Final Project - Data Science with Python

Goal: Detect **three** interesting, nontrivial, and somewhat unexpected finding in a real-world data set of your choice. Nice data sets can be found on www.kaggle.com and www.data.gov. This project must be done in group (of two or three people depends on class size). It is meant to be a significant effort to learn by practicing what you are learning to a real-world data science problem.

- **Submission**: **One submission into Camino per group**. (**three files)**
    - The data set that you used (in **zipped** format. If after zipped, still more than 100M, use google drive and share the file)
    - The write-up of your final project in the form of a Jupyter Notebook. Make sure that your notebook runs without any error, assuming that notebook and data set are in the same folder.
    - If you use some 3rd party interactive graphic tools(for ex: plotly) to generate graphs, you need to submit a html version of your Jupyter Notebook
- **Deadline**: You must submit your files before due date at midnight.
- **Format and grades (out of 100 pts)**: your notebook should have the following format:
    - **(5 pts.) Data set description**. Since everyone is using different data, please make sure to clarify what is in your data set (rows & columns). Points are awarded based on clarity and brevity.
    - **(25 pts.) One section on data clean-up/preparation**. You will surely need to create new columns, or clean existing ones, … etc. So in this part you will report all the code of the cleaning/preparation process. Your data preparation must be entirely performed in the notebook (and not, for example, in Excel before being imported in the notebook). Points are awarded based on clarity, brevity, and showing that you can use the techniques learned in class.
    - **(60 pts. = 20 pts per finding) Quality of Findings**:
        - **(5 pts per finding) A brief summary** of the finding (ideally, only one sentence) addressed to the management. Points are awarded based on clarity and interestingness of the finding
        - **(10 pts per finding) A set of tables and/or charts** that support the validity of your finding. You must include at least one chart. Points are awarded based on how convincing your evidence is.
        - **(5 pts per finding) Managerial insights**: answer the "So what?" question. That is, convince the reader that your finding can be used to improve operations and increase

profit. Points are awarded based on clarity, brevity, and how actionable your finding is.

- **(10 pts) The readability, e.g., clarity/format/layout/attention to details of the whole Jupyter Notebook**

Number of findings:

If you submit more than 3 findings, you will be graded only on the first three.

SOME EXAMPLES OF SECTIONS

Note: these examples briefly illustrate the type of content expected in some of the sections of the .ipynb notebook. Most of the examples below are not finished, do not contain any code, and would not necessarily a high score.

Description of data set:

The data set has:

- One row for each purchase made at a large electronic retailer
- The following columns: Price, Product ID, Customer ID, Purchase Date, Customer's Gender and age, and a binary attribute RETURN which indicates whether the purchase was later returned to the store

Summary of the finding:

The higher the price or the product purchased, the more likely the customer is to return the product

Validity of the finding:

Price of product purchased ($)	Return Probability	Number of purchases in that price range
0 – 50	8%	12,000
50 – 100	10%	8,000
100 – 150	15%	7,000
150 – 200	20%	2,000
200+	28%	3,000

(you'll need to include a chart)

Managerial insights:

The finding is explained by the fact that the more expensive the product is, the less willing customers are to accept a poor fit between their needs and the product characteristics. We could use this finding as follows: whenever someone purchases an expensive product, we could give them a 10% discount to waive their right to return the purchased product.