UDACITY

< Return to Classroom

Investigate a Dataset

REVIEW

HISTORY

Requires Changes

4 specifications require changes

Very impressive work ! your project reflects your hard work and I have to congratulate you for that 😄 Your code is very solid as well, you only need some modifications order to continue. Good luck in your next submission !

Don't hesitate to reach your help in Student Hub, we are here to help you succeed 🏆

Code Functionality

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

- 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.

Good work
Suggestion

Here are a few Pandas buiDeep Learning Specializationlt-in methods that are very useful for exploring variables in this project:

• Boolean-Indexing
• Group-by
• Value-Counts

• Series.map
• Working-with-text-data

- **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.**

Excellent job! solid code and well documented 😃

Quality of Analysis

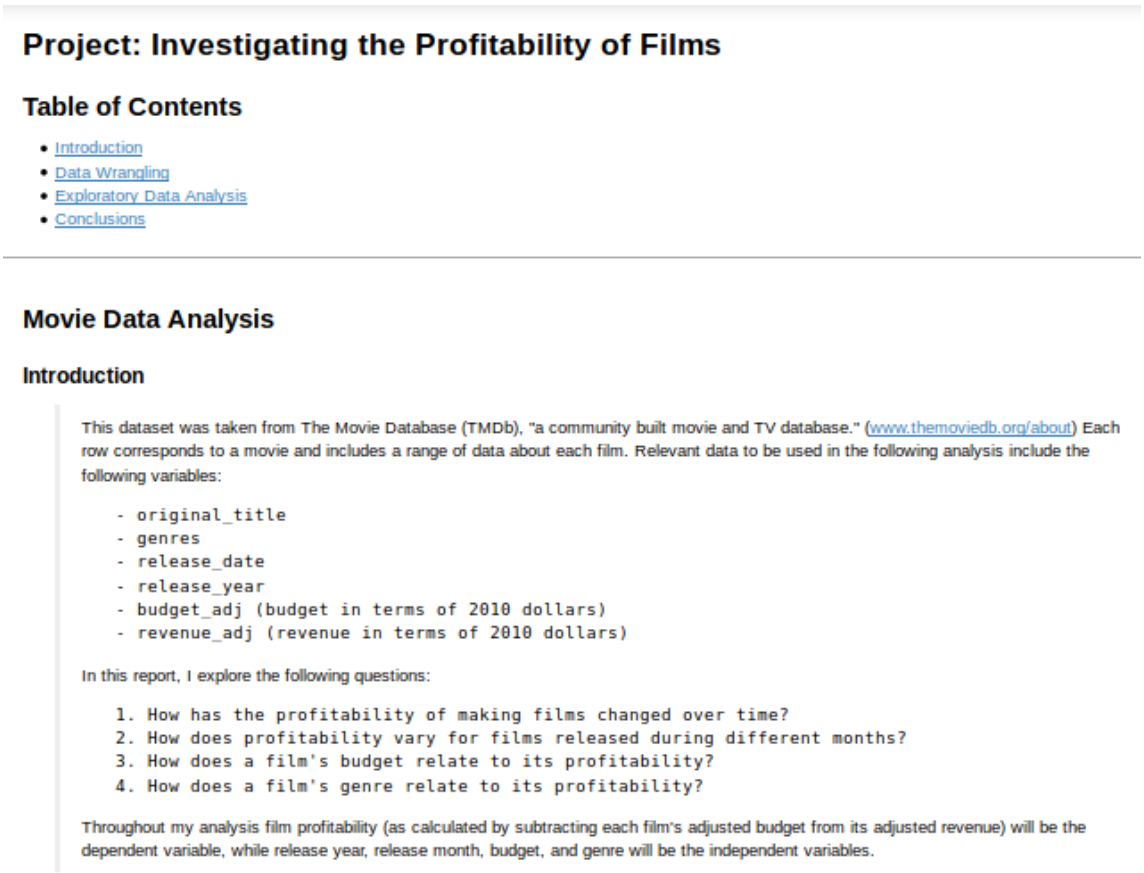
- **The project clearly states one or more questions, then addresses those questions in the rest of the analysis.**

