Midterm Project

Exploratory Data Analysis

Due: 3/24

EDA Midterm Project Purpose

The course midterm project is meant to assess your understanding of the first half of the course. In this assignment, you will find explain and analyze a dataset based on a question of interest of your choosing. You will create a Python notebook that includes the code an analysis of your work. You will also create a short breakdown of important aspects and insights of the project in a written "blog post." There is no word requirement for the blog post, but some examples will be given. You can keep the blog post brief, as long as you hit all the required aspects.

Project Breakdown

Business Understanding

- Your project should start with a question you want to learn more about.
- Provide a background into this question. Why did you choose it? What is interesting about it?
- **This should be written in your final written blog post**
- **Important Note**: Your overall question of interest should be very broad (and dataset dependent). How has honey production changed in the past 10 years? What are the major trends that result in hotel bookings? What are some trends found in people that survived or didn't survive the Titanic? Within your analysis you may find there is a more specific or interesting trend you want to analyze further, there also may not be. Both are fine. Start broad and narrow your report, it makes it easier to find a dataset and create your analysis

Dataset Selection

- Select a dataset online related to your project purpose to help you answer your question above (hint: it is easier to select your dataset beforehand)
- Provide the source of your dataset, including any information you have about how it was collected etc
- Provide a sentence for why this dataset is related to the question you want to answer
- **This should be written in your final written blog post**
- **Important Note**: The dataset you collect must have at least 6 features

Data Preparation and Understanding (Cleaning and Univariate Analysis)

- Use pandas modules to explore your dataset in your python notebook
- Perform needed Data Cleaning in your python notebook
    - Remove any un-needed features
    - Create any new columns
    - Clean Up Column Names
    - Clean Up Column Values
- Describe any data cleaning steps in your written report
- In your written blog post, provide each features
- With your final features decided, perform Univaritate Analysis on each of your features on your features **in your python notebook**
- **Provide a brief description of your univariate analysis in your written blog post**. If there is anything more interesting, you can spend more time in the notebook looking at that

Multi-Variate Analysis - In Python Notebook

- Once you have your final features decided, look more into your dataset to perform multi-variate analysis between features
- You need to perform both bi-variate and multi variate analysis in your Python notebook

- In your Python notebook, you should explain some basic analyses that you find in a markdown cell, and why you want to move to your next analysis
- These plots do not need to be beautiful. These are your "working plots." They just need to be set up in a way that makes sense.

Final Plots - Written Report

- After finishing your analysis, you want to look over your final Python notebook and grab at least 3 plots (but no more than 6 - discuss with the Professor if you feel you need more) that help provide major takeaways from your project. These can be the simplest plots you have, or the worst plots.
- Explain these plots in your blog post report. Provide basic insights about the graphs (such as trends, outliers, skew if histogram distributions) as well as problem specific insights (what does this say about the question you are asking.
- These final plots should have a title, x-axis, y-axis and readable legend (if needed).

Take Away and Future Work

- At the end of your blog post, summarize your findings in bullet points
- Provide at least two directions of future work possible (using this original dataset, combining additional outside data, and/or a new problem you want to explore based on these findings)

Project Deliverable

Written Report

The written report should help summarize your goal for your analysis, the data used, and overall insights discovered. Any key issues you ran into or interesting findings you discovered on the way of finalizing your analysis should also be included. The blog post is informal. Think of the post as if you ran into someone in the hallway and they asked you to explain your project to them. What is it about? Where did you get the data? What problems did you run into? What did you discover about the data?

Python Notebook

The Python notebook for the project should be well organized. It should follow the break down above with legible breaks in your Python Notebook.

EDA Midterm Project Rubric

Overview

This rubric is here to help you understand the expectations for the analysis you create. It is the same rubric that will be used to evaluate each project. You should look at the rubric before you begin working on your analysis and before you submit it.

