

# Investigate\_a\_Dataset

June 14, 2020

**Tip:** Welcome to the Investigate a Dataset project! You will find tips in quoted sections like this to help organize your approach to your investigation. Before submitting your project, it will be a good idea to go back through your report and remove these sections to make the presentation of your work as tidy as possible. First things first, you might want to double-click this Markdown cell and change the title so that it reflects your dataset and investigation.

## 1 Project: Investigate a Dataset (Replace this with something more specific!)

### 1.1 Table of Contents

Introduction

Data Wrangling

Exploratory Data Analysis

Conclusions

## 2 what is the relation between release year and vote average?

## 3 what is the relation between the release year and the budget of the movie?

### ## Introduction

**Tip:** In this section of the report, provide a brief introduction to the dataset you've selected for analysis. At the end of this section, describe the questions that you plan on exploring over the course of the report. Try to build your report around the analysis of at least one dependent variable and three independent variables. If you're not sure what questions to ask, then make sure you familiarize yourself with the dataset, its variables and the dataset context for ideas of what to explore.

If you haven't yet selected and downloaded your data, make sure you do that first before coming back here. In order to work with the data in this workspace, you also need to upload it to the workspace. To do so, click on the jupyter icon in the upper left to be taken back to the workspace directory. There should be an 'Upload' button in the upper right that will let you add your data file(s) to the workspace. You can then click on the .ipynb file name to come back here.

```
In [1]: # Use this cell to set up import statements for all of the packages that you
#       plan to use.

# Remember to include a 'magic word' so that your visualizations are plotted
# inline with the notebook. See this page for more:
# http://ipython.readthedocs.io/en/stable/interactive/magics.html
import pandas as pd
```

## ## Data Wrangling

**Tip:** In this section of the report, you will load in the data, check for cleanliness, and then trim and clean your dataset for analysis. Make sure that you document your steps carefully and justify your cleaning decisions.

### 3.0.1 General Properties

```
In [2]: # Load your data and print out a few lines. Perform operations to inspect data
#       types and look for instances of missing or possibly errant data.
df = pd.read_csv('tmdb-movies.csv')
df.head()
```

```
Out[2]:
```

|   | id     | imdb_id   | popularity | budget    | revenue    | \ |
|---|--------|-----------|------------|-----------|------------|---|
| 0 | 135397 | tt0369610 | 32.985763  | 150000000 | 1513528810 |   |
| 1 | 76341  | tt1392190 | 28.419936  | 150000000 | 378436354  |   |
| 2 | 262500 | tt2908446 | 13.112507  | 110000000 | 295238201  |   |
| 3 | 140607 | tt2488496 | 11.173104  | 200000000 | 2068178225 |   |
| 4 | 168259 | tt2820852 | 9.335014   | 190000000 | 1506249360 |   |

|   | original_title               | \ |
|---|------------------------------|---|
| 0 | Jurassic World               |   |
| 1 | Mad Max: Fury Road           |   |
| 2 | Insurgent                    |   |
| 3 | Star Wars: The Force Awakens |   |
| 4 | Furious 7                    |   |

|   | cast  | \ |
|---|---|---|
| 0 | Chris Pratt Bryce Dallas Howard Irrfan Khan Vi... |   |
| 1 | Tom Hardy Charlize Theron Hugh Keays-Byrne Nic... |   |
| 2 | Shailene Woodley Theo James Kate Winslet Ansel... |   |
| 3 | Harrison Ford Mark Hamill Carrie Fisher Adam D... |   |
| 4 | Vin Diesel Paul Walker Jason Statham Michelle ... |   |

