BikeShare Data Analytics

Josina Joy, Charu Arora, Dharmam Buch and Sreenivasan Shobhana Venkatraman

***Abstract*—Bike-sharing systems are becoming increasingly ubiquitous in urban environments. They provide a low-cost, environmentally-friendly transportation alternative for urban cities and down town areas. The data available at these systems can be used to analyze the trends in demand/supply, predict demands based on various factors to assist the smooth functioning of the system. In this paper, we analyze the BikeShare data from Bay Area, visualize trends, predict demands and depict the current scenario with live Data.**

Index Terms: **BikeShare, Spark, Real Time Data, Machine Learning**

I. INTRODUCTION

THE Bay Area Bike Share system allows users to rent bicycles for short journeys between stations throughout the city. Users can be annual members or short term (1 or 3 days). The system is completely automated for users. There are 69 stations across 5 cities in the Bike Share system, with an average of 17 docks per station[4] .

II. DATASET

The dataset contains historical data for 3 years: August 2013 - August 2014, September 2014 - August 2015, September 2015 - August 2016. The dataset contains:

- Status Data: Information about the number of bikes and docks available at a given station at a given time.
- Station Data: Information about a station that includes location details and capacity
- Trip Data: Information about the start and end stations, duration, member subscription, bike details.
- Weather data: Information about the weather which includes temperature, humidity, wind speed, events for a particular zip code and date.

The Bikeshare community also provides us the Live feed of data for 2 cities: San Francisco and San Jose[1]. The feed contains information about stations in the two cities – location details, capacity, availability of bikes and docks, operational status.

III. DATA PREPROCESSING

The datasets are 3.74 GB in size. The data was preprocessed to aggregate relevant data for analysis and operations. Spark was utilized for the same. The RDDs were created from the datasets. The data was filtered to discard entries with no data for relevant attributes. Join operations were carried out to associate Weather Data to Status Data, Status Data to Station Data, Trip Data to Weather Data. Date Time field formats were inconsistent across datasets which had to be standardized. The results were loaded in DataFrames and then stored as CSVs.

IV. APPROACH

A. Historical Data Analysis

The archived data is analyzed to see the trends in the bikeshare system. This was done using pyspark. Station data and status data were joined to obtain the average percentage utilization of each station from 2014 to 2016 by averaging the number of docks available for each station throughout the year. The busiest hour when bikes were not available at stations was calculated by filtering the 'bikes_available' attribute from the status data with the value '0'.

Apart from this, the number of bikes rented per city was calculated by combining the status data and the station data. Here, the date was used as the key to join the two datasets.

The average trips per day for a given location were linked to the weather prediction for that day to analyze the relationship between the two. For this, the weather data was combined with the Status data to obtain the number of available docks and weather conditions.

B. Building Predictive Models

From the Data Analysis we concluded that there is a correlation between the weather and the

number of bikes rented in a day. Since the bay area has micro-climates we build models to predicted the demand for bikes depending on the weather conditions for a day for a city. The datasets used for building the predictive models were Station data, Status data and Weather data. Station data and Status data were combined to get information about the average demand for a city on a particular day. This previously combined data was joined with weather data to get the weather information for a particular day. The Machine Learning algorithms used to build models are – Decision Trees, Random Forests, Gradient Boosting and Naïve Bayes. The required data for the model input was prepared using Spark programs written in PySpark and Scala.[2] Spark ML libraries from the mllib suite were used to build the models .

C. *Real-Time Data Analysis*

The realtime data feed was used to depict the current scenario of the bikeshare system. The data from the feed is in JSON format. It is streamed using Kafka. A dedicated topic channel 'BDProject' was created to stream the data. The producer sends the data on this channel one record at a time. On the consumer side an Elastic Search object is created which is used to create an index for the incoming data – 'bikedata-index'. As each stream arrives, the document is inserted into the index[4] . The data is visualized in Kibana[5] . Saved searches are creating on records to aggregate records belonging to the two cities – 'San Francisco' and 'San Jose'.

V. ANALYSIS AND RESULTS

A. *Visualizations*

After analyzing the historical data[3] , some of the major trends noticed were as following:

- The bikes are mostly rented by annual members, basically subscribers, for the purpose of travel to and from work. In 2016, approximately 21 million bikes were rented at 8 am and 19 million at 5 pm, with an average being around 6 million at a particular hour of the day.

