CALIFORNIA STATE UNIVERSITY, FRESNO
DEPARTMENT OF COMPUTER SCIENCE

February 29, 2020

Class:	Big Data Analytics (CSCI 191T)	Semester:	Spring 2020
Points		Document author:	Madhuri Pyreddy
		Author's email:	madhuripyreddy @mail.fresnostate.edu
		Assignment number:	3

1 Statement of Objectives

The objective of Assignment 3 is to retrieve datasets online from different websites like kaggle, visualize the datasets using python libraries like Matplotlib and Pandas, and analyze their trends while making visualizations like bar graphs and pie charts. Also, using the data from the data set, we will be able to perform descriptive analysis for movie dataset. Descriptive analysis is a type of analysis that summarizes the data, broken into measure of tendency and measure of variability. Measure of tendency includes mean, median, and mode. Measure of variability includes standard deviation and variance. We will be able to analyze the findings from the 2 graphs that visualize the relationship between Rating and movies and the relationship between hour and count. For Pandas, we will learn how to manipulate the data using different commands to clean or remove unnecessary data. For MathPlotLib library, we will plot data using bar graph and pie chart. From these types of data, it would be easier to analyze trends to view the relationships between the two variables plotted on each graph.

2 Experimental Procedure

For this report, I had to download multiple data sets like Movies, Ratings, and Hour Data Cases. For the Movies data set, I downloaded as a csv file and wrote some code to download csv file. Then, I performed descriptive analytics to find the dispersion and central tendency of the quantitative measure of budget. I found the mean and variance of the budget. After cleaning the data, I realized that I could find the mean and variance of the budget. I plotted a bar graph and pie chart to visualize the relationship between rating and movies. I did the same thing for the Hour data set to visualize the relationship between hour and count.

2.1 Procedure

In this procedure, I need to find the information for visualizing and downloading data sets like movie set, Ratings, and Life expectancy

A) Movie metadata
1) Go to the website on Kaggle
2) Download the data set on .csv file
3) Perform descriptive analysis using libraries like pandas, numpy, and statistics for .
4) Find mean and variance of budget
5) Clean the data and perform step 3 again

B) Ratings

1) Go to Kaggle website
2) Download the data set on .csv file
3) Perform descriptive analysis using libraries like pandas, numpy, and statistics for variables like ratings and movies.
4) Find rating's score distribution and check if it is Gaussian, Power Law, or Exponential.
5) Use MatplotLib to plot bar graph and pie chart for ratings

C) Data

1) Go to Kaggle
2) Download the data set on .csv file
3) Perform descriptive analysis using libraries like pandas, numpy, and statistics for variables like hours and count .
4) Find hour's score distribution and check if it is Gaussian, Power Law, or Exponential.
5) Use MatplotLib to plot bar graph and pie chart for data set of hours

3 Analysis

Movie Budget.csv

The movie dataset includes important variables like budget, genres, revenue, runtime, status, voting average, and voting count. The movie dataset includes 24 columns and 45467 rows. I performed descriptive analysis to see the mean and variance of the budget. I noticed that when performing the mean of the budget, it did not execute because the value of the budget needed to be converted to float. When I converted budget to float, I noticed that the mean and variance of the budget was different. The mean of the budget is 4.22e+06 and the variance of the budget is 3.03e+08.

Ratings small.csv

The ratings dataset includes 4 baseline variables: userID, movieID, rating, and timestamp. The first data set has 100005 rows and 4 columns. In this data set, I am visualizing and analyzing the data to see which movies are most likely to get higher ratings or lower ratings. In the ratings bar graph that visualizes the relationship between ratings and movieIDs, it is shown that as a Gaussian distribution. The reason is that the trend displays that from from 0 to 3 ratings, the movies increase up to 5,000,000 movies. From 3.5 to 4 ratings, the movies increase up to 7,000,000 ratings. From 4 to 4.5 ratings, the number of movies decreased to 2,000,000. The movies are more likely to receive a rating of 7 and movies are less likely to receive the lowest rating of 0.5. T Most likely, there is a strong correlation between both ratings and movies respectively. This graph is shown as a Gaussian form because the graph shows that the ratings increases and decreases at some point through different movies. In the ratings pie chart, it spreads out the percentage of each movie rating. A majority of the movies received a rating of 4 (around 26.9 percent) and a minority of the movies received a rating of 0.5 (around 1.6 perent).

