OMIS 114: Data Science with Python
Santa Clara University Leavey School of Business
Fall 2024

Final Group Project

Objective: The objective of this project is to answer **a meaningful, data-driven research question** by curating **three interesting, nontrivial, and somewhat unexpected findings** from a real-world dataset. You will choose a dataset (e.g. as suggested on Storytelling with Data's Public Data sources, data.world, UCI's Machine Learning Repository, Kaggle, data.gov) and work as a group to explore, analyze, and present your findings. You are also welcome to partner with local organizations that have data available for analysis, if they are willing to share such data.

This project is designed to be a significant effort, allowing you to practice the data science skills learned in class by solving a real-world problem. Your project will be evaluated on the quality and relevance of the findings to the central research question and on the contributions of each group member.

Submission: Your group will submit the following:

1. **Dataset:** The dataset(s) that you used (in **zipped** format). If the zipped file is still more than 10MB, share the file with the instructor using Google Drive.
2. **Presentation Deck:** Your slide deck used for for your in-class final presentation.
3. **Final Write-up (Jupyter notebook):** Make sure that your notebook runs without errors, assuming that notebook and data set are in the same folder. Each student should clearly label the sections they contributed to.
    - Optional HTML version: If you use third-party interactive graphic tools (e.g., Plotly) to generate graphs, submit an HTML version of the Jupyter notebook.

    Exploring new packages, commands in Pandas which we did not cover in class, etc. is encouraged; just make sure to describe what they are, if you use them.

Format and grades (out of 150 pts): Your notebook should cover items 1-4, preferably in order. Note the other items (5 onward) in informing your group's efforts.

1. **(20 points) Research Question & Hypotheses + Data Description**
    a. **Formulate a Central Research Question:** Your project should revolve around a central research question/theme (e.g., "What factors influence employee retention?" "How can we reduce customer churn?") Points awarded on clarity and relevance.
    b. **Hypothesis Development:** Develop 3 hypotheses that you aim to test using the data. Each hypothesis should be logically connected to the research question for the above questions.
    c. **Explain your dataset(s):** What is the dataset about? What are the variables of interest? Provide a data dictionary to define each variable's meaning.
    d. **Define the unit of analysis**: Clarify the unit of analysis in your dataset (what does each observation represent)? You may need to do this repeatedly if you transform your data.
    Points are awarded based on clarity and brevity.

2. **(20 pts.) Data Preparation**.
    a. Report all code related to cleaning and preparing the data. This may include creating new columns, cleaning existing ones, handling missing data and/or filtering data. Your data preparation should be performed directly in the notebook.

Points are awarded based on clarity, brevity, and showing that you can use the techniques learned in class.

3. **(60 pts. = 20 pts per finding) Three findings relevant to your research question**
   Present three findings that help answer your research question/ test your hypotheses. Each finding should be nontrivial.
   For each finding:
   a. **(5 pts per finding)** A brief summary of the finding (ideally, only one sentence) addressed to management. Points are awarded based on clarity and interestingness of the finding.
   b. **(10 pts per finding) A set of tables, graphs, and/or models** that support the validity of your finding. You must include at least one chart. Points are awarded based on how convincing your evidence is.
   c. **(5 pts per finding) Managerial insights:** answer the "So what?" question. That is, convince the reader that your finding can be used to improve business operations and strategy.
   Points are awarded based on clarity, brevity, and how actionable your finding is.

4. **(10 pts) Presentation and Readability**
   a. Ensure the Jupyter Notebook is well-organized, with appropriate headings, clear explanations, and attention to detail.
   b. Formatting and Layout: Use clean formatting and provide sufficient explanations to make your work easy to follow.

5. **(25 points) Peer Evaluation by Group Members**
   a. Each student must take responsibility for a specific section of the project, such as data preparation, analysis, visualization, or writing. Label at the top of the notebook which part(s) of the notebook each group member was responsible for.
   b. Each group member will provide detailed feedback on the contributions of others, focusing on effort, collaboration, and quality of work.
   Points are awarded based on the quality and detail of the peer evaluations.

6. **(5 points) Progress Checkpoints**
   a. Project milestones must be submitted on Camino at their requested times.

7. **Peer Evaluation by Other Groups (10 pts)**
   a. Each group will review another group's project, focusing on clarity, relevance of findings, and insights for operations/strategy/policymaking.

Deadline: You must submit your files before the due date/ time (end of the finals exam time period)

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

How do you come up with a theme and hypotheses? Pick out questions in a related industry/cause that draw interests to you. Then, think about potential solutions/drivers that attenuate the issue at hand. In terms of the question asking, you can be creative and start with the question and think about the related data needed to answer the question; on the other hand, you may have a great dataset and ask questions based on what is available.

OMIS 114: Data Science with Python
Santa Clara University Leavey School of Business
Fall 2024

A Short Example

Note: This example briefly illustrates the type of content expected in some of the sections of the .ipynb notebook. The example below is not necessarily finished/ does not contain code, and would not necessarily receive a high score.

Research Question: Understanding Drivers of Customer Returns for An Online Retailer
Hypothesis: Higher prices are associated with higher return probabilities.

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

(show other relevant summary statistics, data visualizations, etc. as desired)

Managerial insights:

The analysis suggests the more expensive the product is, the more likely one may return the product. This may be explained by customers becoming less likely to accept a poor fit between their needs and the product characteristics for more expensive products.

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.