Spatiotemporal and Demographic Analysis of Ford GoBike© Data in the San Francisco Bay Area

Shayan Monshizadeh
Princeton University
shayanm@princeton.edu

Kim Sha
Princeton University
ksha@princeton.edu

Steven Takeshita
Princeton University
st14@princeton.edu

Sami Belkadi
Princeton University
sbelkadi@princeton.edu

Abstract

Ford GoBike is a bike-share service operating in the San Francisco Bay Area. They have published anonymized trip data from June 2017 to present. The goal of this paper is to cluster the trip data into prototypical trips and users to gain insights that can guide business decisions. We also predict the age of a user, given all of the ride details. Probabilistic imputation is used to fill in missing gender and birth year. Feature engineering on the start and end time are used to enrich the data set. K-means clustering and Latent Dirichlet Allocation are utilized to cluster the trips. Ordinary least squares, LASSO, and ridge regression are used to predict the user's age for a given trip. It is shown that Latent Dirichlet Allocation discovers clusters with easily interpretable features while K-means clustering is used to visualize geographical movement of users. Ridge regression best predicts the age of a rider for a trip in terms of mean squared error.

1 Introduction

Ford GoBike operates in San Francisco, East Bay, and San Jose. In order to improve their service, Ford GoBike records information about each trip and publishes it on their website on a monthly basis. Through analysis of the datasets, a variety of trends and visualizations can be leveraged to gather insights into how the service is being used. By understanding how people are using the service, Ford GoBike can gain insight into how to advertise to a target demographic, select ideal locations for new stations and anticipate demand depending on the time of day and the weather. Each of these insights improves the company's revenue stream. Understanding the latent structure in the bike-share data can also help Ford GoBike make their service easier to use.

This paper outlines methods that can be used to discover latent structure and trip patterns in the dataset. In Section 1.1, the main ideas from related studies are introduced, in order to place this study in context. A description of the data follows in Section 1.2. The methods used in this analysis are described in depth in Section 2, including data preprocessing, feature engineering, clustering and regression methods. Section 3 outlines ideal hyperparameters, results and analysis for each method. A discussion of these results is included in Section 4.

1.1 Related Work

Past studies have done similar analyses on San Francisco bike-share data. In one study, Venketeswaran et. al showed that logistic regression was a better predictor of user type (subscriber or customer) than SVMs [5]. Specifically, they used trip duration and the presence of holidays to capture customers, and found that subscribers travelled largely on weekdays.

Another study attempted to cluster bike stations, as opposed to bike trips, in France under the Velib service [2]. Their algorithm leveraged the Poisson distribution to group eight sets of station. The

resolved clusters represented features such as docks near railway stations or docks near parks. A similar study on the same bike-share service in Paris instead used K-means and hierarchical clustering [3]. They found clusters of stations near residential and business areas based on the number of available bikes at the station. The resulting availability of bikes at these stations were intuitive, for example an increase in bikes near business areas at the beginning of work days and a decrease at the end of the day. This prior research clusters the stations rather than trips themselves. We take a different approach by exploring the inherent structure of trip data.

1.2 Data Description

Ford GoBike publishes monthly data logging all rides and users' demographics from 2017 to the present [1]. Each trip is anonymized and parsed into rows that include:

- **Trip Information:** Trip Duration (in seconds); Start/End Date and Time; Bike ID.
- **Station Information:** Start/End Station IDs, Names, and Coordinates.
- **User Information:** Type (Subscriber / Customer); Year of Birth; Gender.

The compiled data set has 12 feature columns and 1,338,864 sample rows. It corresponds to data logged between 06/28/2017 and 06/30/2018. This data was merged with a data set [4] containing hourly meteorological data for San Francisco. The data consists of temperatures and descriptions, such as "sky is clear", "mist", "broken clouds", "light rain", and "haze".

