Big Data Analysis of Taxi Services in Chicago and New York

Himani Batra, Katia Kermoyan, Nishant Shristiraj, Srihitha Reddy Sivinnagari
Department of Information Systems, California State University
Los Angeles
E-mail: hbatra@calstatela.edu, kkermoy@calstatela.edu, nshrist@calstatela.edu, ssivann@calstatela.edu

Abstract: With the increase in population and traffic, the usage of taxis has also increased in metropolitan cities like Chicago and New York. This paper carefully studies data about existing services and derives meaningful insights and relationships between the extent of taxis used in Chicago and New York.

1. Introduction

For this project, we used two data sets of total size 2.4GB about taxi services in Chicago and New York from kaggle.com.

Kaggle is a platform for predictive modeling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users.

Both datasets include similar columns like total amount of fare, time taken for each ride, pick up and drop off locations, distance travelled and company/ vendor. The difference between them is that the Chicago dataset contains 12 month-data about Chicago taxi rides in 2016, while the New York dataset contains data about New York taxi rides in January and February of 2014.

We created a cluster in IBM BlueMix cloud, and used tools like Hadoop, Hive, Pig, Excel, Tableau, and Azure ML to perform analysis and draw insights into these two datasets.

2. Specifications of Datasets

2.1 Dataset 1

Includes 12 month-data about Chicago taxi rides in 2016.
URL: https://www.kaggle.com/chicago/chicago-taxi-rides-2016

- File size: 2.02GB
- Number of files: 12
- File Format: csv

2.2 Dataset 2

Includes data about New York taxi rides in January and February of 2014.
URL: https://www.kaggle.com/kentonnlp/2014-new-york-city-taxi-trips

- File size: 489MB
- Number of files: 1
- File Format: csv.gz

3. H/W Experimental Specifications

- **Location:** IBM Cloud – BigInsights for Apache Hadoop
- **Version:** IOP 4.2
- **Number of Nodes**: 1 Data Node
- **Memory Size:** 24GB RAM
- **Number of CPUs:** 4 VCPU
- **Data Storage:** 1TB SATA
- **Number of Nodes**: 1 Management Node
- **Memory Size:** 48GB RAM
- **Number of CPUs:** 12 VCPU

4. Tools

Figure 1. Tools used

4.1 IBM Bluemix

Was used to create a cluster within the Apache Hadoop framework.

4.2 Hadoop Distributed File Sysytem (HDFS)

Was used in the IBM cluster for storing the data downloaded from Kaggle.com.

4.3 Pig and Hive

Were used to build tables, queries and analyze the data.

4.4 Tableau and Excel

Were used to draw insights from the analysis.

4.5 Azure ML

Was used for predictive analysis on total_amont(New York) and trip_total(Chicago).

5. Flowchart of Data Analysis

Figure 2 briefly describes the flowchart used for this project. For creating visuals, we are extracting data from our dataset using Hive and Pig queries. The size of this data to be visualized ranges from 1KB to 200Kb, exported in Excel as CSV or loaded directly in Tableau as text file.

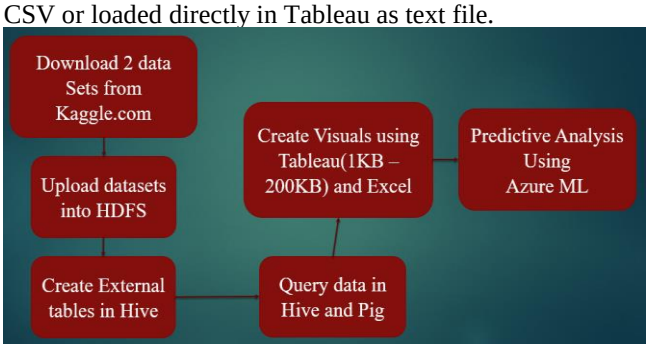

Figure 2. Project Flowchart

6. Data Analysis

6.1 Highest Earning Companies in Chicago

Figure 3 shows the companies doing the maximum business. Hive query was grouped by company ids to find the sum of

their total income throughout 2016. From the visualization, Company 107 covers more than 40% of the market in Chicago.

Figure 3. Highest Earning Companies in Chicago

6.2 Taxi IDs[1] with highest fare and respective number of rides in Chicago

Figure 4 shows two different sections in the analysis. It compares the taxi ids with the amount of money earned and the respective number of rides they give. From the analysis it is clear that there is no correlation of amount one makes with the number of rides.

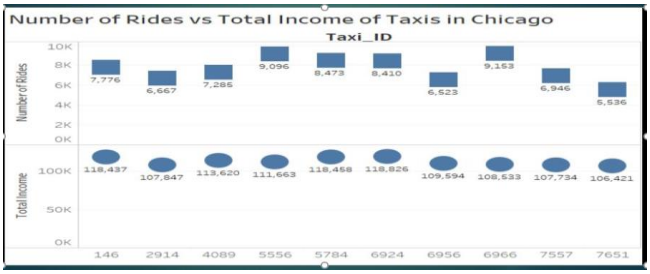

Figure 4. Taxi-ids[1] that generate the highest fares in Chicago with the respective number of rides.

6.3 Comparison of Payment Modes in Chicago

The analysis in Figure 5 compares the credit and cash payments for taxis for each month. It is clear that credit card payment are more in each month than cash payments.Anyone willing to start a new taxi company can collaborate with credit card companies to offer benefits to customers using those cards.

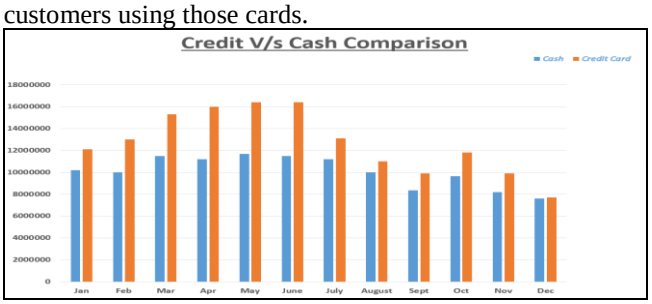

Figure 5. Total monthly comparison between Credit and Cash payment in Chicago.

6.4 Total Monthly Pickups in Chicago

It is clear from the analysis in Figure 6 that there are more number of rides in Spring and Summer than in Winter. Chicago experiences snowfall in winter, which might be the leading reason for this drop in number of rides.

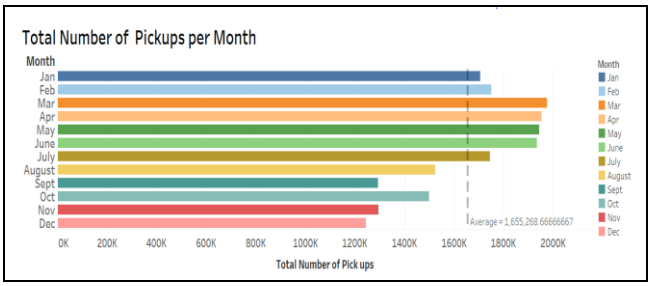

Figure 6. Total number of monthly pick-ups in Chicago.

6.5 Communities[2] with highest frequency of taxis in Chicago

Figure 7 shows the communities which enouter highest numbers of taxi frequency. Communities 8 and 32 are downtowns of Chicago and 77 is the business district. The analysis shows that taxis have more miles travelled in these ares than any other area in Chicago.

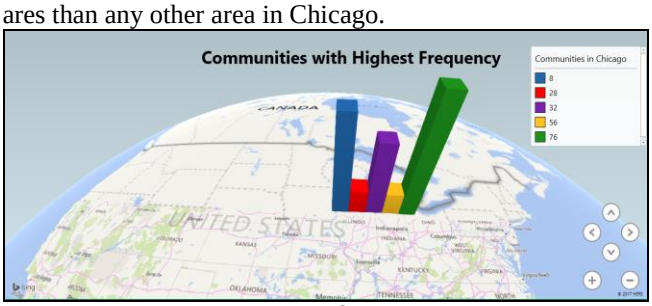

Figure 7. Chicago communities[2] that have the maximum miles travelled by taxi.

6.6 Communities[2] with Highest Pickup and Dropoff in Chicago

Communities 8 and 32 have the highest number of users getting picked up and droped off, as shown in Figure 8. As discussed in section 6.5, these communities are downtowns of Chicago and expected number of users are higher in these areas than any other areas.

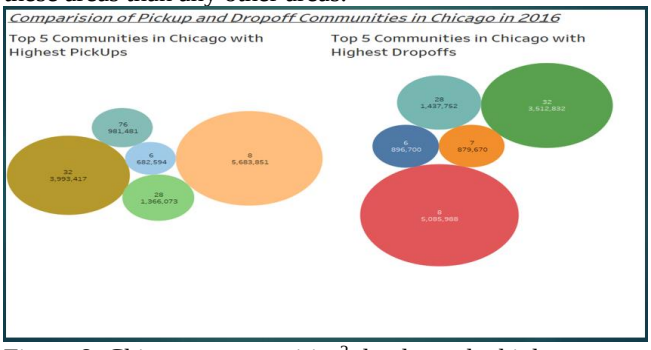

Figure 8. Chicago communities[2] that have the highest demand for pick-ups and drop-offs.

