**Team members:**

**فادي مالك ميخائيل 20201470332**

**سامر عادل كامل 20201498752**

**عبدالرحمن صلاح إبراهيم لبيب 20201444840**

**كريم محمد حسن اسماعيل 20201474079**

**Data Description**

**What can we say about the success of a movie before it is released?** Are there certain companies that have found a consistent formula? Given that major films costing over $100 million to produce can still flop, this question is more important than ever to the industry. Film aficionados might have different interests. Can we predict which films will be highly rated, whether they are a commercial success?

***IMDb*: It is an online database that contains information about cinema and television films, film stars and serials of all countries and periods on earth**

We choose the TMDb movie data set for data analysis. This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue. we would like to find other intresting patterns in the dataset.

**Data cleaning**

1. Graphical user interface, text, application

   Description automatically generatedwe have to remove irrelevant attributes.
2. We need to drop the duplicates data.A picture containing application

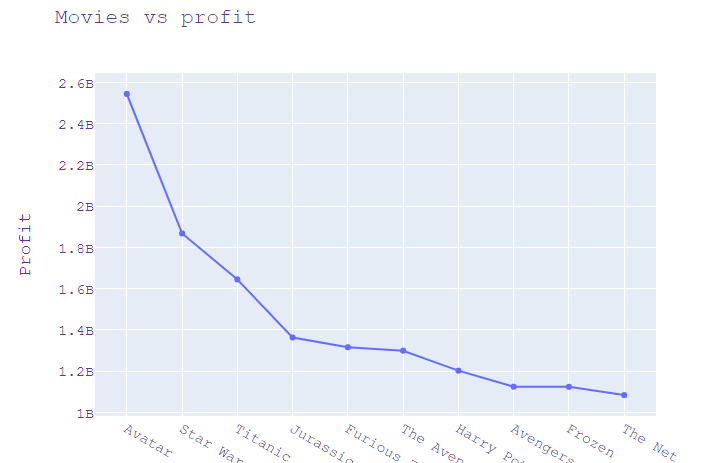
   Description automatically generated
3. Graphical user interface, text

   Description automatically generatedwe will create a method to extract the data from multi valued cells

**Questions:**

Which Movie Has the Highest Profit?Graphical user interface, text

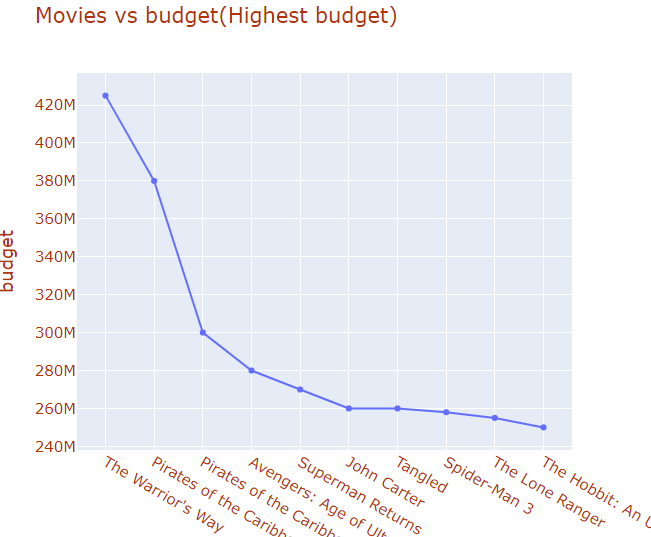
Description automatically generated



* Idiom: Scatter plot
* Marks: points and lines connects between them
* Channels: Horizontal and vertical position
* The highest profit is Avatar then Star Wars: The Force Awakens

### we used scatterplot with lines between them to see the increasing and decreasing in values

# Movie with Highest and Lowest Budget?

Text

Description automatically generated

Chart, line chart

Description automatically generated

* Idiom: Scatter plot
* Marks: points and lines connects between them
* Channels: Horizontal and vertical position
* The highest budget is The Warriors way and the lowest is Love,Wedding,Marriage

# Which movie made the highest revenue and lowest as well?

Graphical user interface, text, application

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Avatar is the highest profit and mallrats with Shattered Glass is the lowest profit

Graphical user interface, text, application

Description automatically generatedWhich Movie has shortest and longest runtime?

