An Investigation into New York City's Bike Share Program for the Year 2015

Ryan C. Donovan
Data Analytics
NCI
Dublin, Ireland
Email: x16104269@student.ncirl.ie

***Abstract*—CitiBike for 2015 is investigated by using MapReduce, mySQL, HBase, and R. MapReduce is used in ETL, subsetting and some rudimentary statistics. Sqoop utility is used to import data from mySQL to HBase with limited success. HBase is used to store data in HDFS in order for MapReduce to make subsets to be sent to R for analysis. kNN and association rules machine learning algorithms are run to predict missing values and patterns in the data with limited success. Temporal and distance analysis of the bike rentals are main focus**

I. Introduction

Bike sharing has become very popular and as of 2016 there are roughly 1,000 cities in the world that have them [1]. It can have the benefit of reduced traffic and even help people's fitness. CitiBikes was launched in New York City at the end of 2013 and has been well received by the citizens of New York. They offer 2 services to customers, being yearly subscriber or 1 and 3 day passes. The subscriber pays a flat fee and are allowed to use a bike for 45 minutes before additional fees are succumbed. 1-Day and 3-Day passes pay additional fees after 30 minutes. CitiBikes release data about each bike rental along with certain demographics of renters. Analyzing this data could be could show where and when rentals occur most often which could be used to increase bike capacity of a station or even shut down one with low usage. The data could also be used to find out which genders or ages use the bike most often, along with a myriad of other bits of information. This investigation seeks to answer the question:

> *Can the Citibike bikeshare program better allocate its resources using Hadoop's MapReduce paradigm along with statistical and machine learning techniques?*

Each year's dataset is over a GB so MapReduce (MR) will be used as it is good for processing high volume data. The databases mySQL and HBase will be used to store data and R will be used for analytics. MR is used for ETL, data subsetting, and capture of rudimentary statistics. Related work will be discussed and then followed by the methodology. Next, evaluation of results will be discussed and followed by conclusions and ideas for future work.

[1] http://www.seattletimes.com/seattle-news/transportation/will-helmet-law-kill-seattles-new-bike-share-program/ [Accessed April 25, 2017]

II. Related Work

Two key pieces of software are used to get manageable data sets for analysis in R: MapReduce and HBase. These will be reviewed along with other studies involving BikeSharing programs.

MapReduce (MR) is a programming model that operates on top of HDFS that started in 2004 with Google's implementation of Bigtable to process big data by using parallel processing [1] ,[2]. MR can be used with languages like Java, Python, C++ and others but this study uses Java [3]. Google's Bigtable can be used to store data at Petabyte scale, be scaled linearly as demand dictates, and also be stored on multiple low-cost hardware [4]. MR can also perform these same tasks as it follows same model. MR has been used in physics in the Large Hadron Collider to process data at the Petabyte scale [5]. The data set for all years in CitiBike up to 2017 is roughly 6GB so it would be perfect for using MR to process. Data stored in HDFS affords fault-tolerance and multiple back ups [2] so data loss shouldn't be a problem. This investigation uses pseudo-distributed mode (one cluster) so it doesn't represent most real world use cases of MR as it doesn't parallele process using other compute

Most databases traditionally used have been row oriented, but HBase has created a column-oriented database which is open source and sits on top of HDFS [6], [3], and [7]. [3] further state that it is good for fast access for data in HDFS and is great for scale as it can just add new data storage nodes to its framework. If this experiment was using real-world hadoop clusters it would be great to analyze all the years of the bike share because it would be quicker according to the literature.

Bikesharing has been implemented about many cities in the USA and [8] notice that high population centers and employment have predicted higher usage. [9] and [10] investigated London's bikeshare program which discuss analyze the time and distance of rentals. [11] state that age and station location has a strong influence on usage.

III. Methodology

A. Data Exploration

Data for 2015 used can be found at [2]. There are 12 CSV files (1 per month) totalling 1.73GB. The 15 attributes it contains are shown in Figure 1. stoptime was not used in this investigation.