2 Methods

2.1 Data Processing

Several anomalies were detected in the data. For example, some users stated they were born in 1886, making them 131 years old in 2017. In response, we capped the maximum user age at 100, removing rows with users born before 1918 (0.7% of total trips). Moreover, there were extreme outliers with respect to trip duration, which ranged from as little as 1 minute to 23 hours. Thus, we restricted our analysis to trips lasting between 3 minutes and 3 hours. Several anomalies were also encountered in the station-specific data. First, 2061 of the trips contained null values for start and end station names and ID. Those coordinates were also far outside the Bay Area. Since these rows were $< 0.15\%$ of total trips, they were removed. Finally, duplicates for slightly different station names (e.g. 'Shattuck Ave at 55th Ave' vs. 'Shattuck Ave at 55th St') were replaced with unique names.

2.2 Feature Engineering

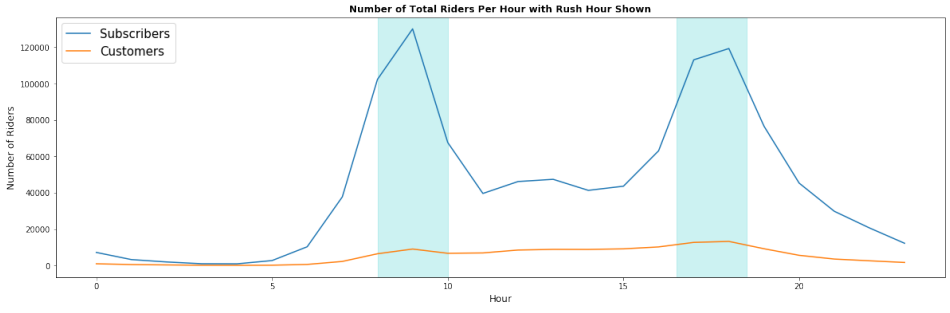


Figure 1: Daily use of Subscribers and Customers during a weekday. Highlighted regions represent peak rush hours, between 8-10am and 5-7pm.

The original dataset has 12 features. We expanded the feature space to extract more information about the date and time of rides, which we expect to demonstrate trends across seasons and time of day. Figure 2 and Figure 3 show the distribution and variation of trips over the course of a week and year. The data was sliced several ways: by time-stamp in 24-hr time, by weekday or weekend, by season, and by the presence of holidays. With the expectation that weather would have a large impact on trips, an hourly meteorological dataset for San Francisco was merged with the

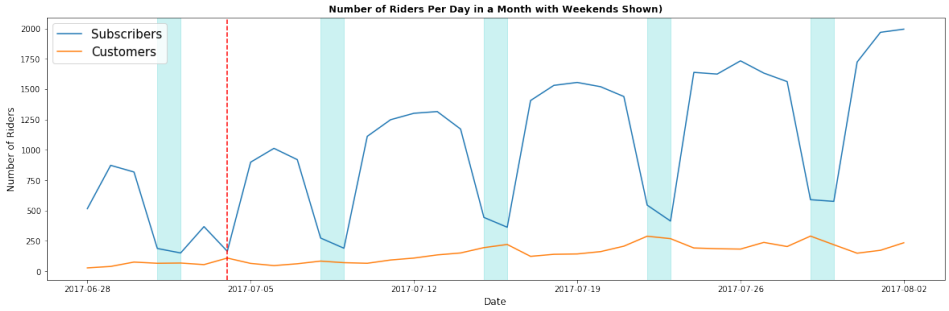


Figure 2: Bike use of Subscribers and Customers on a weekly basis. The highlighted regions represent weekends and the dashed red line is July 4th.

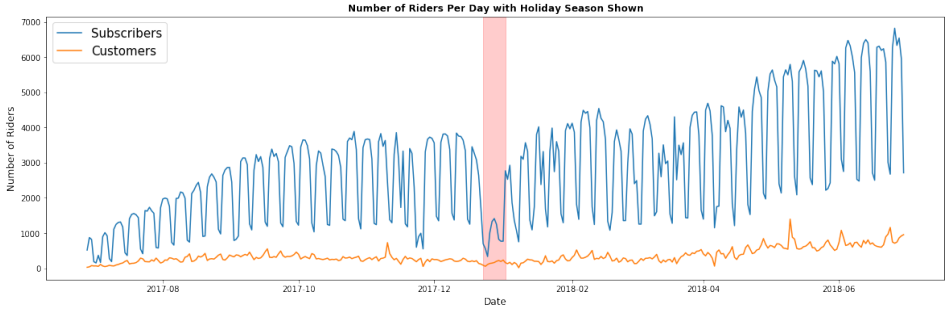


Figure 3: Bike use of Subscribers and Customers throughout a year. The highlighted region represents winter break.