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

* Idiom: Bar chart
* Marks: lines
* Channels: length to express quantitative attributes
* The highest runtime is The story of film: An odyssey and the lowest is Bambi meets godzilla

Which Year Has the Highest Profit Rate?

Text

Description automatically generatedGraphical user interface, chart, histogram

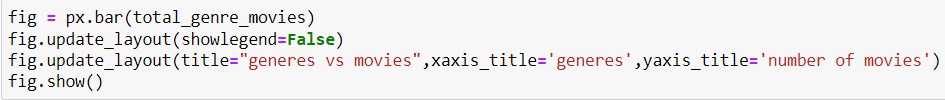
Description automatically generated

* Idiom: Line chart
* Marks: points and line connecting between them
* Channels: aligned lengths to express quantitative value,

separated and ordered by key attribute into horizontal regions.

* year 1977 he most profitable years And the profit was very low between the years 1960 and 1979.(1966 to be exact)

Which Genre Has the Highest Release of Movies?



Chart, bar chart, histogram

Description automatically generated

* idiom: Line chart
* mark: lines.
* channels: lengths to express quantitative value and

spatial regions: one per mark separated horizontal and aligned vertically ordered by quantitative attribute

* Genre which has the highest releases of movies is Drama

we will use bar chart because we will compare categorical values based on quantitative values

another way:

Text

Description automatically generated

A picture containing text

Description automatically generated

* idiom: word cloud
* mark: text
* channels: Font Size to express the frequency of categorical attribute

What kinds of properties are associated with movies that have high revenues?



Chart, treemap chart

Description automatically generatedA picture containing chart

Description automatically generated

idiom: Heatmap

* mark: squares.
* channels:
  + color by quantitative attribute.
  + Number to express value of correlation.
* Budget vs Revenue:

There is a good possibility that movies with higher investments result in better revenues as Correlation = 0.69