Data.csv

The data dataset includes 2 baseline variables: Hours and Count. The first data set has 1048576 rows and 24 columns. In this data set, I am visualizing and analyzing the data to see which trips are more likely to take long depending on the number of people (count). This data graph visualizes the relationship between number of people and hours. In this graph, it is shown that from 0 to 7 hours, the number of people that make a trip goes up to 800,000. From 7 to 10 hours, the number of people that make a trip goes down to 400,000. At 15 hours, the number of people that make a trip is at a highest peak of 1,200,000 people. After 15 hours, it slowly decreases to below 200,000 people. More likely, people take 7 hours and 16 hours to make a trip to a city. People that take a trip are less likely to take 20 hours to reach to a city. This graph is shown as a Gaussian form. In the data pie chart, it spreads out the percentage of people who are more likely to take long for a trip to a city. 12.8 percent of the people take 17 hours to travel to a city and 2.6 percent of people take 20 hours or more to travel to a city as well.

4 Encountered Problems

The problem with looking at the data sets is that I needed to plot both bar graphs and pie charts for 2 different datasets. For instance, in the life expectancy data set, the names in the x axis kept overlapping each other. Also, another problem was cleaning the data for the budget because there were too many Nan's and 0's so I had to use a specific command to clean the data for the budget in the movie data set. Making a bar graph and pie chart for ratings data set and hours data set was also difficult because I had to find the right libraries to create bar graphs and pie charts for the specific data sets.

5 Conclusions

In this assignment, I learned how to visualizing data from ratings and life expectancy data sets. In order to visualize data, I managed to visualize different data sets using bar graphs and pie charts. For the movies metadata dataset, it is shown that once you clean the data, it makes a huge difference in finding the mean and variance of the budget in the movies. Once I finished visualizing graphs, I decided to analyze the trends for each of the data sets. There are positive and negative trends in each of the graphs, allowing me to draw conclusions based on the data shown. In the Rating vs movieID graph, the trend shows as a Gaussian distribution. This hour and count graph shows that as the hours go by, the number of people increases and then decreases.

6 References

1) Pandas Tutorial
2) Kaggle
3) MatPlotLib

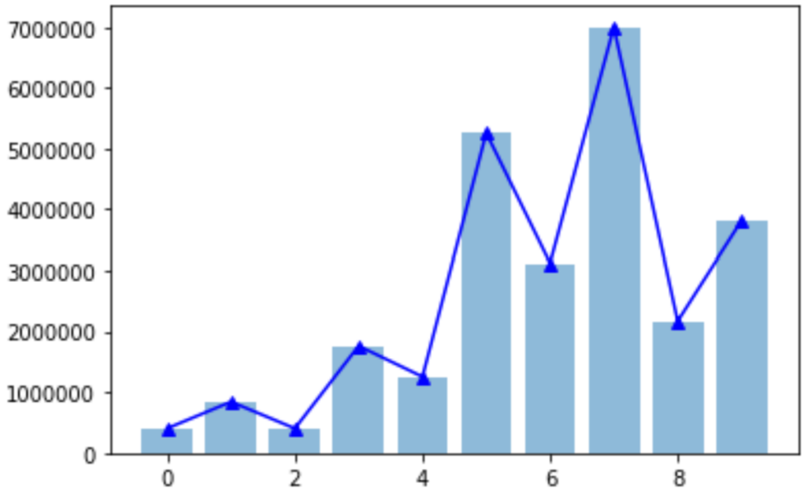

Figure 1:

