Analytics and Application 2022/23 Group 14

Seminar Paper

Author 1: Jan Richter (Student ID: 7399138)
Author 2: Christian Dalsgaard (Student ID: 7405550)
Author 3: Leonardo Gorelli (Student ID: 7405579)
Author 4: Patrick Beeck (Student ID: 7393670)
Author 5: Oliver Sauren (Student ID: 739374)
Supervisor: Univ.-Prof. Dr. Wolfgang Ketter
Co-Supervisor: Nastaran Naseri

Department of Information Systems for Sustainable Society
Faculty of Management, Economics and Social Sciences
University of Cologne

February 1, 2023

Eidesstattliche Versicherung

Hiermit versicheren wir an Eides statt, dass wir die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist uns bekannt, namentlich die Strafandrohung gemäß § 156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß § 161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.

Team 14

Köln, den 01.02.2023

1 Information

Link to GitHub repository: `https://github.com/OliverSauren/AA_22_Team_14`

Task	Patrick	Jan	Christian	Leonardo	Oliver
Data Cleaning	x				x
Weather Cleaning		x	x		
Research further data		x	x	x	
Temporal			x		x
Geo Demand	x	x			
KPIs	x			x	
Clustering		x		x	x
Prediction	x		x		
Report	x	x	x	x	x

Executive Summary

The goal of the project is to perform data analysis for bike rental in Boston in 2017 to find useful information that will help the bike rental company to optimize its business activities. The data set used contains information about the time, duration as well as start and end station of each individual bike rental. This information was used to analyze temporal and geographical demand patterns. Furthermore, the analysis considers three key performance indicators and a clustering part as well as a predictive part. The highest demand for bicycles is at 8 a.m. and 5 p.m. It can be concluded that commuters in particular use bicycles. The total amount of rides decrease from November to March. This shows a higher demand in months with a higher average temperature. Additionally, the trip duration is higher on the weekend, which could have been caused by day trips. The geographical demand analysis shows that MIT at Mass Ave / Amherst St. is the most popular start and end station. The key performance indicators are the average utilization rate, shares of customers and subscribers, and the ratio of customers to subscribers. The average utilization rate in Boston in 2017 is 16:48 min. The highest average utilization was recorded in April at 24.46 min, while the lowest average utilization was recorded in February 2017 at only 11:36 min. The general share of subscribers is 84,1% and of customers 15,9%. The data analysis shows that especially in summer months the ratio of customers to subscribers per ride is higher than in other months. For the cluster analysis, we used two clusters, representing the user type of either subscriber or customer. The predictive model predicts the bicycle demand regarding time and weather. The results could be used to increase bike offers on key days and decrease the offer for repair and maintenance.

Contents

1 Information

2 Detailed Report **1**

2.1 Problem description: Business and data mining goal 1

2.2 Data description 1

2.3 Data preparation 1

3 Data analysis **2**

3.1 Temporal Demand Patterns and Seasonality 2

3.2 Geographical Demand Patterns 4

3.3 Key Performance Indicators 5

3.4 Cluster analysis 7

3.5 Predictive analytics 8

3.5.1 ARIMA model 9

3.5.2 Random forests 9

3.5.3 Regression 9

4 Conclusion **10**

A Appendix **12**

List of Figures

1 Hourly repartition 2
2 Weekly repartition 3
3 Monthly repartition 3
4 Most frequented Stations 4
5 Least frequented Stations 5
6 Average Monthly utilization 6
7 Development of the monthly user types' shares 6
8 Total monthly revenue 7
9 Percentage of bikes used at least once a day 8
10 Revenue divided to user type 12

2 Detailed Report

2.1 Problem description: Business and data mining goal

For the task at hand, we received a data set from the Boston-based bike sharing company 'Bluebikes' for the year 2017. The goal is to use the available data to make recommendations for optimizations for the company. To achieve this goal, we first check temporal and geographical demand patterns. After that, we create KPIs and cluster the user groups, and finalize with a prediction on the demand.

