UDACITY

‹ Return to "Data Analyst Nanodegree" in the classroom

Investigate a Dataset

REVIEW

HISTORY

Meets Specifications

Congratulations...!!! 🎉

You current submission fulfill all the requirements mentioned by the previous reviewer..... Well Done 👍

- This was a great implementation and I congratulate you for passing all rubric items with this submission.
- The submission does a really good job in many phases and I appreciate your hard work in achieving that. I have given some useful tips and suggestions with links to help you out in your future assignments.. I suggest you to check them out in your free time..!!
- It was delightful reviewing your work as it was well-thought-out.
- I encourage you to keep up the good work as it will make you a great Data Analyst. Way to go! 👏

All the best for your upcoming projects..!!! 😄

Code Functionality

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

Good Work..!!! 💯

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

- I appreciate that you have organised your code and have taken care of markdown cell and code cells as per relevance. This is a good portrayal of a planned and organised submission!!

TIPS: ⚡ ⚡

- It is always recommended that you handle your errors by segregating the erroneous block of codes into the singular ones run them line by line to pinpoint the main issue. This is frequently suggested and practised by top-notch coders.
- Jupyter notebook is a very powerful tool to document your codes and comments alongside.
- It helps you to segregate different blocks of code for better error handling along with suitable headings, comments and conclusions in different types of cells. This provides a focused approach and helps to establish a better connection with your audience.

`You have truly developed this skill and the submission portrays it clearly..!! Well done` 😃

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.

Nice work in using Pandas library to facilitate the work for this submission!!!👍

TIPS and SUGGESTIONS:

- Pandas is a very handy and powerful python library for handling data frames and various tedious tasks as hand.
    - Link1
    - Link2
- Here's are two links on a number of tips and tricks which we can use when using pandas....!! I encourage you to check it out in your free time! 😃

Learning Notes 📃

Some important Pandas built-in functions:

- Value-Counts
- Indexing and Selecting data
- Apply, Map
- Group-by

The code makes use of functions to avoid repetitive code. The code contains good comments and variable names, making it easy to read.

The project nicely avoids many of the repetitive blocks of code by using pre-defined functions in the submission!!

- I also appreciate that you have organised your code and have taken care of markdown cell and code cells as per relevance. This is a good portrayal of a planned and organised submission!! 😄
- Comments and appropriate variable names are essential for a good coder.
- These not only guide the viewer through the code but also helps in understanding it easily. You have portrayed these skills well... Keep up this good work in future too...👍

Quality of Analysis

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

Nicely framed..!! The submission rightly states and addresses multiple insightful questions in the analysis.. 💯

- All the questions are relevant and have been addressed to in the analysis and relevant visualisations have been framed.... 😃
- I also appreciate that you have included all the questions in the introduction section too. This certainly helps in getting pre-idea of what will be investigated ahead in the project. Good work..!

> # Introduction
>
> TMDB-Movie Dataset was genrated by 'The Movie Database API' https://www.kaggle.com/tmdb/themoviedb.org that contains metadata on ~5000 movies like genre, popularity, cast, budget, revenue etc of each film. By exploring this data, we can draw the insights or find answers to some interesting questions like:
>
> - Which movies earned the highest profit?
> - Which celebrity appeared the most in the movies?
> - Which movies received the highest ratings by the viewers? Did the movies that earned the highest profit received the highest ratings by the viewers?
> - Which movie genre earned the highest profit?

Data Wrangling Phase

The project documents any changes that were made to clean the data, such as merging multiple files, handling missing values, etc.

The submission contains a separate section of data wrangling and proper steps have been taken to identify missing values, duplicates or non - important columns and resolving them relevantly. Good Job...!! 👏

Suggestions and comments:

Data Wrangling is aimed at cleaning the data and also transforming it into a state which can be easily analyzed.

Note that, having uncleaned data could invalidate the analysis or provide inaccurate results. These are a few steps to take before analysis.