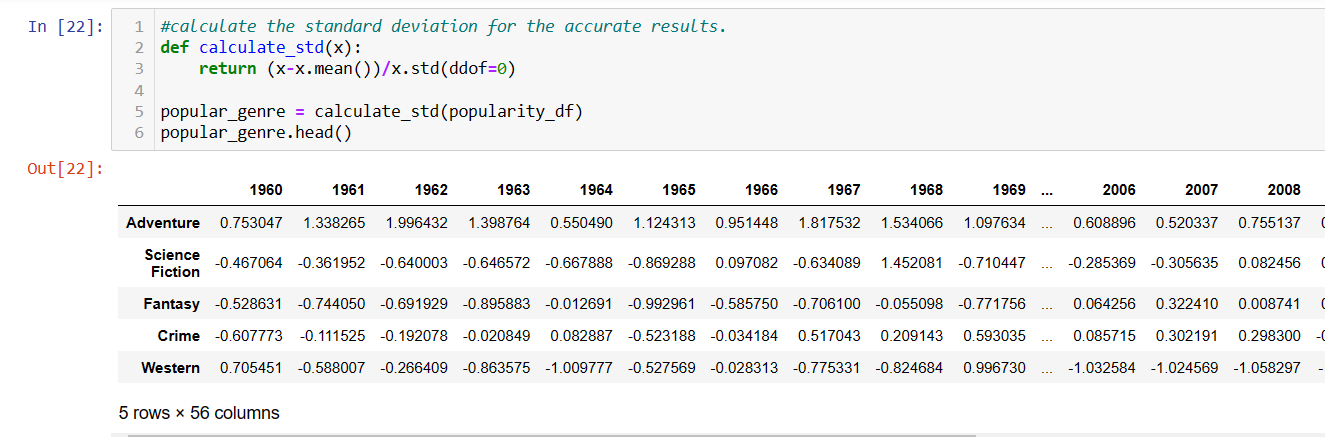
* Popularity vs Revenue

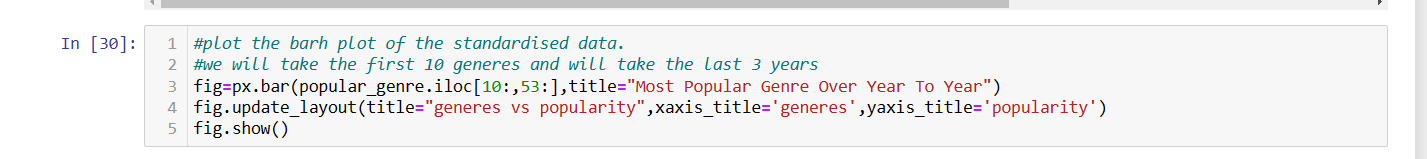
The revenue seems to be increasing with popularity. We can say that if the popularity of movie is high then the revenue of the movie may be high as Correlation = 0.63

* Vote Average vs Revenue

The correlation between revenue and vote average is 0.21. So, vote average is not highly related to the revenue

**Which genres are most popular from year to year?**

****



Chart

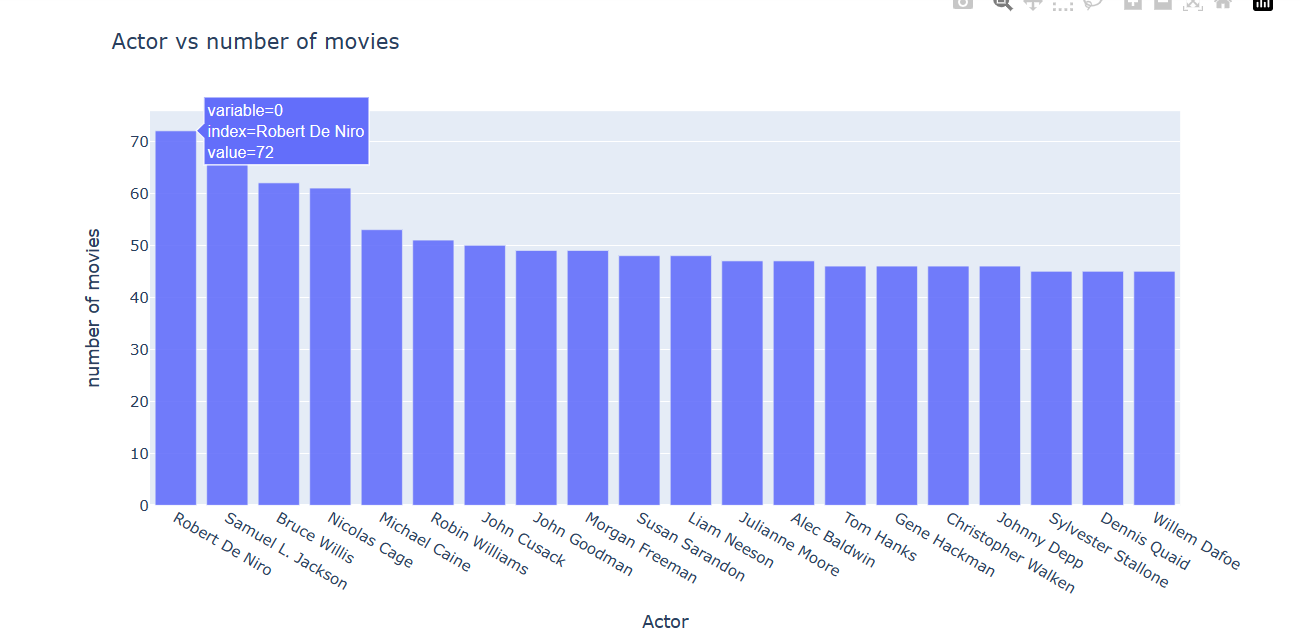
Description automatically generated

* The answer: we see Adventure is increasing.

