DBA4811 ANALYTICAL TOOLS FOR CONSULTING
Group Report

DBA4811 GROUP 7

Name	Matriculation Number
Andreas Lukita	A0221743M
Kueh Choon Hwa	A0218415J
Goh Shi Yun Eugina	A0221380U
Lee JiaHui Joyce	A0189094N
Vaddadi Divya Asritha	A0189519L

Submitted On: 5th November 2022

Table of Contents

1. Introduction

Ziro is a ride-sharing company established in 2019, with a mission of being the first socially responsible ride-sharing company dedicated to treating drivers fairly, charging riders less, and preserving the planet. It competes with Uber, Lyft, Wingz, and Flywheel for its piece of the $61 billion rideshare market (Helling, 2022). Like its competitors, it has ventured into the bike-sharing segment as well in its state of origin, San Francisco. Given that its competitors have also seen success in their expansion to bike-sharing services, Ziro's management team decided that they would need to expand to other states in an effort to capture market share before its competitors completely dominate the different states. As such, Ziro identified Chicago as a suitable location to enter.

2. Objective

Ziro aims to achieve successful market expansion to get a hold of the Chicago bike-sharing market. This grand plan would require Ziro to commit huge investment in capital in terms of building its bike-fleet. In order to ensure such a plan is successful, Ziro must pay careful attention to its operation and marketing strategy to ensure its financial resources are fully and efficiently utilized. Ziro approaches our team of analysts to perform an in-depth analysis of the Chicago bike-sharing market to understand if Chicago is indeed the market to invest in. Such market analysis requires insights in terms of bike-users behavior as well as demand forecasting of bike usage on a daily basis.

The first part of this report focuses on the high level overview of the Chicago bike-sharing market. This would provide more confidence for Ziro to conduct a pilot test at the top 20 stations before launching a full-scale operational effort in Chicago. Such strategy helps in minimizing the risk associated with high investment in capital. Our analysts would also recommend the best time period to conduct the pilot test to ensure its effectiveness and show Ziro a glimpse of the predicted total number of daily bike-users in Chicago.

The second part of this report aims to provide Ziro with a focused analysis on these top 20 stations that our team recommends, and helps Ziro to predict the expected demand at one of these 20 stations per day given the associated weather condition.

3. Data Preprocessing

3.1 Cleaning

To understand the bike-sharing market in Chicago, we leveraged on existing data from one of the existing players, Divvy (2019). The initial data (trip and stations) obtained were in separate CSV files and subsequently merged together based on station IDs (Appendix A). Additionally, we had also scrapped the hourly weather data of Chicago (Wunderground) (Appendix B) to add another dimension of factors that could affect demand for bike sharing before dropping unnecessary columns (trip_id & bike_id). As

inputs for customer usertype is optional, there are some missing values in these fields, to fill in the missing values we had to use the existing data to extrapolate the distribution and spread of these variables and repopulate it into the missing fields. In addition, we drop records with extremely low probabilities such as birth years before 1939. As some of the analysis will also require the data to be aggregated by date and / or from station name to determine the demand, additional aggregate datasets were developed to meet these needs.

3.2 Feature Engineering

Several new features related to time were created from the merged dataset to give a more comprehensive analysis (Appendix C). Of which, We had also taken into account the type of holidays (School breaks, weekends & public holidays) that Chicago has as this could affect the demand.

4. Exploratory Data Analysis

To develop a sound understanding of the riders' consumer behavior, we segmented our analysis into 2 parts. Firstly, we conducted an overall analysis on their usage patterns, demographics and the popular locations they frequented. Next, we drilled deeper by analyzing the user behavior among the top 20 stations.

4.1 Overall Analysis of Customers Across All Stations in Chicago

In this section, we will be highlighting key trends on the riders' usage patterns, demographics and locations they frequent across all stations.

4.1.1 Usage Patterns: Popular Times, Days & Demographics

Finding 1: Weekdays, especially Wednesday, are most popular during peak hours, which are 8am and 5pm. Comparatively, demand for weekends is more distributed and less frequent from 11am to 6pm.

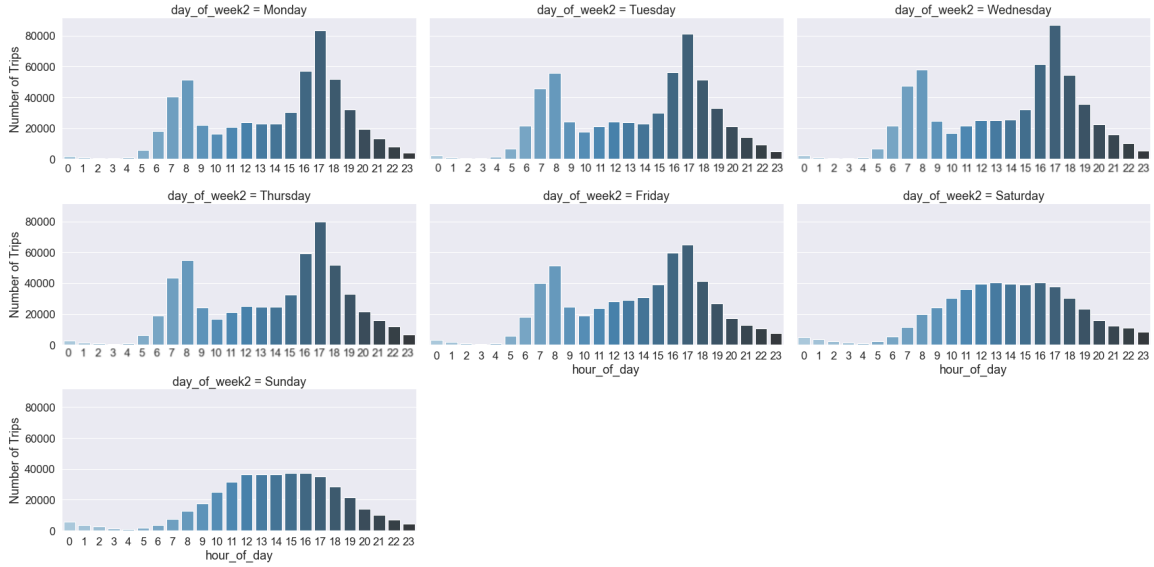


Figure 1: Bar Graphs of Frequency of Trips against Day of Week and Time of Day

The peak hours on weekdays follow the frequency of commute at 8am and 5pm as shown in Figure 1. The trend shows that on average, more trips are made in the evening than mornings on the weekdays. An interesting thing to note is that riders take rides more frequently on Monday, Tuesday, Wednesday and Thursdays, but Wednesday 5pm is the most popular period. One possible rationale for this finding is that people are not in a rush to commute back home due to the next day, Saturday, typically being a rest day rather than a workday. On the other hand, for the weekends, the demand for bike rentals is more distributed and less frequent, where the peak periods happen between 11am to 6pm. What this could suggest is that compared with weekdays, where the use is more geared towards a necessity to get to and fro work, weekend usage centers around having a more ease mode of traveling (leisure) than rushing from point to point.

Finding 2: Shorter rides are taken during peak periods

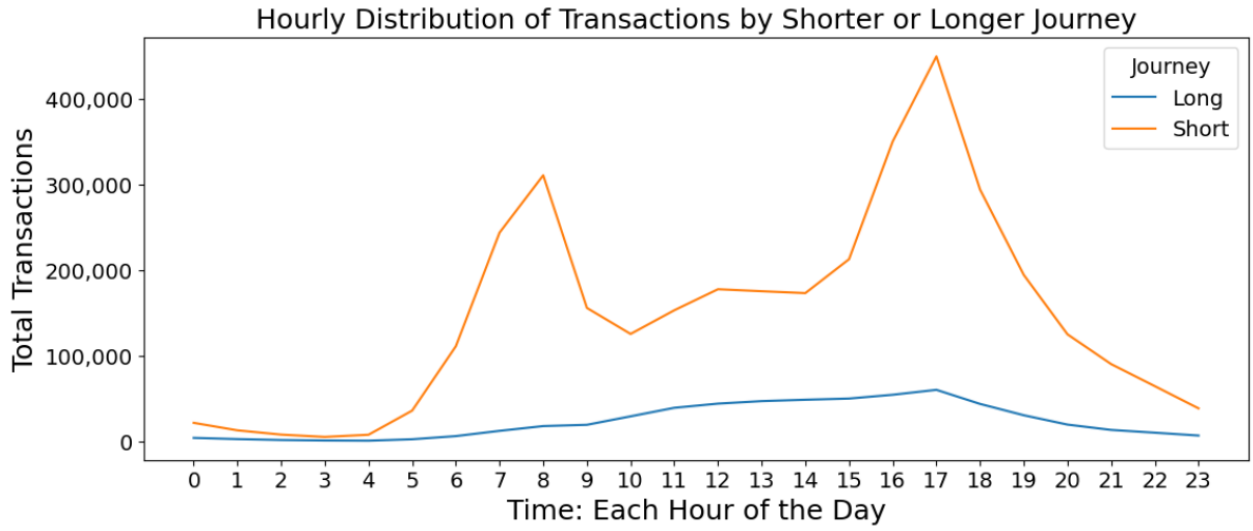


Figure 2: Line Graphs of Frequency of Trips vs Duration of Trips

Next, we analyzed the trip duration by categorizing trips under 2 buckets: Short (less than 30 minutes) and Long. For longer journeys, there are no distinct peak periods as compared to short journeys. This shows that during the peak periods, our users utilize our service for short distance transportation. This finding, to a certain extent, reinforces the notion that demand during peak periods is spanning from people who are rushing from point to point.

