UDACITY

< Return to Classroom

Investigate a Dataset

REVIEW

HISTORY

Meets Specifications

Congratulations 🎓
You've made it!!
This is a fantastic submission!

You took a lot of variables from the dataset and did an excellent job of systematically exploring it and coming up with some interesting findings. I've really liked the way you've structured your project! Please do check specific comments.

Be a lifelong learner.
Stay Safe and Stay Udacious!

Further Readings for the curious YOU:
Exploratory Data Analysis
A topic that is neglected in Data Science Projects
Value of Exploratory Data Analysis
Why Exploratory Data Analysis is important
Top 10 Data Analysis Tools for Business

Code Functionality

- **All code is functional and produces no errors when run.**
- **The code given is sufficient to reproduce the results described.**

The submitted code works well as it doesn't produce errors when run. Also, it's sufficient to reproduce the results described

results described.

The coding structure and logical flow of your coding practices are impressive. Keep up the good work.

- **The project uses NumPy arrays and Pandas Series and DataFrames where appropriate rather than Python lists and dictionaries.**
- **Where possible, vectorized operations and built-in functions are used instead of loops.**

You have shown a good understanding of built-in functions and the correct application thereof.

Learning Notes
Here are a couple of links with useful pandas and numpys built-in functions and methods:
Pandas Cheat Sheet
Numpy Cheat Sheet

- **The code makes use of at least 1 function to avoid repetitive code.**
- **The code contains good comments and meaningful variable names, making it easy to read.**

Things done well

- You have given a brief introduction about the dataset you will be working on. That's a really good practice for a reader (or yourself when you will revisit the project in future) to get an idea about the dataset and it's realted fields.
- You have also posed the questions you will be exploring at the beginning of the project. This is good practice as it lays the ground work for the rest of the project and give you a direction to think and analyze the dataset even before delving deep into it.
- The structure of your notebook is clean and has a logical flow. Different sections are clearly shown for each one of the steps of the data wrangling process.
- Your code is properly commented and contain good variable names which is making your code easy to read. You have incorporated many markdowns and in-code comments. Markdowns are very important as this allows the reader to follow along with the intentions of the author. Good job!

Suggested Readings

Six steps to more professional data science code
Reviewing Data Science Projects

Quality of Analysis

- **The project clearly states one or more questions, then addresses those questions in the rest of the analysis.**

Things done well

You have done a great job with the questions you have posed! Also, each of these questions are explored thoroughly.

In EDA, great questions help you focus on relevant parts of your data and direct your analysis towards meaningful insights. Questions should be measurable, clear and concise. They should be designed to either qualify or disqualify potential solutions to your specific problem or opportunity. You have gone above and beyond in this area :)

Suggested Readings

Your Data Won't Speak Unless You Ask It The Right Data Analysis Questions
How Do Data Scientists Ask the Right Questions?

Data Wrangling Phase

- **The project documents the steps that were taken to clean the data, such as merging multiple files, handling missing values, etc.**

The structure of your notebook is clean and has a logical flow. Comments are used to clearly identify each one of the steps of the data wrangling process. You have correctly identified issues regarding the datasets.

Comments:

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. It is about targeting a field, row or column in a data set, and applying an action such as joining, parsing, cleansing, consolidating, or filtering to create the desired output, which will then be used down the road. Data wrangling involves activities like:

• Remove unused columns.
• Remove duplicate rows.
• Change data formats (date columns)
• Discard missing values.

Benefits:

• it makes your data useful
• it can be organized into a standardized and repeatable process that moves and transforms data sources into a common format, which can be reused multiple times.

Suggested Readings

The Growing Importance Of Data Cleaning
Data Cleaning Using Python Pandas
Data Cleaning with Python and Pandas: Detecting Missing Values

Exploration Phase

- The project investigates the stated question(s) from multiple angles.

- The project explores at least three variables in relation to the primary question. This can be an exploratory relationship between three variables of interest, or looking at how two independent variables relate to a single dependent variable of interest.
- The project performs both single-variable (1d) and multiple-variable (2d) explorations.

The analysis makes good use of both single (1D) and multiple (2D) variable explorations to investigate different features and the relations between these features in the dataset. Questions are investigated from a single variable perspective as well as a multiple-variable perspective. Good Job!

Learning note:
Below is a table that allows us to see the distinction between 1D and 2D explorations:

Summary: Differences between univariate and bivariate data.