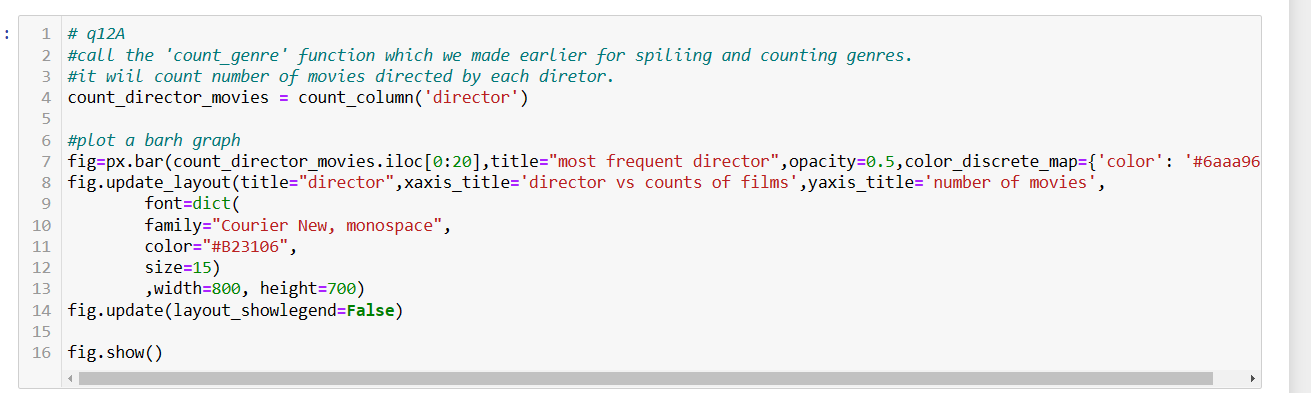
Idiom: Stacked Bar plot.

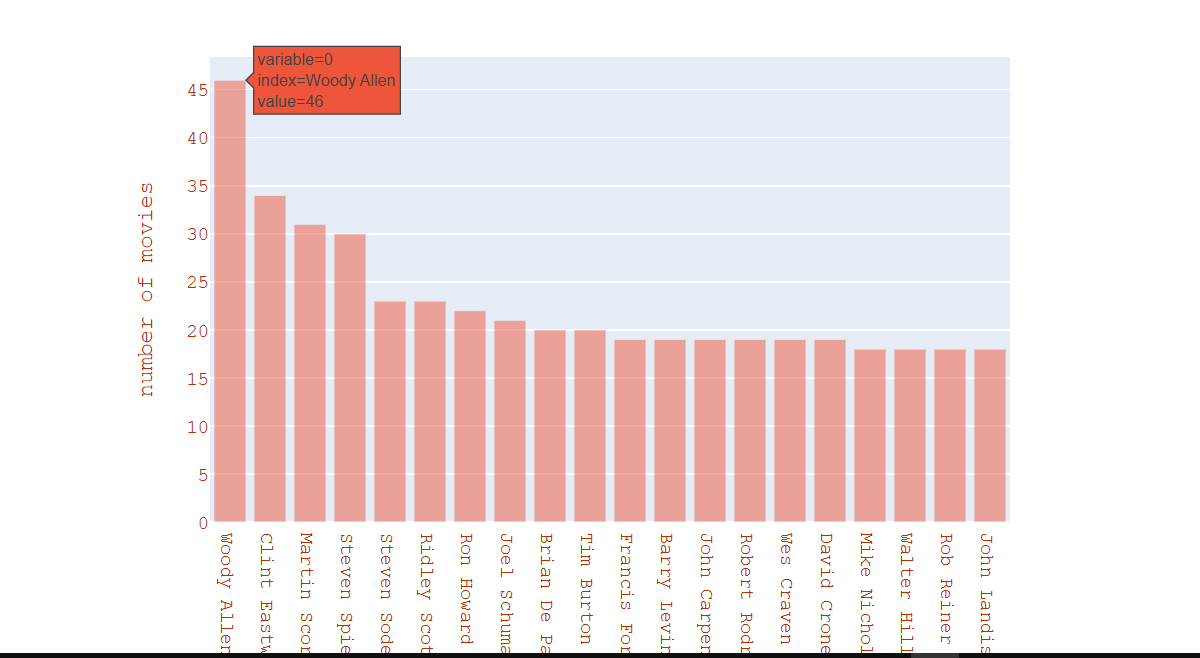
* Mark: Lines
* Channels
  + length and color hue
  + spatial regions: one per glyph

**Most Frequent star cast?**



The answer: from that graph we can see that Robert De Niro and Samuel L. Jackson is the most frequent actor then Bruce Willis and Nicolas Cage.