Attributes from 12 CSVs for 2015		
trip duration	start latitude	end longitude
start time	start longitude	bike id
stop time	end station id	user type
start station id	end station name	birth year
start station name	end latitude	gender

Fig. 1. Attributes in CitiBike dataset that are released publicly

Initial investigation on January 2015 data was done in R to get a feel for the data through graphing, acquiring basic statistics and locating potential data problems like missing values to help in writing the MR programs for ETL.

The following briefly discusses attributes used. Trip duration is total length of a rental in seconds. The time attributes give timestamp to minute granularity. id and name uniquely identify rental stations. Similarly, bike id identifies each bike. Latitude and longitude represent the locational information of each station. User type indicates whether a customer is a yearly subscriber or 1/3-day pass holder. Birth year represents the birth year of a customer but it will be transformed to age for better understanding. Gender represents the gender of a customer and has three values: unknown, male, and female.

In the beginning there were issues in loading data. Missing values were given the term ”empty” in the initial load to HBase. Many mismapping occurred where data was spread over multiple rows rather than one. This was happened because Java’s ”split” method used wrong parameters. Empty values were skipped when using the parameter ”(,)”. For example, only 4 rows were picked up in a row like: ”1,23,Subscriber,,543” when should’ve been five with one missing value. Changing parameter in split method to ”(”,”,-1)” solved this issue. A MR was run to get missing value counts for all attributes and birth year was only one having them accounting for 13% of rows. There was a wide variety in missing values for birth as shown in Figure 2.

A customer’s age was calculated by taking difference of 2015 and the birth year. The age only gives the floor value of person’s real age as didn’t have the exact DOB. It was noticed that there were people with ages over 110 when running a MR for counts of each age. It is very unlikely these renters were this age and this study will only take into account people under 60 years of age in case the data is corrupted. People over 60 only account for 4% of the data. All stations weren’t

[2]https://s3.amazonaws.com/tripdata/index.html [Accessed April 25, 2017]

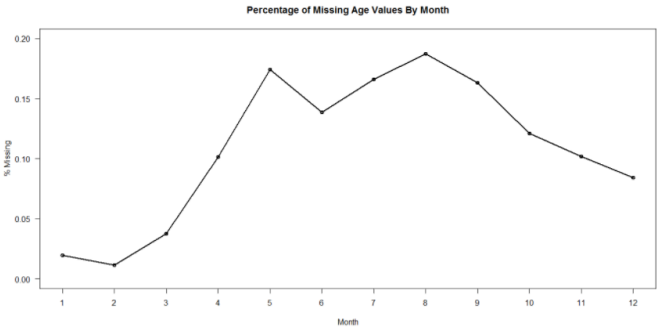

Fig. 2. Shows missing values in percentage for age

operational throughout the year so a MR was run to find all the stations that were for 2015 (66% in total). When comparing stations only all year stations will be analysed. A MR was used to get unique station id and name combinations to see if all stations were unique. Some had slightly different spelling and this was fixed.

B. Data Loading

	Host Machine	Virtual Machine
OS	Windows 10 (64-bit)	Ubuntu (64-bit)
CPU	Intel(R) Core(TM) i3-5005U @ 2GHz	2 Processors @ Unkown GHz
RAM	8.00 GM	5264 MB
Storage	1.00 TB	50.48 GB

Fig. 3. Specification for Host and VirtualBox Machine

One laptop with a single node Hadoop cluster in pseudo-distributed mode was used and run on a VirtualBox machine. Compute specification are shown in Figure 3. Data was initially loaded into mySQL, followed by using Sqoop to load data into HBase. Most data being processed is processed together so only one column family was selected for HBase table. The computer crashed every time when using Sqoop to load the entire data set. An initial workaround was to load 1 CSV at a time into mySQL and then Sqoop this data to HBase. This would be done 12 times and was very time consuming. This only worked when using 3-4 attributes and crashed when more were used.

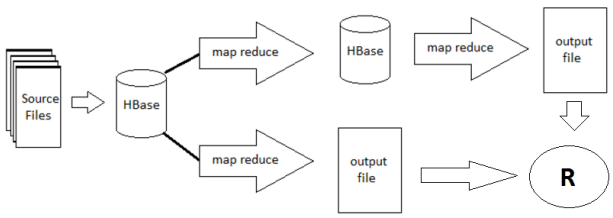

Fig. 4. Data Flow of the System

The final solution to solve the loading of entire data set was to upload the 12 source files directly to hdfs and then use a self-written MR to load the data into HBase. This worked and the rest of the data flow process used is shown in Figure 4.