Finding 3: Bike renting services seem to be more popular among those aged 25 - 40

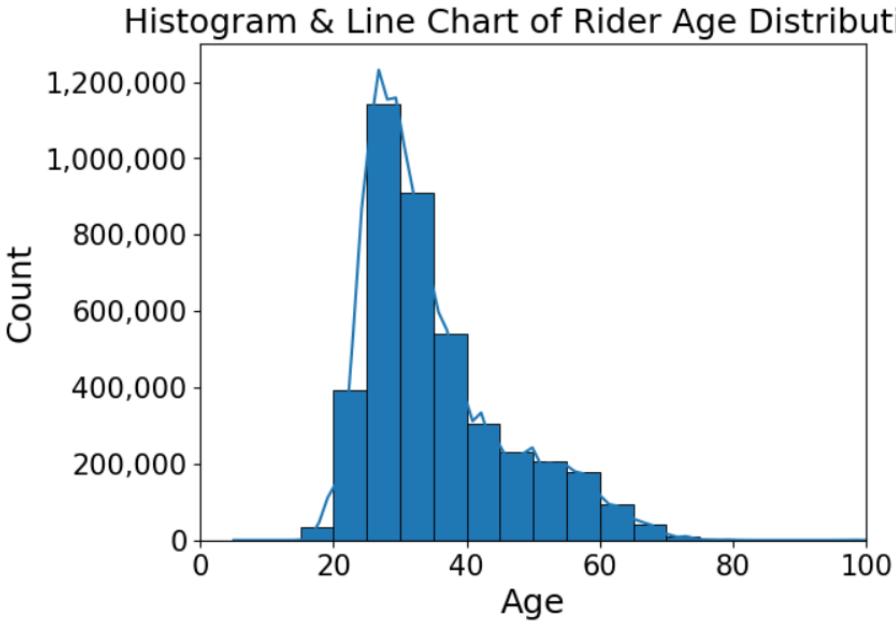


Figure 3: Histogram of Total Trips across Age of Riders

As shown in Figure 3, Riders are generally aged between 25 - 40 and more concentrated around the ages of 30 - 35. This signifies that the majority of riders are working adults.

Across Figures 1 - 3, the majority of riders who tend to use bike renting services are millennial working adults who tend to live within the vicinity of their destination (i.e. office) or using it as their last mile transportation method.

This is due to 3 reasons:

1. 8 - 9am and 5 - 6pm seem like the timing when people start and end their day at work.
2. Shorter trips are taken during peak hours which means the distance is short.
3. They are mostly aged 25 - 40.

Furthermore, since the demand for weekends is lower, this means riders might perhaps prefer renting bikes for commute more than leisure. Marketing efforts should hence be more concentrated over the weekdays, especially Wednesday evenings.

4.1.2 Usage Patterns: Seasonality

Finding 4: Historically, Summer is the most popular season while Winter is the least.

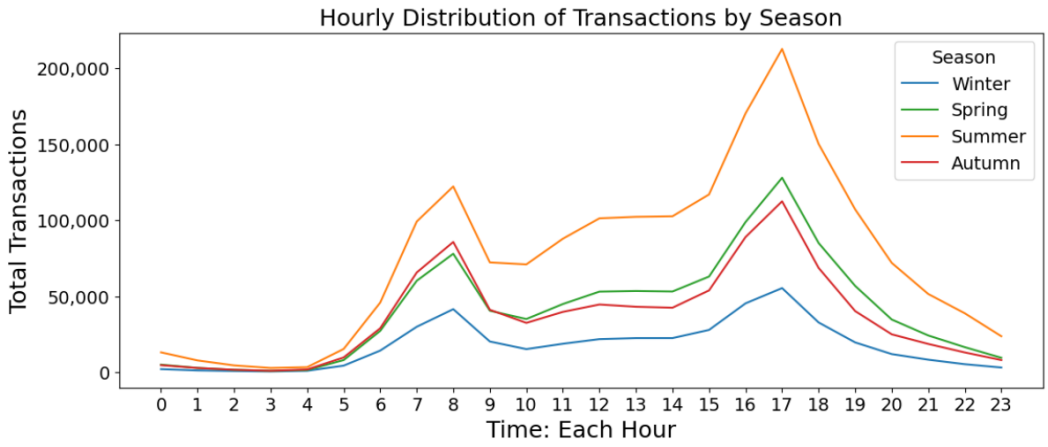


Figure 4: Line Charts of Frequency of Trips against Time of Day among the 4 Seasons

Throughout all seasons, the peak periods (8am and 5pm) are common as shown in Figure 2. However, the total demand for bike rental services differentiate throughout the seasons. Summer catches the highest demand followed by autumn, spring and finally winter. Hence, the most crucial time of the year to ensure that the business operations and marketing efforts are at its optimal service level would be during summer between June to August. Higher marketing budget could be allocated to those 3 months as it has the highest potential in subscriber gains.

Whereas over the winter, the low demand could allow the company to recall bikes for inspection and maintenance or upgrades as capacity required to fulfill demand is significantly lower, hence minimizing any loss in sales while performing continuous improvements for our users. It could also be a good chance for pilot experiments given any strategy revamping as it could be tested on a smaller consumer group, minimizing any potential loss.

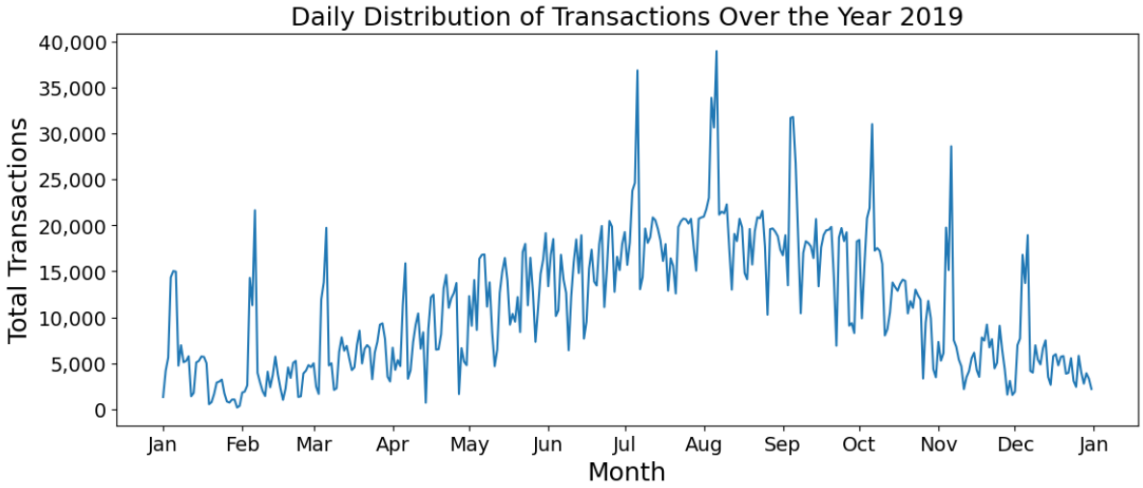


Figure 5: Line Distribution of Daily Demand over the Year 2019

As seen in the total daily distribution patterns, the total transactions peaks from July to September, aligning with the seasons with increased demand.

4.1.3 Usage Patterns: Holidays

Finding 5: Bikes are rented more among non-work days. Among holidays/weekends, those that fall during summer, spring and winter tend to record a higher number of rides.