2.2 Data description

The columns of the data are the following: A start and an end time are in two separate columns (datatype: object), which describe the time a bike was rented. There are data on the start and end station of a ride, represented by both a station name (datatype: object) and an ID (datatype: int) for the start and the end station. Furthermore, the ID (datatype: int) of the bike used for a particular ride is given, and finally, also data on the user type (datatype: object). For this, it can be differentiated between subscribers and customers. Additionally, there are weather data being used containing values (datatype: float) for the maximum and the minimum temperature as well as an indication of whether or not there was precipitation (0.0 or 1.0)

2.3 Data preparation

For the data preparation, there were different data retrieved. First, a set of Bluebike's data for the year 2017 was given and already pre-edited as mentioned in the assignment. First, we checked for null values. We found that four entries contained null values ("NaN"). We also found odd data which we then deleted. This contained the stations "8D OPS 03", "8D QC Station 1" and "8D QC Station 02", which we thought are likely test data stations. The data also contained six entries that had an end time of a bike ride time before the start time. This only occurred on the 5th of November, due to an issue with daylight savings.

3 Data analysis

3.1 Temporal Demand Patterns and Seasonality

When doing our analysis of the temporal demand, we start by cleaning the data further, as we only want to work with the start and end times of the rides. From here we add some depth to the data, in which we make variables for which day of the week the ride took place, the month, and the duration of the ride. After this is done, we can start our actual analysis of the data. Firstly, we produce a graph where we show how many rides we had within each hour of the day. For rides spanning the hour mark, we count them from the hour the ride began.

For the hourly repartition in Figure 1, we clearly see two spikes in our data. One at 8 o'clock, and once again at 17 o'clock. This is very likely closely tied to a work day from where you are expected to be at the office at 9 and then leave after 17. This tells us that one of the largest drivers for demand is people working office jobs. Notably also is that the second spike is higher than the first, which could indicate that some people use another mode of transport to get to their job, but then take a bike home when they get off work.

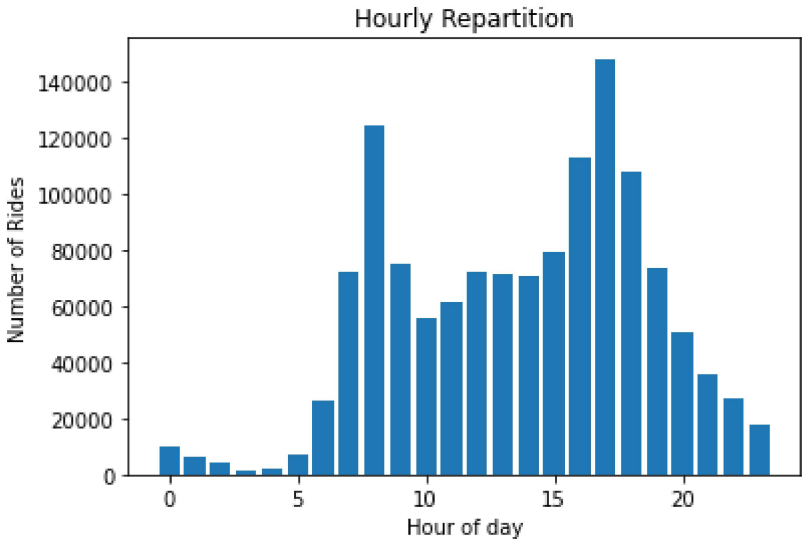

Figure 1: Hourly repartition

Looking at different holidays throughout the year, we find different patterns than those from the entire year. For new years day, we see that demand slopes down as we approach midnight, and customers are at their chosen place of celebration. Notably also, is that we see more activity through the night than when we look at the entire dataset.