There were some transformations that will be discussed in the following. Some text files contained quotes around fields and others didn't of which caused issues so these were accounted for programmatically in MR. Latitudes and longitudes were not mapped explicitly, but were used to calculate distance between starting and ending stations in a rental. This reduces storage as only one field was kept rather than four. The distance only represents the straight line distance between each station and not the distance a person actually travelled on the bike. Distance was calculated using an equirectangular distance approximation used by [3] because it decrease processing time by not using exact distance. All the stations are within 10 miles of each other so this approximation is comparable up to 1/100th of mile (as was checked for 1,000 rows). The start time was not in a structured pattern like "mm/dd/yyyy hh:mm" (i.e. June 1st at 1:00AM would be "6/1/2015 1:00" and not "06/01/2015 01:00"). This issue was taken care of programmatically and was used to map the day, month, year and hour of a rental to HBase. The final HBase table contained 13 attributes and 9,937,950 rows.

C. Rental Times and Distance Investigation

Length of rental times were investigated by getting counts and total summed duration for unique combinations of date and time down to the hour level using MR. This was then sent to R in CSV form to be analyzed as was done for other parts of this study. A MR got summary statistics like mean and standard deviation as well. The subset was graphed for rental counts by different temporal views. This graphing allowed for outliers using boxplots to be located for elimination in analysis (less 1% of subset). Knowing rental lengths could be used to set the time length a customer can use a bike before being incurring additional expenses. Another MR got the average distance of rentals across different temporal attributes to see if there was any patterns surfacing. This could be used in adding new stations at a certain distance from others by knowing the distance most customers travel.

D. Bike Investigation

Bike use was investigated to see if they were all used similarly in terms of times, distances and rental count. A MR got each bike's total distance traveled, time used, and rental counts for the year. R was used for graphing histograms and scatter plots to look for interesting patterns. A MR was done to find out which bikes were used for the entire year so bike use comparison could be unbiased by comparing like with like. Only 45% of bikes were used for entire year. A linear regression was run to see if there was a relationship between the number of rentals and distance a bike travelled.

[3]http://www.movable-type.co.uk/scripts/latlong.html [Accessed April 25, 2017]

E. Station Investigation

Three MRs were written to subset station data for analysis. The first got the counts of start and stop stations to see which stations are most used and vice versa. This could be used to eliminate some low usage stations or add capacity to high used ones. Also, spatially seeing the station usage could give a "bird's eye" view of all the stations. Some distances were 0.0 miles and this could've been an error in coding. A MR was run to check to find all occurrences of bikes being rented from and returned to the same station. There were many, thus confirming there was no error. The third MR run got the most recent and oldest date a station was used to find stations used throughout the year. 66% of the stations have been used in 2015 so when comparing use of stations to each other these 66% were used.

F. Machine Learning Investigations

Two machine learning algorithms used a subset created via MR: kNN and Association Rules. This MR filtered out missing age values and age values over 60 of which will be labeled "bad" values. First the bad values were taken out and then a random selection of rows were taken from each month so that they all had same percentage. August had the most bad values at 22% so other months were fixed programmatically to account for the differences. For example, if another month had 10% bad values, than another 12% needed to be filtered out to make the subset properly stratified. A random number generator was used to randomly select rows to equal the threshold for each month. After the months were dealt with, 25% of this data was subset was sampled using random number generator again. This subset was then sent to R via CSV for machine learning to take place.

Gender values were either male, female or unknown in the dataset so kNN was used to see if these unknown values could be filled in for better gender analysis for entire dataset. Only the numerical attributes were chosen (Age, Distance, and rental length) because kNN needs all numerical or all categorical attributes to run. Histograms of the 3 attributes were done to search for outliers that could possibly bias the algorithm. Distance and time both had outliers and these were filtered out (only accounted for less than 1% of data). The square root of the row count was used for initial k = 1,324. A max-min normalization was also done so that the 3 attributes will have equal contribution to the model [12]. R didn't run fast enough at first because there were 1.75 million examples so the data set was reduced to 10%. k-values from 1 to square root of number examples at intervals of 5 were then tested.