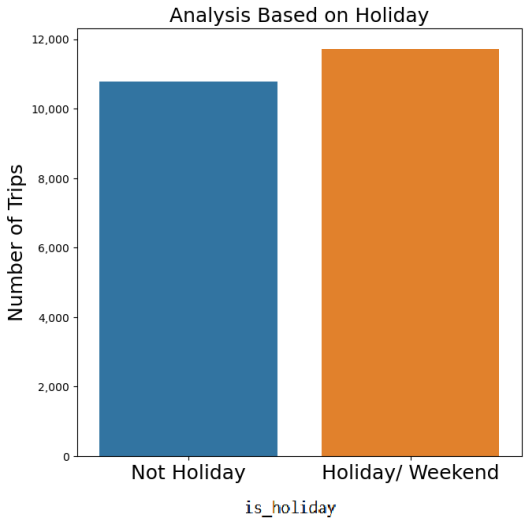


Figure 6: Bar Chart of Average Daily Number of Trips Among Holidays & Non-Holidays

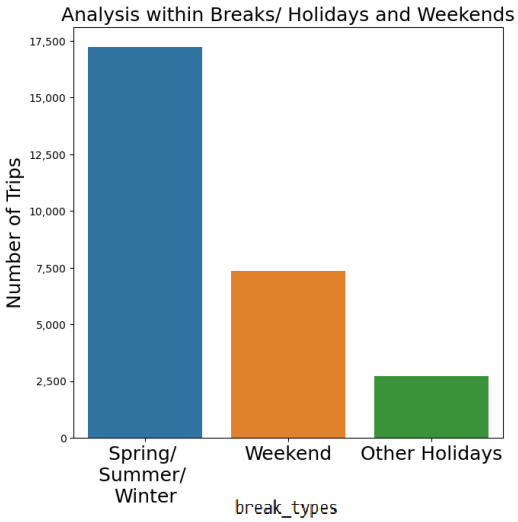


Figure 7: Bar Chart of Average Daily Number of Trips Among Holidays & Weekends.

Surprisingly, the number of trips over the holidays/weekends is not very far apart from the number of trips over weekdays. Among the holidays and weekends, spring/summer/winter vacations record the highest trips which means riders tend to rent bikes for leisure purposes when holidays are longer. From this, we can understand that the behavior of renting bikes is generally used more for commutes than for leisure which supports the findings in section 4.1.2 with the exception of when the school vacation days period.

This finding has highlighted that aside from the summer season vacation periods are also another high demand stage for bike-sharing. With that, Ziro could consider their entry to the market at a time where it is able to capitalize on both the summer season and vacation breaks, i.e. summer vacation period. By doing so, it will help Ziro to garner greater brand awareness as demand could find their way to Ziro's services.

Finding 6: Females are more likely to indulge in leisure rides during long holidays.

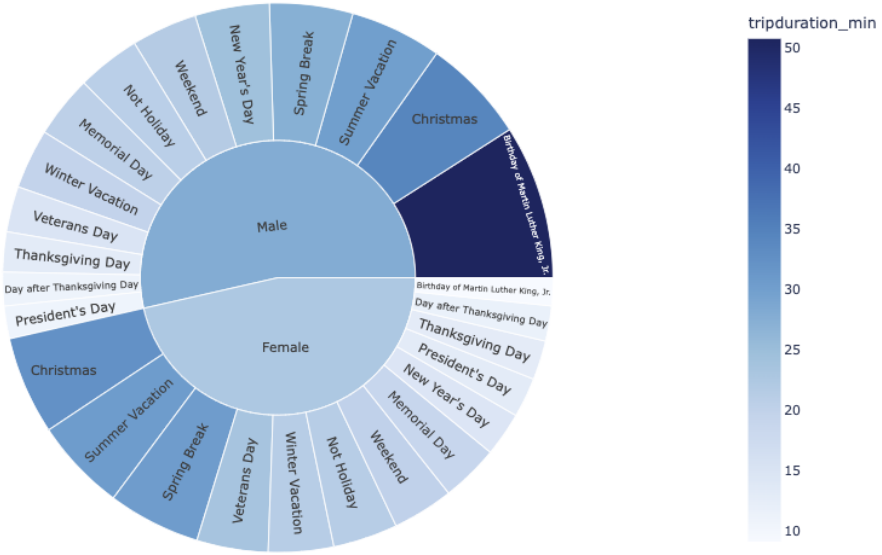


Figure 8: Sunburst Chart of Average Trip duration by Gender and Type of Day

As shown in Figure 8, Male users clock a longer trip duration compared to females on an average. However, the top 3 types of days for longer rides among males are: (1) Birthday of Martin Luther King Jr (2) Christmas (3) Summer Vacation. Meanwhile for females: (1) Christmas (2) Summer Vacation (3) Spring Break. On an average, females clocked longer durations for Summer Vacations and Spring Breaks which might indicate that females are more likely to enjoy leisure rides. Surprisingly, Male users clocked a longer duration for the Martin Luther King Jr holiday. Perhaps, there could be an outlier or a cycling event on that day.

4.1.4 Popular Locations of Users

Finding 7: Chicago's central area is the most heavily populated in terms of stations which is an indication of a higher demand.

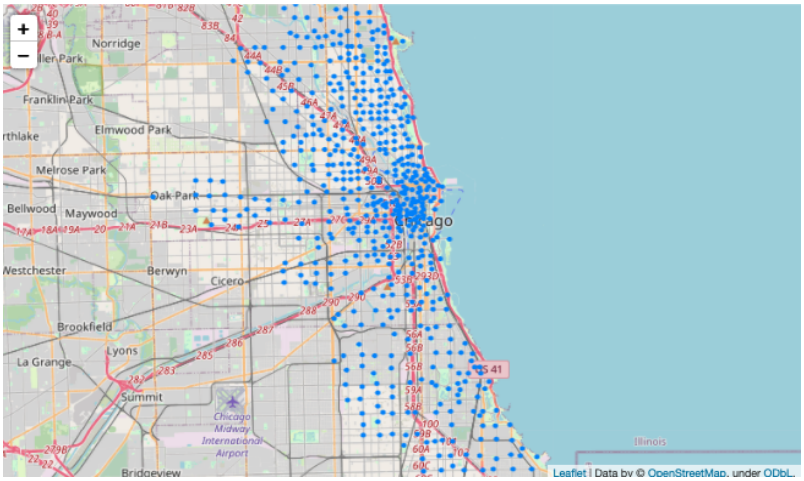
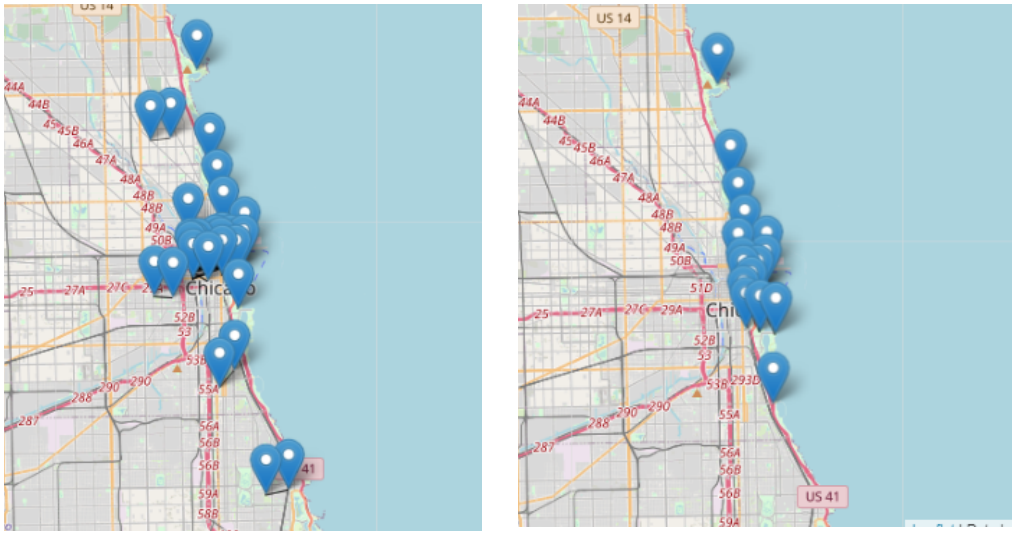


Figure 9: Geospatial Map of Bicycle Stations

By plotting all the bike stations on the map, Figure 9 shows that the bike stations are quite evenly spread out in the city, with the area in the middle being more densely packed. This might be the strategy that is used by the competitor company to ensure maximum consumer reach across the entire city and also at the same time allow their service to be conveniently accessible. As we investigate further, the middle area that is densely packed with stations is actually the Central Business District (CBD) area of Chicago. This area also has schools like University of Illinois and Northwestern University (Medical School) nearby. This might be able to explain the dense clustering of stations in that area as there is a higher probability of it facilitating higher foot traffic which increases the chance of new riders and increase in usage of bike-sharing services.

Finding 8: The top 20 stations during weekdays and weekends act as a guide for Ziro to set up their pilot stations.



Weekdays *Weekends*

Figure 10: Geospatial Maps of Top 20 Stations on Weekdays and Weekends

To deep-dive into the above point on rider demand, we identified the top 20 busiest stations for weekdays and weekends. As seen from Figure 10, the top 20 stations are clustered around the middle area which was identified above. This area and these stations can be a guide to where we should set up our first batch of bike stations to ensure an adequate consumer reach.

However, it can be seen that the popular stations are not exactly the same for weekdays and weekends. This shows the different usage pattern and riders' route on different types of the day and thus different usage purposes, as supported by previous sections. Thus, is_weekday might be an important factor that affects demand on different days and at different bike stations.

We selected the overall top 20 busiest stations through the count of trips at each station regardless whether it is a check-in or check-out (Appendix E). In the next section, we will be deep diving into data related to the top 20 stations selected. We will investigate the customer demographic and usage patterns to further understand our target audience. This can give us a direction in devising a relevant business strategy and recommendation.