6.7 Busiest days of a month and peak hours of a day in New York

The visualization in Figure 9 shows that the demand is higher on weekends.

---

[1] Taxi-ids are identification cards issued to taxi drivers.
[2] The city of Chicago is divided into 77 communities that are officially recognized by the City of Chicago by numbers.

While Figure 10 shows that, in an average day, demand for pickups are low in the mornings and the highest around taxis is 6:00PM and 7:00PM.

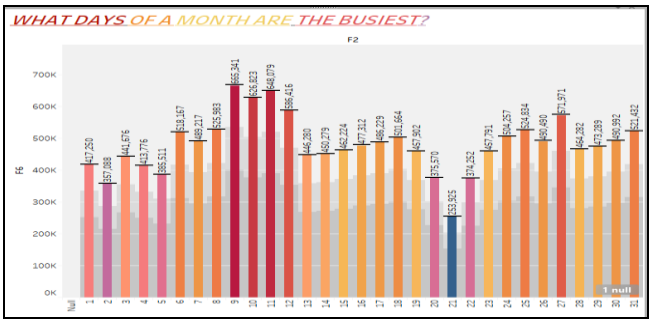

Figure 9. Daily pick-up trends of an average month in New York.

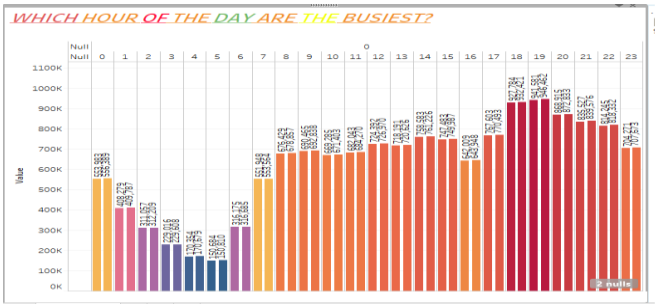

Figure 10. Hourly pick-up trends in a day in New York.

6.8 Regions of Highest Demand and Maximum Miles in New York

Figure 11 shows the specific areas of New York which has the highest demand for taxis. Downtown New york and Central region have the maximum pickups. The longest mile travelled by a taxi is 100 miles.

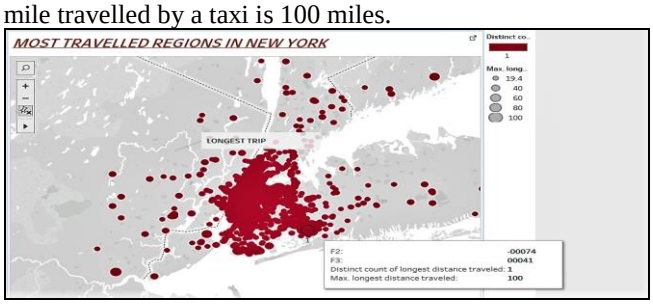

Figure 11. New York regions with highest demand and longest distance travelled.

6.9 Highest Earning vendors in New York

Figure 12 shows that there are only two vendors in New York city taxi market. VTS is the vendor with highest income among the two vendors which is $7,634,893.

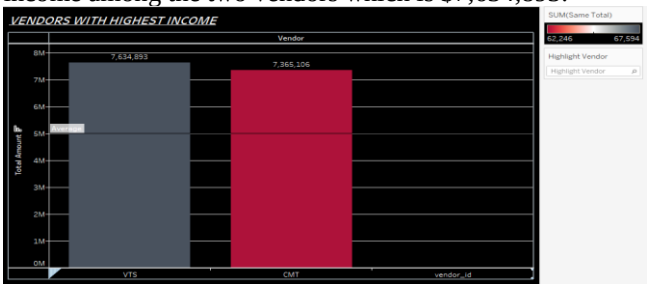

Figure 12. Highest earning vendor in New York.

8. Machine Learning in Azure

Azure Machine Learning is a cloud predictive analytics service that makes it possible to quickly create and deploy predictive models as analytics solutions. You can work from a ready-to-use library of algorithms, use them to create models on an internet-connected PC, and deploy your predictive solution quickly.