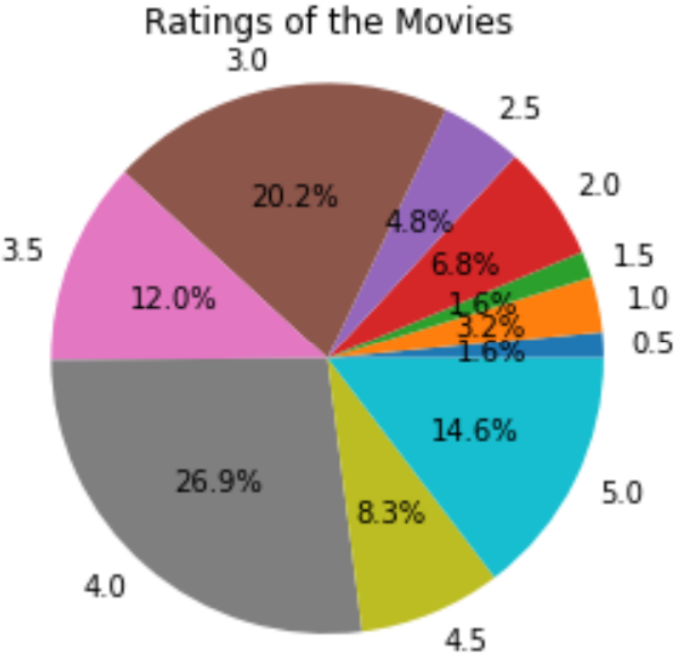

Figure 2:

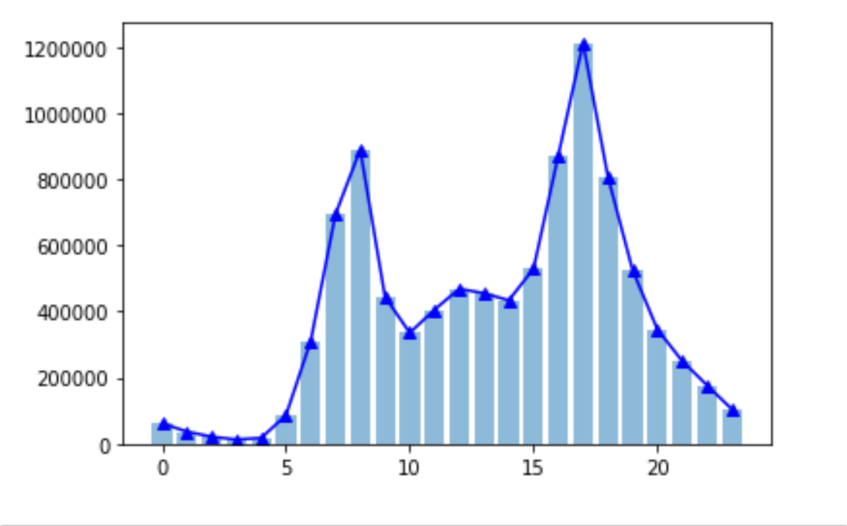

Figure 3:

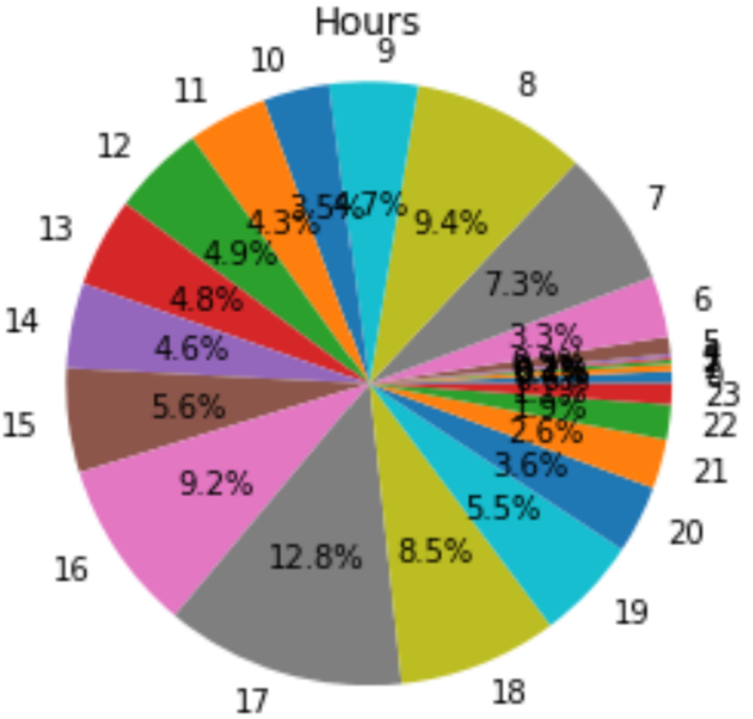
Hours
9
4.7%
8
9.4%
7
7.3%
6
3.3%
5
23
1.9%
22
2.6%
21
3.6%
20
5.5%
19
8.5%
18
12.8%
17
9.2%
16
5.6%
15
4.6%
14
4.8%
13
4.9%
12
4.3%
11
3.5%
10

Figure 4: