UDACITY

‹ Return to "Data Analyst Nanodegree" in the classroom

DISCUSS ON STUDENT HUB

Investigate a Dataset

REVIEW

HISTORY

Meets Specifications

Congratulations Student....!!! 🎉

This was a great implementation and I congratulate you for passing all rubric items with this submission.
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! 👏🏽

I am always interested in improving my reviews so please can you give comments on the review and rate for the same? Thanks in advance!

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.

Awesome Work..!!! 👍👍

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

- 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` 😃

SUGGESTION: 📰

- `It is usually a good practice to rerun complete jupyter code file before converting it to .html file since it becomes easy to view .`
  It appears that the HTML file was converted without completely running certain blocks of code in .ipynb. 😄

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.

Excellent work 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
- Group-by

It is really interesting to see that the submission includes above-mentioned practices...Good Job!! 👍

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

Great Work in avoiding repetitive code by using a pre-defined `drop` function .....!!

- 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, docstrings 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...
  Well Done..!!👍

SUGGESTION 📃

- For a below attached block of code, a user-defined function would be more appreciated and shows the good practice of coding.

Looks good. Now, I'll repeat the process with the other columns with multiple values: cast and director. However, I will create copies of the original df and apply these separately so the processing power is not slowed too much.

In [29]:
```
df_split_cast = df.copy()
split_cast = df_split_cast['cast'].str.split('|').apply(pd.Series, 1).stack().reset_index(level=1, drop=True)
split_cast.name = 'cast_split'
df_split_cast = df_split_cast.drop(['cast'], axis=1).join(split_cast)
df_split_director = df.copy()
split_director = df_split_director['director'].str.split('|').apply(pd.Series, 1).stack().reset_index(level=1, drop=True)
split_director.name = 'director_split'
df_split_director = df_split_director.drop(['director'], axis=1).join(split_director)
```

In [30]:
```
df_split_genre.head()
```

```
def split_funct(df,col_name, nm):
 split=df[col_name].str.split('|').apply(pd.Series, 1).stack().reset_index(level=1, drop=True)
 split.name=nm
 df=df.drop([col_name], axis=1).join(split)
 return df

df_split_cast = split_funct(df,'cast','cast_split')
```

Similarly this function cann be called for other columns

Quality of Analysis

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

Excellent Work in stating and addressing multiple insightful questions in your analysis.. Good Job!!!👍🎉

All the questions are relevant and have been addressed to in the analysis and relevant visualisations have been framed.... 😃

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 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..!! 👍

Learning notes 📰

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?

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.

Data visualization can also:

• Identify areas that need attention or improvement.
• Clarify which factors influence customer behaviour.
• Help you understand which products to place where.
• Predict sales volumes.

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... !! 👍

Learning Notes 📃

Limitations are the challenges that you personally face during this project or could have faced in the long run. To give you an idea to address the limitations or challenges you personally faced while doing this project, here are some factors to consider and can be addressed:

1) Was the data sufficient to prove your findings?
2) Was there any hindrance such as missing values, missing data?

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 and the limitations of that analysis...!!

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. 👏🏽🎉

Awesome! The plots are well labelled with appropriate comments and easy to interpret. 👍🏽

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review