|   | homepage  | director         | \ |
|---|---|------------------|---|
| 0 | http://www.jurassicworld.com/                     | Colin Trevorrow  |   |
| 1 | http://www.madmaxmovie.com/                       | George Miller    |   |
| 2 | http://www.thedivergentseries.movie/#insurgent    | Robert Schwentke |   |
| 3 | http://www.starwars.com/films/star-wars-episod... | J.J. Abrams      |   |
| 4 | http://www.furious7.com/                          | James Wan        |   |

```

tagline      ... \
0      The park is open.      ...
1      What a Lovely Day.      ...
2      One Choice Can Destroy You      ...
3      Every generation has a story.      ...
4      Vengeance Hits Home      ...

overview runtime \
0      Twenty-two years after the events of Jurassic ...      124
1      An apocalyptic story set in the furthest reach...      120
2      Beatrice Prior must confront her inner demons ...      119
3      Thirty years after defeating the Galactic Empi...      136
4      Deckard Shaw seeks revenge against Dominic Tor...      137

genres \
0      Action|Adventure|Science Fiction|Thriller
1      Action|Adventure|Science Fiction|Thriller
2      Adventure|Science Fiction|Thriller
3      Action|Adventure|Science Fiction|Fantasy
4      Action|Crime|Thriller

production_companies release_date vote_count \
0      Universal Studios|Amblin Entertainment|Legenda...      6/9/15      5562
1      Village Roadshow Pictures|Kennedy Miller Produ...      5/13/15      6185
2      Summit Entertainment|Mandeville Films|Red Wago...      3/18/15      2480
3      Lucasfilm|Truenorth Productions|Bad Robot      12/15/15      5292
4      Universal Pictures|Original Film|Media Rights ...      4/1/15      2947

vote_average  release_year  budget_adj  revenue_adj
0      6.5      2015      1.379999e+08      1.392446e+09
1      7.1      2015      1.379999e+08      3.481613e+08
2      6.3      2015      1.012000e+08      2.716190e+08
3      7.5      2015      1.839999e+08      1.902723e+09
4      7.3      2015      1.747999e+08      1.385749e+09

```

[5 rows x 21 columns]

**Tip:** You should *not* perform too many operations in each cell. Create cells freely to explore your data. One option that you can take with this project is to do a lot of explorations in an initial notebook. These don't have to be organized, but make sure you use enough comments to understand the purpose of each code cell. Then, after you're done with your analysis, create a duplicate notebook where you will trim the excess and organize your steps so that you have a flowing, cohesive report.

**Tip:** Make sure that you keep your reader informed on the steps that you are taking in your investigation. Follow every code cell, or every set of related code cells, with a markdown cell to describe to the reader what was found in the preceding cell(s). Try to make it so that the reader can then understand what they will be seeing in the following cell(s).

### 3.0.2 Data Cleaning (Replace this with more specific notes!)

```
In [3]: # After discussing the structure of the data and any problems that need to be
#       cleaned, perform those cleaning steps in the second part of this section.

##dropped columns
df.drop(['imdb_id', 'revenue', 'director', 'tagline', 'homepage', 'keywords', 'overview'],
df.head(1)

##null rows
df.isnull().sum()
# drop rows with any null values
df.dropna(inplace=True)

df.isnull().sum().any()
```

Out[3]: False

```
In [4]: print(sum(df.duplicated()))
# drop duplicated data
df.drop_duplicates(inplace=True)
```

1

#### ## Exploratory Data Analysis

**Tip:** Now that you've trimmed and cleaned your data, you're ready to move on to exploration. Compute statistics and create visualizations with the goal of addressing the research questions that you posed in the Introduction section. It is recommended that you be systematic with your approach. Look at one variable at a time, and then follow it up by looking at relationships between variables.

### 3.0.3 Research Question 1 # What is the relation between release date and vote average?

```
In [5]: # Use this, and more code cells, to explore your data. Don't forget to add
#       Markdown cells to document your observations and findings.
# low = min(df_01.vote_average)
# high = max(df_01.vote_average)
# plt.ylim([math.ceil(low*(high-low)), math.ceil(low+0.5*(high-low))])

import matplotlib.pyplot as plt
import numpy as np
import math

# First I got the vote average of each year.
df_01 = df.groupby('release_year').vote_average.mean()
# Then We take the sample of df_01 group by using DataFrame 7 rows with columns release_
df_01 = pd.DataFrame({'release_year': ["1960", "1970", "1980", "1990", "2000", "2010", "2015"],
```

```

                                'vote_average':[6.325000,6.417073,6.160256,5.993939,5.876000,5.979
low = min(df_01.vote_average)
high = max(df_01.vote_average)
plt.ylim([math.ceil(low*(high-low)), math.ceil(low+0.5*(high-low))])
plt.bar(df_01.release_year, df_01.vote_average)
plt.title("Movies Evaluation From 1960 to 2015")
plt.xlabel("Year")
plt.ylabel("Vote Average of Movies")
plt.show()

<matplotlib.figure.Figure at 0x7fe01ee2f6d8>

```

3.0.4 The voting average for movies decreases as years increases..

3.0.5 Research Question 2 (Replace this header name!)

## 4 what is the relation between the release year and the budget of the movie?

## Conclusions

**Tip:** Finally, summarize your findings and the results that have been performed. Make sure that you are clear with regards to the limitations of your exploration. If you haven't done any statistical tests, do not imply any statistical conclusions. And make sure you avoid implying causation from correlation!

