Final Project Guideline Document

Data Analytics INFO 8066

- **Marks** = 100
- **Project Grade Weight** = 35% (25% Group + 10% Individual Contribution)
- **Project Group size** = minimum 4,maximum 5 per group
- **Presentation Time** = 13-15 mins (max). **Marks will be deducted if the presentation is over time.**
- **Presentation Date:** Week 14 – Week 15
- **Final deliverable Due :** December 17th, 2023 (11:59pm)
    - **Must include:** Report(.doc), ppt presentation, Python file(.ipynb), Optional(Excel File)

Description: Your final project is meant to expose you to a full life cycle of a data project. You are required to understand the domain and present key insights from the real-world dataset keeping CRISP - DM framework in mind. The goal is to perform a comprehensive analysis of the chosen dataset, from data preprocessing to visualization, and create a project report and presentation summarizing their findings.

Real word open datasets can be found using any of the following links.

- Kaggle.com,
- https://datasetsearch.research.google.com/
- https://data.ontario.ca/dataset

You are not allowed to use datasets that have already been used during the course (ie: Tips.csv, Titanic.csv)

Cleaning and manipulation of data outside python or excel is not permitted. You can also create a prescriptive analytics model in EXCEL. Please note you are not limited to the above resources and are free to bring your own datasets if it satisfies the copyright laws.

Final presentations need to be in PowerPoint

Use of Tableau or PowerBI is not permitted

Key Project Components *(Ensure all components are covered in your ppt and/or report)*

- **Dataset Selection:** Students must choose a dataset from Kaggle that interests them, with a brief justification for their choice.
- **Data Acquisition:** Download the selected dataset and import it into both Excel and Python.
- **Data Preprocessing:** Clean and preprocess the data in both Excel and Python, addressing issues like missing values, outliers, and necessary transformations.
- **Exploratory Data Analysis (EDA):** Conduct EDA using Python, exploring and visualizing the dataset's key characteristics and patterns. Utilize Python libraries such as Pandas, Seaborn, and Matplotlib for this step.
- **Data Visualization:** Create a variety of visualizations in Python using Seaborn and Matplotlib, highlighting important

- **Linear Regression Analysis:** Incorporate linear regression analysis to model relationships within the dataset. Perform regression analysis using Python to understand the linear relationships between variables. insights and trends.
- **Data Analysis in Excel(Optional):** Utilize Excel to complement the analysis by performing additional calculations, creating pivot tables, and generating charts to support the insights gained from Python.
- **Excel What-If Analysis:** Apply Excel's What-If analysis tools to perform scenario analysis, sensitivity analysis, or other relevant analyses related to the dataset. This step can involve creating Excel models to understand how different variables impact the data.
- **Report and Presentation:** Prepare a project report that includes data summaries, visualizations, key findings, and insights. Create a visually appealing presentation to communicate the project's results effectively.
- **Documentation:** Document the entire project process, including data sources, preprocessing steps, EDA findings, and Excel What-If analysis. Ensure that both Excel and Python code are well-documented.

Project Report Outline Example (Add more as needed)

Project Title:
Briefly provide a title to your data analysis final project.
For example:

- "Analyzing customer churn in ABC phone company"
- "Identifying the Characteristics of Successful Kickstarter Campaigns"
- "Analyzing the Effects of Climate Change on Global Food Production"
- "Analyzing 5000 Movie ratings"
- "Determining the Drivers of Customer Satisfaction"

Group Member(s): Write the name of your group members

Introduction: Introduce your dataset and your reason for choosing it. Provide background research on the topic.

Client: Mention the Ideal End user of your analysis findings (please be specific and define your end user at an operational level)
For example, instead of simply writing "the end user is Netflix organization", write, "the end user of our analysis is the Chief Revenue Manager of the Movie streaming giant Netflix, who is looking for studying user preferences with the ultimate goal of increasing her company's revenue"

Operational Analytics Problem Statement: Include client problem (something you must make up based on the dataset your group chose): This includes problem background and how is this problem curtailing company's growth or operations. For example, a project that involves analyzing a company's online sales, you can define the problem as: A large online retailer is not able to keep up with the competition in terms of online sales and as such has been losing market share to its competitors over the past few years. This company is also not able to attract new customers and retain its existing customer base. Hence this company has hired a team of data professionals to investigate their customer demographics and online transactions to find insights and ultimately recommend key decisions to alleviate the problem.

Project Deliverable:
Explain at a high level the components of your project's data analysis findings.
For example;
Key metrics:

- Number of users
- Number of sessions
- Average session duration
- Bounce rate

Insights:

- User behavior
- What users are doing on the site
- What areas need improvement

Recommendations:

- Changes to improve the user experience
- New features or functionality to consider

Consider this as a paid consultant project, your deliverables are binding terms with client and will dictate the overall project completion and quality.

How does your solution provide value?

Please explain clearly how your solution can help solve the institution's operational problem. For example, if a health care institution wants to reduce patient wait times, you might discuss how your data analysis report can help identify areas where wait times are the longest and where process improvements can be made. This will help the institution to focus on those areas in order to reduce patient wait times.