Very deep and important questions 🏆 to meet the expectations you have to make an introduction about the dataset at the beginning of your report, describe the event and the variables it contains. Also enumerate the questions so that the reader can easily get an idea of what your project will be. Do not forget to use a markdown cell for this: smile:

As always we want you to go above and beyond, here are some suggestions of interesting questions for this dataset:

1. How is popularity trending over time?
2. How are revenues trending over time?
3. How is runtime trending over time?
4. Do top ratings movies always generate big revenue?
5. Do higher budget movies always generate big revenue?
6. Is there any impact of vote count on revenue?
7. Can we provide a list of the most popular directors based on ratings?
8. Can we provide a list of directors that generates big revenue?
9. What are typical runtimes for directors? Is there a duration preferred by directors?

10. Is there a relation between popularity and revenue for directors? etc.



Project: Investigating the Profitability of Films

Table of Contents

Movie Data Analysis

Introduction

This dataset was taken from The Movie Database (TMDb), "a community built movie and TV database." (www.themoviedb.org/about) Each row corresponds to a movie and includes a range of data about each film. Relevant data to be used in the following analysis include the following variables:

```
- original_title
- genres
- release_date
- release_year
- budget_adj (budget in terms of 2010 dollars)
- revenue_adj (revenue in terms of 2010 dollars)
```

In this report, I explore the following questions:

```
1. How has the profitability of making films changed over time?
2. How does profitability vary for films released during different months?
3. How does a film's budget relate to its profitability?
4. How does a film's genre relate to its profitability?
```

Throughout my analysis film profitability (as calculated by subtracting each film's adjusted budget from its adjusted revenue) will be the dependent variable, while release year, release month, budget, and genre will be the independent variables.

Data Wrangling Phase

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

Good work in implementing a Data Wrangling Phase
Suggestion

The most important aspect of Data Wrangling is to clean or transform the data preparing it for analysis.

One main issue is having missing data while conducting analysis, which can provide skew/bias results. Luckily there are a few methods that Pandas provide to deal with these issues:

- The first thing to do is to always Identify the missing values within the dataset. The few steps after this explain how to deal with the missing data
- If there are columns with a few rows of missing data the Dropna method could be used to drop the missing rows.
- If there are rows with missing data the Fillna-method can be used instead of dropping them

completely (This method can vary with the data and the project)
- The final option is if there are way too many missing values within a column it is best to drop the column completely using the Drop-column-method

Data Wrangling does not only involve Identifying and dealing with missing values but also involves in transforming the data to a more effective state to target the analysis. Here are other wrangling methods:

- Binning or Cutting Groups continuous or numerical values into smaller groups or 'bins'
- Pandas-Dummies Transforms categorical data into dummy/indicator variables

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.

Please include 1-D explorarions, here are some tips to improve in this item:

This link summarises the difference between bivariate and univariate data.