Association Rules was run to see if any interesting patterns would surface. The attributes age, distance, gender, month, tripduration and user type were all used with same subset used in kNN. Variables were changed to categorical variables to reduce processing time because the numerical variables had a wide range of values and the matrix might suffer from the curse of dimensionality. The apriori algorithm from arules package in R was used with multiple values of support and confidence.

IV. Evaluation and Results

A. Temporal and Distance Results

Interesting patterns were found by graphing the counts of rentals by various date and times. Figure 5 shows the total rentals for each day of the year. The values had an up and down pattern that made more sense when viewed by average rentals for weekday and weekend for each week as in Figure 6. The average weekday had higher rentals averages than weekends on most weeks. Maybe some higher rental weekends were special events or weather related? Figure 7 shows the rental distribution for every hour for each day of the week. There are rentals spike at 8am and 5pm for weekdays which is most likely people working using the bike for transportation. The peaks on Saturdays and Sundays are from 12 to 4 in the afternoon.

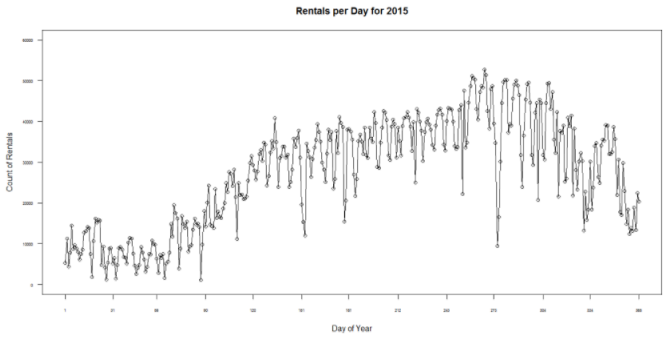

Fig. 5. Shows the rental count per day for all of 2015

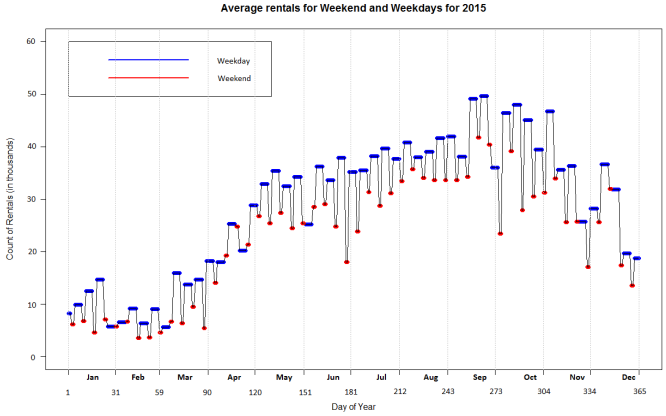

Fig. 6. Shows the difference between average rentals for weekday and weekends

Figure 8 below shows the average amount a time a bike was rented less than 60 minutes because 98.7% of data fell under this amount. Bikes were used most often at around 7 minutes.

Distances over 4 miles only account for 0.08% of the data so the plot below only shows data from 0 to 4 miles. Most people seemed to rent bikes and only ride them about 0.7 miles as seen in Figure 9. Note that this is straight line distance, so this is average is under the actual average a person actually

