UDACITY

‹ Return to Classroom

Investigate a Dataset

REVIEW

HISTORY

Meets Specifications

Dear Student,

This is some very good work done by you. You have

- created a functional and error free code
- used numpy and pandas wherever possible and put it to good use
- did good data cleaning and manipulation
- stated clear questions and did a good analysis of the same
- mentioned the summary of the dataset and the solution in general and also mentioned the limitations

Some key aspects that you can ponder over is

- Try to include probabilistic and quantitative aspects while making the analysis. Visualisation is a good tool but it does not depict a clear picture in terms of what are the exact numbers or probability of some event happening. Check out this link for better understanding https://www.analyticsvidhya.com/blog/2017/02/basic-probability-data-science-with-examples/

This is a very good project

I wish you all the very best

Code Functionality

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

Great work here. Very well done. All the code is functional and working with no errors. Also, check out this link to level yourself up with using Jupiter notebooks https://www.dataquest.io/blog/advanced-jupyter-notebooks-tutorial/

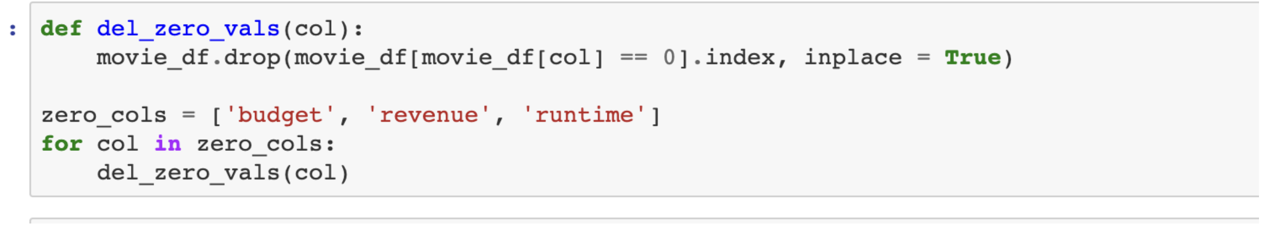
- The project uses NumPy arrays and Pandas Series and DataFrames where appropriate rather than Python lists and dictionaries.
- Where possible, vectorized operations and built-in functions are used instead of loops.

This is some very good work done by you. You have appropriately made use of Numpy and Pandas functions wherever possible. Some links to learn some advanced techniques for the libraries

- Advanced pandas https://towardsdatascience.com/30-examples-to-get-you-from-a-novice-to-an-advanced-pandas-user-e6eb4e8750b7
- Advanced Numpy https://medium.com/analytics-vidhya/advanced-numpy-218584c60c63

- 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.

Good work. However, remember that any repetitive code should be made modular to have an efficient code. Although, I believe you can still make use of loops to iterate over the columns in table used for any exploration or cleaning.



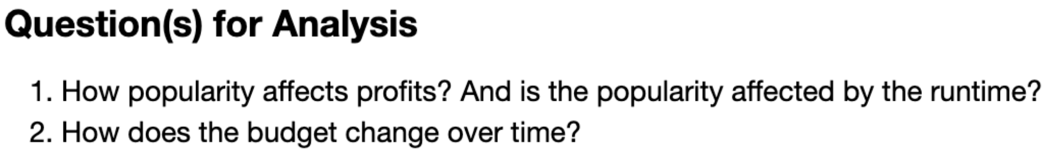
```
:  def del_zero_vals(col):
       movie_df.drop(movie_df[movie_df[col] == 0].index, inplace = True)

   zero_cols = ['budget', 'revenue', 'runtime']
   for col in zero_cols:
       del_zero_vals(col)
```

Quality of Analysis

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

This has been very well thought out where you try to analyse and compare some features with respect to No Show feature. You can also try a combination of features to do a multivariate analysis where you can combine neighbourhood and age and try to do some feature engineering and check analyse your results. You can check what feature engineering is - https://www.kaggle.com/learn/feature-engineering