Finally, as seen in Figure 2 we look at days and months, to see if we detect any pattern there. For days, we see that the most usage takes place during the

week, with a drop in the weekends. This could further tell us that the primary user of these bikes are customers who use it to commute to and from work.

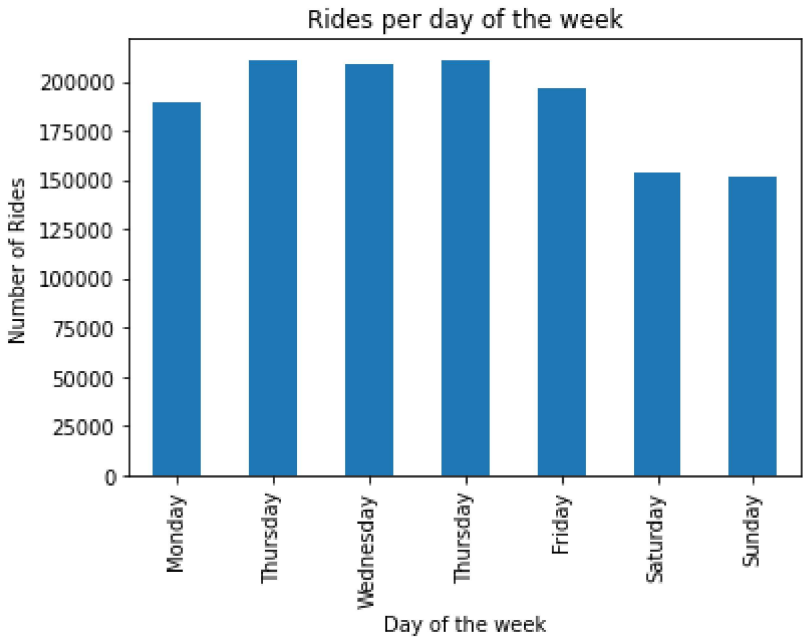

Figure 2: Weekly repartition

Looking at the months in Figure 3, we see a noticeable difference between the seasons. Interestingly, this bar chart lines up very well with the average monthly temperature in Boston, which tells us that, above all, demand is dictated by temperature.

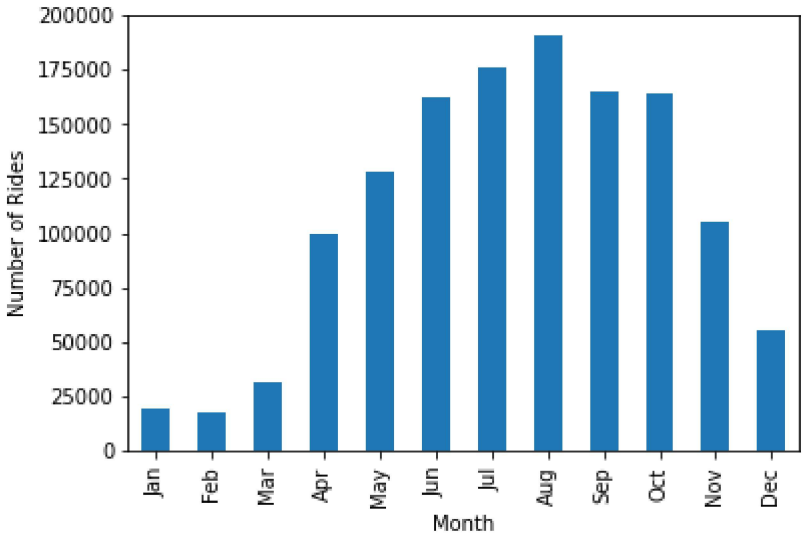

Figure 3: Monthly repartition

The average duration of a ride on a daily basis depending on the starting hour peak at night times at 3 am. For the days of the week, the average duration is

significantly higher on the weekends at approximately 25 minutes and around 17 minutes on weekdays. The average duration on the monthly basis shows a peak in April and an overall higher duration in the summer months.

