CALIFORNIA STATE UNIVERSITY, FRESNO
DEPARTMENT OF COMPUTER SCIENCE

February 29, 2020

Class:	Big Data Analytic (CSCI 191T)		Semester:	Spring 2020
Points		Document author:	Leonardo Yoshida	
		Author's email:	yoshida_leoy@mail.fresnostate.edu	
		Assignment number:	3	

1 Statement of Objectives

The objective of this assignment was to perform descriptive analysis on two datasets, one personal and provided. The main significance of this assignment was to interpret the visualizations created using matplotlib and determine the type of distribution each dataset had for a specific variable.

2 Experimental Procedure

For this experiment, the tools utilized were python, pandas, numpy and matplotlib . The experiments was performed in a windows machine with intel-i7 processor and 16GB of RAM and on a linux machine with intel-i7 and 8GB of ram. Both yielded the same results. The datasets used were about movie metadata, movie ratings and US accidents and their impact onf traffic.

2.1 Procedure

The first step of the experiment was to load the datasets into the python program. To do that I used the read_csv function provided by the pandas library. Once the datasets were loaded, I used the info, head and describe functions also provided by pandas to get a basic understanding of the datasets. For the movies metadata dataset, i had to remove rows that had a budget value of 0, since they would interfere with the analysis. To do that, first i converted the values of the row to numeric using the to_numeric function provided pandas. The function converted the strings into numbers with the exception of some specific values that could not be converted. To deal with does values I added the parameter errors='coerce' to the function and this transformed those values into NaN. Once the data was transformed to numeric and NaN, i used the dropna function to remove all rows that had budget equal to NaN and then took a subset of the data removing rows with budget equal to 0. Once that was complete i utilized the mean and var functions to find the mean and variance of the budget column. For the movie ratings and US accidents I did not need to go through the cleaning step since the columns I was going to analyze were already clean. For these datasets, I simply selected the column I was going to analyze and used the function value_counts() to create a set of key value pairs where the key was the variable and the value was the number of occurrences. Once I had that, Iu used the matplotlib library to plot a pie chart and a histogram to visualize their distribution.

3 Analysis

Discuss and justify the experimental results. Insert the chart or graph of your results to back up your analysis.

3.1 Movies MetaData

After analyzing the budget column of the movies metadata dataset, I was able to find its mean and variance. Those were equal to 21604277.457480315 and 1177219164512387.5 respectively. By looking at these numbers we can clearly see that most data points in the dataset seem to be far different from the mean, showing there is a big variation when it comes to a movies budget.

	budget	revenue	runtime	vote_average	vote_count
count	8.890000e+03	8.890000e+03	8880.000000	8890.000000	8890.000000
mean	2.160428e+07	5.466837e+07	105.383108	6.013273	466.748256
std	3.431063e+07	1.365862e+08	28.292072	1.247683	1030.294482
min	1.000000e+00	0.000000e+00	0.000000	0.000000	0.000000
25%	2.000000e+06	0.000000e+00	91.000000	5.400000	20.000000
50%	8.000000e+06	3.220762e+06	101.000000	6.200000	94.000000
75%	2.500000e+07	4.479118e+07	116.000000	6.800000	406.750000
max	3.800000e+08	2.787965e+09	705.000000	10.000000	14075.000000

```
print("Mean:", movies_meta['budget'].mean(), "Variance:", movies_meta['budget'].var())
```

```
Mean: 21604277.457480315 Variance: 1177219164512387.5
```

Figure 1: Descriptive analysis of the data

3.2 Movie Ratings

By looking at both the pie chart and histogram on Figure2, we can see most of the ratings are between 3 and 4. If we look at the shape of the histogram, I believe the most probable distribution that can be applied to this graph is a Gaussean distribution. However, because of technical problems with python, I was not able to plot the distribution line.

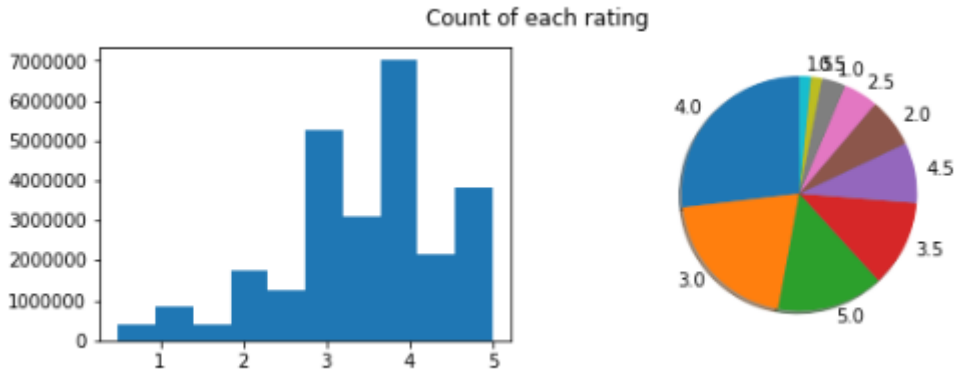

Figure 2: Histogram and pie chart showing distribution of movie ratings

3.3 Personal Dataset: US Accidents

By looking at both the pie chart and histogram on Figure3, we can see most of the accident types are of type 2. If we look at the shape of the histogram, I believe the most probable distribution that can be applied to this graph is a Exponential distribution since the amount of accidents seem to decrease exponentially the higher the type. However, because of technical problems with python, I was not able to plot the distribution line.

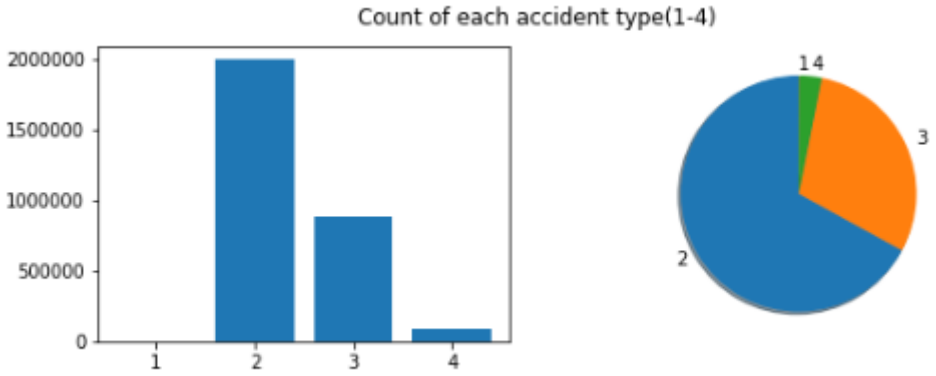

Figure 3: Histogram and pie chart showing distribution of US accidents

4 Encountered Problems

As mentioned before I had some trouble with the distribution line. My problem was mostly on how to calculate it. I researched several possible libraries to help me solve this problem such as scipy and seaborne however none of the solutions I found helped me

5 Conclusions

In conclusion, by doing this experiment I learned several things. First, i got better used to plotting graphs using matplotlib and pandas. By plotting different types of graphs, i got a better understanding of the different graphs that can be created using the library. Second, I got a better understanding of descriptive analysis by performing it on different datasets. Finally I grasped a better understanding on how to identify distribution types by analyzing means, variances and the shape of graphs.

6 References

https://www.kaggle.com/rounakbanik/the-movies-dataset/version/7
https://www.kaggle.com/sobhanmoosavi/us-accidents