bike-share data to include the temperature and a description of the weather for each trip. Through feature engineering, we increased the feature space of the data from 12 to 20 features. Through the statistical visualization presented above, we can see that the number of subscribers is growing at a quicker rate than customers. This means that commuters in San Francisco are likely the target customers to whom Ford GoBike should tailor their service.

2.3 Treatment of Categorical Variables and Standardization

A number of the variables in the dataset are categorical variables (start and end station name, member gender, user type, weather description etc.). To allow the models to interpret these categorical variables, we create new dummy/indicator variables. This increases the number of columns in the dataset from 20 to 841. Moreover, features must be compared and used for regression on a similar scale. Thus, standardization of the dataset is necessary by computing the z-score for each value using StandardScalar.

2.4 Clustering

We cluster the data into prototypical trips using the following methods:

- *K-means clustering:* measures similarities between sample rows using the Euclidean distance metric to generate representative trip profiles. The number of clusters is determined by the silhouette method.
- *Latent Dirichlet Allocation (LDA):* more complex dirichlet distribution to model the prototypical rides and determine important features for each cluster. The optimal number of components was selected by finding the setting in which log-likelihood is maximized.

Figure 4 outlines the series of steps taken to tune hyperparameters for the k-means and LDA algorithms. Parameter tuning for k-means initially resolved 2 clusters for a silhouette score of 0.82.

A closer look at cluster membership reveals that the data is clustered unevenly into trips that take around 11 minutes ($> 99\%$ of the data) and trips that extend into the range of 10 hours ($< 1\%$ of the data). This latter cluster likely corresponds to users that neglected to return the bikes, resulting in outlier effects that skew the k-means algorithm. For the purposes of this study, these outliers are eschewed in favor of the next highest silhouette scores (0.57) corresponding to 7 clusters.

2.5 Regression

We split the data using an 80/20 split for train and test data and measured accuracy using mean squared error. We evaluated the following regressors:

- *Ordinary Least Squares:* (OLS). Baseline method for regression.
- *Ridge Regression:* with α selected using cross validation. $L2$ regularization to avoid over fitting.
- *LASSO:* with α selected using cross validation. $L1$ regularization to avoid over fitting and encourage sparsity.

3 Results

3.1 Hyperparameter Tuning

Since the entire dataset was too large to run cross-validation on, a representative sample was taken from the dataset (10% of rows) and used for testing in tuning LDA's hyperparameter. Results are shown in Table 1 with curves shown in Figure 4.

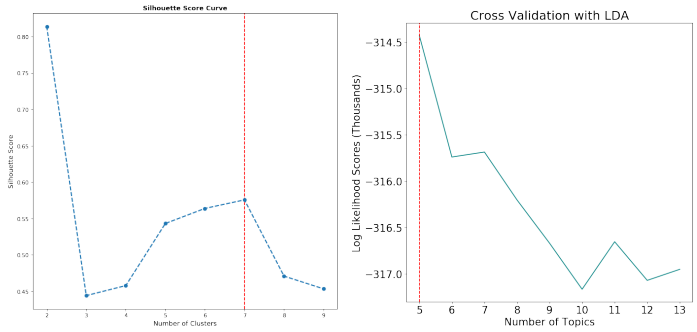


Figure 4: K-means parameter tuning using silhouette scores (Left) and LDA tuning using the log-likelihood metric (Right). 7 clusters are chosen for k-means; 5 topics are chosen for LDA.

Method	Number of Components/Clusters	Scoring Method: Value
K-means	7	Silhouette Score: .57
LDA	5	Log Likelihood: -314000, Perplexity: 113.5

Table 1: Clustering methods with scores of optimal number of components/clusters.