Azure Specifications:

- Cloud platform
- 10GB Memory

8.1 Linear Regression Model for New York Dataset

On observing our datasets which had continuous values, we decided to get an insight that we couldn't get from the existing data. We chose linear regression model algorithm to get predictive values of total amount that a vendor charged for a taxi ride. One row in the datasets represents one trip. The labels chosen in both datasets are the total amount which are the summation of fare charged, tips given to the drivers, tolls, extras and tax.

8.1.1 Selection of Labels and Features

We selected a few columns which are most appropriate in predicting the labels. Our initial analysis on the two datasets helped us decide of the columns which were best suited to be featured. We chose total_amount in New York data set and trip_total in Chicago dataset as labels on which we want to predict. There are several columns in both data sets like pickup_latitude, pickup_longitude, dropoff_latitude, dropoff_logitude, vendor, communities as features in the datasets.

8.1.2 Model Construction Process

The datasets were loaded into azure ML in the form of a .csv file. The columns that were not being used for our predictive analysis were eliminated using the select column module. The datasets are then cleaned by deleting the rows which have zero values. The columns which are to be included in the analysis are selected using the select column module. The filtered-out columns are then sent to the split model module to split the dataset to a ratio of 7:3 where 70% of the data is sent to the train model and the rest 30% of the data is used up for testing for both datasets. Then we train our selected label using three linear regression models. Score models are included to check the predicted values. We then evaluate model using by comparing the predicted value against the actual value. Evaluate model also gives us the accuracy, precision and recall values after comparison.

8.1.3 Linear Regression Model

Regression is a machine learning used to predict a numeric outcome. Linear regression attempts to establish a linear relationship between one or more independent variables and an outcome, or *dependent variable*. Multiple inputs are used to predict a single numeric outcome, also called Linear Regression Model.

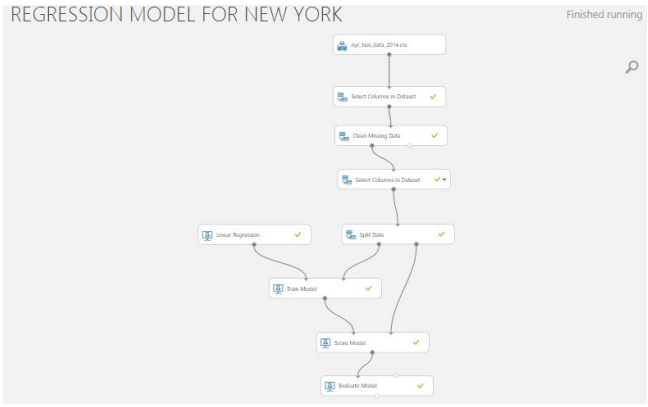

Figure 13. Model Construction with Linear Regression for New York

When we visualized the evaluate module, the metrics shown for errors metrics showed the difference between the total_amount in the dataset and the predicted values of the total_amount column.

Metrics

Mean Absolute Error	2.131045
Root Mean Squared Error	5.26648
Relative Absolute Error	0.283662
Relative Squared Error	0.186548
Coefficient of Determination	0.813452

Figure 14. Error Values of Evaluate Model for New York

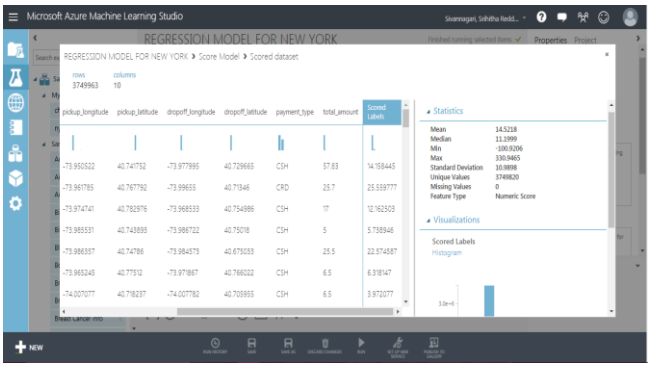

Figure 15. Predicted Score Value Label for New York

8.2 Linear Regression Model for Chicago Dataset

We carried out the same model construction for Chicago dataset. Selected the columns that were we wanted excluded from the analysis. Cleaned the dataset by removing the rows which have zero. We then selected the columns that we include in the analysis. The data which is cleaned with all the features and labes is sent to split data and is divided in 25% and 75%. We the chose the linear regression module and train the data. The trained model is sent to score model. When we visualize this module, we got the scored label which gave the predicted values for the trip_total in Chicago. We evaluated this model to find the accuracy.