- Identify missing values in the dataset ✅
- Decide what to do with missing values ✅

- Identify fields which are relevant to the analysis and eliminate any fields that will not be useful in the analyses. ✅
- Identify data fields which do not have proper data types and decide better data types for these columns. ✅
- Make sure to check the data before and after the data wrangling is applied to make sure any changes have been done. ✅

Good work in looking into all the above points!!

Some Helpful documentation and blogs:

- Pandas.isnull
- Pandas.dataframe.info
- Dropna function to drop any rows with missing values
- Fillna function to fill missing values
- pandas.DataFrame.drop to drop whole column
- Handling missing values in dataset

Exploration Phase

The project investigates the stated question(s) from multiple angles. At least three variables are investigated using both single-variable (1d) and multiple-variable (2d) explorations.

The questions were thoroughly investigated from various angles, and both 1d and 2d explorations were used for several variables investigated..!! Well done..!! 👍

COMMENTS: 💬

Exploratory Data Analysis (EDA) is an approach for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set. The graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

- 1) Plotting the raw data such as histograms, histograms, probability plots, lag plots, block plots, scatter plots.
- 2) Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.

Remember: 💡

- What is very important when you analyze data is to stay focused on your questions. Build plots or

statistical summaries which answer your questions, and not just because they are nice.

Below are the key differences between univariate and bivariate analysis:

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?

Below are some useful links for your refrence..

Link1

Link2

The project's visualizations are varied and show multiple comparisons and trends. Relevant statistics are computed throughout the analysis when an inference is made about the data.

At least two kinds of plots should be created as part of the explorations.

GOOD JOB!!! 👏

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. The project rightly builds descriptive visualizations using multiple types of plots...!! 👍

COMMENTS

- Data visualization is the presentation of data in a pictorial or graphical format. It enables decision-makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns.
- Because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports. Data visualization is a quick, easy way to convey concepts in a universal manner – and you can experiment with different scenarios by making slight adjustments.

https://seaborn.pydata.org/tutorial/categorical.html

Conclusions Phase

The results of the analysis are presented such that any limitations are clear. The analysis does not state or imply that one change causes another based solely on a correlation.

Good work presenting the results of the analysis while showing its limitations clearly... !! 👍🏽

- I appreciate that you have taken care of the suggestion given by the previous reviewer

Learning Notes 📃

- A description of limitations typically identifies either a shortcoming of the dataset that has caused difficulty (e.g.missing data) or a shortcoming of the methods of analysis (e.g. a statistical approach which may not be ideal given the characteristics of the data set).

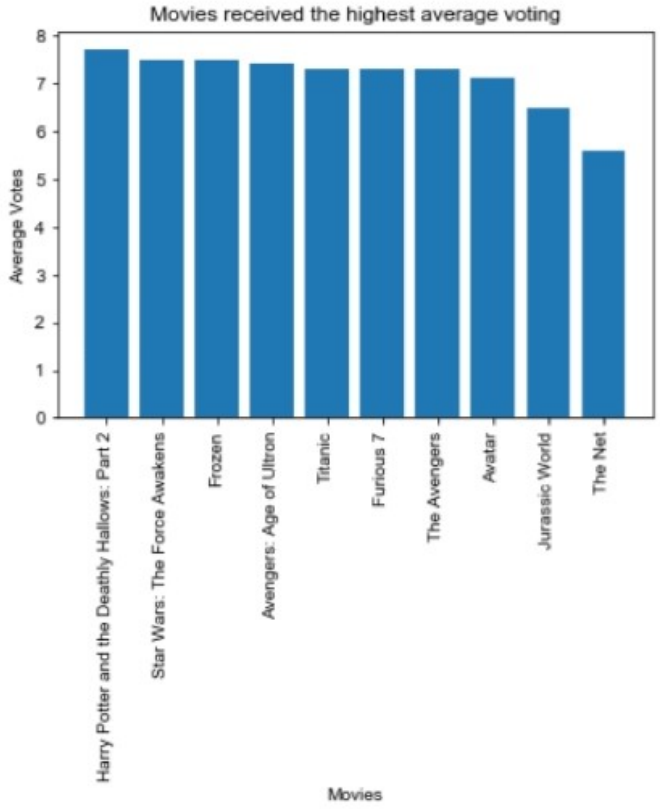
Communication

Reasoning is provided for each analysis decision, plot, and statistical summary.