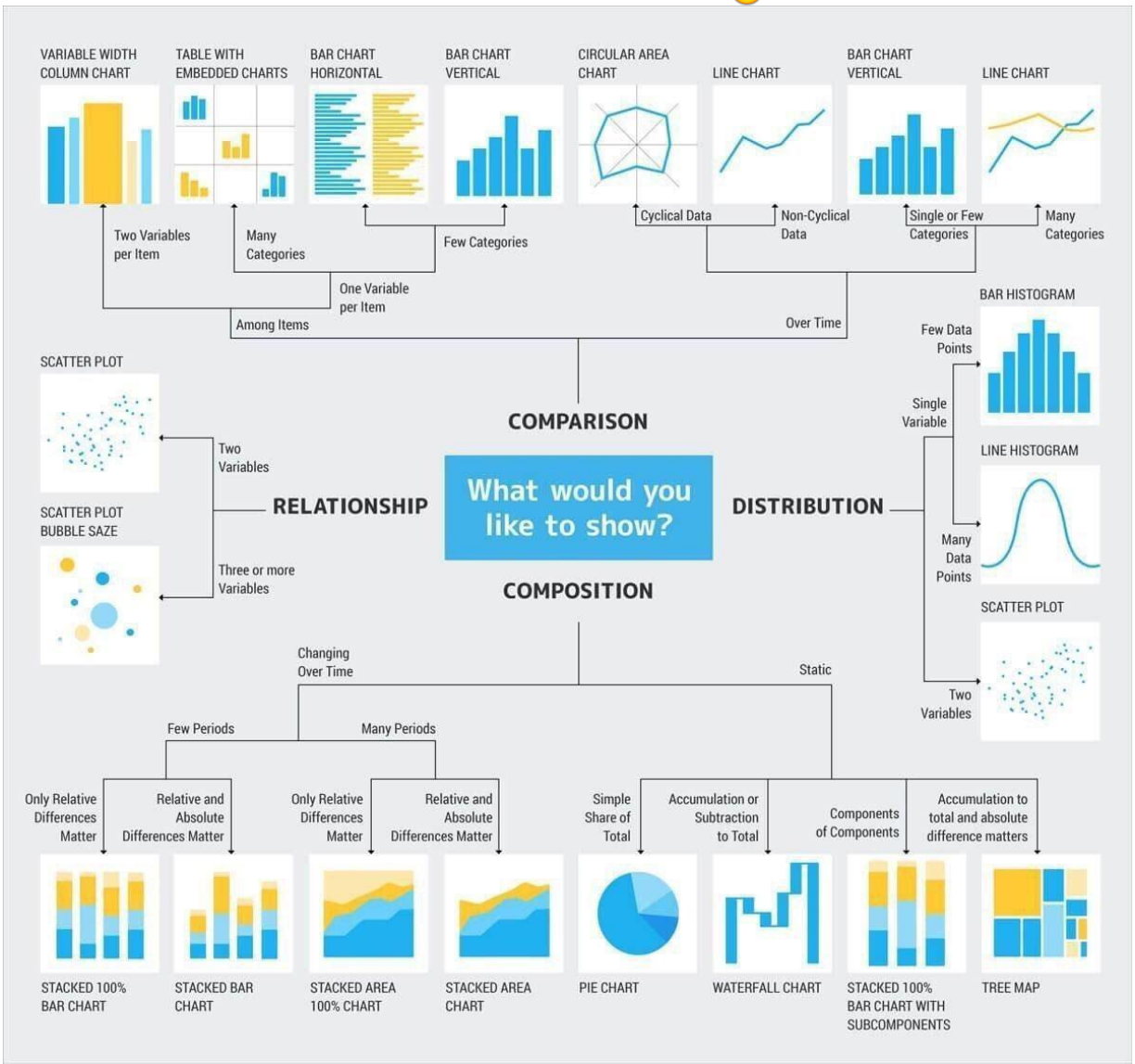
Univariate Data	Bivariate Data
• involving a **single variable**	• involving **two variables**
• does not deal with causes or relationships	• deals with causes or relationships
• the major purpose of univariate analysis is to describe	• the major purpose of bivariate analysis is to explain
• central 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?

And here's some more inspirations for you: https://seaborn.pydata.org/tutorial/categorical.html

- 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.

This is something you can improve, your report will improve as soon as you add more variable graphics, let me suggest the following tool to decide which one to use in each case 😉



please take a look at the following link and add different types of graphics, your report would look more professional with that.

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.

correlation.

There should be a separate subsection inside the conclusion section called 'Limitations' where you would have to discuss the limitations of this dataset which might have adversely affected your analysis. Examples

would be null or missing values, whether these samples are an effective representation of the population or not or maybe that you could dive deeper into your analysis with additional specific information.

The conclusions and limitations have the following structure:

Conclusions

In the first section I examined the popularity of Western movies over the decades. I made my analyzation based on the values of 'released_year' and 'popularity'. I could not find any correlations between the numbers and the assumptions but I found it by taking into account the numbers of released movies.

