable and multiple-variable analyses that can be done with the information.

**Question:** What are some ways to find and understand groups in data? Give an example of how to use the information with cluster analysis.

**Question:** Explain different ways to show info that isn't a number. Use the facts to give an example.

**Report Preparation**

Create a complete report that summarizes your responses to each question. Include Python code, visuals, and extensive explanations. Ensure that the report is well-structured and simple to read.

**Final Deliverable**

## Your final submission will include a written report, 10 or more captioned "quick and dirty" visuals presenting your most key findings. The dashboard should include numerous charts to explain your results on an ongoing basis, presuming you will continue to gather data over time: assume you are a data analyst developing a dashboard for your company's CEO, who will be able to review critical indicators (related to your hypothesis) every month. In your written portions, concentrate on the answers to your original queries while also describing surprises and problems found along the journey, such as data quality concerns. Each visualization picture should include a title and a brief description (<2 words). Provide enough description for each caption so that anybody reading your report may comprehend your results. Feel free to annotate the images to highlight certain characteristics of the data while keeping in mind the visual concepts we've learnt so far.

## **Grading**

The assignment has a maximum score of 25 points. I will provide grades based on the breadth and depth of your research, if visualizations follow the expressiveness and effectiveness standards, and how well-written and synthesized your ideas are.   
Below are our grading criteria:   
1. Poses relevant queries for the specified dataset.   
2. Evaluate and convert data quality appropriately.   
3. Provided a comprehensive analysis that addressed various questions.   
4. Provide thorough analysis with relevant follow-up questions.   
5. Created compelling graphics to answer analytical problems.   
6. Captions should be clear and easy to read, conveying key ideas.

## **Submission Details**

## In groups of six, choose a dataset of interest and do exploratory analysis in Python to understand the data structure, examine hypotheses, and produce early findings. Create a PDF of the report using the following template outline: Include a collection of ten or more visualizations that demonstrate your results, one summary "dashboard" visualization, and a write-up of your methodology and what you learnt. Submit the assignment, which includes the report, Python code, explanation, dataset, and results, to GitHub by September 20, 2024. Also, provide the GitHub URL on or before: for review. Be prepared for a presentation on the day that you will be told.

## **Recommended Data Sources**

* [Find Open Datasets and Machine Learning Projects | Kaggle](https://www.kaggle.com/datasets)
* [dataset · GitHub Topics · GitHub](https://github.com/topics/dataset)
* [NYC Open Data](https://opendata.cityofnewyork.us/): data on NYC trees, taxis, subway, citibike, 311 calls, land lot use, etc.
* [data.gov](https://www.data.gov/): everything from hourly precipitation, fruit & vegetable prices, crime reports, to electricity usage.
* [Dataset Search by Google Research](https://datasetsearch.research.google.com/): indexes public open datasets.
* [Stanford Open Policing Dataset](https://openpolicing.stanford.edu/data/)
* [Physician Medicare Data](https://data.medicare.gov/data/archives/physician-compare)
* [Civil Rights Data Collection](https://ocrdata.ed.gov/)
* [Data Is Plural newsletter’s Structured Archive](https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit#gid=0): spreadsheet of public datasets ranging from curious to wide-reaching, e.g. “How often do Wikipedia editors edit?”, “Four years of rejected vanity license plate requests”
* [Yelp Open Dataset](https://www.yelp.com/dataset)
* [U.S. Census Bureau](https://www.census.gov/data.html): use their Discovery Tool
* [US Health Data](https://data.cdc.gov/): central searchable repository of US health data (Center for Disease Control and National Center for Health Statistics), e.g. surveys on pregnancy, cause of death, health care access, obesity, etc.
* [International Monetary Fund](http://www.imf.org/en/Data)
* [World Bank](https://data.worldbank.org/)
* [IPUMS.org](https://www.ipums.org/): Integrated Census & Survey Data from around the World
* [Federal Elections Commission](http://www.fec.gov/finance/disclosure/ftpdet.shtml): Campaign Finance & Expenditures
* [Stanford Mass Shootings in America Project](https://library.stanford.edu/projects/mass-shootings-america): data up to 2016, with pointers to alternatives
* [USGS Earthquake Catalog](https://earthquake.usgs.gov/data/data.php)
* [Federal Aviation Administration](https://www.faa.gov/data_research/)
* [FiveThirtyEight Data](https://github.com/fivethirtyeight/data/): Datasets and code behind fivethirtyeight.com
* [ProPublica Data Store](https://www.propublica.org/datastore/): datasets collected by ProPublica or obtained via FOIA requests, e.g. [Chicago parking ticket data](https://www.propublica.org/datastore/dataset/chicago-parking-ticket-data)
* [Machine Learning Repository](http://archive.ics.uci.edu/ml/) - large variety of maintained data sets
* [Socrata Open Data](https://opendata.socrata.com/)
* [17 places to find datasets for data science projects](https://www.dataquest.io/blog/free-datasets-for-projects/)
* [Awesome Public Datasets](https://github.com/awesomedata/awesome-public-datasets) (github): topic-centric list of high-quality open datasets in public domains
* [Open Syllabus](https://opensyllabus.org/): 6,059,459 syllabi