4.2 Analysis of Customers Usage Pattern Across the Top 20 Stations in Chicago

4.2.1 Usage Patterns: Popular Times, Days & Demographics

Finding 1: Weekdays' peak hours are 7 - 8am and 4 - 5pm while weekends' are from 2 - 3pm. Riders are mainly working adults.

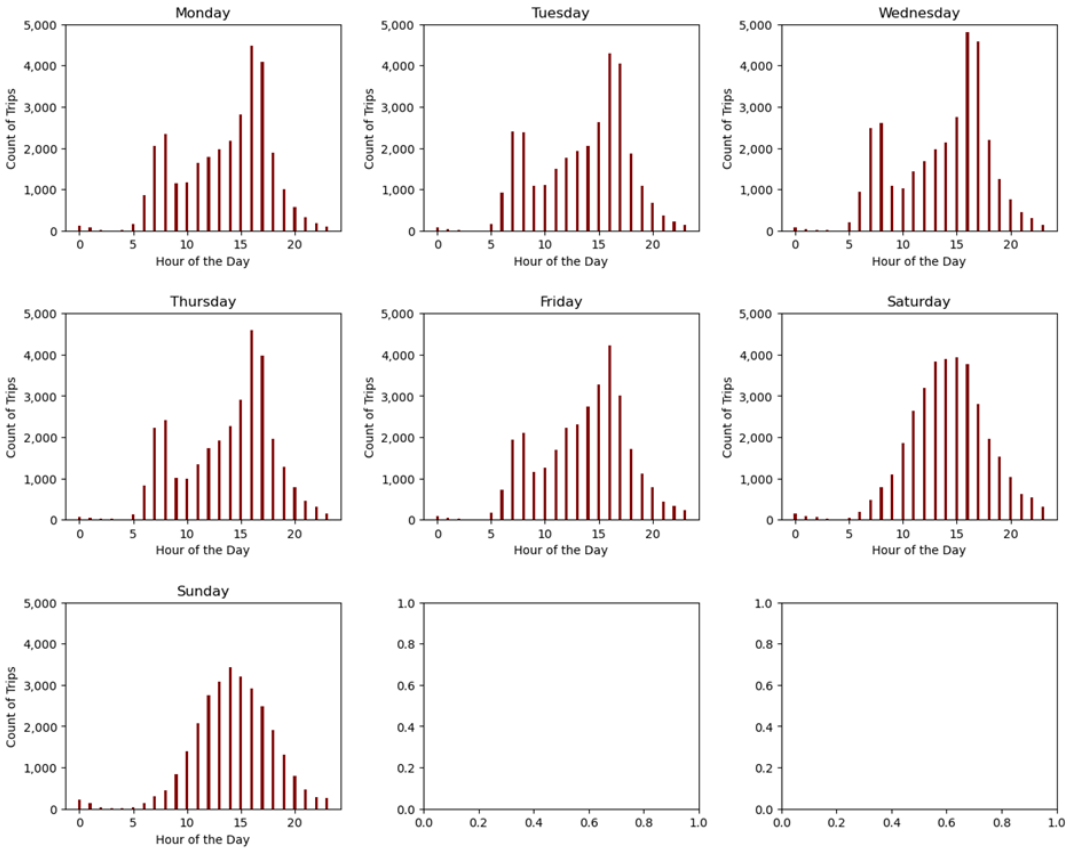


Figure 11: Bar Graphs of Frequency of Trips against Day of Week and Time of Day of Top 20 Stations

Looking at the usage pattern of the top 20 bike stations, it can be seen that usage during weekdays has a distinct peak at 7 - 8 am and 4 - 5pm whereas for weekends, it has a smoother curve, peaking at 2 - 3pm. This shows that the bulk of bike usage is during the period when people start or end their work or school which implies that our main users might be working adults or students. From the graph, it can also be seen that the evening peak is higher than the morning peak for all weekdays.

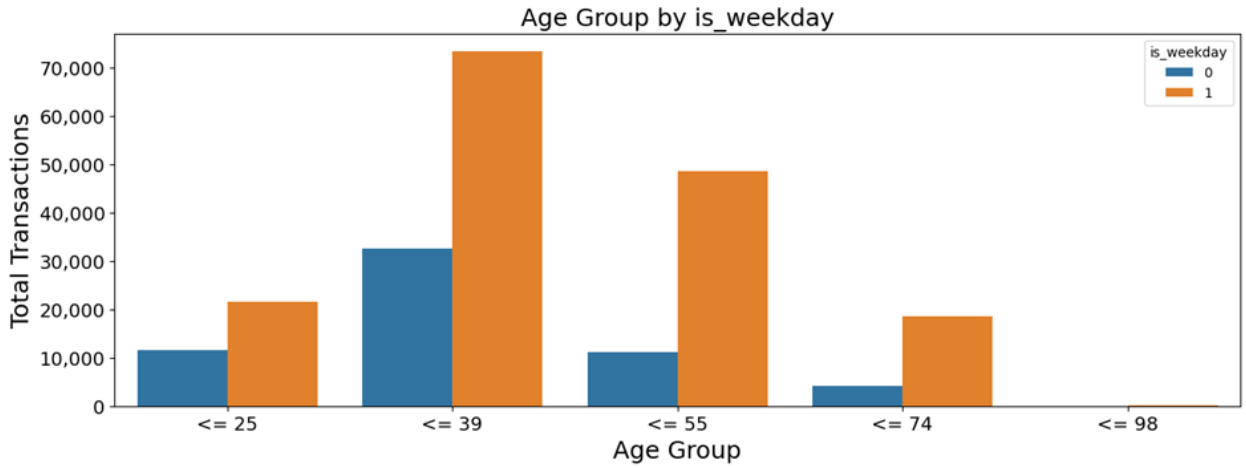


Figure 12: Demand of Top 20 Stations by Age Group & Is_weekday

This is further supported by our users' age demographic analysis for the top 20 stations identified, showing that riders of age between 25 to 39 contributed most to the bike usage count regardless of weekday or weekend. Thus marketing campaigns can be targeted at boosting usage during non peak hours and attracting more riders from other age groups.

Finding 2: More males rider usage compared to women rider usage

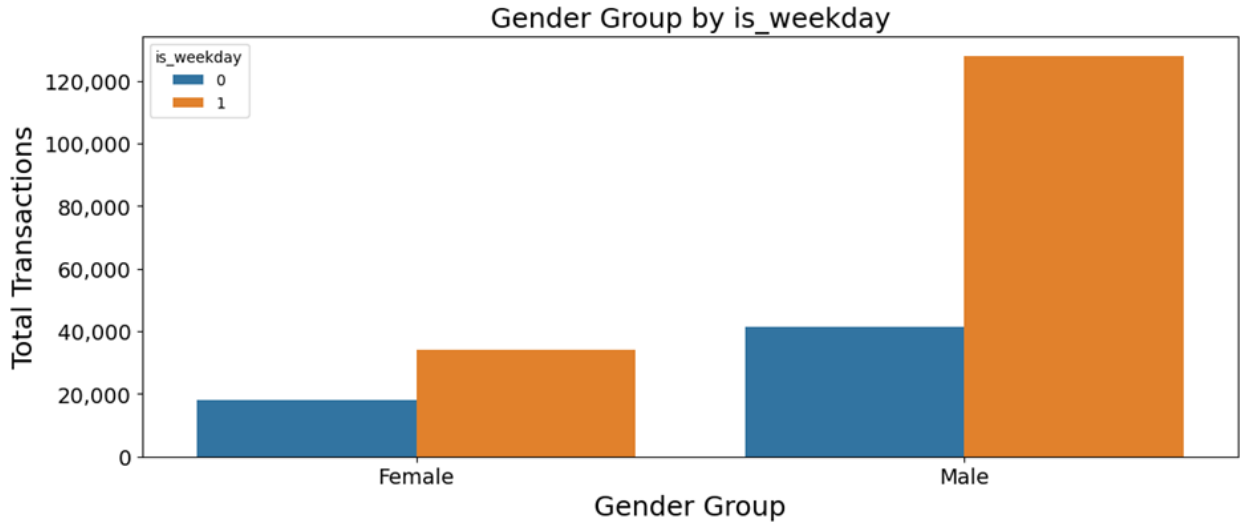


Figure 13: Demand of Top 20 Stations by Gender & Is_weekday

Furthermore, it can be seen that there is significantly more bike usage by males compared to females for both day types. According to Chicago's population statistics, females make up 51.5% of the population (Infoplease, 2022). This implies that the bike sharing industry has yet to capture enough female users.

Finding 3: Trip duration on weekdays are generally shorter than on weekends

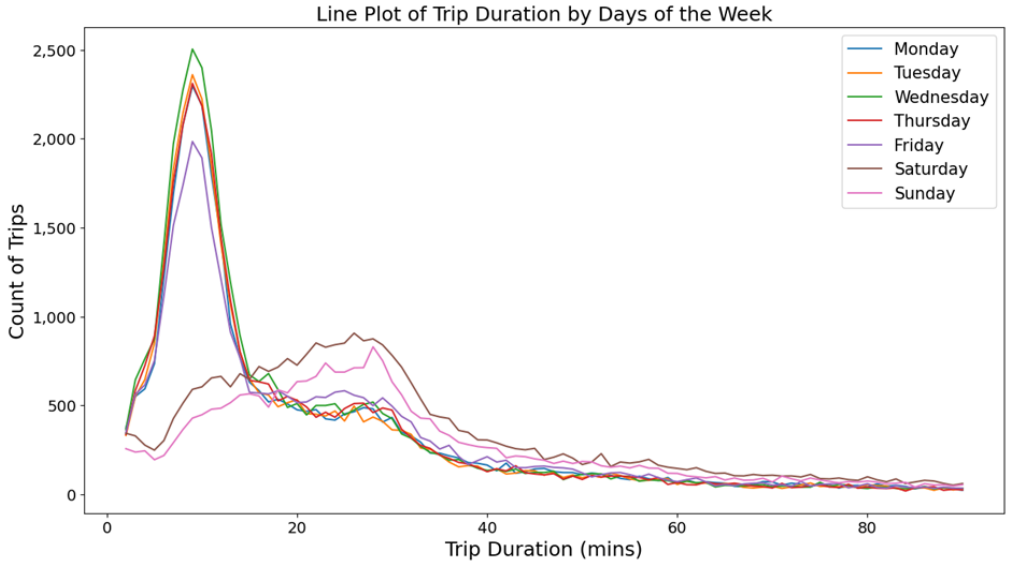


Figure 14: Distribution of Trip Duration (Minutes) of Top 20 Stations on by Day of Week

For the usage duration of the bikes, it can be seen that the mode trip duration on weekends is longer compared to weekdays. Trip durations are mostly around 25 - 35 minutes on weekends while on weekdays it's around 10 minutes. We can interpret this trend as riders on weekdays using the bike services as a form of last mile transportation to office from home or other public transport whereas riders on the weekends might be using it for leisure rides.

5. Models Used

For our predictive modeling, the predicting variable is demand for bike-sharing services. The modeling is done on the aggregated daily and/ or hourly demand as well as demand regarding only the top 20 stations. In our models, we had opted to drop the tripduration variable as it is an outcome of demand and it will not yield any insight from conclusions like, longer tripduration will lead to a larger demand.

Firstly, our team acknowledges that there are some limitations in terms of its support when using a multilinear regression model. Similarly, since we are trying to predict the demand for bike-sharing services, our response variable cannot be a continuous variable nor non-negative value. This leads us to two possible models, namely Poisson Regression and Negative Binomial Regression. We decided with the latter as the data set fails to obey the fundamental assumption of the Poisson Distribution which requires the variance to be equal to the mean. For this model, we mainly took into account the average weather parameters into fitting it with the exception of uvHigh and solarRadiationHigh variables. We had also dealt with instances of multicollinearity and created dummies for categorical variables before accounting for variable transformation and variable interaction to derive the best model using this method. Additionally, a time series forecasting was implemented to be able to account for 3 variables (Level, Trend

and Seasonality). For that, we opted for the triple exponential smoothing with additive trend and multiplicative seasonality.

Aside from those two models, various other regression models from the scikit-learn python package were used (Appendix J) and mean square error (MSE) and adjusted R-square (R^2) values were used to determine the performance of the models. As such, our evaluation will mainly focus on the best model selected given those metrics where the lower the MSE or higher the R^2 would indicate a more comprehensive model in projecting demand. Likewise, prior to fitting the model, we created dummy variables from categorical variables. To prevent any overfitting of the models, a train test split is done with a test size of 20%. To tune some of the parameters of our models, we had decided to utilize grid search or a randomized search technique with a cross validation feature to make efficient use of our data and to get a more accurate estimate in the event of an out-of-sample happening.

6. Evaluation

6.1 Overview of Demand

Demand prediction is done on the dataset as a means to determine if the state of Chicago has a substantial market demand for bike-sharing for Ziro to measure if it is a viable option for it to start their bike-sharing venture in. From the Negative Binomial model which has an adjusted R^2 of 0.495, it suggests that compared to January, the months from June to October have association with demand (Appendix F). This finding coincides with our analysis of Figure 5.

The time series forecasting was able to generate a pseudo R^2 value of 0.523 (Appendix G). It has also allowed us to forecast the following 7 days of demand where the estimated daily market demand is generally around 3,700 for weekdays and a relatively lower demand of 2,500 during the weekends (Appendix G). Before Ziro opt to invest funds, they should consider the relative market size of the state that they would be serving and their ability to compete with their existing rivals in the market.

Lastly, from our other regression models, the XGBoost Regressor, with an ability to explain 69.7% of the response variable (Appendix J), is the one that is able to explain demand patterns the best by combining pre-existing models by allocating weights to the input models which will help to mitigate the risk of some models with a more precise estimate of another.

6.2 Demand for Top 20 Stations

By narrowing our scope of stations to the set of locations where we would be allocating the bike stations, we will be able to focus more on factors that mainly affect the areas where the proposed locations will be at than on locations that will not be served by Ziro.

6.2.1 Negative Binomial

The best model that was developed from this method is one consisting of 19 variables which resulted in an adjusted R^2 value of 0.507 (Appendix K). The top few significant variables that are associated with demand are related to months of the year, specifically during the months in the third quarter (July to September) which all have a positive effect on demand of above 1 from the benchmark month of January. Aside from these variables, variables regarding weather conditions seem to have a slight negative association with the demand of bike-sharing services and that month around the third quarter is also having a positive relationship with demand but at a lower magnitude than that of July to September compared to January.

Therefore, it would suggest that it will be better for Ziro to enter the market during these higher demand months of the year if possible as it will help Ziro to increase their customer reach and grow their user base at a faster rate than in the lower demand months like January as seen from Figure 4 and 5.

6.2.2 Time Series Forecasting (Triple Exponential Winters Holt)

Likewise, we forecasted the next 7 days (2019-12-31 to 2020-01-07) period demand and proceeded to compute the pseudo R^2 value $corr(Y, \hat{Y})^2$ to evaluate our model performance and obtain a value of 0.55 (Appendix L). From this, we can expect the amount of demand Ziro should strive to achieve in the first seven days of 2020 is approximately 180 during the weekdays and 120 on the weekends (Appendix L).

Therefore, Ziro should consider if this size of demand is still profitable as it generally depicts one of the lowest expected demand points in the year, near the winter season. If the daily revenue is able to outweigh the daily operating cost (variable cost), Ziro should opt for the venture into Chicago provided that it is still able to cover some of the fixed cost incurred from profits in the more popular times of the year.

6.2.3 Best Model - Voting Regressor

The best regression model from the top 20 stations is the voting regressor (Appendix O). It is able to explain 73.7% of the target variable Demand as observed from the data. This is unsurprising considering that Voting is one of the simplest way of combining the predictions from multiple machine learning algorithms that our team has conducted, which include Decision Tree (70.8%), Random Forest (70.7%), Support Vector Regressor (66.5%), Multilayer Perceptron Neural Network (55.6%), RF Bagging (68.6%), XGBoost(62.0%) and AdaBoost (50.2%).

7. Implications/ Recommendations/ Actions

We categorized our recommendations into 3 different objectives:

1. Operational Efficiency: Recommendations for Ziro's marketing team to launch and sustain marketing efforts
2. Capture Market Share within Chicago: Recommendations to entice users from existing competitors
3. Capitalize on Under-Addressed Market Potential: Recommendations to meet unaddressed target audiences' needs that current competitors aren't targeting

7.1 Operation Efficiency

Recommendation #1: Pilot bike renting service should be launched during the summer and targeted at working adults on weekdays.

As highlighted in Section 4.1, as well as, our first predictive model, the demand for bike renting is the highest during the summer period i.e. June - August. Furthermore, since working millennials are the major key adopter of bike renting services, this age group should be targeted according to their behavior. Hence, Ziro should target the working millennials during weekdays, especially the peak hours of commute, 8pm and 6pm to ensure a higher rate of adoption. Additionally, this period also coincides with students' summer vacation days where Ziro will also be able to leverage on to quickly grow their user base as there is a relatively higher demand during this type of holidays compared to others.

7.2 Capture Market Share within Chicago

Recommendation #2: Attracting potential users

From the insight found about usage patterns of bike users, strategic marketing campaigns can be used to capture greater market share. Since we know that bike users usually rent bikes during peak hours, as a new company entering a market, we should try to attract as many bikers from the competitors as possible. One way to do so is to promote our company during those peak timings when we have the most potential users around that area. Ziro should leverage on competitive pricing in the first few months of the launch to make their services more lucrative to the potential rider than other competitors. An example of competitive pricing would be exclusive promotions where special rental pricings can be given to subscribers during peak hours. Since the peak bike usage period differs with the day type, as seen in the data analysis above, weekends and weekdays should have different special deal timings, following the peak period that was found in the analysis. This method captures consumers who have low brand loyalty towards the competitor brand. This can help us gain an adequate amount of users and exposure during the starting phase of the business.