After that I analyzed the ratings of the most and least expensive movies and I found out that the more expensive movies got higher votes than the cheaper ones.

Limitations

In the first section - although the literature details the phenomenon - I could not find any correlation between 'popularty' and 'release year'. It would be good to know more about what is behind the value 'popularity' and what popularity means here. Just to name a few... How was it calculated? Which criterias and values were measured exactly to get these numbers? It could be caculated based on ticket sales? Or based on audience appraisal? However, I found correlation between my assumptions and the number of released western movies but I would not name it causation without a much more detailed further analysis.

In the second section, I made my calculations based on the values of budget adjustment to take the fluctuations into account, I found this really useful. But there were more missing values in the 'budget_adj' column. During the cleaning process I replaced the missing values with the average, but it still can distort the result (for instance, there would be other movies among the most expensive 200 movies).

Suggested limitations:

1. We have used TMBD Movies dataset for our analysis and worked with popularity, revenue and runtime. Our analysis is limited to only the provided dataset. For example, the dataset does not confirm that every release of every director is listed.
2. There is no normalization or exchange rate or currency conversion is considered during this analysis and our analysis is limited to the numerical values of revenue.
3. Dropping missing or Null values from variables of our interest might skew our analysis and could show unintentional bias towards the relationship being analyzed. etc.

Communication

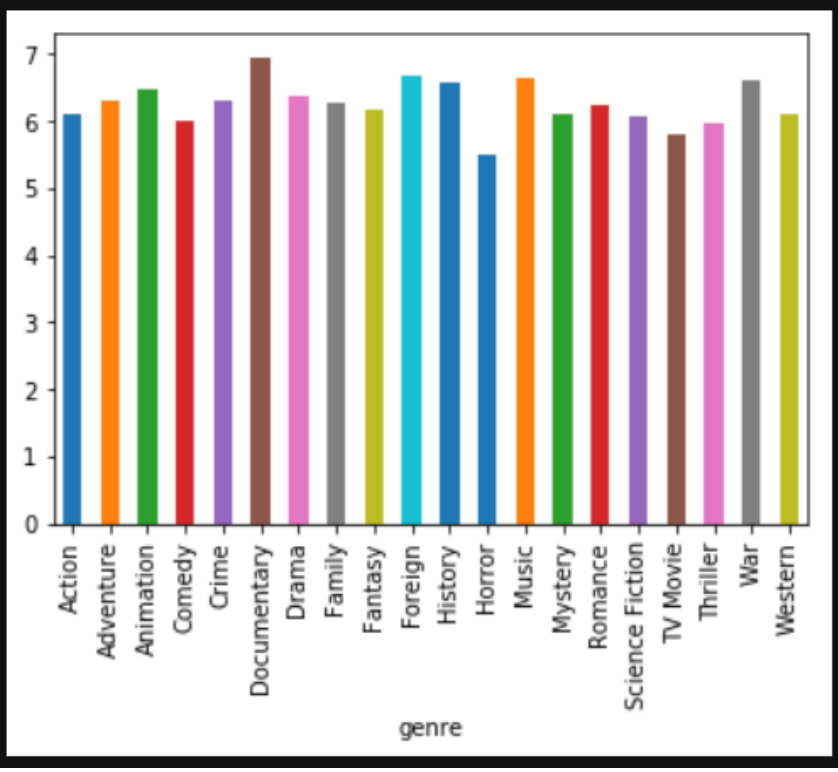
- The code should have ideally the following sections: Introduction; Questions; Data Wrangling; Exploratory Data Analysis; Conclusions, Limitation.
- Reasoning is provided for each analysis decision, plot, and statistical summary.
- Interpretation of plots and application of statistical tests should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

Fantastic 👍

- 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.