### Additional Data Sources

To get up and running quickly with this assignment, I recommend exploring one of the following provided datasets:

* [World Bank Indicators, 1960–2017](https://github.com/ZeningQu/World-Bank-Data-by-Indicators). The World Bank has tracked global human developed by indicators such as climate change, economy, education, environment, gender equality, health, and science and technology since 1960. The linked repository contains indicators that have been formatted to facilitate use with Tableau and other data visualization tools. However, you're also welcome to browse and use the original data [by indicator](https://data.worldbank.org/indicator) or [by country](https://data.worldbank.org/country). Click on an indicator category or country to download the CSV file.
* [Chicago Crimes, 2001–present](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g) (click Export to download a CSV file). This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.
* [Daily Weather in the U.S., 2017](https://vis.csail.mit.edu/classes/6.894/data/weather.csv.gz). This dataset contains daily U.S. weather measurements in 2017, provided by the [NOAA Daily Global Historical Climatology Network](ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/). This data has been transformed: some weather stations with only sparse measurements have been filtered out. See the accompanying [weather.txt for descriptions of each column](https://vis.csail.mit.edu/classes/6.894/data/weather.txt).
* [Social mobility in the U.S.](https://opportunityinsights.org/data/). Raj Chetty's group at Harvard studies the factors that contribute to (or hinder) upward mobility in the United States (i.e., will our children earn more than we will). Their work has been [extensively featured](https://www.nytimes.com/search?query=raj%20chetty) in The New York Times. This page lists data from all of their papers, broken down by geographic level or by topic. We recommend downloading data in the CSV/Excel format, and encourage you to consider joining multiple datasets from the same paper (under the same heading on the page) for a sufficiently rich exploratory process.
* The [Yelp Open Dataset](https://www.yelp.com/dataset) provides information about businesses, user reviews, and more from Yelp's database. The data is split into separate files (*business*, *checkin*, *photos*, *review*, *tip*, and *user*), and is available in either JSON or SQL format. You might use this to investigate the distributions of scores on Yelp, look at how many reviews users typically leave, or look for regional trends about restaurants. Note that this is a large, structured dataset and you don't need to look at all of the data to answer interesting questions. In order to download the data you will need to enter your email and agree to [Yelp's Dataset License](https://s3-media2.fl.yelpcdn.com/assets/srv0/engineering_pages/e926cc12796d/assets/vendor/yelp-dataset-license.pdf).

If you want to investigate datasets other than those recommended above, here are some possible sources to consider. You are also free to use data from a source different from those included here. If you have any questions on whether your dataset is appropriate, please ask me.

* [data.boston.gov](https://data.boston.gov/) - City of Boston Open Data
* [MassData](https://opendata.digital.mass.gov/) - State of Masachussets Open Data
* [data.gov](https://www.data.gov/) - U.S. Government Open Datasets
* [U.S. Census Bureau](https://www.census.gov/data.html) - Census Datasets
* [IPUMS.org](https://www.ipums.org/) - Integrated Census & Survey Data from around the World
* [Federal Elections Commission](http://www.fec.gov/finance/disclosure/ftpdet.shtml) - Campaign Finance & Expenditures
* [Federal Aviation Administration](https://www.faa.gov/data_research/) - FAA Data & Research
* [fivethirtyeight.com](https://github.com/fivethirtyeight/data/) - Data and Code behind the Stories and Interactives
* [Buzzfeed News](https://github.com/BuzzFeedNews)
* [Socrata Open Data](https://opendata.socrata.com/)
* [17 places to find datasets for data science projects](https://www.dataquest.io/blog/free-datasets-for-projects/)

Additional Tools

Your dataset almost certainly will require reformatting, restructuring, or cleaning before visualization. Here are some tools for data preparation:

* Graphical Tools
  + [ggplot2](http://ggplot2.org/) library
  + Python [Jupyter notebooks](http://jupyter.org/) with libraries eg.  [Matplotlib](http://matplotlib.org/)
  + open source tool for working with messy data.
* Programming Tools
  + [Pandas](http://pandas.pydata.org/) data table and manipulation utilites for Python.