Recommendation #3: Leverage predicted demand to conduct flash marketing deals

Based on the prediction of demand in Section 6, the team should establish a benchmark of demand Ziro needs to achieve each day. If the predicted demand is lower than the benchmark, Ziro's marketing team can actively push out flash deals on the app to invite riders to rent bikes and increase their usage. Dynamic discounting is a cost-saving

marketing solution which only requires the company to invest into giving discounts during periods of low usage, instead of always giving discounts for users that are already using it when there is no reduced price. This is to ensure that they are able to use the least possible cost to generate the greatest profit.

7.3 Capitalize on Under-Addressed Market Potential

Recommendation #4: Attracting more female and student users

Based on the insights found, there are more male users than females users even though the Chicago population has a higher percentage of females. This shows that the female market is an untapped market potential for Ziro to capitalize on. Furthermore, the frequency of rentals by users below 25 years old was ranked 4th among the five age groups. This shows another market gap that Ziro can capitalize on. Hence, more can be done to attract females and younger adults because logically, they are the ones who are more active thus they would be a profitable segment to the bike-sharing business. A suggestion to attract students is to locate bike-sharing stations within the campus and offer student deals. Meanwhile for women, Ziro could implement deals that attract couples where we leverage on men to bring along their female companions to ride a bike.

Recommendation #5: Attracting more older riders (>40 years old)

To increase ridership from riders of older age, Ziro can conduct collaborations with the government. Similarly to Singapore's HPB Steps challenge, a possible suggestion is to collaborate with Chicago's Department of Health in providing a rewards system to encourage bike riding as an initiative of healthy living. This collaboration would be a win-win situation as it can help the authorities promote a healthier population.

8. Potential Areas of Improvement

Determining how to reallocate the bike from one station to another through the shortest possible route to ensure serviceability at all times is an important field in operations research. At certain points of the day, demand for bikes at some stations could fluctuate heavily which results in some stations not having sufficient supply of bikes. This could be detrimental to bike-users' satisfaction level of our service and potentially lead them to switch to other bike-operators and our company losing out on valuable customers. Through non-linear optimization methodology, our analysts could find the optimal number of bikes to allocate at each station whilst minimizing the cost of reallocation simultaneously.

9. Conclusion

To conclude, with the goal of hitting break-even within the shortest timeframe and maximizing profits while achieving cost reduction, Ziro's strategies will be guided with the exploratory insights and predictive models to optimize key performances. From the understanding of current market situation, future demand prediction, optimal bike allocation to marketing efforts, Ziro will successfully establish itself in this new expansion.

10. References

Divvy. (2019). *Index of bucket "divvy-tripdata".* Retrieved October 4, 2022, from: https://divvy-tripdata.s3.amazonaws.com/index.html

Helling, B. (2022, June 16). *Ziro Rideshare: Car options, availability, pricing and more. Ridester.com.* Retrieved October 4, 2022, from: https://www.ridester.com/ziro/

Infoplease. (2022). *Chicago, IL Demographic Statistics*. Retrieved November 1, 2022, from: https://www.infoplease.com/us/census/illinois/chicago/demographic-statistics

11. Appendice

Appendix A: Divvy Chicago Dataset (Column Description)

Columns	Description	Columns	Description
Dataset 1 (Trips Data)			
trip_id	Unique identifier of trip	*from_station_name*	Name of start station
start_time	Time the trip starts	*to_station_id*	Unique identifier of end station
end_time	Time the trip ends	*to_station_name*	Name of end station
bike_id	Unique identifier of bike	*usertype*	Subscriber or Customer
trip_duration	Total time the trip took (in seconds)	*gender*	*Only available for subscribers
from_station_id	Unique identifier of start station	*birthyear*	*Only available for subscribers
Dataset 2 (Stations Data)			
id	Unique identifier for stations	*longitude*	Coordinates
name	Name of the station	*dpcapacity*	The total number of docks at the station
latitude	Coordinates		

Appendix B: Scrapped Chicago Weather Dataset (Column Description)

Columns	Description	Columns	Description
Scrapped Dataset 1 (Weather Data)			
date	The date parameter is used to call the specific date request, using format "YYYYMMDD"	*windspeedLow*	Lowest Wind speed of the period
time	The timing at which the PWS takes the hourly reading	*windspeedAvg*	Wind speed average of the period
start_hour	The hour of the day which it takes the reading at	*windgustHigh*	Highest Wind gust of the period
end_hour	The end hour of the day which it takes the reading at	*windgustLow*	Lowest Wind gust of the period
tz	Time zone of PWS	*windgustAvg*	Wind gust average of the period
obsTimeLocal	Time observation is valid in local apparent time by timezone	*dewptHigh*	Maximum dew point of the period
lat	Latitude of PWS	*dewptLow*	Minimum dew point of the period
lon	Longitude of PWS	*dewptAvg*	Average dew point of the period
solarRadiationHigh	Highest Solar Radiation of the period	*windchillHigh*	High Windchill temperature of the period
uvHigh	Highest UV Index of the period	*windchillLow*	Low Windchill temperature of the period
winddirAvg	Wind direction average of the period	*windchillAvg*	Windchill average of the period

humidityHigh	Highest Humidity of the period	*heatindexHigh*	Heat index high temperature of the period
humidityLow	Lowest Humidity of the period	*heatindexLow*	Heat index low temperature of the period
humidityAvg	Average Humidity of the period	*heatindexAvg*	Heat index average of the period
qcStatus	Quality control indicator: -1: No quality control check performed 0: This observation was marked as possibly incorrect by our quality control algorithm 1: This observation passed quality control checks	*pressureMax*	Highest Barometric pressure in defined unit of measure of the period
tempHigh	High Temperature of the period	*pressureMin*	Lowest Barometric pressure in defined unit of measure of the period
tempLow	Low Temperature of the period	*pressureTrend*	Pressure tendency over the preceding period
tempAvg	Temperature average of the period	*precipRate*	Rate of precipitation - instantaneous precipitation rate. How much rain would fall if the precipitation intensity did not change for one hour
windspeedHigh	Highest Wind speed of the period	*precipTotal*	Accumulated Rain for the day from midnight to present in defined unit of measure

* PWS - Personal Weather Station

Appendix C: Features Engineered (Column Description)

Columns	Description	Columns	Description
Features Engineered			
day_of_week	The day of the week	Year_Date	The year of the date
is_weekday	Determine if the day is a weekday	Month_Date	The month of the year
hour_of_day	Hour of the day	Day_Date	The date of the month
y_m_d	Date in the format of “YYYY-MM-DD”	is_holiday	Determine if the day is a holiday (which includes school vacation breaks as holidays)
season	The season of the year which the date falls on: 1: Winter 2: Spring 3: Summer 4: Autumn	break_types	0: Not holiday 1: Spring break, Summer break or Winter break 2: Weekend 3: Others
rider_age	The age of the rider in 2019 (2019 - birthyear)		

Appendix D: Additional Analysis conducted on the Usertype Breakdown

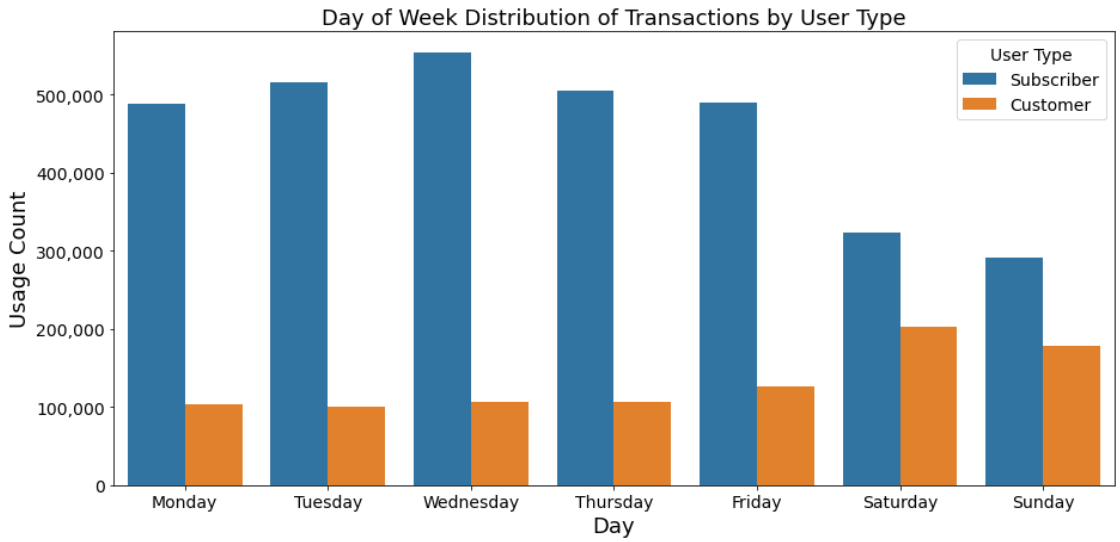


Figure 15: Bar Chart of Daily Demand by User Type

From Figure 15, it can be seen that generally throughout the week demand by subscribers are more than those of customer user type. However, looking deeper, demand by subscribers is high during the weekdays relative to the weekends and the reserve can be said about the customer user type. This would suggest that subscribers are using the rides as a means of transport during their working day more than on their rest day and customers use bike-sharing services as a leisure activity on their rest days.