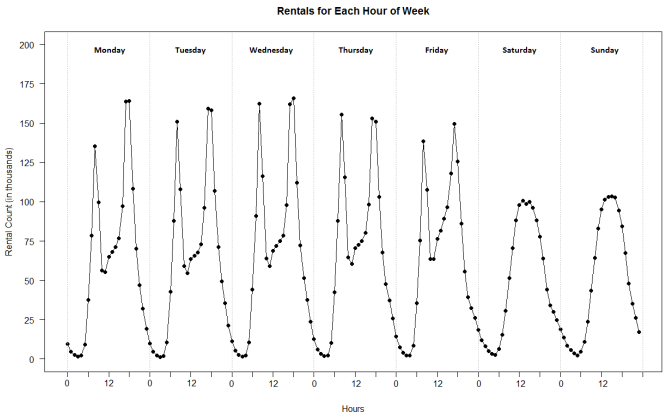

Fig. 7. Shows the count of rentals for hours for each day of week for 2015

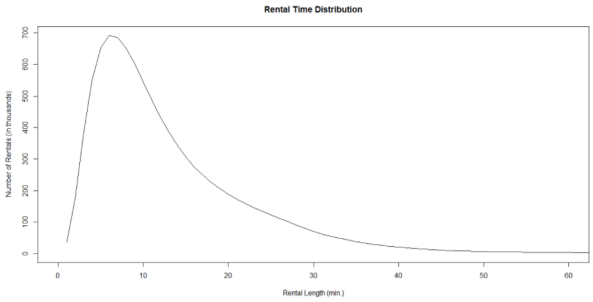

Fig. 8. The count off rentals per Day

rode the bike. This could be used to make sure each station is within 0.7 miles of each other.

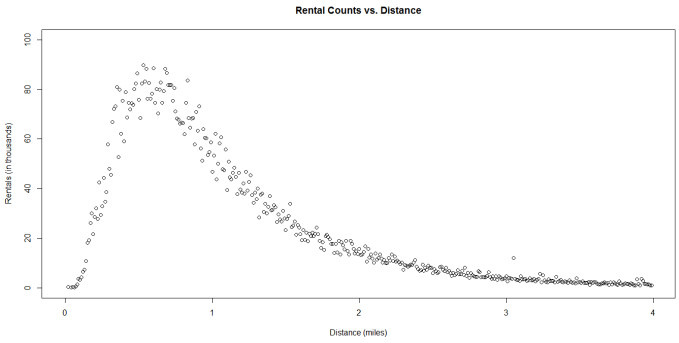

Fig. 9. The counts of rentals by distance for 2015

B. Stations Investigation

Stations that were compared with ”like” stations were split into categorical count ranges ranges as seen in Figure 10. The plot shows each stations spatial location with categorical count of rentals for the year. This plot is close to reality because it is a small geographical location and the roundness of earth doesn't affect it greatly. There were only 6 stations that had over 78,763 rentals for the year and these should definitely be

kept. All the small green dots represent under 26,662 rentals and could be candidates for reduced capacity or cancellation.

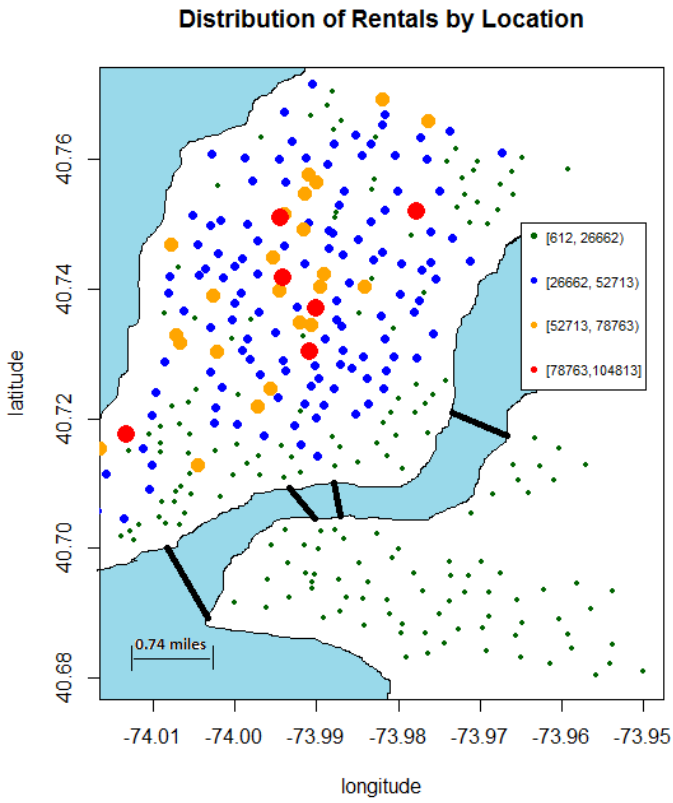

Fig. 10. Shows the relative counts for stations in the NYC area by latitude and longitude