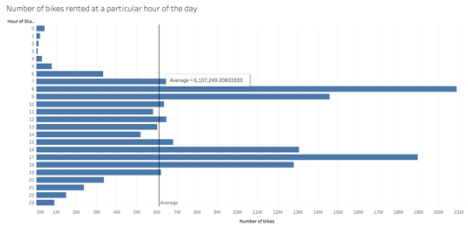


Figure 1: Bar chart 1

Throughout the years, each station was utilized a minimum of 33.33%. This shows that none of the stations were under-utilized. The station 'Cyril Magnin St at Ellis St' was the busiest stations in 2016 with the percentage of docks used being 82.85%. Percentage utilization for some of the stations such as 'Townsend at 7th', 'Santa Clara at Almaden', 'San Antonia Shopping Center' has been consistent throughout the three years. Also, the performance of the station 'California Ave Caltrain Station' has decreased from 40% to 33.33% and the performance of the station 'SJSU 4th at San Carlos' has increased from 36.84% to 57.89%, from 2014 to 2016, with the number of docks being consistent throughout the three years for both the stations.

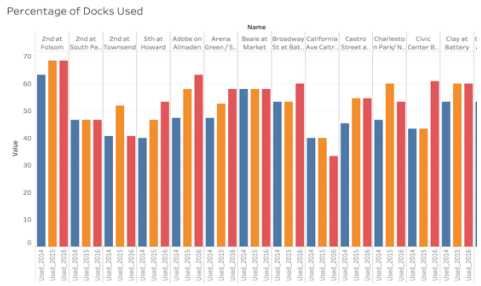


Figure 2: Bar chart 2

- Bikes that were rented by customers, for 1-3 days, were mostly rented on weekends and on holidays for instance, while the average bikes rented for 2015 was 116. The number of bikes rented shoot up on holidays, like to 334 on the Valentine's day, and 343 on the Independence Day.
- Trend for Bikes rented by subscribers were mostly seen from Monday to Friday.

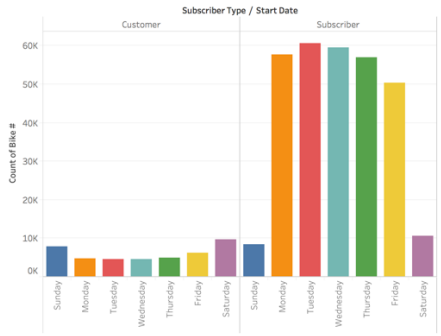
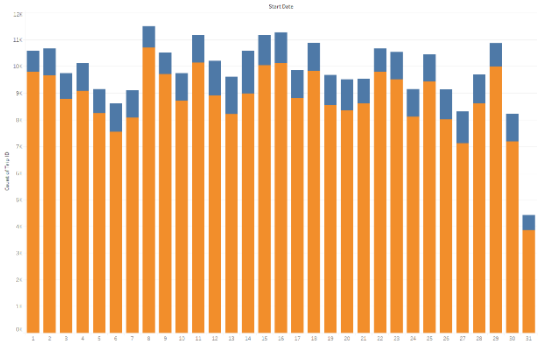


Figure 3: Lines chart 1

- As expected, there is no correlation between bikes rented and the day of the month. Except on 31th which only 7 months have, the rides are more or less equally distributed on the other days.



- Figure 4(a): Lines chart 2

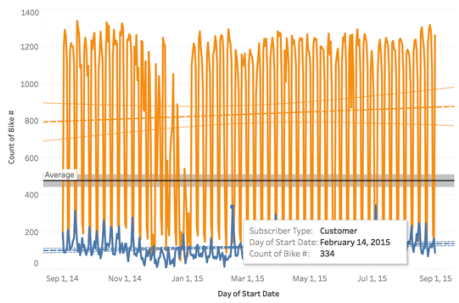


Figure 4(b): Lines chart 3

- The following figure shows the color coded distribution of stations according to usage and also the station to station ride usage heatmap.

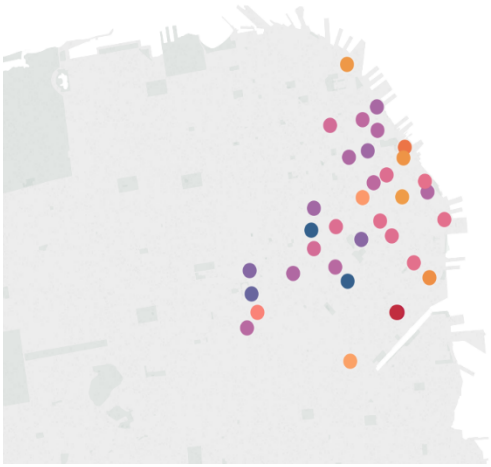


Figure 5: Map

Figure 6: Heat Map

- Correlating the number of trips with weather, it was noticed that bikes rented had a direct relationship with the events such as rain, temperature and dew point of the location. However, no relation was seen with the wind speed, humidity and sea level of that location.