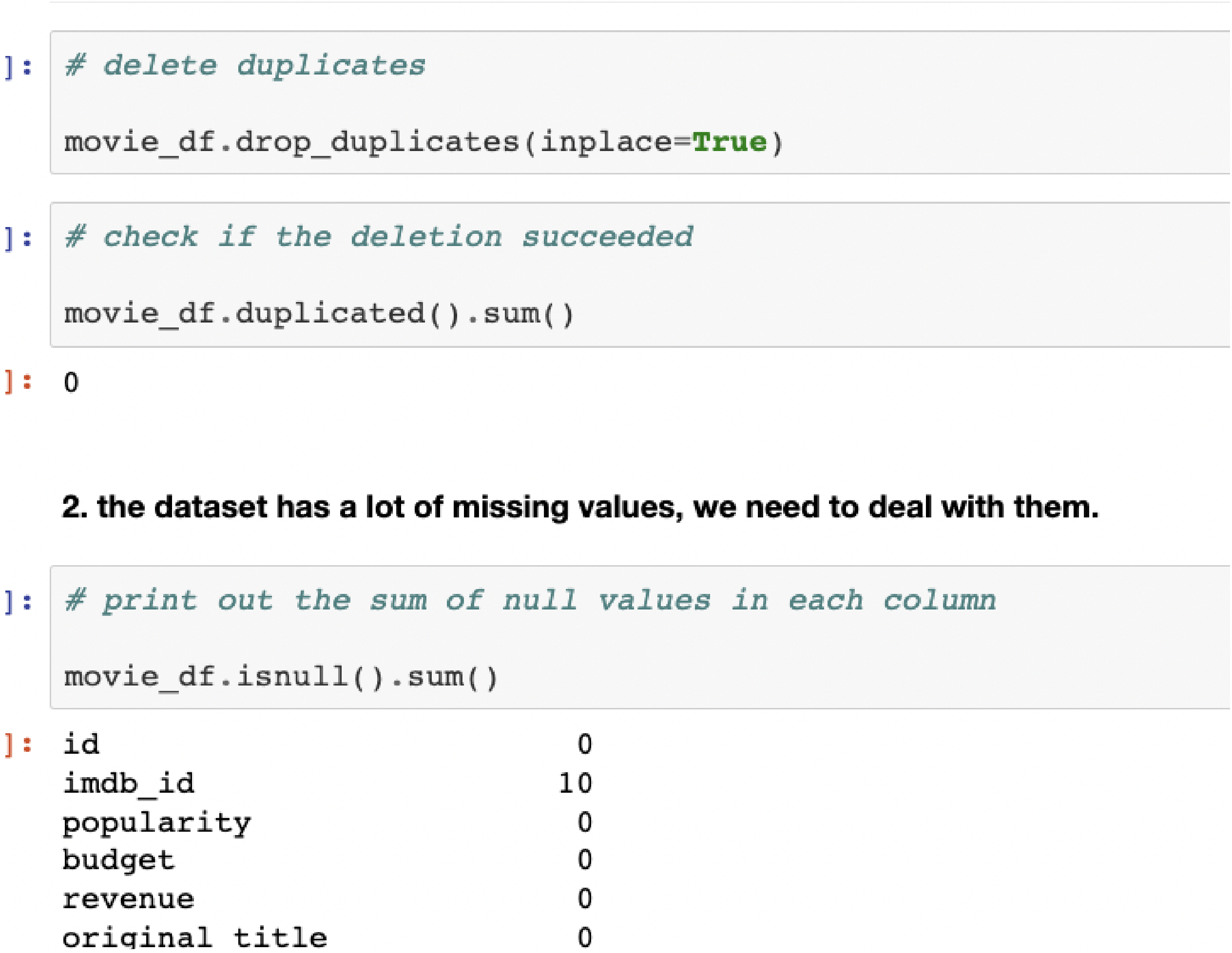
Question(s) for Analysis

1. How popularity affects profits? And is the popularity affected by the runtime?
2. How does the budget change over time?

Data Wrangling Phase

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

Great work here.



```
]: # delete duplicates

   movie_df.drop_duplicates(inplace=True)

]: # check if the deletion succeeded

   movie_df.duplicated().sum()

]: 0
```

2. the dataset has a lot of missing values, we need to deal with them.

```
]: # print out the sum of null values in each column

   movie_df.isnull().sum()

]: id                        0
   imdb_id                  10
   popularity                0
   budget                    0
   revenue                   0
   original_title            0
```

```
cast                      76
homepage                7929
director                  44
tagline                 2824
keywords                1493
overview                   4
runtime                    0
genres                    23
production_companies    1030
release_date               0
vote_count                 0
vote_average               0
release_year               0
budget_adj                 0
revenue_adj                0
dtype: int64
```

director, keywords, genres, and production_companies columns have null values!