3.2 Geographical Demand Patterns

To analyze the geographical distribution of the customer's demand, it is necessary to aggregate the rides of 2017 and summarize them with regard to the start station, and end station respectively. After determining the number of rides that started and ended at a certain station separately by grouping by the station IDs, the results are combined to get an overview of all stations and the total number of rides related to each station, whether it was the start or the end station of a ride. The appendix contains extracts of the lists resulting from the analysis. Among the most frequented stations are two stations near the Massachusetts Institute of Technology ("MIT at Mass Ave / Amherst St", "MIT Stata Center at Vassar St / Main St"). Nearby is the station "Central Square at Mass Ave / Essex St" which is as the name already indicates a central position within the city. Furthermore, there are to find bus and subway stations at this place when looking at the area on Google Maps. The same holds more or less for the station "Kendall T". It is close to the station "MIT Stata Center at Vassar / Main St" and according to Google Maps, there is at least a subway station to find, too. The third most frequented station is "South Station - 700 Atlantic Ave" which is a bit distant from the others and seems to be right to a train station. Combining this information, it seems that many people, especially students, commute to and from MIT and include the shared bikes to get there or back home. The bikes are probably used to get to the train station and then from the train station of their arrival to the MIT facilities or the city center / central square.

	station_name	total_count
67	MIT at Mass Ave / Amherst St	84762
80	MIT Stata Center at Vassar St / Main St	58049
22	South Station - 700 Atlantic Ave	54823
68	Central Square at Mass Ave / Essex St	52647
189	Kendall T	46495

Figure 4: Most frequented Stations

The least used station is "Four Corners - 157 Washington St" which is quite far away to the south from the city center of Boston and there is at least one bus line stopping close by. Furthermore, there are two other Bluebikes stations very close and two more a bit more distant but still close by. An explanation for the low frequency could be the offer of alternative stations or services to use that bring

more value to customers due to, e.g. convenience or lower distance to go and use. When searching the station "Faneuil St at Market St" on BlueBikes' website, it appeared that this station could not be found after searching for "Faneuil" in the names of start and end stations, there was listed "Faneuil Hall - Union St. at North St." as a station as well. The assumption is that the station "Faneuil St at Market St" was not accessible over the whole year. As it was used on the 30th of December for the last time, the station was probably not moved from "Faneuil St at Market St" to "Faneuil Hall - Union St. at North St." (permanently). "Columbia Rd at Ceylon St" is close to "Four Corners - 157 Washington St". The suggestion of alternative stations or services in the area of these two stations is strengthened by the low number of uses. This area seems to be a quieter part of the town. Looking at Google Maps it seems to be a residential area with residential buildings, some restaurants, and shop as well as a park, all reachable by foot. This reduces the need of taking a bike. The same aspects apply to the city part "Huron Ave. At Vassal Lane" is placed in. For the station "18 Dorrance Warehouse" the look at Google Maps reveals that the surrounding area probably is an industrial park as there are especially companies located around the bike-sharing station.

	station_name	total_count
232	Four Corners - 157 Washington St	5
207	Faneuil St at Market St	98
203	Columbia Rd at Ceylon St	325
1	18 Dorrance Warehouse	336
181	Huron Ave. At Vassal Lane	383

Figure 5: Least frequented Stations

3.3 Key Performance Indicators

As key performance indicators for the bike-sharing business of Bluebikes the average duration of usage, the shares of customers' and subscribers' related rides, the estimated revenue as well as average utilization rate. The average duration of usage is depicted in the figure below. A rented bike was used for almost 19 minutes (18.91 minutes = 1,134.89 seconds) on average per trip over the year 2017. On a monthly basis, the average duration of usage was 17.51 minutes with changes during the year. In general, from April to October the average usage was higher than the monthly average. Especially between April and June, it was noticeably higher ($>$ 20 minutes). The highest value was reached in April 2017 with 24.46 minutes, while the lowest average usage duration was recorded in February 2017 with only 11.36 minutes.