3.2 Clustering Trips with Augmented Weather Data

Integrating the SF weather data, we see interesting distributions of features across the prototypical trips (Figure 5). Weather description and temperature are well resolved with clear distributions of the top features for each. However, duration is not as well resolved with many trips lasting similar lengths, which is consistent with the fact that most trips are relatively the same length. Latent Trip 2 is associated with smoke, colder temperatures, and shorter trips, which are all unfavorable conditions for biking. Intuitively, this makes sense as, due to the California forest fires from 2 years ago, there were days when biking would be harmful to health, resulting in lower duration rides. On the other hand, Latent Trip 3, the prototypical trip most associated with rain does not correspond with shorter trips, but rather average length trips. Although not intuitive, this could result from the fact that even when raining, a bike is a user's only means of transportation, and therefore they will ride regardless

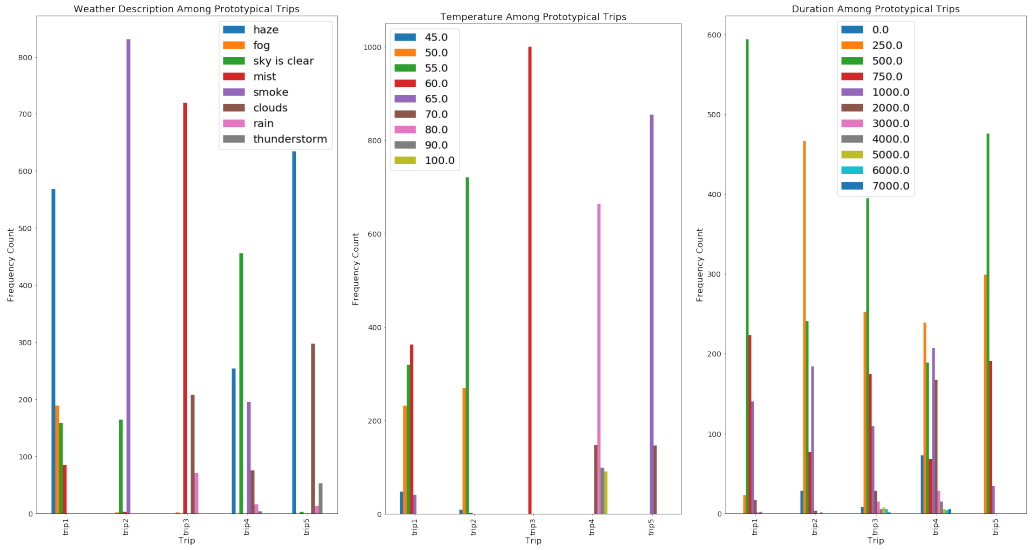


Figure 5: Prototypical trips with their distributions of weather descriptions, temperature and duration of trip

Rank	Trip 1	Trip 2	Trip 3	Trip 4	Trip 5
1	Subscriber	Female	Not Holiday	Weekday	Not Holiday
2	Weekday	Weekday	Start Time 5 PM	Not Holiday	Male
3	Male	Customer	End Time 5 PM	Female	February
4	Winter	Subscriber	Male	Subscriber	Spring
5	Summer	Male	Subscriber	Male	Subscriber
6	Spring	Weekend	Female	Birth Year 1991	Weekday
7	Female	Start Time 2 PM	Start Time 4 PM	Start Time 9 AM	Female
8	Start Time 9 AM	End Station Embarcadero At Sansome	Customer	End Time 9AM	Birth Year 1980
9	End Time 9 AM	Start Time 1 PM	End Station Caltrain Station	Start Time 6 PM	Birth year 1990
10	Fall	Birth Year 1969	Birth Year 1985	Spring	End Time 3

Table 2: Most Important Features of Each Latent Trip

of weather. Moreover, rain is not as harmful as smoke to a rider and therefore will not have as strong of a correlation with shorter rides.

3.3 Interpreting LDA Clusters

Table 2 contains the prototypical trips found by LDA and their top distinct features. LDA found very interesting and intuitive clusters of trips in the data. For example, the first cluster can be labelled as "morning commuters". This can be seen in the top features of Latent Trip 1, which describe a user as a subscriber who rides mostly on weekdays, in all seasons, around 9 AM.

