MIE1624: Introduction to Data Science and Analytics – Winter 2025

Assignment 1
Due Date: 11:59 pm, February 14, 2025
Submit via Quercus

Background:

For this assignment, you are responsible for answering the below questions based on the dataset provided. You are required to submit (only) two files. The first should be an IPython Notebook (.ipynb file) containing all the code of the analysis you performed to answer the questions. Please ensure that this notebook is not saved in any other formats, such as PDF or HTML. The second file is a 5-page report (PDF file) in which you will briefly describe what you have done and present the results of your analyses. In your report, you should use visual forms to present your results. How you decide to present your results (i.e., with tables/plots/etc.) is up to you, but your choice should make the results of your analysis clear and obvious. Please interpret your findings within the broader context of the dataset and the underlying research questions, ensuring your analysis clearly aligns with the problem at hand.

Dataset:

Kaggle, a platform known for its data science competitions and datasets, annually conducts the Stack Overflow Annual Developer Survey, renowned as the most extensive ongoing survey of software developers and others involved in coding worldwide. The purpose of this is to gather and analyze comprehensive insights about the trends, technologies, and demographics in the software development community. More information on this survey can be found on:
https://www.kaggle.com/datasets/berkayalan/stack-overflow-annual-developer-survey-2024

The **survey_results_public_2024.csv** dataset, available both on the mentioned website and on Quercus, contains the survey results. The results from roughly 65,500 participants are shown in 112 columns, representing survey questions. Not all questions are answered by each participant, and responses contain various data types.

The original dataset has been cleaned for you and now contains fewer rows and columns, with rows containing null salaries removed. In the cleaned dataset provided for this assignment (**clean_kaggle_data_2024.csv**), the last column ("ConvertedCompYearly") contains a numerical target variable, the annual salaries of the participants in USD. **You should work with the clean dataset for this assignment.**

Questions:

The objective of this assignment is to explore the survey data to understand **(1)** distinctions between remote and hybrid job modes in technology and information sectors and **(2)** the effects of education on income level. The following tasks should be completed:

1. **[3pts]** Perform Exploratory Data Analysis (EDA) to analyze the survey dataset and to summarize its main characteristics. Present **3 graphical figures** that represent meaningful trends in the data. For example, you might choose to provide a figure for education level vs. salary, a figure for the distribution of education levels or ages of the participants, or a figure demonstrating average job satisfaction by job title. You are recommended to consider exploring features that could potentially

explain the observed salary difference between **remote** and **hybrid** workers (which you will analyze in Question 2). For instance, you might examine how the salary distributions vary across different geographical regions or countries (e.g., United States of America, Canada, India, etc.).

Note: You are not restricted to the suggested features. Feel free to explore and include other features in the dataset that you think might provide valuable insights. Use your visualizations to tell a coherent story and identify patterns or relationships that could aid in feature engineering for the subsequent questions. Moreover, you are encouraged to use a variety of chart types (e.g., bar plots, histograms, scatter plots, box plots) and identify which features might be useful for understanding the observed differences in salaries.

2. **[6pts] Estimating Salary Differences Between Hybrid and Remote Job Modes**

   The job modes are listed in the fifth column of the cleaned dataset ("RemoteWork").

   a. **[0.5pts]** Compute and report descriptive statistics for the two groups of **hybrid** and **remote** job modes (remove missing data and outlies if necessary). Provide your rationale for either keeping or removing outliers and missing values.

   b. **[1.5pts]** Perform a two-sample t-test to compare average salaries between hybrid and remote job modes.

      Note: First, perform a manual calculation using Python to compute the two-sample t-test statistic, including detailed step-by-step calculations of the means, variances, and pooled standard deviation for the two groups. Next, use Python's built-in functions to perform the t-test and compare the results with your manual calculations, discussing any discrepancies found between the two methods. You may need to check whether the assumptions required for the statistical testing are satisfied.

   c. **[1.5pts]** Bootstrap your data for comparing the mean salary for the two groups of hybrid and remote workers. Note that the number of instances you sample from each group should be relative to its size. Use 10,000 replications. Present your findings in three distinct figures: one demonstrating the two bootstrapped distributions (for hybrid and remote job modes), another depicting the distribution of the difference in means, and a third showcasing the standardized distribution of the difference in means.

   d. **[0.5pts]** Perform a two-sample t-test with a significance level of 0.05 threshold using bootstrapped data. Compare your results with those from 2.b and explain your findings.

   e. **[1pt]** Compute 95% confidence intervals (CIs) for the difference in mean salary between the hybrid and remote job-mode groups using two methods: (a) Formula-Based Approach (e.g., a t–distribution–based method) using the original data. (b) Bootstrap Approach using your previously generated bootstrap data. Finally, compare both sets of confidence intervals and discuss whether they reach the same conclusion about the difference in mean salaries. If they differ, explain possible reasons for these discrepancies.

   f. **[1pt]** In addition to comparing mean salaries, examine the median salaries of the two groups by performing a Mood's median test. Check if the results from the nonparametric test (i.e., Mood's p-value) and from bootstrapping (i.e., the two-sided p-value for the differences in bootstrapped medians) are close to each other. Discuss any notable similarities or differences.