Univariate Data	Bivariate Data
• involving a **single variable**	• involving **two variables**
• does not deal with causes or relationships	• deals with causes or relationships
• the major purpose of univariate analysis is to describe	• the major purpose of bivariate analysis is to explain
• central tendency - mean, mode, median • dispersion - range, variance, max, min, quartiles, standard deviation. • frequency distributions • bar graph, histogram, pie chart, line graph, box-and-whisker plot	• analysis of two variables simultaneously • correlations • comparisons, relationships, causes, explanations • tables where one variable is contingent on the values of the other variable. • independent and dependent variables
Sample question: How many of the students in the freshman class are female?	**Sample question:** Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

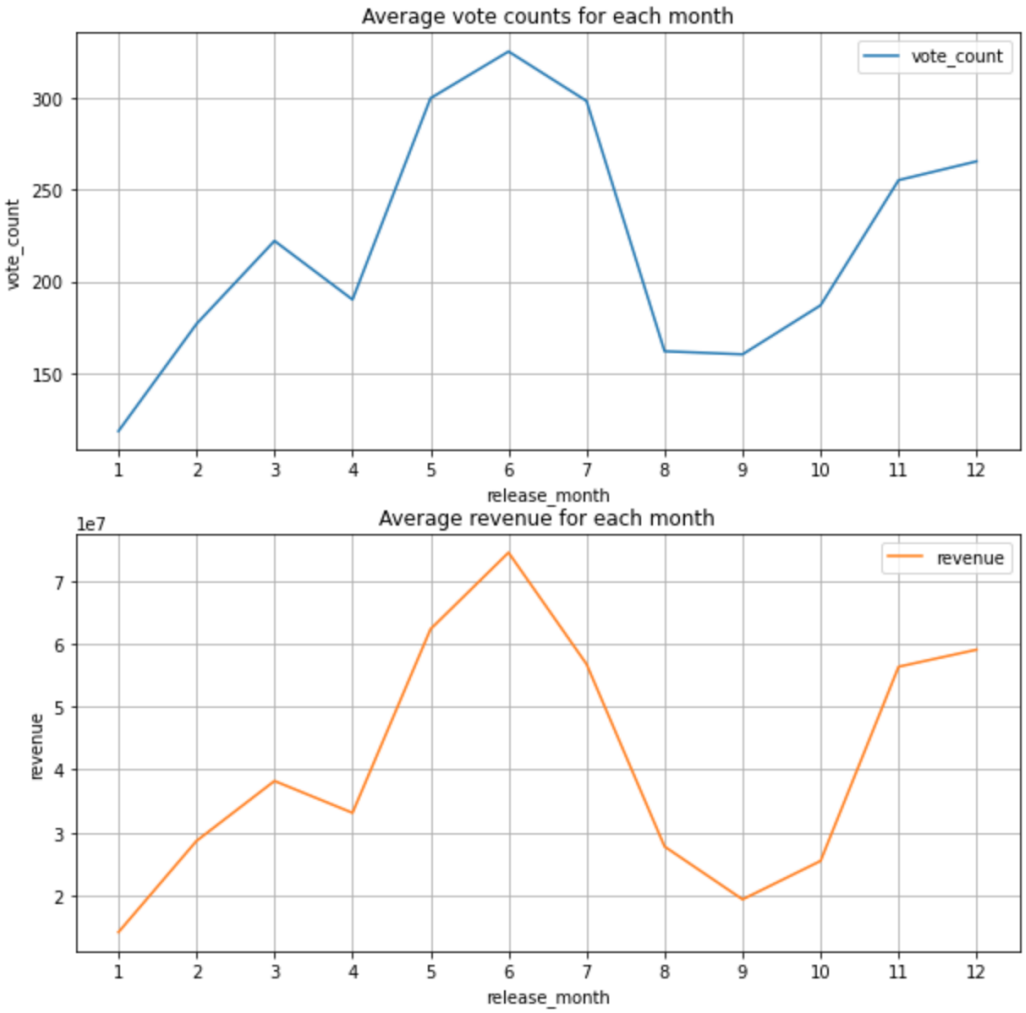
Suggested Readings

1D and 2D Variable Explorations
A Comprehensive Guide to Data Exploration

- The project's visualizations are varied and show multiple comparisons and trends.
- At least two kinds of plots should be created as part of the explorations.
- Relevant statistics are computed throughout the analysis when an inference is made about the data.