C. Bike Investigation

99.8% of bike were used for an average distance between 0.9 miles and 1.2 miles with values outside this range being outliers and disregarded. The mean was 1.07 miles and median was 1.06 miles. The standards deviation was 0.05 miles so it appears that the bikes have all been used similarly in terms of distance. There was also a correlation, R = 0.99, between the distance a bike has traveled over its lifetime and how many times it was used which is further proof that each bike is used the same. Linear regression gave an equation of "LifetimeDistance = 1.065 * RentalCount + 5.81", where the y-intercept had an error of 2.11 and the slope an error of 1.07. The residual error was 52.62 with df = 8,475, F-Statistic = 375,800, and p-value < 0.01. Figure 11 shows the data mapped with linear fit line that fits quite well. The thickness in Figure 12 at about 1200 is because most bikes in study have been used for this amount as can be seen in Figure 12.

D. Other Results

90.1% of rentals are from subscribers. Males use differs from females as Figure 13 shows the percentages of rentals by males as compared to females for ages 16 to 60. They appear to follow similar distributions but differ in size as males clearly rent bikes more often than females.

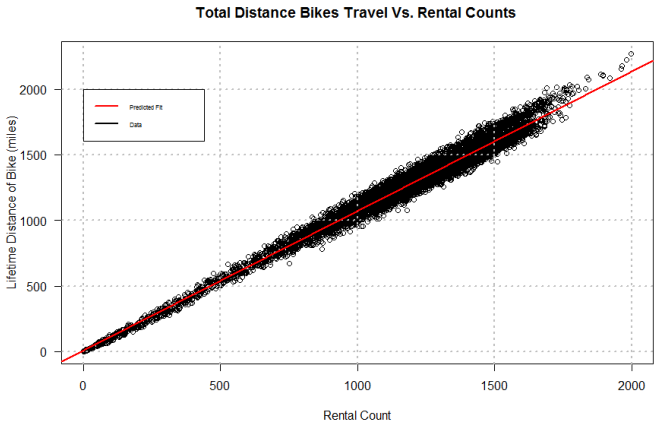

Fig. 11. Shows a bikes lifetime distance travelled by rental counts with a linear fit line overdrawn

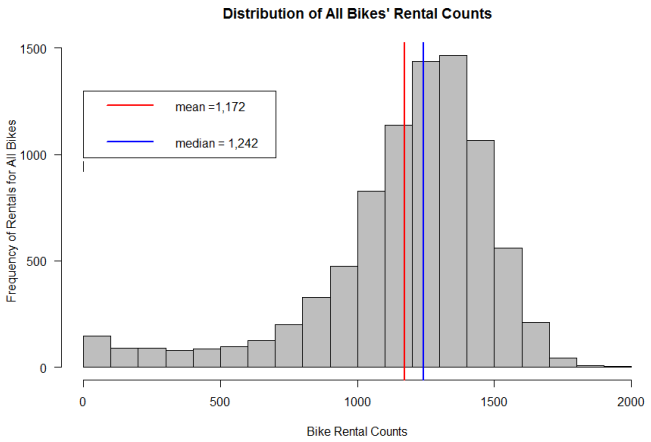

Fig. 12. Shows the event frequency of bikes being rented

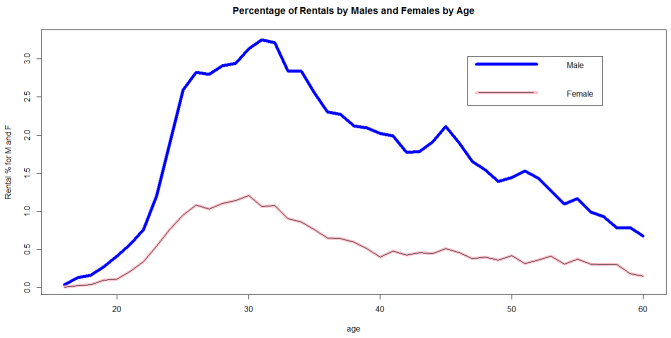

Fig. 13. Shows the rental percentage for 2015 by gender and age

E. Machine Learning

When using k = 1,324, males were predicted with 75% precision and 99.99% recall but females had 75% precision but 0.01% recall. This maybe happened because of class imbalance because the gender class label had 3 times as many males than females. Thus, another subsample was made