In contrast, Latent Trip 3 users can be described as "evening commuters" as shown by the fact that they are mostly riding around 4-5 PM and are mostly subscribers as well. The most interesting trait of this cluster is the fact that many users end their trip at the San Francisco Caltrain Station, which is the gateway to the lower Bay Area from the city. As shown in Figure 6, the majority of these rides are coming from the Division at Potrero Station and the Financial District Station, both of which are central business hubs. Thus, Latent Trip 3 can be described as users who commute to the lower Bay Area via Financial District and Mission.

Latent Trip 2 is the only cluster that is non-commuter focused. The top traits of this trip are customers (one time users) who travel on the weekend around 1-2 PM. This is the only latent trip that occurs on the weekends as opposed to rush hour times. Latent Trip 2 users also appear to be older, with users typically around 60 years old. The most interesting spatial information of our LDA clustering can be found in this cluster, with an end station of The Embarcadero at Sansome St being a very important trait. This station is in fact the closest point to Pier 39, San Francisco's most tourist-heavy destination, and thus it corresponds with the "weekend tourist" attributes of Latent Trip 2 users. All trips to this station between 1-2 PM can be seen in Figure 6. Most of the trips come from the Mission and SoMa districts of San Francisco, as well as some from the Golden Gate Park region, which are all very tourist-heavy destinations as well.

The fact that LDA clustering matches up very well with routes between stations is of value to the city and advertisers looking to determine which streets are travelled by which demographic of people. Targeted ads and boutique brands along these routes can bring in greater revenue and show potential for further investigation through the use of more advanced machine learning techniques.

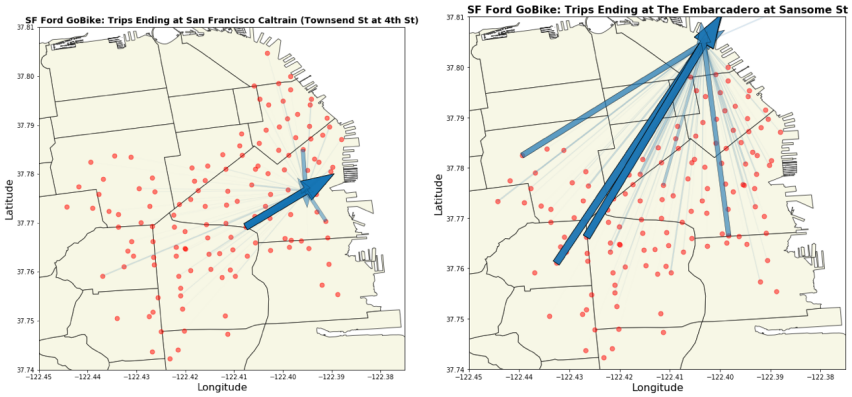
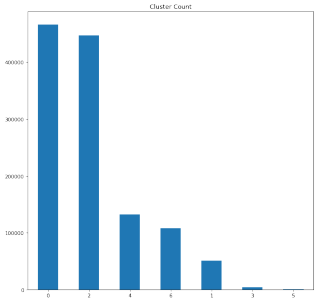


Figure 6: All rides ending at San Francisco Caltrain Station between 4-6 PM on weekdays (Left), along with all rides ending at The Embarcadero at Sansome St at 1-3 PM (Right).

3.4 Analysis of K-means

The cluster centroids generated by the k-means algorithm delineate representative trips that are conducted across the San Francisco Bay Area. Figure 7 outlines the distribution of data across the 7 clusters. Note that most of the data is grouped into clusters 0 and 2, representing trips that last around 11.5 minutes starting around 5PM and 9AM, respectively.



Cluster	Duration (s)	Hour of Day
0	699	5PM
1	652	2PM
2	685	9AM
3	16162	12PM
4	662	6PM
5	59383	4PM
6	628	9AM

Figure 7: Distribution of data samples across k-means clusters (Left), with the corresponding trip duration and hour (Right). The 2 highlighted rows represent the most dense clusters.