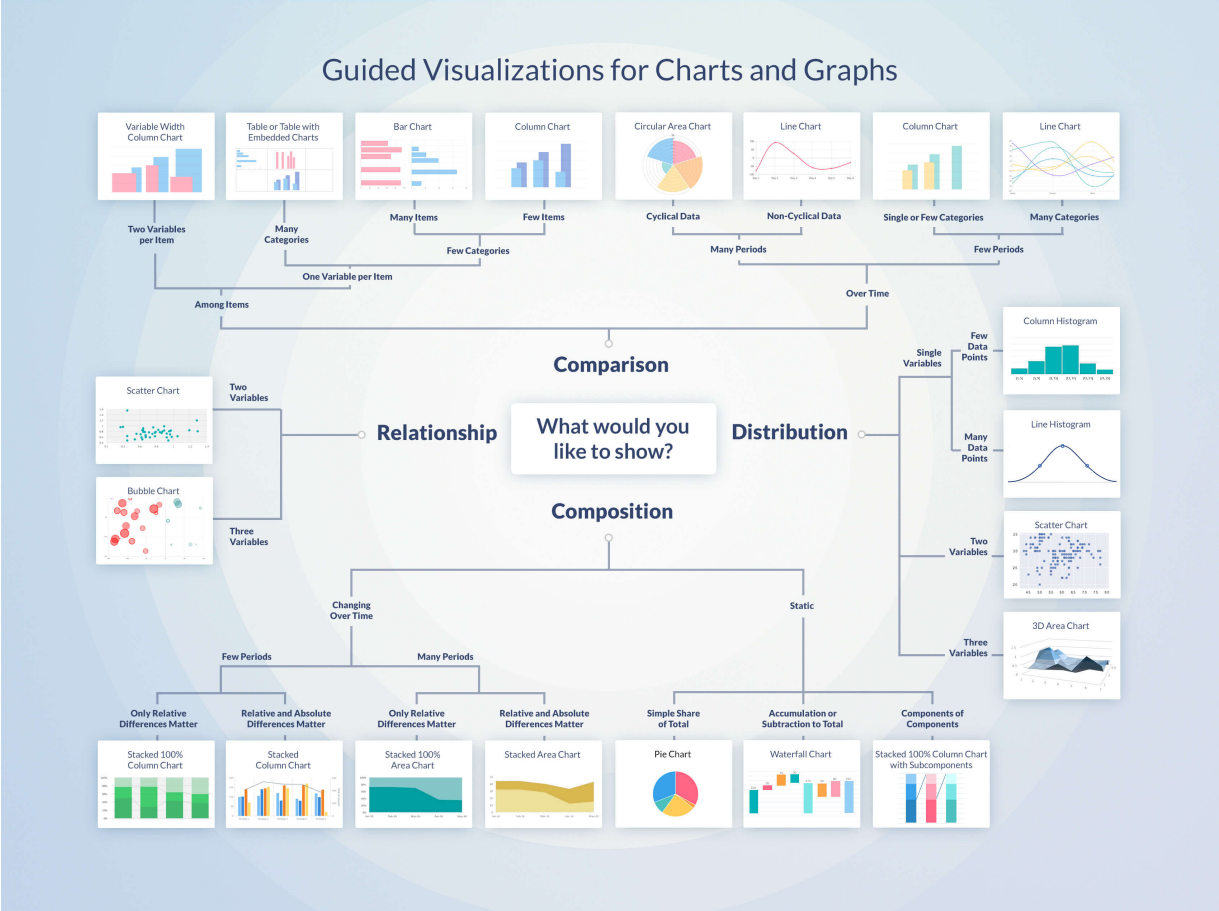
The visualizations appropriately support the investigation of the data set.

Visualizing data requires a lot of patience and determination because it's not easy selecting the best visualization to match with a given data type. Well enough, the project rightly builds descriptive visualizations using a variety of plots. I personally found this visualization really insightful.



Learning note:

Below is the screenshot I have provided in relation to plotting data. I have found this very useful when deciding what types of plots to use. I hope this will be useful for you in future too:



Suggested Readings

A step-by-step guide to Data Visualizations in Python
Data visualization with different Charts in Python

Conclusions Phase

- The Conclusions have reflected on the steps taken during the data exploration.
- The Conclusions have summarized the main findings in relation to the question(s) provided at the beginning of the analysis accurately.
- The project has pointed out where additional research can be done or where additional information could be useful.
- The conclusion should have at least 1 limitation explained clearly.
- The analysis does not state or imply that one change causes another based solely on a correlation.

Good Job adding Conclusions and Limitations to your project summarising your findings and figuring out what issues this dataset lacked or what you dropped affecting the overall analysis

Communication

- The code should have ideally the following sections: Introduction; Questions; Data Wrangling; Exploratory Data Analysis; Conclusions, Limitation.
- Reasoning is provided for each analysis decision, plot, and statistical summary.
- Interpretation of plots and application of statistical tests should be correct and without error.
- Comments are used within the code cells.
- Documented the flow of analysis in the mark-down cells.

Well done!

It is very important to communicate the results adequately; however, it is also very important to describe each activity, analysis, or graph. This will allow your audience to understand what you are doing and how you are doing it. Moreover, the reasoning makes your work organized, formal, and sophisticated.

- Visualizations made in the project depict the data in an appropriate manner (i.e., has appropriate labels, scale, legends, and plot type) that allows plots to be readily interpreted.

Visualization presented clearly depict the data and represent the questions posed. The plots are well structured and easy to interpret. The plots have clearly represented titles and labels. Good job.

Comments:

One of the most important steps in creating an impactful visualization is making sure all of its elements are labeled appropriately. The text components of a graph give your reader visual clues that help your data tell a story and should allow your graph to stand alone, outside of any supporting narrative.

DOWNLOAD PROJECT