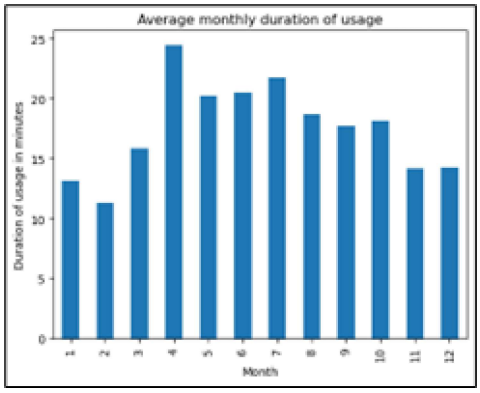

Figure 6: Average Monthly utilization

Another KPI contains the ratios of the different user types. For this, the shares of customers and subscribers are calculated. Although customers probably spend more money per minute due to the higher costs of single trips, the subscribers are important because they are the source of a steady monthly revenue stream. In 2017, there were 84.1% of the rides related to subscribers, whereas, in turn, only 15.9% were related to customers. The shares change from month to month. In particular, the highest percentage of subscribers-related rides was recorded in December (96.40%), while the lowest percentage was in July (77.07%). The following figure depicts the distribution of the rides with respect to the two different user types in absolute numbers as well as in percentages. Looking at it, one can find that during the time from April to October the percentage rates of customer-related rides are significantly higher than during the rest of the months. The simple conclusion behind this is probably that the subscribers are e.g., people commuting and also riding during the winter. The single-trip users seem to prefer higher temperatures to using a bike from time to time.

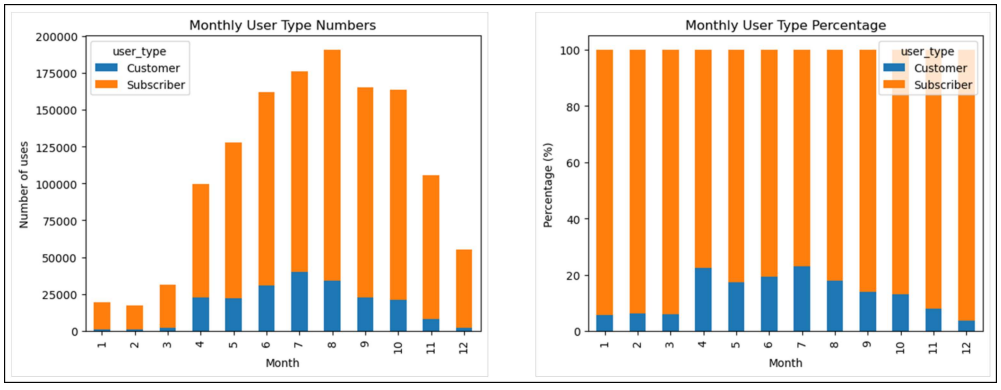

Figure 7: Development of the monthly user types' shares

The revenue estimates are based on two components: Ride-related revenues on the one hand and the monthly fee paid by subscribers on the other hand. For calculating the revenues resulting from the rides there were several conditions to

consider. Subscribers have an unlimited contingent of 45-minute rides. They only have to pay for rides longer than 45 minutes, which amounts to $2.5 per additional 30 minutes. Customers instead pay per ride with rides up to 30 minutes for $2.95. Above that, they are charged an extra $4 per additional 30 minutes. Based on these conditions, the duration of each ride was checked in consideration of the respective user type associated with the ride. If the duration in fact was higher than 45 or 30 minutes for subscribers or customers the extra fees were calculated for this time. After adding the fixed fees for customers' rides and the monthly subscription fees, the highest revenue could be recorded for July 2017 at $501,199. The lowest revenue was in February ($198,479). This is probably due to the fact that according to Bluebikes' website, the season only opened on 27th February (https://www.bluebikes.com/system-data). However, there are still several rides recorded between the 1st of January and the end of February. The figure below shows the total monthly revenues. In the appendix, there can be found a list with the exact revenue divided according to the user types.