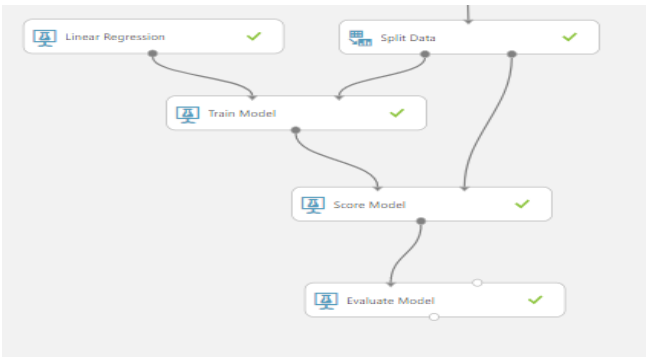

Figure 16. Model Construction with Linear Regression for Chicago

When we visualized the evaluate model, the metrics showed the Error value which the difference between the original values in the dataset and the predicted values.

Metrics

Mean Absolute Error	5.956004
Root Mean Squared Error	42.751226
Relative Absolute Error	0.667546
Relative Squared Error	0.955534
Coefficient of Determination	0.044466

Figure 17. Error Values of Evaluate Model for Chicago

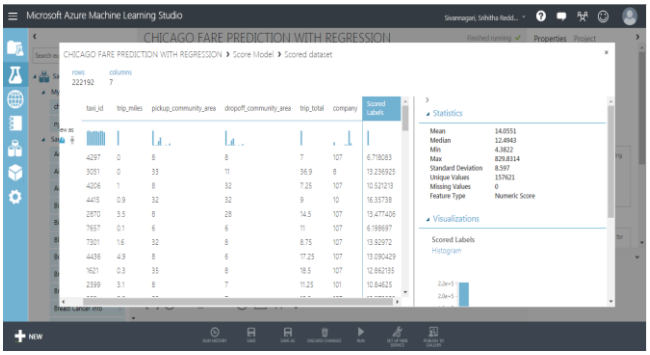

Figure 18. Predicted Score Value Label for Chicago

9. Comparison between Two Cities

On careful execution of predictive analysis, score label shows that the total earning in the coming years for both the cities are going down. For any profitable business, it is very important to generate cash flow.

Figure 19 shows the comparison between the total amount different vendors are making. These are the top vendors in Chicago and New York. The top vendors in New York are VTS and CMT, which are making approximately $12 million, where as the top vendor in Chicago by the id 107, makes close to $80 million. The other vendors not make as much as 107 in Chicago, but they are comparably earning more than the vendors in New York. So, it can be inferred that Chicago is a better place for starting a new business in taxi services.

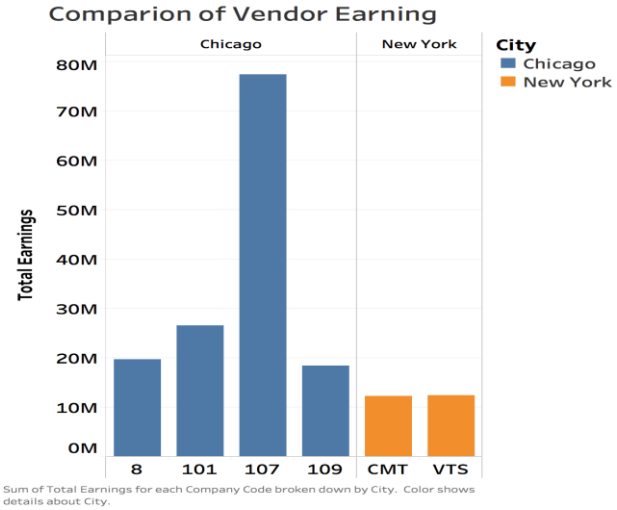

Figure 19. Comparison of Earning for Different Vendors

10. Conclusion

The analysis we have carried out on the existing data together with the predictive analysis helped us get a comparison between New York and Chicago data set.
The values we got after Prediction analysis we carried out in both the datasets were decresing linearly.

References

[1] Garyericson. "A Simple Experiment in Machine learning Studio." A Simple Experiment in Machine Learning Studio | Microsoft Docs, docs.microsoft.com/en-us/azure/machine-learning/studio/create-experiment

[2] "Pig Function Cheat Sheet." Qubole, www.qubole.com/resources/pig-function-cheat-sheet/.

[3] "Apache Hive Guide." Cloudera, 22 Nov. 2017, www.cloudera.com/documentation/enterprise/5-8-x/PDF/cloudera-hive.pdf.

GitHub
https://github.com/srihithasreddy/cis5200-project