I think it is better to drop the null values.

```
]: movie_df.dropna(how = 'any',inplace = True)
```

```
]: # check if all null values are deleted

   movie_df.isnull().sum()
```

```
]: id                      0
   imdb_id                 0
   popularity              0
   budget                  0
   revenue                 0
   original_title          0
   cast                    0
   homepage                0
   director                0
   tagline                 0
   keywords                0
   overview                0
   runtime                 0
   genres                  0
   production_companies    0
   release_date            0
   vote_count              0
   vote_average            0
   release_year            0
   budget_adj              0
   revenue_adj             0
```

```
dtype: int64
```

Exploration Phase

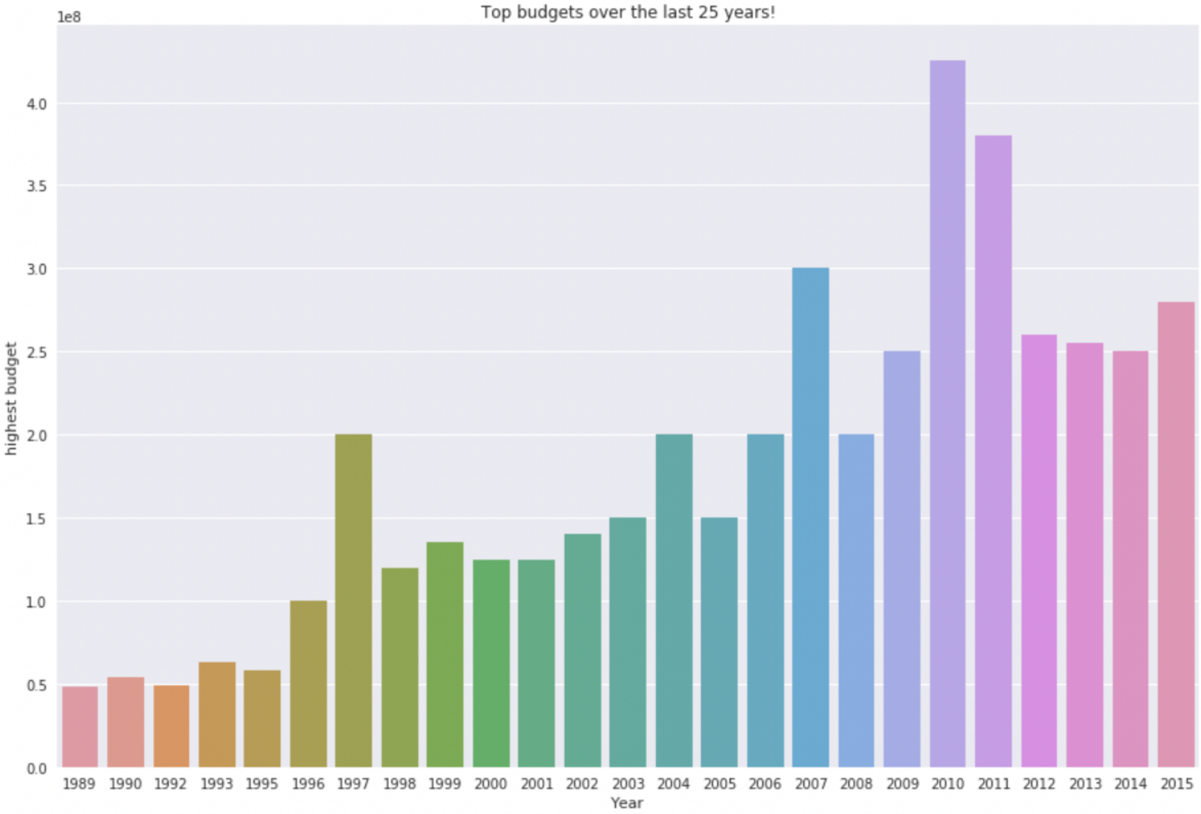
- The project investigates the stated question(s) from multiple angles.
- The project explores at least three variables in relation to the primary question. This can be an exploratory relationship between three variables of interest, or looking at how two independent variables relate to a single dependent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