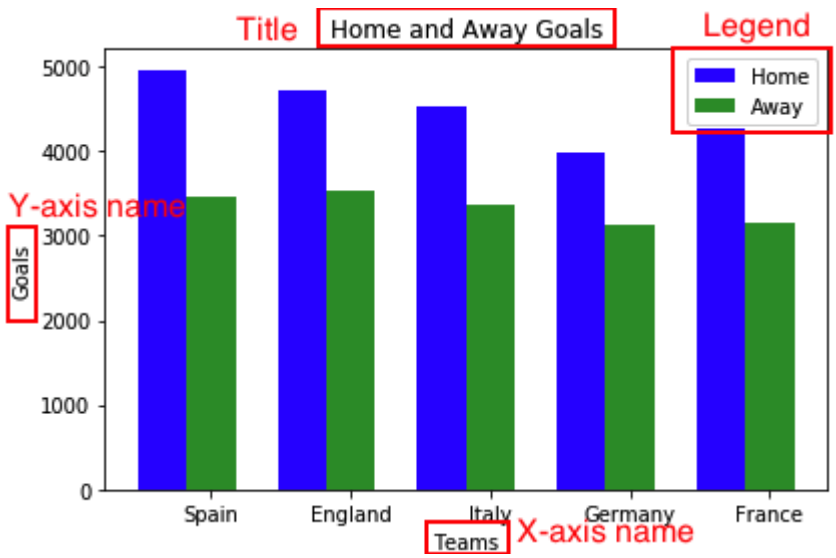
```
[433]: df_vote_average.plot.bar()

[433]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff7a7696898>
```



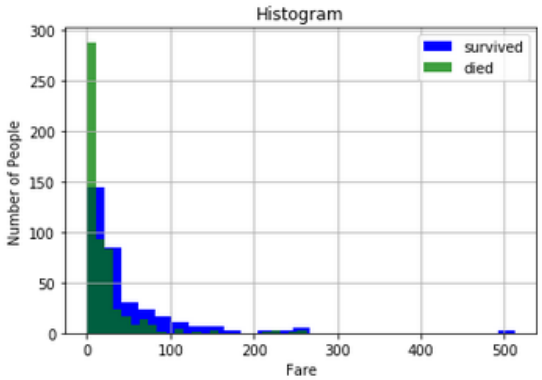
Please make sure that each graph has the following three characteristics:

1. A title
2. Names of the axes (in the X and Y axis)
3. Labels



Here are a Few samples of how to do it:

In [19]:
```
plt.hist(df.Fare[df.Survived == True], 25, facecolor='b', alpha=1, label='survived');
plt.hist(df.Fare[df.Survived == False], 25, facecolor='g', alpha=0.75, label='died');
plt.legend()
plt.xlabel('Fare')
plt.ylabel('Number of People')
plt.title('Histogram')
plt.grid(True)
```



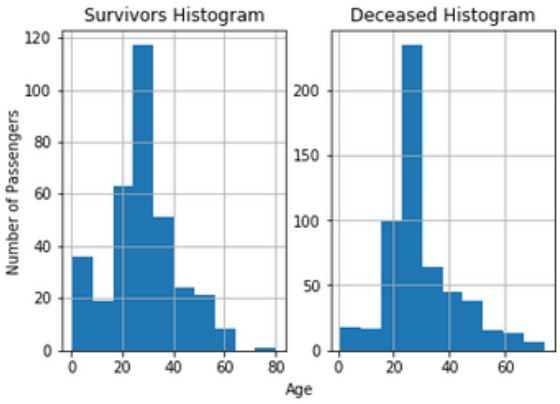
Is passenger age associated with survival?

In [15]:
```
fig, axes = plt.subplots(1, 2)

df.Age[df.Survived == True].hist(label='survived', ax=axes[0])
df.Age[df.Survived == False].hist(label='survived', ax=axes[1])

axes[0].set_title('Survivors Histogram')
axes[1].set_title('Deceased Histogram');

fig.text(0.5, 0.02, 'Age', ha='center');
fig.text(0.04, 0.5, 'Number of Passengers', va='center', rotation='vertical');
```



RESUBMIT

DOWNLOAD PROJECT

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

RETURN TO PATH

Rate this project

START