* Marks: Lines
* Channels:
  + 1-length to express quantitive value.
  + 2-Spetial length one per mark
* Separated horizontal and aligned vertically.
* Ordered by quantitative attribute.
* Top 20 Director Who Directs Maximum Movies ?

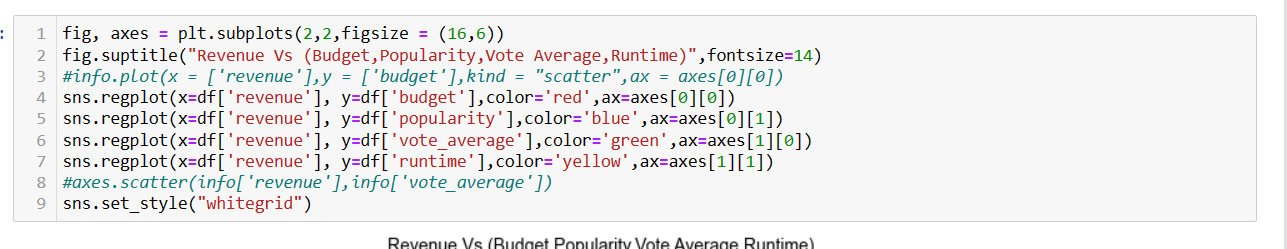


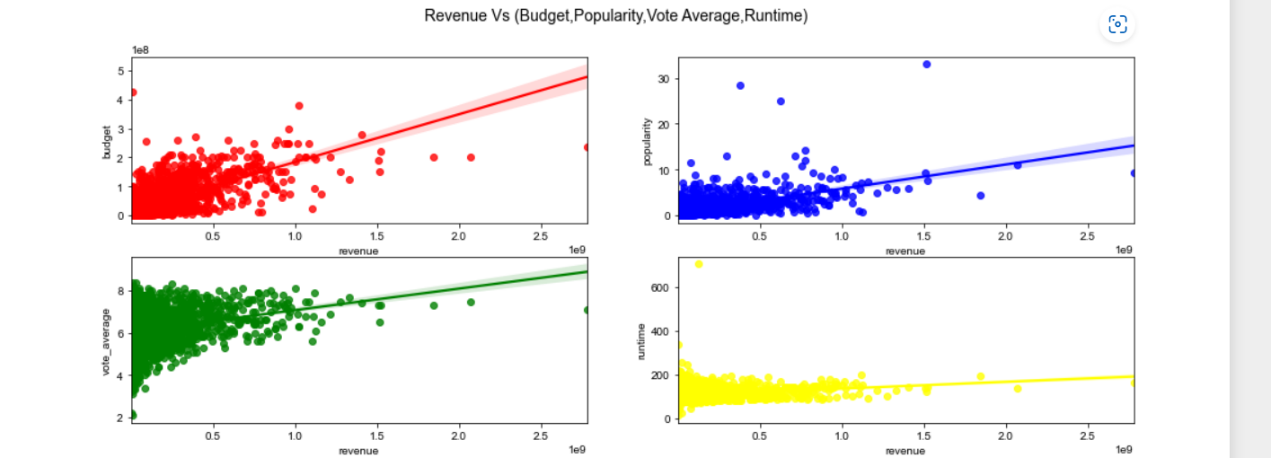
The answer: the most director made movies is Woody Allen.

Idiom: line chart.

* Marks : Lines
* Channels:
  + 1-length to express quantitative value.
  + 2-Spetial length one per mark
    - Separated horizontal and aligned vertically.
    - Ordered by quantitative attribute.