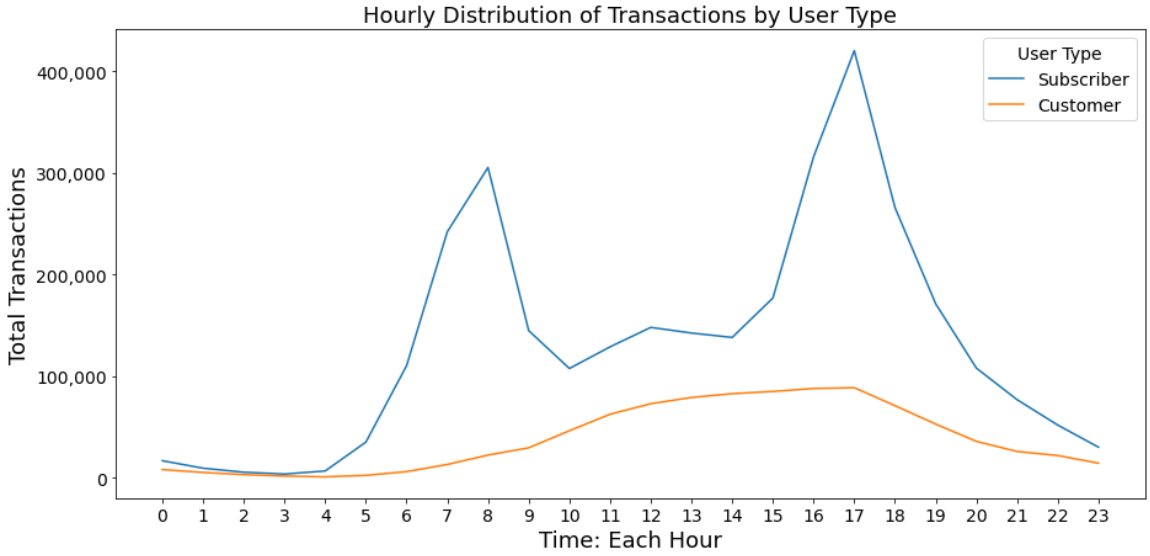


Figure 16: Line Chart of Hourly Demand by User Type

By looking into the hourly transaction pattern by user types, differentiating peak hours for subscribers and customers is noticed. For subscribers, the peak hours tend to be between 7 - 8am and 5 - 6pm. Whereas for customers, the surge in demand happens between 12 - 5pm. This further reinforces the theory proposed above.

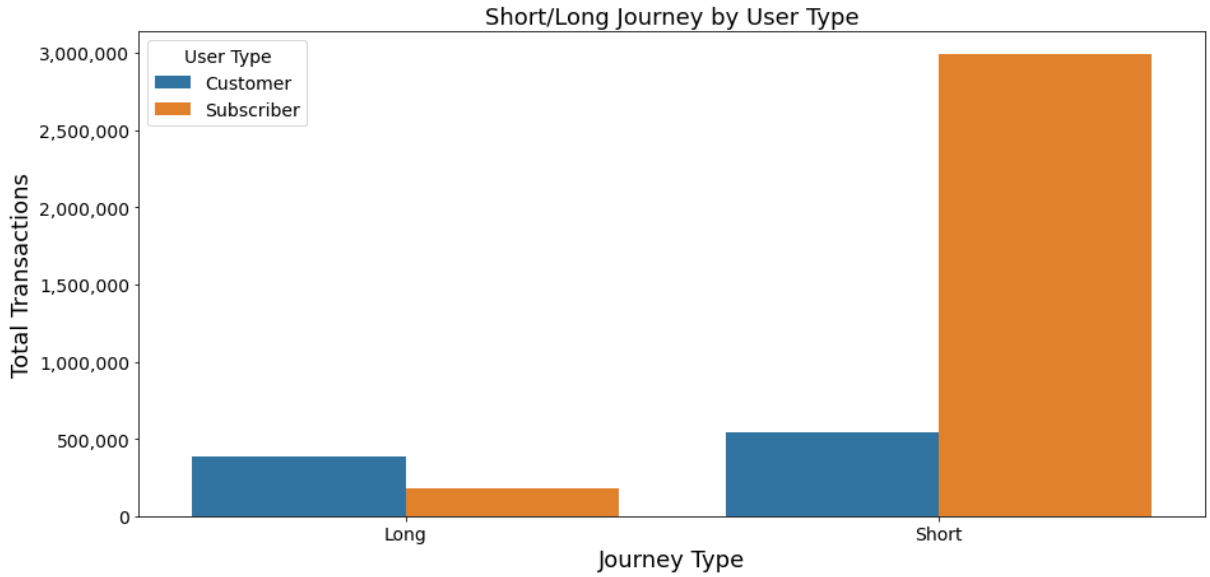
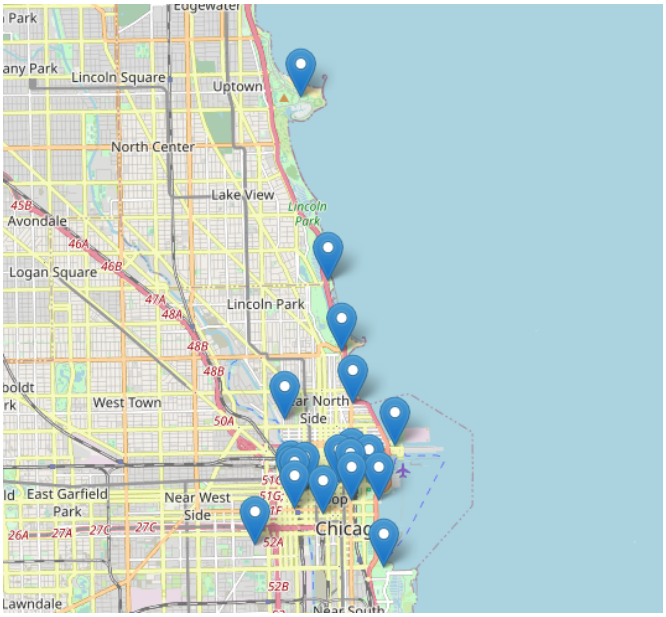


Figure 17: Bar Chart of Demand by Journey Type & User Type

When categorizing, trips into short or long journeys with short journeys being within 30 minutes. As such, Figure 17 illustrates that more subscribers are prone to use our service for shorter journeys (< 30mins) as compared to subscribers. Whereas more customers are prone to use our service for longer journeys (> 30mins) as compared to subscribers.

Appendix E: Name & Locations of the Top 20 Stations



Station Name			
1.	streeter dr & grand ave	11.	canal st & madison st
2.	lake shore dr & monroe st	12.	montrose harbor
3.	michigan ave & washington st	13.	clinton st & madison st
4.	michigan ave & oak st	14.	lasalle st & jackson blvd
5.	canal st & adams st	15.	michigan ave & lake st
6.	millennium park	16.	larrabee st & kingsbury st
7.	theater on the lake	17.	loomis st & lexington st
8.	columbus dr & randolph st	18.	morgan st & polk st
9.	clinton st & washington blvd	19.	lake park ave & 56th st
10.	shedd aquarium	20.	university ave & 57th st

Appendix F: Result Table of Negative Binomial Model on Overall Dataset

Best model has 17 Xs (Adjusted R^2 = 0.4953417307813692 , rank deficiency = 0):