**Tip:** Once you are satisfied with your work here, check over your report to make sure that it satisfies all the areas of the rubric (found on the project submission page at the end of the lesson). You should also probably remove all of the "Tips" like this one so that the presentation is as polished as possible.

### 4.1 Submitting your Project

Before you submit your project, you need to create a .html or .pdf version of this notebook in the workspace here. To do that, run the code cell below. If it worked correctly, you should get a return code of 0, and you should see the generated .html file in the workspace directory (click on the orange Jupyter icon in the upper left).

Alternatively, you can download this report as .html via the **File > Download as** sub-menu, and then manually upload it into the workspace directory by clicking on the orange Jupyter icon in the upper left, then using the Upload button.

Once you've done this, you can submit your project by clicking on the "Submit Project" button in the lower right here. This will create and submit a zip file with this .ipynb doc and the .html or .pdf version you created. Congratulations!

```

In [24]: # Continue to explore the data to address your additional research
        # questions. Add more headers as needed if you have more questions to

```

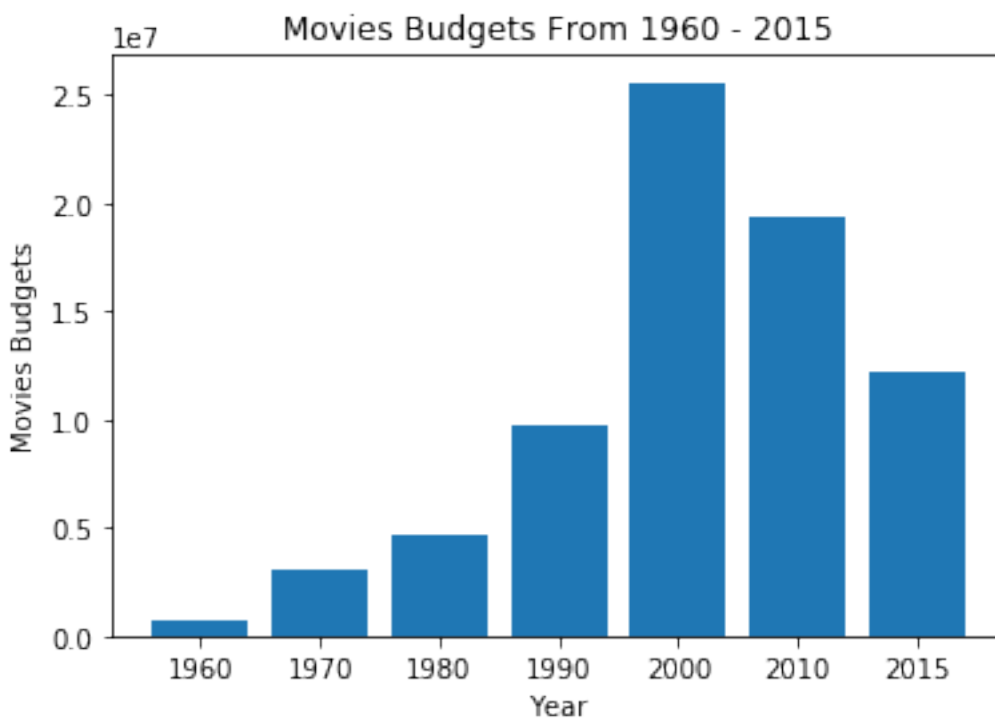
```

# investigate.
# First I got the vote average of each year.
df_02 = df.groupby('release_year').budget.mean()

# df.groupby(['release_year']).count()['id']

df_02 = pd.DataFrame({'release_year': ["1960", "1970", "1980", "1990", "2000", "2010", "2015"],
                      'budget': [6.892796e+05, 3.096755e+06, 4.647436e+06, 9.772137e+06, 2.5e+07, 1.9e+07, 1.2e+07]})
plt.bar(df_02.release_year, df_02.budget)
plt.title("Movies Budgets From 1960 - 2015")
plt.xlabel("Year")
plt.ylabel("Movies Budgets")
plt.show()

```



In [ ]: *## 2000 budgets average is the highest ..*

```

In [7]: from subprocess import call
        call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])

```

Out[7]: 0