**What kinds of properties are associated with movies that have high revenues?**





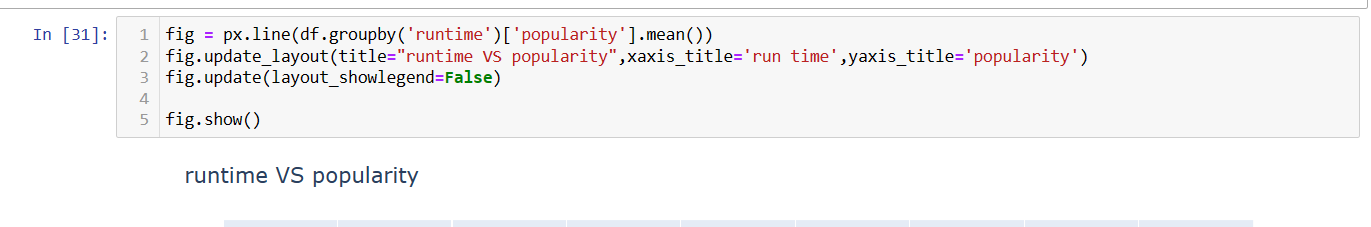
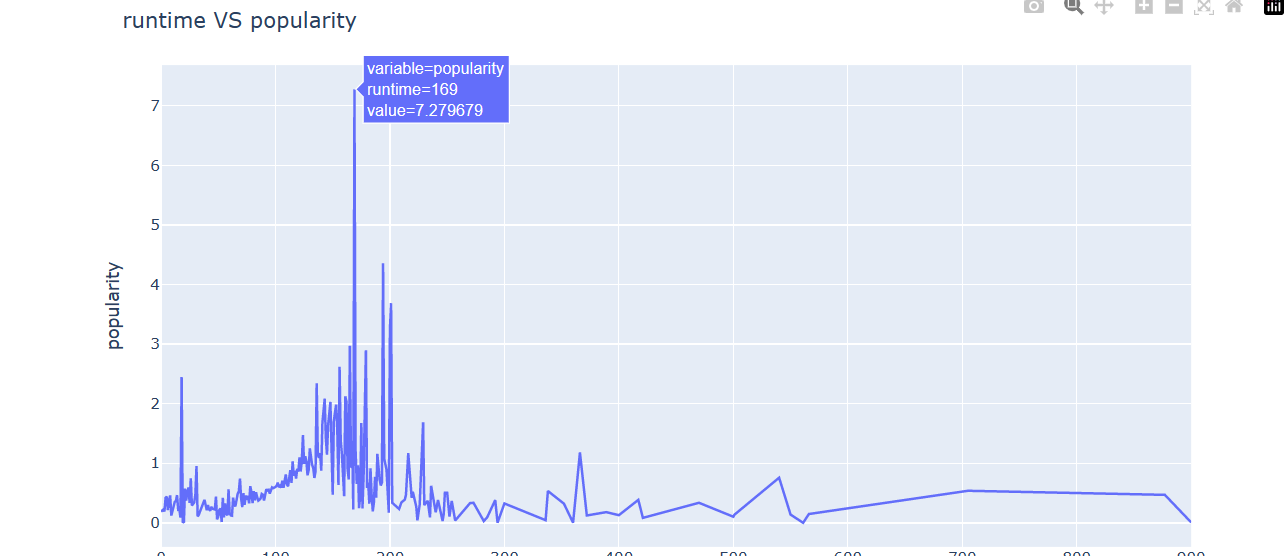
* Idiom: reg plot (Scatter plot).
* Marks: point and line.
* Channels: horiz + vert position +colors.

The answer:

* Budget vs Revenue  
  There is a good possibility that movies with higher investments result in better revenues.
* Popularity vs Revenue  
  The revenue seems to be increasing with popularity.

To conclusion: Movies with higher budgets have shown a corresponding increase in the revenues

Which length movies most liked by the audiences according to their popularity?



* Idiom: line chart
* Marks: points AND line connection marks between them.
* Channels:

1-aligned lengths to express quant value

2- separated and ordered by key attribute into horizontal regions.

The answer: According to the plot we can say that movies in the range of 100-200 runtime are more popular than other runtime movies.

