CSC 635 Data Mining

## Assignment #1 Report

### Submitted to:

### Dr. Jamil Saquer

### Author(s):

### Rafail Islam

**HW1 Report**

**Introduction**

In this assignment, I did data analysis on a provided data set. This data set contains information about a restaurant over a period of time in 1990. Among several important attributes, tips for the waitresses given by the customer is important as it covers a big portion of the salary of waitresses. From this data set, we want to know several questions such as average tips given my male or female customers, what is the busiest day of that restaurant. We will also try to conclude if it has any pattern that says if customers tend to give more tips on a specific day.

For the second part, while selecting the data set, I made my mind to answer a question what makes America superior in term of Education and Technology. I found one interesting answer. The USA has a lot of startup companies which became a pioneer in their field such as Amazon, eBay, Facebook, Google, SpaceX, etc. To answer my question, I found a small yet interesting data set of 500 US companies [1] that were founded in different state from 1900 to 2015. This data set has 22 attributes such as company name, founding year, categories of company, state, city, employee, etc. I found these stated attributes important to answer my question – the raise of companies in the USA. I showed how many companies have been founded in the last 15 years in different states of the US. I also showed which state has had a high raise in funding new companies in the last 15 years. Among all the categories of the company, I noticed Data/ Technology related company has a very high raise in some specific state. Similarly, few cities have more newly founded companies in recent years.

**Background**

Data analytics gives us the process of analyzing raw data to find trends and answer questions. In python, data analytics is even more fun as it has a lot of tools, open-source libraries that make data analysis, complex mathematical computation, finding relations among data, and vitalizing data or relationships much easier. In this assignment, I used different python libraries like pandas for data prepossessing and computing several complex computations. Matplotlib has been used to visualize the relationship and trends among data.

I used pandas dataframe to handle the row data and also used several pandas mathematical tools such as mean, count, sum. I also used pandas 'groupby' operation to build more complex relations among specific attributes. Several plot styles have been used to visualize trends. I also used a box plot to show a statistical summary for a given attribute.

**Implementation**

This assignment is implemented as instructed in the document. Firstly, I read the data file “tips.csv” with pandas read\_csv function. Along with some basic operation in pandas dataframe, our first question was “How much tip do customers usually give on average”. To find the answer of this question, I used the mean() function. I did a statistical analysis of tips given by all customers by using a boxplot. I answered if male customers give more tips than female customers by using pandas ‘groupby’ function. This groupby operation is done by the attribute ‘sex’ then count mean to know the average tips given by a particular ‘sex’ customers. After that, I used some basic operations to visualize dataset by using a bar and scatter plot. Our next question was “what day is the busiest day of that restaurant”. To answer this question, I did a 'groupby' operation bases on ‘day’ attribute and this operation will show how many customers were served on a particular day. The last question in part I was drawing a pie chart that will show the percentage of all tips given by the customers by the days were served.

In part II, I choose a data set from Kaggle [1] which is about 500 US companies that were founded from 1799 to 2015. In this part, I want to know three answers: i) which state has the most newly formed company in the last 15 years, ii) which state has most ‘Data/Technology’ types companies in the last 15 years, and iii) what categories of the company have been founded in last 15 years ( from 2000-2015). To answer this question, I did a few ‘groupby’ operations and plot them using a bar plot.

**Experimental Setup and Results**

I used jupyter notebook for this assignment with python 3.8. Mostly, I used pandas and matplotlib libraries for all the operations I did here. In part I, after doing all the operation, I came to know that male customers usually give more tips than female customers in Fig.1. I also came to know that Saturday is the busiest day of that restaurant in Fig.2. My result also shows that Saturday is the highest (35.9%) tips giving day followed by Sunday (33.82%) in Fig.3.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Fig.1 : Average tips given by male and female customers | Fig.2 : Busiest day of the restaurant | Fig.3 : Percentage of tips per day |

In part II, I got to know that California has the highest number of the company founded in the last 15 years, followed by New York in Fig.4. It also shows that most of the newly formed companies are Data/Technology and Finance related in Fig.5. Our next investigation was about which state has the most tech companies. It shows that CA has the most Data/Technology companies in Fig.6.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Fig.4 : Company per state | Fig. 5: Categories of newly founded companies | Fig.6 : Highest number of Data/Technology company per state |

**Conclusion**

In this assignment, I learned how to analyze a dataset and draw a useful conclusion based on the data. I also learned a few styles of data visualization techniques and statistics of any given attributes. In a few words, I learned how data analysis help to answer the question easily and effectively.

**References**

[1]. Damarla, Rishi. “Data about 500 US Companies.” Kaggle, 20 Sept. 2020, www.kaggle.com/rishidamarla/data-about-500-us-companies.

**Code**

I have attached the source code as a separate file that can be executed in Jupyter Notebook.