3. **[3pts]** Select the education level column from the cleaned dataset ("EdLevel") and repeat steps **a** to **d**, but this time use Analysis of Variance (ANOVA) instead of a t-test for hypothesis testing to compare the mean salaries across three education groups: Bachelor's degree, Master's degree, and Professional degree. For the ANOVA, limit your analysis to survey participants from North America (i.e., the US and Canada). [**0.5pts** for **a; 0.5pts** for **b; 1.5pts** for **c; 0.5pts** for **d**].

   Note*:* You do not need to perform the ANOVA manually; using a built-in function is sufficient. Additionally, Similar to question 2.c, question 3.c requires you to submit three figures: one demonstrating three bootstrapped distributions, another showing pairwise distributions of the difference in means, and a third depicting standardized distributions of the difference in means.

4. **[1 Bonus Point] Complete the Assignment using a Large Language Models (LLMs)**

   For extra credit (without exceeding the maximum total of 12 points), you are encouraged to complete all parts of this assignment (questions 1–3) entirely through an LLM such as ChatGPT, Claude, Gemini, etc. Provide samples of the prompts you used to elicit answers, code, and suggested figures from the LLM. If the LLM's initial responses contained errors or required verification, show how you validated and corrected those. If you are using ChatGPT, you could provide the public link to your chat. Finally, reflect on how prompt engineering and iterative feedback helped improve the final output.

Submission:

1) Produce a 5-page report (including figures) explaining your response to each question for the given dataset and detailing the analysis you performed. When writing the report, make sure to explain each step, what you are doing, why it is important, and the pros and cons of that approach.

Note: The report submission is strictly limited to five pages, inclusive of figures, explanations, and all other content. While this may seem challenging, part of this exercise is to cultivate the skill of presenting complex information in a concise and clear manner. Try allocating 4 pages to questions 1-3 and the remaining 1 page to the bonus part.

2) Produce a Jupyter (IPython) Notebook detailing the analysis you performed to answer the questions for the given data set.

What to submit:

1. Submit via Quercus a Jupyter (IPython) notebook containing your implementation and motivation for all the steps of the analysis with the following naming convention: **lastname_studentnumber_assignment1.ipynb**.

   Make sure that you **comment** on your code appropriately and describe each step in sufficient detail. Respect the above convention when naming your file, making sure that all letters are lowercase, and underscores are used as shown. **A program that cannot be evaluated because it varies from specifications will receive zero marks. Also, to receive the marks, your code must be executed without any errors.**

2. Submit a report in PDF including the findings from your analysis. Use the following naming conventions **lastname_studentnumber_assignment1.pdf**. **Page size: US Letter (8.5"x11"), 1-inch margin, font size: 11 pt.**

Late submissions will receive a standard penalty:

- up to one hour late - no penalty
- one day late - 15% penalty
- two days late - 30% penalty
- three days late - 45% penalty
- more than three days late - 0 mark

Tools:

- **Software:**
  - **Python Version 3.X** is required for this assignment. All libraries are allowed but here is a list of the major libraries you might consider: Numpy, Scipy, Sklearn, Matplotlib, Pandas.

    Specify `!pip install <library_name>` commands in the first cell of your notebook for all the 3rd party libraries you use. Make sure that your Jupyter notebook runs on Google Colab cloud.
  - No other tool or software besides Python and its component libraries can be used to touch the data files. For instance, using Microsoft Excel to clean the data is NOT allowed.
  - Read the required data file from the same directory as your notebook: for example, pd.read_csv("clean_kaggle_data_2024.csv").

- **Required data files:**
  - **clean_kaggle_data_2024.csv**: survey responses with yearly compensation.
  - The data file cannot be altered by any means. The Jupyter notebook will be run using a local version of this data file. Do not save anything to file within the notebook and read it back.

Other requirements:

1. A large portion of marks is allocated to analysis and justification. Full marks will not be given for code alone.
2. Output must be shown and readable in the notebook. The only files that can be read into the notebook are the files posted in the assignment without modification. All work must be done within the notebook.
3. The notebook should be presentable, do not show large amounts of raw output.
4. Ensure the code runs in full before submitting. Just before you submit, rerun the entire notebook (navigate to Kernel => Restart Kernel and Run all Cells). Ensure that there are no errors.

Best of Luck!