eliminating male examples randomly so both would have same amount. Now using, k = 296 females were predicted better but males prediction was worse. Multiple k's were attempted from k = 1 to 300 by steps of 10 and all the results were similar. The best accuracy achieved was only 0.59. The best recall for predicting male was 0.56 and 0.63 for female. Best precision of for males was 0.60 and females 0.59. These numbers were all over 50% but not enough to justify for filling in unknown values.

The association rules didn't produce an significant rules that were not obvious. An example of a rule was (0-1 miles, Male) => (0-15 minutes) with sup = 0.40, conf = 0.94 and lift = 1.31 meaning it happened more than average. Maybe the lack of input attributes hindered the results in being of any value.

V. Conclusion and Future Work

A big part of the bikeshares business happens during the week and is highly influenced by the working crowd as the hours of most rentals happens near the usual start times of Americans at 9AM and 5PM. Most bike trips roughly happen in the 9-14 minute zone so maybe they could lower the no-fee time for both subscriber to lower amount to make more money without affecting the majority of bike riders. The average distance traveled is about 0.8 miles so stations should be within this distance from eachother as bike riders generally travel this distance. By seeing the plot of stations rental counts over latitude and longitude there could be candidate stations for removal or expansion. Bikes travel the same mileage by rental but it would be nice to investigate the mileage a bike is retired so bikes can be predicted when to be retired. Future work could use other years of data starting from the end of 2013, but with his study was not possible because the HDFS setup and low processing power of the system. Maybe more attributes could be added to association rules to see if there were more promising results. More investigation to gender and user types could also add value in bike sharing analysis.

References

[1] D. Miner and A. Shook, *MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems*. "O'Reilly Media, Inc.", 2012.

[2] R. C. Taylor, "An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics," *BMC bioinformatics*, vol. 11, no. 12, p. S1, 2010.

[3] T. White, *Hadoop: The definitive guide*. " O'Reilly Media, Inc.", 2012.

[4] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," *ACM Transactions on Computer Systems (TOCS)*, vol. 26, no. 2, p. 4, 2008.

[5] L. Wang, J. Tao, R. Ranjan, H. Marten, A. Streit, J. Chen, and D. Chen, "G-hadoop: Mapreduce across distributed data centers for data-intensive computing," *Future Generation Computer Systems*, vol. 29, no. 3, pp. 739–750, 2013.

[6] M. Wasi-ur Rahman, J. Huang, J. Jose, X. Ouyang, H. Wang, N. S. Islam, H. Subramoni, C. Murthy, and D. K. Panda, "Understanding the communication characteristics in hbase: What are the fundamental bottlenecks?" in *Performance Analysis of Systems and Software (ISPASS), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 122–123.

[7] M. N. Vora, "Hadoop-hbase for large-scale data," in *Computer science and network technology (ICCSNT), 2011 international conference on*, vol. 1. IEEE, 2011, pp. 601–605.

[8] R. B. Noland, M. J. Smart, and Z. Guo, "Bikeshare trip generation in new york city," *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 164–181, 2016.

[9] R. Beecham, J. Wood, and A. Bowerman, "A visual analytics approach to understanding cycling behaviour," in *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, 2012, pp. 207–208.

[10] ——, "Studying commuting behaviours using collaborative visual analytics," *Computers, Environment and Urban Systems*, vol. 47, pp. 5–15, 2014.

[11] E. Fishman, S. Washington, N. Haworth, and A. Watson, "Factors influencing bike share membership: an analysis of melbourne and brisbane," *Transportation research part A: policy and practice*, vol. 71, pp. 17–30, 2015.

[12] B. Lantz, *Machine learning with R*. Packt Publishing Ltd, 2013.