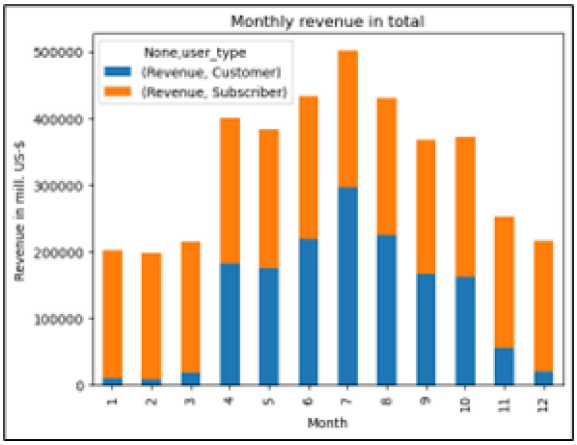

Figure 8: Total monthly revenue

Finally, the utilization rate was computed as the last KPI. The highest rate was recorded on the 5th of July when 71.25% of the bikes were used at least once during that day. In contrast to this, on the 13th of February, only 1.84% were used, but the season-opening times have to be kept in mind at this point. The following graph shows the percentage of bikes which were used during the day at least once. Note that the rate of bikes used is quite high between April and September and the peaks during this time are above 60%, even heading towards 70%.

3.4 Cluster analysis

We used a PCA to try to reduce noise and dimensionality. But by reducing the dimensionality by one, we only keep 87% of the variance. Thus we did not use it

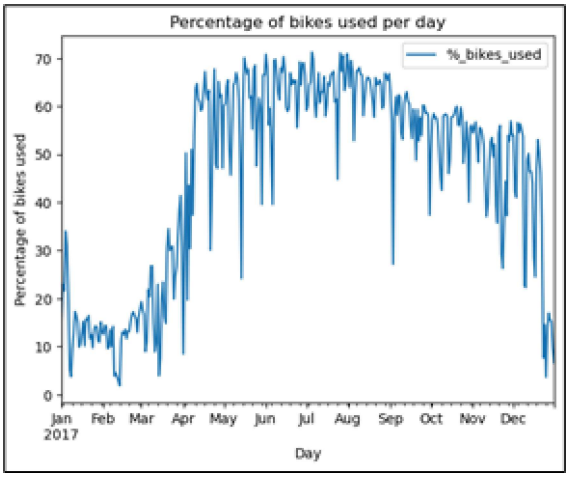

Figure 9: Percentage of bikes used at least once a day

further. Also, we applied the k-mean clustering algorithm which revealed to us that we use two clusters, for the user groups subscriber and customer. Cluster two has a larger amount of data with 748.933 instances while cluster one has 564.824 instances. The analysis of the two clusters shows that the two user groups behave similarly, especially for the average day of the week and the average month. Only the average start time is quite different. For the first cluster, it is in the early morning at 9:31 and 17:46 for cluster two. Also, the average trip length is 2 about two minutes shorter for cluster two with 17.28 minutes, while it is 19.28 minutes for cluster one.

	Cluster 0	Cluster 1
Instances	564824	748933
Average Start	9:31	17:46
Average day of week	2.771	2.871
Average month	7.543	7.476
Average trip length	19.28	17.28

3.5 Predictive analytics

In our predictive analytics part, we start by importing the data and collecting the rides within each hour of each day, e.g. 13 rides between midnight and 1 o'clock on the 1st of January. The weather data is then imported and paired with our existing data. After this, we add dummy variables for holidays or weekends, so we know not to expect the two daily peaks from commuters. After this, we then begin with our prediction models where we spilt our data into test and training data. The first three quarters of the year are used as training data, and the last quarter is used as test data. In our ADF test, we find the data to be stationary, however, we improve on this with decomposing. From here we use ACF and

PACF to estimate the parameters of the ARIMA model.