Well done.. 😄👏🏽

You have done a great job describing every analysis decision, and plot stating the results obtained from that analysis...!!!

- Including the conclusions after the visualisations give the reader on the spot clearance of the findings.



Findings: Surprisingly enough, all of the top 10 movies that earned the most profit have an average vote of less than 8. The highest profit earned movie 'Avatar' has an average vote of 7.1. Moreover, the 10 movies that earned the highest voting are not in the list of top 10 movies that earned the highest profit.

Points to remember:

- While making documentation it is important to view your submission from the audience perspective so as to make it more and more indulging.
- One should never hesitate to mention the thought process of any finding or steps in analysis via markdown cells where appropriate so as to establish a connection with the audience.

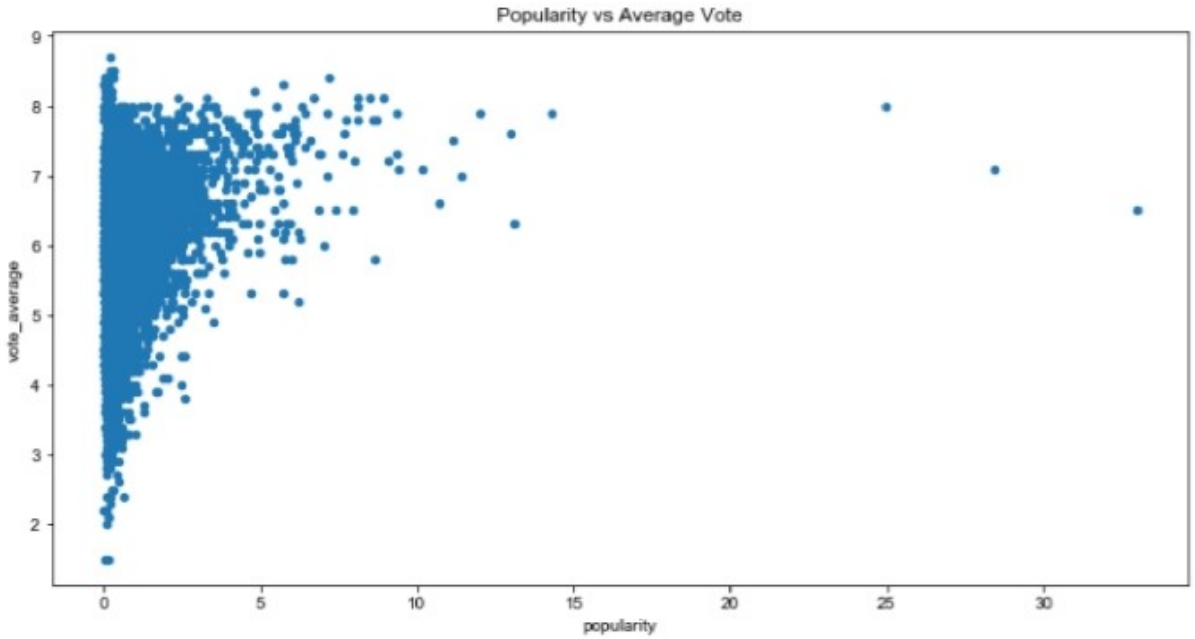
Visualizations made in the project depict the data in an appropriate manner that allows plots to be readily interpreted.

The analysis and visualizations throughout the report are well drafted.

- The chart contains a clearly represented title that explains the details of the presented graphs ✅

- The chart contains a clearly represented title that explains the details of the presented graphs. ✅
- Both axes have suitable titles with good naming conventions. ✅

This attention to detail really goes a long way to help communicate your results to an audience. Good work..!!👍👍



DOWNLOAD PROJECT

RETURN TO PATH

Rate this review