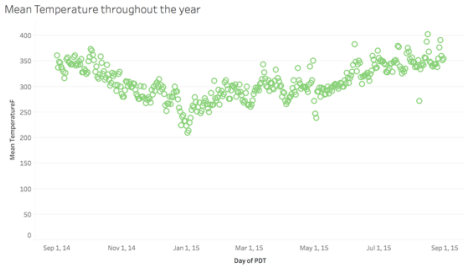


Figure 7(a): Circle view chart 1

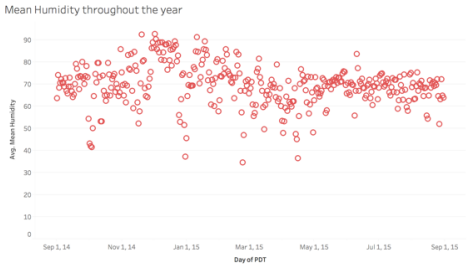


Figure 7(b): Circle view chart 2

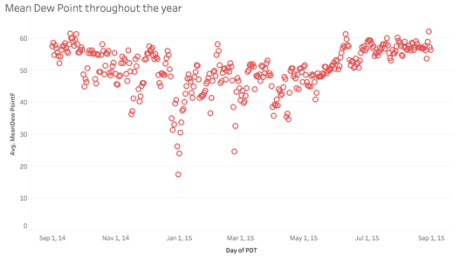


Figure 7(c): Circle view chart 3

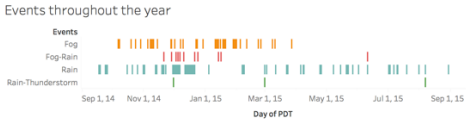


Figure 7(d): Table chart

B. *Predictive Models*

The accuracy of the predictive models is as given below:

Model	Accuracy
Decision Tree	96.96
Random Forest	97.3
Gradient Boosting	80.51

Decision Tree and Random Forests gave the highest accuracy. The output variables are categorical variables with a size of 80.

C. *Live Data Analysis*

A dashboard was created in Kibana for visualizing the Live Data.

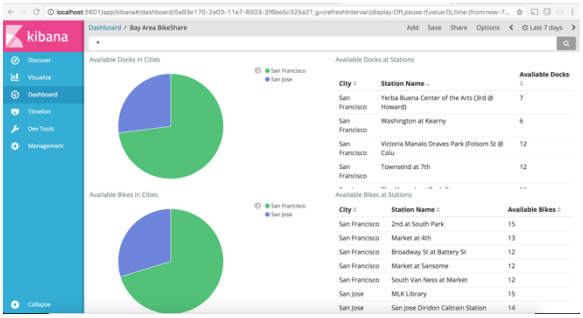


Figure 8: Dashboard – Part1

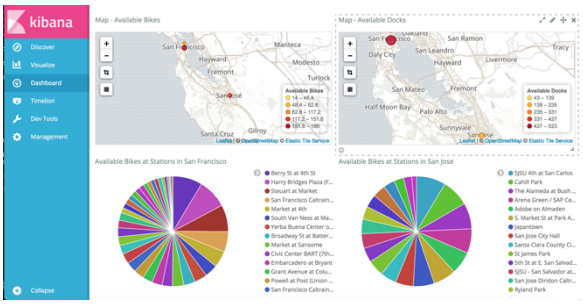


Figure 9: Dashboard – Part2

The various visualizations are explained as follows:

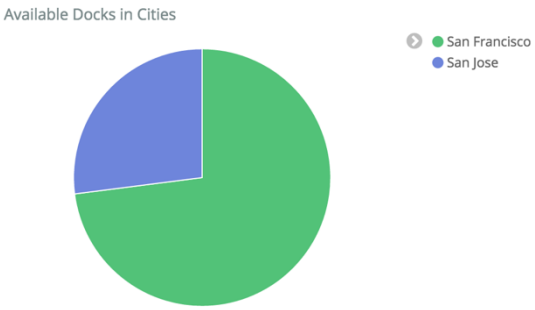


Figure 10: Pie-Chart 1

This depicts the total number of docks currently available in the two cities of San Francisco and San Jose.

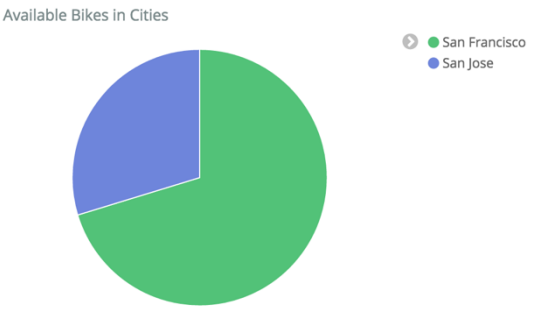


Figure 11: Pie-Chart 2

This depicts the total number of bikes currently available in the two cities of San Francisco and San Jose.