3.5.1 ARIMA model

We encountered problems during coding, which made us unable to get a MAE. Fortunately, we managed to find results for the other models

3.5.2 Random forests

We use 1000 estimators to train our model, which results in Mean Absolute Error of 27,37. This is the lowest MAE we find, however, it is possible this could be further improved by pruning the trees. Due to the nature of the model, there is a risk of overfitting, which we can only comment on if we had multiple data set to test our model on.

3.5.3 Regression

In our simple polynomial regression, we find that 2 degrees are best, resulting in MAE of 90,57. In the Ridge regression, we find an alpha of 1e-10 to be best, resulting in MAE of 68,28 bikes. Finally, in the Lasso regression, we find the same alpha, giving us a MAE of 67,25.

4 Conclusion

The analysis of Bluebikes' rides data for 2017 formed a solid basis for a first evaluation of a bike rental service in Boston. The business has a steady revenue stream due to its subscribers. However, during the time from April to October the share of revenues from customers is much higher compared to the time between November and March. It has to be noted that 2018 was the first year with a year-round service offering. Before that, including 2017, there was a season specified. In 2017, in particular, the 'off-season' was from the 1st of January until the 27th of February. This has to be considered on top of the colder temperatures due to the winter season when analyzing the data. Nevertheless, it could be advantageous for Bluebikes to look into possible measures to increase the number of single trips and share of customers from November to March. Protection against worsening weather conditions could be one (e.g., rain covers, bigger tires during the winter season). As another pricing model to add, there is the idea of offering fixed prices for pre-defined popular, frequented, or cinematic routes through and around the city to attract single users, such as visitors, tourists, occasional users, etc., even more. A deeper look into the usage durations could support more sophisticated pricing models which include e.g., an adjustment of the additional time slots of 30 minutes users pay for an extra fee if needed. Furthermore, Bluebikes could seek cooperation with the city administration to provide benefits to people using the bike-sharing service and by that contributing to a 'greener' (less carbon dioxide) and 'healthier' (people keep fit) city which, in turn, leads to decreased costs for the municipality and improves living quality. Also allowing people to use Bluebikes' service could be invoked to help design or re-shaping parts of the city/city part they live in. As Bluebikes is owned by several cities of which the city of Boston is one, such cooperation is not only likely but probably creates a well-working symbiosis. In general, the extension of the dataset to more than one year, especially to the following years (2018 and following), could result in a value-added with regard to the analysis and data evaluation. This way a proper time series analysis can be conducted and the prediction results could be improved, cause right there lies one of the main limitations of this analysis. Using data from January to September to forecast the demand from October to December noticeably limits the relevance of the prediction results. Using two or more years to predict the demand of another year allows a much better analysis of patterns. More detailed information on the users would be beneficial as a deeper analysis of usage clusters is possible. A more advanced and detailed version of the weather data would allow differentiating between different weather conditions and their impact on the demand.

A Appendix

		Revenue
start_month	user_type	
1	Customer	9884.95
	Subscriber	191706.00
2	Customer	8183.00
	Subscriber	190296.00
3	Customer	18471.75
	Subscriber	196828.50
4	Customer	182751.10
	Subscriber	217833.50
5	Customer	175449.40
	Subscriber	207298.50
6	Customer	219621.30
	Subscriber	213211.00
7	Customer	295608.00
	Subscriber	205591.00
8	Customer	224746.25
	Subscriber	204716.00
9	Customer	166352.00
	Subscriber	201403.50
10	Customer	162846.80
	Subscriber	209241.00
11	Customer	55436.20
	Subscriber	196773.50
12	Customer	20401.00
	Subscriber	195901.00

Figure 10: Revenue divided to user type