INFO8066 – DATA ANALYTICS

FINAL PROJECT - DATA VISUALIZAITON

DUE: END OF WEEK 15

WORTH: 35%

Late Submissions will not be accepted

Introduction:

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.

General Requirements:

This section contains information about the general requirements that your final project must meet. *Please read all requirements carefully before you start.*

- You must do the analysis in Jupyter Notebook
- Part of this assignment will include a written report, this must be in a PDF format.
- Please ensure that your submission follows the naming rules specified: INFO8066_Group#_Final_Project.ipynb
- Submission Requirements: **.ipynb file + .pdf report + .ppt file**
- Use of Tableau or PowerBI is not permitted
- Excel What – if analysis is optional
- You are not allowed to use datasets that have been used during the course(ie: Tips.csv, Titanic.csv, any other)

Scrum Meeting (2%)

All members of the group are required to attend the in-person scrum meeting scheduled for Week 14.

Deliverable I : Jupyter Notebook .pynb

Task 1: Data Selection (2%)

Students must choose a dataset from Kaggle that interests them, with a brief justification for their choice including a personal connection to the topic.

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

Task 2: Problem Statement (2%)

For this section, you are tasked with defining the business problem statement that your data analysis aims to address. This section is crucial as it sets the foundation for the entire project, guiding your analysis towards a meaningful and actionable outcome.

Task 3: Data Pre-Processing (2%):

In this section, you will focus on cleaning and preprocessing the raw data to make it suitable for analysis. Data preprocessing is a critical step that involves handling missing values, outliers, and performing necessary transformations to ensure the quality and integrity of your dataset. You will be required to utilize Python, leveraging a selection of **2 - 5+** Python pandas functions taught in class to accomplish this task effectively.

Task 4: Exploratory Data Analysis (2%)

In this section, you will conduct an exploration of the dataset using a variety of EDA methods. Your goal is to understand the underlying patterns, relationships, and characteristics of the data before arriving at your final conclusions. You are encouraged to go beyond the scope of the initial requirements by incorporating additional research from other sources to enhance your understanding and explanation of the findings.

Task 5: Data Visualization (5%)

In this section, you will focus on creating a series of compelling visualizations to effectively communicate insights derived from the data. Your goal is to produce visualizations, exploring a variety of marks and encodings to convey meaningful information. Each visualization should not only be expressive but also effectively designed to enhance understanding and facilitate interpretation. Additionally, you are required to provide rich captions that describe the visualizations and contextualize the insights within the broader analysis.

Task 6: Predictive Model (2%)

In this section, you will incorporate predictive modeling techniques, specifically linear regression analysis, to model relationships within the dataset. Your goal is to perform regression analysis using Python to understand the linear relationships between variables and derive insights and trends from the data. Additionally, you are encouraged to create visualizations to demonstrate the linear regression found and provide a clear justification for using predictive modeling in this context.

Deliverable II : Report (4%)

Write your report using a word doc and save it as a PDF file called INFO8066_Group#_Final_Project.pdf (It must be in a PDF format). Your report should adhere to strict formatting and writing standards to ensure clarity, professionalism, and adherence to APA citation guidelines.

Your report should be ~10 – 15 pages (including visualizations). The report should contain the following sections. If you want you can use a different structure but ensure all the necessary sections are included.

Running header on the top left corner of the report will be “INFO8066: Data Analytics – [Title]”

1. **Title:** The title of your project
2. **Team Members:** Provide the names, student numbers, and email addresses of all your team members
3. **Group Number:** Mention the group number assigned.
4. **Course Instructor:** Enter your instructor’s name. Ensure correct spelling.
5. **Introduction:** Introduce the dataset and provide background research on your chosen industry. Discuss the rationale for choosing this topic and dataset, including any personal connections or interests.
6. **Business Value/Problem Statement:**
7. **Data :** Describe the dataset you are using for the project
8. **Data Pre Processing :** Detail the preprocessing steps undertaken to clean and prepare the dataset for analysis. Include python code and output screenshots from .ipynb file
9. **EDA**
10. **Data Visualization Results:** Report the qualitative and quantitative results.
11. **Predictive Modelling :** Discuss the performance and evaluation of the predictive models
12. **Discussion / Analysis:** Provide recommendation based on the analysis. Discuss the implications of the findings and potential strategies for the target client

Deliverable III : PowerPoint (4%)

Ensure your PowerPoint presentation reflects your understanding of the topic with original content. Maintain design consistency, proofread for errors, use animations effectively, and aim for a business professional appearance.

Individual Contribution – Overall Presentation Skills (10%)

Criteria include clear articulation, appropriate volume, professional attire, in-depth knowledge, engagement through eye contact, and accurate answers to questions. Practice and preparation are key to success in this area.

Success Criteria:

This assessment's overall weight can be found on the Instructional Plan (IP) and reflected within the eConestoga grade book. Any student that discovers a conflict existing between the IP and grade book shall notify their course faculty member.

For specific evaluation standards, students shall consult the associated assessment rubric found in the Rubrics secton of eConestoga.

- Failure to submit an assessment by the specified end date, to the correct dropbox, will result in a grade of zero (0).
- No opportunity will be provided to make up for an unsubmitted deliverable.
- It is the student's responsibility to ensure that their work has been submitted through eConestoga, on time, to the correct course and in the correct folder.
- Be aware that Conestoga College's Academic Offense policy will be enforced.