# Pie chart

A pie chart is a circular analytical chart, which is divided into region to symbolize numerical percentage. In px. Pie, data anticipated by the sectors of the pie to set the values. All sector are classify in names. Pie chart is used usually to show the percentage with next corresponding slice of pie. Pie chart helps to make understand well because of its different portions and color codings

Text

Description automatically generated

Chart, diagram

Description automatically generated

* Idiom: Pie Chart
* Marks: interlocking area
* Channels: 2D area varies

separated & ordered radially, uniform height

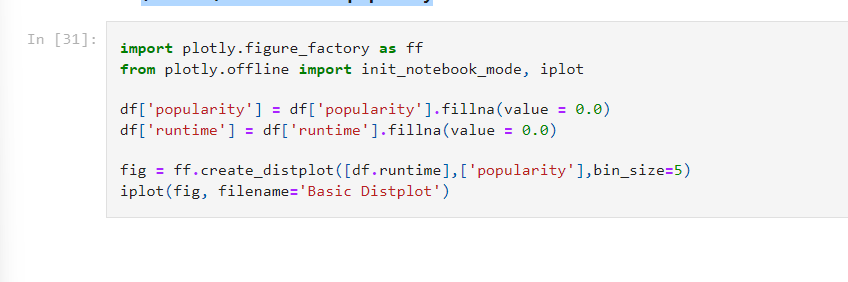
**– task: part-to-whole judgements**

**Conclusion :** united States has the biggest budget with 55% of all budget

# Distplot

**Distplots are used to plot a univariate distribution of observations. This basically plots a histogram and fits a kernel density estimate(kde) and rug plot on it**

* this plot dispaly the relation between populariy and runtime **the films with runtime range (100-200)min have most popularity**

 A picture containing chart

Description automatically generated

Chart

Description automatically generated

* **Conclusion :** the films with runtime range (100-200)minutes have most popularity

# Box Plots

**Median (50th percentile) = middle value of the data set. Sort and take the data in the middle. It is also called 50% percentile that is 50% of data are less that median(50th quartile)(quartile)**

* 25th percentile = quartile 1 (Q1) that is lower quartile
* 75th percentile = quartile 3 (Q3) that is higher quartile
* height of box = IQR = interquartile range = Q3-Q1
* Whiskers = 1.5 \* IQR from the Q1 and Q3
* Outliers = being more than 1.5\*IQR away from median commonly.
* trace = box
* y = data we want to visualize with box plot
* marker = color

Graphical user interface, text

Description automatically generated

Chart

Description automatically generated

Graphical user interface, application

Description automatically generated

Conclusion : most revenue of the film in 2015 are between(0 – 10 ) Million dollar

Plotly provides high degree of interactivity by use of different controls on the plotting area – such as buttons, dropdowns and sliders etc. These controls are incorporated with **updatemenu** attribute of the plot layout. You can **add button** and its behaviour by specifying the method to be called.

There are four possible methods that can be associated with a button as follows −

* **restyle** − modify data or data attributes
* **relayout** − modify layout attributes
* **update** − modify data and layout attributes
* **animate** − start or pause an animation

The restyle method should be used when modifying the data and data attributes of the graph. In the following example, two buttons are added by Updatemenu() method to the layout with restyle method.

Value of type property is buttons by default. To render a dropdown list of buttons, change type to dropdown. A Box trace added to Figure object before updating its layout as above. The complete code that renders boxplot and violin plot depending on button clicked, is as follows

Text

Description automatically generated Chart, pie chart

Description automatically generated Chart, bar chart, histogram

Description automatically generated

**Conclusions**

* 'Avatar', 'Star Wars' are the most profitable movies and judgening by that people tends to like CGI which refers to the 3D computer graphics used to create characters, scenes, and other special effects.
* runtime range (100-200) minutes have most popularity so its better not to produce a a movie with that runtime range.
* May,june,november and december are most popular month for releasing movies, if you want to earn more profit.
* Revenue is directly connected to the budget.
* It is better that the production company be one of those that are like Warner Bros, Universal Pictures or Paramount Pictures production companies earn more life time profit than other production companies as they would not be short on the budget.
* It is better to have a suitable budget (do not try to cut down in costs) as Movies with higher budgets have shown a corresponding increase in the revenues.
* It is better if the movie genre is one of the highest genres that people likes, ex: drama action, and comdy.
* It is better that the director have made many movies that people liked alot