One may expect clusters 0 and 2 to be representative of trips conducted during evening and morning rush hours, respectively. To test this, the corresponding start/end coordinates of these two clusters are overlaid on top of a map illustrating the frequency with which unique trips occur during rush hours. As seen in Figure 8, the direction of the red arrows roughly match those of the most pronounced

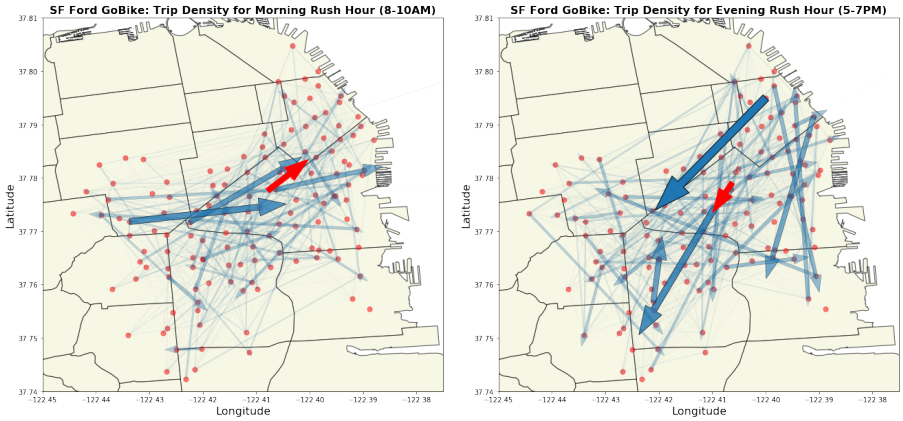


Figure 8: Trip density plots corresponding to morning (Left) and evening (Right) rush hours. The width and transparency of each blue arrow is scaled by the relative frequency with which that *unique* trip occurs. Red arrows are the k-means cluster centroids.

blue arrows (thicker and more opaque). It makes sense that the highest frequency trips drive the ultimate direction of the clustered centroids. In addition, the short distance of these centroid trips are indicative of a multitude of distinct underlying short trips that are not resolved in Figure 8. Thus, one can say that these clusters are representative of trips conducted during rush hours.

3.5 Predicting Age

Three regressors with reasonable performance to predict age, given the ride data, were identified. See Table 3 for the regressors and their associated hyperparameter and mean squared error.

Regressor	**Hyperparameter**	**Mean Squared Error**
Ordinary Least Squares	N/A	103.4772112
LASSO	$\alpha = .001$	103.4770736
Ridge	$\alpha = 0.1$	103.4772148

Table 3: Performance of regressors to predict user age. The best performance among the regressors is highlighted.

3.5.1 Important Features to Predict Age

As LASSO is the best regressor, we investigated the coefficients to understand the relative importance of the features. See Table 4 for the top five features. The 2nd and 5th most important features, though seemingly unrelated, can be correlated and mutually explained. If the user is a customer (2nd feature), they are less likely to be a commuter going to work because commuters purchase subscriber membership to save money on rides. The customer is more likely to be riding the bikes for leisure rather than as a means for getting to a destination. Moreover, we know from the original data analysis that most rides occur on weekdays, coinciding with workdays. If a trip occurs on a weekday (5th feature), the trip is most likely to be a commuter heading to work. Knowing that a rider is a commuter indicates that the rider will most likely be older as presumably children and tourists account for a much smaller portion of the rides. Therefore, whether or not a rider is a commuter (correlated with the 2nd and 5th feature) will be highly correlated with predicting age.

Feature	**End Station Longitude**	**User Type Customer**	**Is Female**	**Start Station Longitude**	**Is Weekday**
Coefficient	1.175	0.980	0.596	-0.430	-0.341

Table 4: Top five important features for LASSO regression.