How to Use: Before you Begin

1. Look at the bold headings under the criteria column to understand what we will be looking for in your project
2. Go through each criteria item in more detail
3. Familiarize yourself with what is required in the process

How to Use: Before you Submit

1. Once your analysis is complete, go through each criterion and do your best to honestly evaluate where you think your project fails
2. If you think your project “does not meet specifications”, then you can make sme changes
3. Once you are confident, go ahead and submit

Criteria	Does not meet specifications	Meets specifications	Exceed specifications
Code Readability			
Is there a runnable .pynb notebook?	There is no .ipynb notebook submitted	There is a working .ipynb notebook submitted	NA
Does the code use proper formatting techniques (indents, spaces, line breaks) to help with readability?	The code does not use formatting techniques or formatting techniques do not improve readability. Some lines are longer than 80 characters	The code uses formatting techniques and they improve readability. All lines are shorter than 80 characters	The code uses formatting techniques in a consistent and effective manner to improve code readability. All complex code is explained with comments. All lines are shorter than 80 characters.
Are there cells in the notebook saved with both markdown and python cells?	There are no markdown cells in the python notebook	There are both python markdown and coding cells	There are both python markdown and coding cells. These cells are well organized and set up in a way that is easy to read and understand.
Code Functionality and Logic			
Does the code work?	Some code is not functional	All code is functional,	Not Applicable
Does the student use good coding practices?	Student sometimes uses repetitive code where a function would be more appropriate. The code uses constants or column numbers to access variables or subsets of data.	Student almost never uses repetitive code where a function would be more appropriate. The code references variables by name instead of using constants or column numbers.	The code is never repetitive and makes use of functions where appropriate and uses sound practices to access variables, subset data, or perform complex operations.
Depth of Analysis			
Is the data set explored in many ways?	The student does not appropriately use univariate,	The student appropriately uses univariate, bivariate, and	The student appropriately uses different plot types to explore

	bivariate, and multivariate plots to explore most of the expected relationships in the data set.	multivariate plots to explore most of the expected relationships of interest in the data set.	expected and unexpected relationships in the data. A variety of leading questions, dead-ends, and alternate approaches are presented.
Are questions and observations included as text throughout the analysis?	Questions or findings are missing in multiple places which make it unclear what the student was thinking or what the student found	Questions and findings are placed in between blocks of R code regularly, so it is clear what the student was thinking throughout the analysis	Not Applicable
Are there a variety of relevant visualizations and statistical summaries?	The student creates less than 10 visualizations. Relevant statistics, such as means, medians, quartiles, or confidence intervals are often not reported	The student creates at least 10 visualizations. The visualizations are varied and show multiple comparisons and trends. Relevant statistics are computed throughout the analysis.	The student creates a variety of visualizations that shw multiple comparisons and trends. Relevant statistics are calculated throughout the analysis are included in (or discussed in a paragraph with) the visualizations
To what extent did the student demonstrate mastery of EDA?	The student does not successfully make use of tools and techniques from the course	Tools and techniques are applied successfully and sometimes creatively	Student has used tools and techniques creatively and with stated purpose
Final Plots and Summary			
Are the final plots varied and do they meet some of the following criteria: ● Draw comparisons ● Identify trends ● Engage a wide audience ● Explain a complicated finding ● Clarify a gap between perception and reality ● Enable the reader to digest large amounts of	The student does not include final plots	The plots are well chosen and the plots fulfill at least 2 of the criteria. The plots are varied annd reveal interesting trends and relationships	Each plot reveals an important and different comparison or trend in the data. The plots incorporate many of the variables from the data set in a way that allows the plots to convey a lot of information while still be interpreted easily. The plots fulfill 4 or more of the criteria

information			
Are the plots explained ?	The reasoning and findings are not explained for each plot or the text about one plot is not descriptive enough to stand alone	The reasonings and findings from each plot are explained and the text about each plot is descriptive enough to stand alone	The reasonings and findings from each plot are explained concisely with appropriate variable transformations, other plot decisions, and/or statistics. The text about each figure is short, descriptive, and adds information that the graphic itself would not easily explain
Are the plots polished ?	One or more plots are missing axes labels, plot titles, axes unitis, or are scaled inappropriately	All plots are labeled appropriately and can be read and interpreted easily	Not applicable
Reflection			
Does the student provide a written reflection of the analysis?	The student does not include a reflection as the last section of the slides. R the student does not communicate struggles, success, and/or ideas for improvement	The student reflects on how the analysis was conducted and reports on the struggles and successes throughout the analysis. The student provides at least one idea or question for future work	The student provides a rich and well-written reflection of struggles, successes, and lessons learned. The student poses ideas or questions for future work. The student explains anny important decisions in the analysis and how those decisions affected the analysis