**INTRODUCTION**

The collaborative project, similar to the individual project, is an EDA on a dataset that you and your partner(s) jointly agree on. Typically, within real-world scenarios as a data analyst, you will be working with other data analysts / data scientists and within this project, we want to simulate that exact environment. Given that different people have vastly different skills, it’s up to your group to decide what the best plan of action is.

During this project, you’ll get another hands-on learning experience in data analytics that not only allows you to apply your skills to uncover insights and tell a compelling story but to also, most importantly, work effectively as a team. By working with real-world datasets, you and your partner(s) will have the chance to demonstrate critical thinking and technical proficiency. Likewise, this project will also allow you to create a portfolio piece that showcases Python (and, optionally, your SQL) ability. Like the individual project, your ability to preprocess, analyze, visualize, and communicate data effectively will be on full display.

As part of this experience, you and your partner will present your project to the broader CodeOp community in a presentation lasting no longer than 45 minutes (see details below). This is your chance to not only share your findings but also refine your storytelling and communication skills in a professional setting.

**PRESENTATION**

These well-rehearsed presentations will be given by all groups on **Thursday (30 January 2025) beginning at 18:30.** Each presentation will be no longer than 45 minutes with no exceptions, which will then be followed by a brief (up to 15 minute) question and answering session.

Like the individual project, the following ***must*** be included in the presentation:

1. Clear objectives and questions that you will be answering throughout the presentation
2. A description of the dataset and how it was obtained
3. Data preprocessing techniques used
4. Most important insights of the dataset primarily given with visualizations
5. Conclusion of the presentation

Note: These requirements do not necessarily need to be in this exact order.

**DELIVERABLES**

1. A single Powerpoint (or similar) presentation in PDF or PPTX format of the group EDA\*
2. A link to the Github (or similar) repository, which contains all Python code
   1. Note: Though the repo should preferably be set to *public*, if you decide to keep it as *private*, please share the link with your instructors

\* = You willl be presenting this to the rest of the CodeOp Community (see Presentation section)

**ADDITIONAL INFORMATION**

**Working in a group**

Every week, you and your partner(s) will **separately** turn in an evaluation form. Within this form, each member of the group will be asked about their own contributions to the project, their partner’s contribution to the project as well as have the opportunity to voice any comments/concerns.

**Important: If, at any time, there is a *significant* uneven distribution of work, the person who chooses to do little or no work will be removed from the team completely and not allowed to finish the project. Because the collaborative project is mandatory to complete the bootcamp, this student will thus not be allowed to receive any certification from CodeOp.**

**Dataset(s)**

You and your partner(s) must agree on the dataset(s) to work on. Keep in mind that these datasets must be approved by the instructor beforehand.

**Usage of SQL**

Similar to that of the individual project, it is completely optional for you and your group use SQL - though doing so may be helpful in the future. In the case that you have a single (or multiple) .csv files, these files could be manually inserted into an SQL database via pgAdmin4 or Postgres. From there, students can then query these tables from this SQL database from Python via *psycopg2* to simulate a real-world scenario of how data would be collected for data preprocessing.

**Usage of Python**

It is required that you and your group use Python. Python must be used to do data preprocessing / feature engineering and to also generate visuals, similar to that of the Module 3 milestone. However, unlike the previous EDA milestone, your group will instead be presenting a Powerpoint. Though this Jupyter notebook will not directly be shown during the presentation, this still must be clean, easy-to-understand and reproducible. After the presentation itself, your instructor will check your code.

**Usage of Additional Libraries**

In the case that you want to use additional libraries / packages to supplement your EDA, this is completely allowed and would actually be preferred. To give you an idea of some *extra* libraries that you might want to consider, here’s a small list:

* Scikit-learn: Arguably the most used machine learning library in Python that provides tools for traditional supervised and unsupervised learning tasks
* Plotly: An alternative to Seaborn but with added functionality / interactiveness
* Boto3/google-cloud/azure: SDKs for Python that enable programmatic interaction with AWS, Google Cloud, and Azure services, respectively
* Statsmodels: Library for statistical modeling and hypothesis testing, including linear regression, time series analysis, and ANOVA.
* NLTK: A library for processing and analyzing human language data, offering tools for tokenization, parsing, and text classification.
* spaCy: A fast, industrial-strength natural language processing library for tasks like part-of-speech tagging, named entity recognition, and dependency parsing.
* Huggingface: A library for natural language processing and machine learning, providing pre-trained models for tasks like text classification, translation, and question answering.

However, if you were to incorporate some type of machine learning algorithm from one of the packages listed, be very careful because you must know exactly what model you are presenting, the details of that model, its advantages and disadvantages, its assumptions, etc..