Available Docks at Stations

City	Station Name	Avaliable Docks
San Francisco	Yerba Buena Center of the Arts (3rd @ Howard)	7
San Francisco	Washington at Kearny	6
San Francisco	Victoria Manalo Draves Park (Folsom St @ Colu	12
San Francisco	Townsend at 7th	12

Figure 12: Table 1

The above table gives the number of available docks at each station and the associated city.

Available Bikes at Stations

City	Station Name	Available Bikes
San Francisco	2nd at South Park	15
San Francisco	Market at 4th	13
San Francisco	Broadway St at Battery St	12
San Francisco	Market at Sansome	12
San Francisco	South Van Ness at Market	12
San Jose	MLK Library	15
San Jose	San Jose Diridon Caltrain Station	14

Figure 13: Table 2

The above table gives the number of available bikes at each station and the associated city.

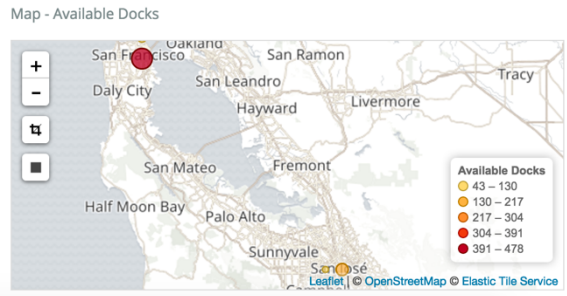


Figure 14: Map 1

A geographical representation of the availability of docks at different locations.

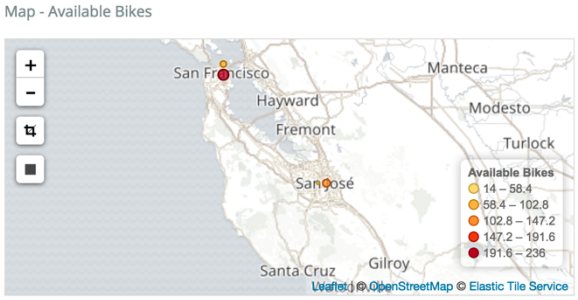


Figure 15: Map 2

A geographical representation of the availability of bikes at different locations.

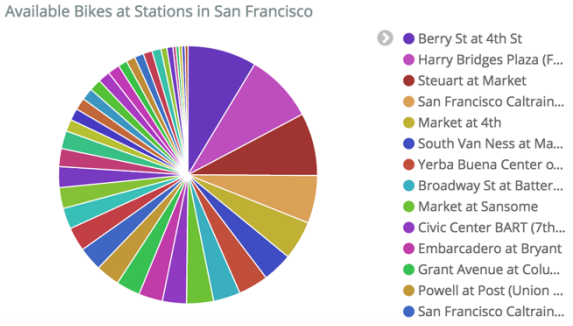


Figure 16: Pie-Chart 3

A pie chart depicting of the availability of bikes at different stations in San Francisco.

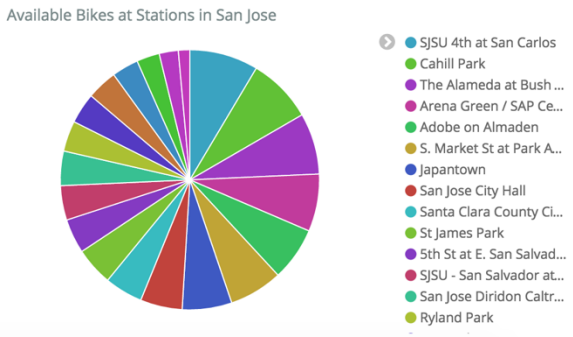


Figure 17: Pie-Chart 4

A pie chart depicting of the availability of bikes at different stations in San Jose.

V. CONCLUSION

In this paper, we did an in-depth analysis of the BikeShare system in the bay area. From the historical data we observed trends like the major customers are subscribers who use bikes to and from workplaces. The customer renting trend was mainly observed on holidays. We were able to find correlations between the weather and demand, hence build machine learning models to predict the same. The Live data analysis gave us a picture of the demand-supply state of the system at any instance.

VI. FUTURE WORK

An efficient algorithm to match demand and supply of bikes among the stations. Enhancing the business model by finding answers to critical questions such as: where a new station should be situated, which station should be shut down, where should the capacity be increased or decreased.

VII. REFERENCES

1. http://www.bayareabikeshare.com/open-data
2. SparkMLlib http://spark.apache.org/mllib/
3. Tableau: https://www.tableau.com/
4. Data Analysis and Optimization for (Citi)Bike Sharing. By: Eoin O'Mahony, David B. Shmoys
5. Elastic Search: https://www.elastic.co/
6. Kibana:https://www.elastic.co/guide/en/kibana/current/visualize.html