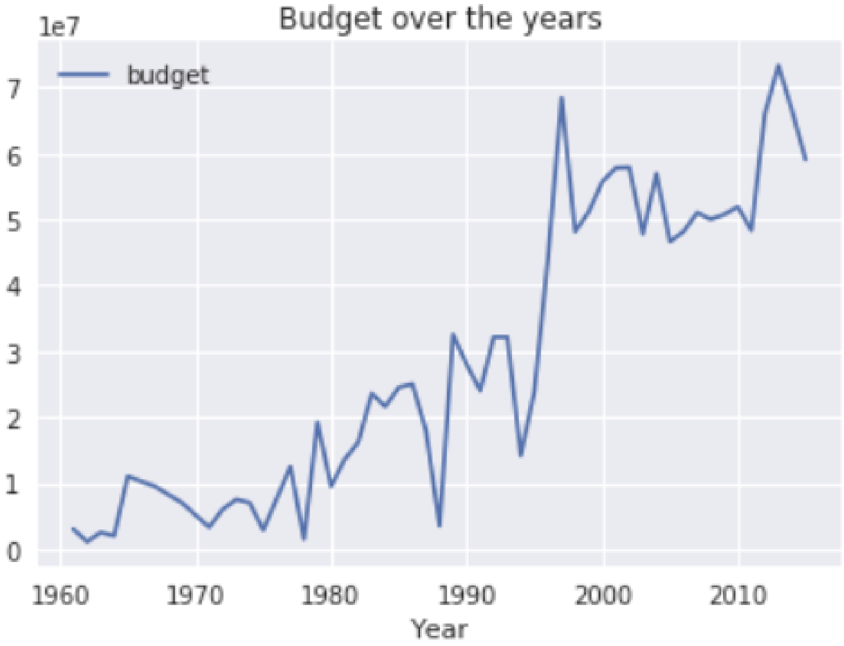
Very well done. Feature engineering will help enormously to answer the questions in a better way

- 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.

You have created interesting plots and these plots give a good idea.

You can check out this link to understand what all graphs you can plot
https://www.analyticsvidhya.com/blog/2021/06/exploratory-data-analysis-using-data-visualization-techniques/
and have a vivid understanding of the analysis



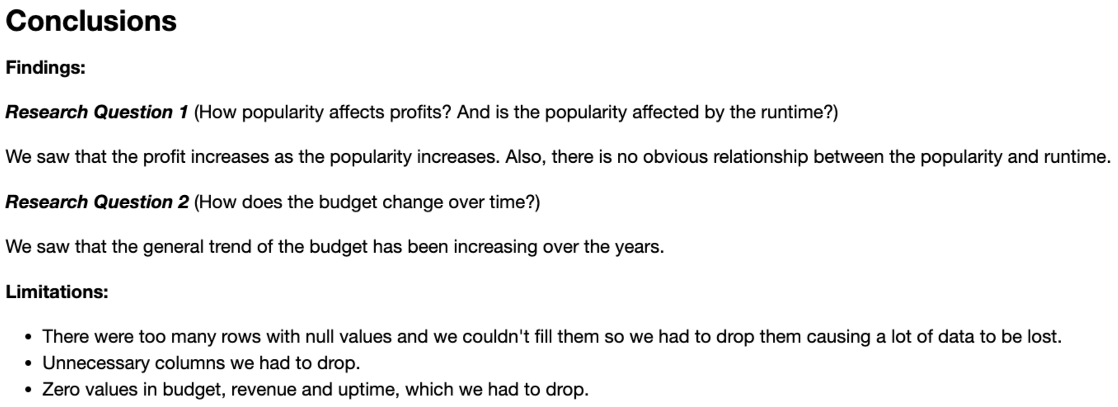


> We can see that the budget has been increasing over the years.

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.

Brilliant work here. The points you have mentioned marks your clear understanding of the data.



Conclusions

Findings:

Research Question 1 (How popularity affects profits? And is the popularity affected by the runtime?)

We saw that the profit increases as the popularity increases. Also, there is no obvious relationship between the popularity and runtime.

Research Question 2 (How does the budget change over time?)

We saw that the general trend of the budget has been increasing over the years.

Limitations:

- There were too many rows with null values and we couldn't fill them so we had to drop them causing a lot of data to be lost.
- Unnecessary columns we had to drop.
- Zero values in budget, revenue and uptime, which we had to drop.

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.

Very well done

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.

Visualisations are appropriate with proper descriptions and labels.

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

START