3.5.2 Severely Mispredicted Ages

Though the average error for prediction was around 10 years off, LASSO severely mispredicted some trips. See Table 5 for two users who were predicted to be much older and much younger. Comparing the two trips, we can see why trip 864125's rider was predicted to be older than trip 912672's rider. All of the important features are roughly the same except when the ride occurred. Trip 864125 occurred on a weekday whereas 912672 was on a weekend. From the above analysis, we know that due to occurring on a weekday and being a subscriber, LASSO predicts trip 864125 to be an older rider due to its assumption that the rider is more likely to be a commuter. By the inverse reasoning, trip 912672 appears to be slightly younger (not travelling on a weekday) even though he is a subscriber. However, by assuming these two features to be indicative of a commuter, LASSO overlooks other instances in which a rider will be a subscriber and travelling on a weekday or weekend. For trip 864125, the rider could be a student travelling to school, since schooldays coincide with workdays. Because commuters account for a major portion of weekday subscribers, LASSO automatically assumes this rider to be older. On the other end, the older rider in trip 912672 could be an elderly man on a weekend bike ride and therefore cannot be lumped in with the non-commuters (tourists) travelling on the weekends. Moreover, we know that the average age of the bike riders is generally much younger than this rider, who is 81 years old, and therefore LASSO will inherently be skewed to predict a younger age by not accounting for these outliers.

Trip ID	Real Age	Predicted Age	End Station Longitude	User Type Customer	Is Female	Start Station Longitude	Is Weekday
864125	2000	1979	-122.444	False	False	-122.417	True
912672	1936	1981	-122.399	False	False	-122.394	False

Table 5: Two trips that were misclassified by LASSO.

4 Discussion

As seen in the results, a variety of conclusions can be drawn from the data. These are capable of guiding decision-making for Ford GoBike, in order to improve the company's growth. Prototypical trip clusters were discerned using LDA such that each cluster shows groups of weather, temperature and trip duration that frequently occur together. These findings confirmed our hypothesis that smoky and cold weather results in shorter trips, and warm and clear weather results in longer trips. LDA also found interesting clusters of prototypical users, providing information on specific demographics of riders and the direction in which they are headed. These clusters include morning commuters, tourists and evening commuters. Since each of these prototypical groups have their own most travelled paths, targeted advertisements could be deduced from the clusters. For example, evening commuters in Latent Trip 3 tend to bike to the Caltrain Station to travel by train to the Lower Bay Area. With this information, Caltrain Ticket Machines could be placed along these routes to make it easier for riders to buy train tickets.

With regards to important features, the user type (customer or subscriber) appears to be an important feature in determining a variety of clusters. Using this feature allowed us to perform simple analysis to understand how the two groups are using the service. The start and end time of each trip are also important, as they can be used to generate other features such as the weekday, weekend and holiday variables. These variables were useful for both LDA and regression analysis, as they helped produce discernible clusters.

Using k-means, we were able to determine the directions of riders going into and out of work areas during morning and evening rush hours. This shows that both LDA and k-means were successful in finding distinct clusters that emphasized different features of the dataset. In terms of predicting age based on trip data, we were moderately successful using LASSO to predict age with a mean squared error of 103. We surmise that the difficulty in prediction arises due to the diversity of the San Francisco riders and the sheer number of different groups of people who use these bikes all over the city on a given date.

Given more time and resources on the project, we would integrate all of the data up to the present (approximately 2 million additional trips) and run our algorithms again to ensure that any trend we saw continues with more data. The computational limits of our computers restricted how much data we could process, but with more computing power and time we would incorporate more features in order to discover more latent structures.

Acknowledgments

Thank you to Matt Myers, Jonathan Lu, Diana Cai, Archit Verma and Barbara Engelhardt of Princetons Department of Computer Science for their support and guidance throughout this project.

References

[1] Ford Corporation. System data. *Ford GoBike*, 2017.

[2] Etienne Cme and Latifa Oukhellou. Model-based count series clustering for bike sharing system usage mining: A case study with the vlib system of paris. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5, 10 2014.

[3] Yunlong Feng, Roberta Costa Affonso, and Marc Zolghadri. Analysis of bike sharing system by clustering: the vlib case. *IFAC-PapersOnLine*, 50(1):12422 – 12427, 2017. 20th IFAC World Congress.

[4] Selfish Gene. Historical hourly weather data 2012-2017, Dec 2017.

[5] Aishwarya Venketeswaran, Brenton Hsu, Sucheta Banerjee, and Wiseley Wu. Analysis of the bayarea bike share data. 2018.