Results: Generalized linear model

Model:	GLM	AIC:	94340.1475
Link Function:	log	BIC:	-54549.1546
Dependent Variable:	Demand	Log-Likelihood:	-47152.
Date:	2022-11-04 16:42	LL-Null:	-50071.
No. Observations:	7008	Deviance:	7346.0
Df Model:	17	Pearson chi2:	8.21e+03
Df Residuals:	6990	Scale:	1.0000
Method:	IRLS		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	5.5632	0.0493	112.8591	0.0000	5.4666	5.6598
solarRadiationHigh_sqrt	0.3023	0.0090	33.4221	0.0000	0.2846	0.3201
solarRadiationHigh	-0.0174	0.0009	-19.8636	0.0000	-0.0191	-0.0157
humidityAvg_sqar	-0.0002	0.0000	-21.6164	0.0000	-0.0002	-0.0001
solarRadiationHigh_sqar	0.0000	0.0000	11.2254	0.0000	0.0000	0.0000
dewptAvg_cbrt	0.0691	0.0083	8.2911	0.0000	0.0527	0.0854
month_date_8	1.0069	0.0477	21.1078	0.0000	0.9134	1.1004
month_date_7	0.9296	0.0488	19.0566	0.0000	0.8340	1.0252
break_types_2	-0.4937	0.0298	-16.5646	0.0000	-0.5521	-0.4353
month_date_9	1.0825	0.0477	22.6735	0.0000	0.9890	1.1761
precipRate_sqrt	-0.1950	0.0208	-9.3741	0.0000	-0.2358	-0.1542
windgustAvg_sqrt	0.1718	0.0204	8.4313	0.0000	0.1319	0.2118
month_date_11	0.4181	0.0465	8.9950	0.0000	0.3270	0.5092
break_types_3	-0.7845	0.0853	-9.1955	0.0000	-0.9517	-0.6173
month_date_6	0.8812	0.0479	18.3961	0.0000	0.7873	0.9751
month_date_10	0.8832	0.0457	19.3434	0.0000	0.7937	0.9726
month_date_5	0.5693	0.0474	12.0029	0.0000	0.4763	0.6622
precipTotal_sqrt	0.1221	0.0114	10.7187	0.0000	0.0998	0.1445

Descending order of 17 X's significance, assuming NegativeBinomial error distribution:

	Coefficient	z-stat
const	5.563214	112.859126
solarRadiationHigh_sqrt	0.302327	33.422074
month_date_9	1.082532	22.673543
humidityAvg_sqar	-0.000150	-21.616425
month_date_8	1.006892	21.107797
solarRadiationHigh	-0.017371	-19.863618
month_date_10	0.883153	19.343372
month_date_7	0.929581	19.056556
month_date_6	0.881201	18.396100
break_types_2	-0.493679	-16.564612
month_date_5	0.569256	12.002888
solarRadiationHigh_sqar	0.000012	11.225425
precipTotal_sqrt	0.122127	10.718748
precipRate_sqrt	-0.194992	-9.374110
break_types_3	-0.784511	-9.195501
month_date_11	0.418076	8.994959
windgustAvg_sqrt	0.171836	8.431319
dewptAvg_cbrt	0.069080	8.291056

Rank deficiency = 0: Df Model (17) is same as number of Xs (17).

Appendix G: Time Series Forecasting Model on Overall Dataset

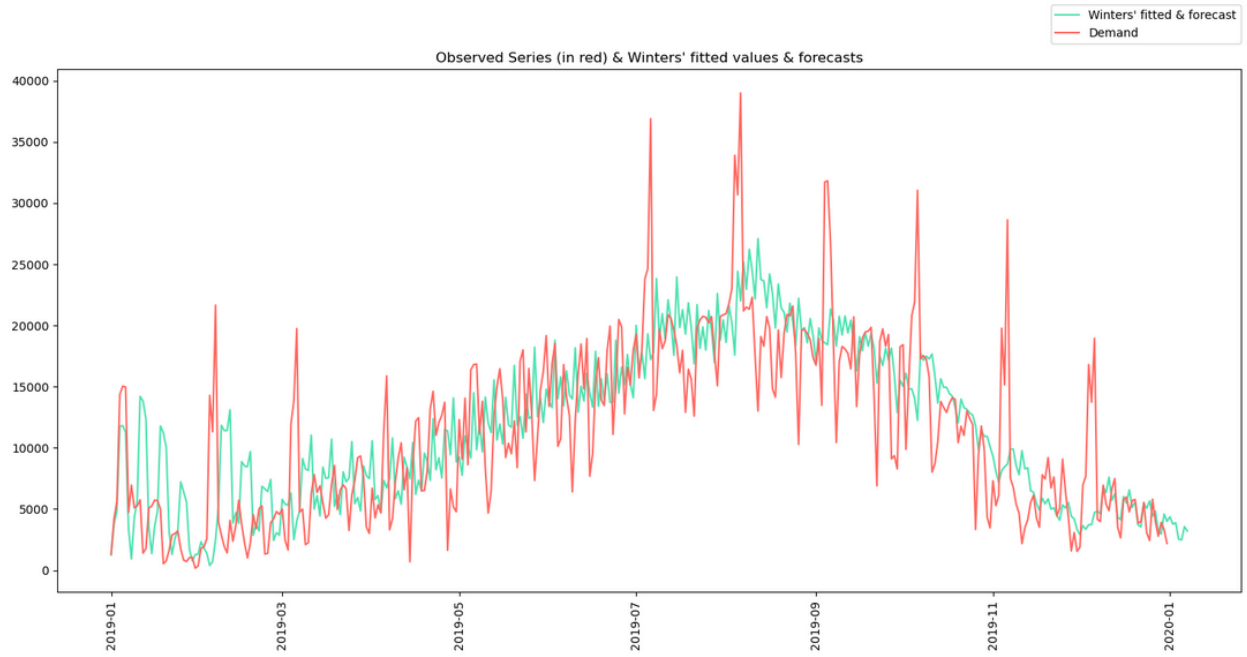
Winters' Triple Exponential Fit:

ExponentialSmoothing Model Results

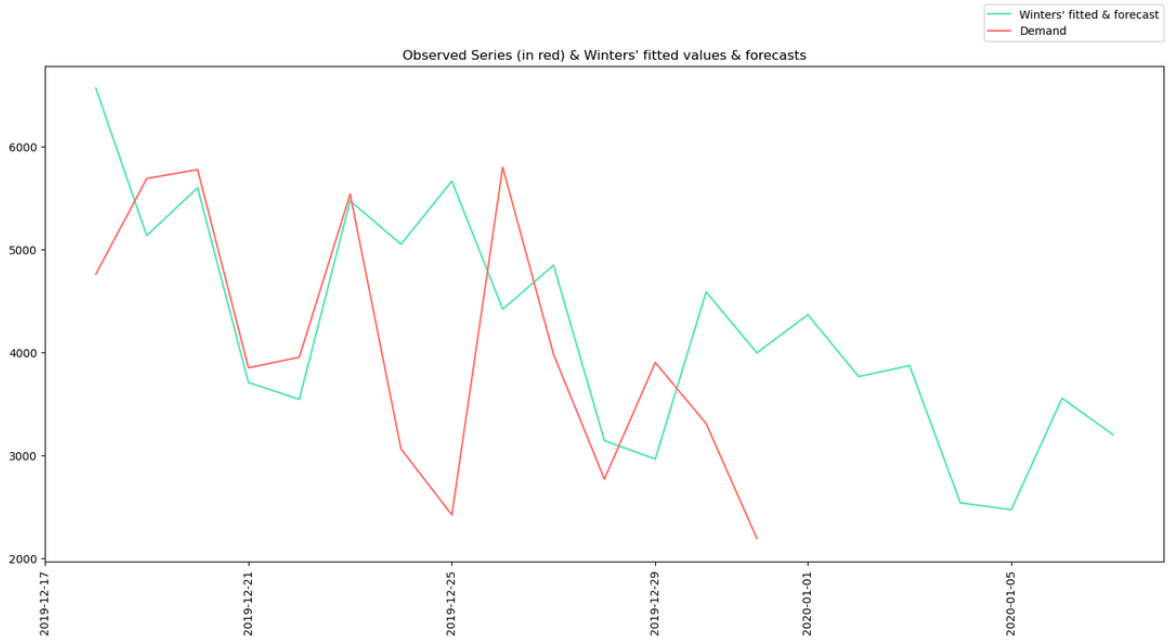
Dep. Variable:	Demand	No. Observations:	365
Model:	ExponentialSmoothing	SSE	9530181860.734
Optimized:	True	AIC	6255.409
Trend:	Additive	BIC	6298.308
Seasonal:	Multiplicative	AICC	6256.446
Seasonal Periods:	7	Date:	Fri, 04 Nov 2022
Box-Cox:	False	Time:	16:42:20
Box-Cox Coeff.:	None		

	coeff	code	optimized
smoothing_level	0.0757143	alpha	True
smoothing_trend	0.0540816	beta	True
smoothing_seasonal	0.0684656	gamma	True
initial_level	11610.717	l.0	True
initial_trend	-590.06122	b.0	True
initial_seasons.0	0.1129991	s.0	True
initial_seasons.1	0.3591509	s.1	True
initial_seasons.2	0.4838633	s.2	True
initial_seasons.3	1.2361855	s.3	True
initial_seasons.4	1.2947521	s.4	True
initial_seasons.5	1.2892399	s.5	True
initial_seasons.6	0.4074684	s.6	True

Last 365 observed (in red) & fitted values, & 7 forecasts:



Last 14 observed (in red) & fitted values, & 7 forecasts:



Last observation & 7 Winters' forecasts:

```
2019-12-31    2189.000000
2020-01-01    4366.903201
2020-01-02    3765.098227
2020-01-03    3871.809465
2020-01-04    2536.787903
2020-01-05    2470.365763
2020-01-06    3554.972728
2020-01-07    3200.010894
Freq: D, dtype: float64
```

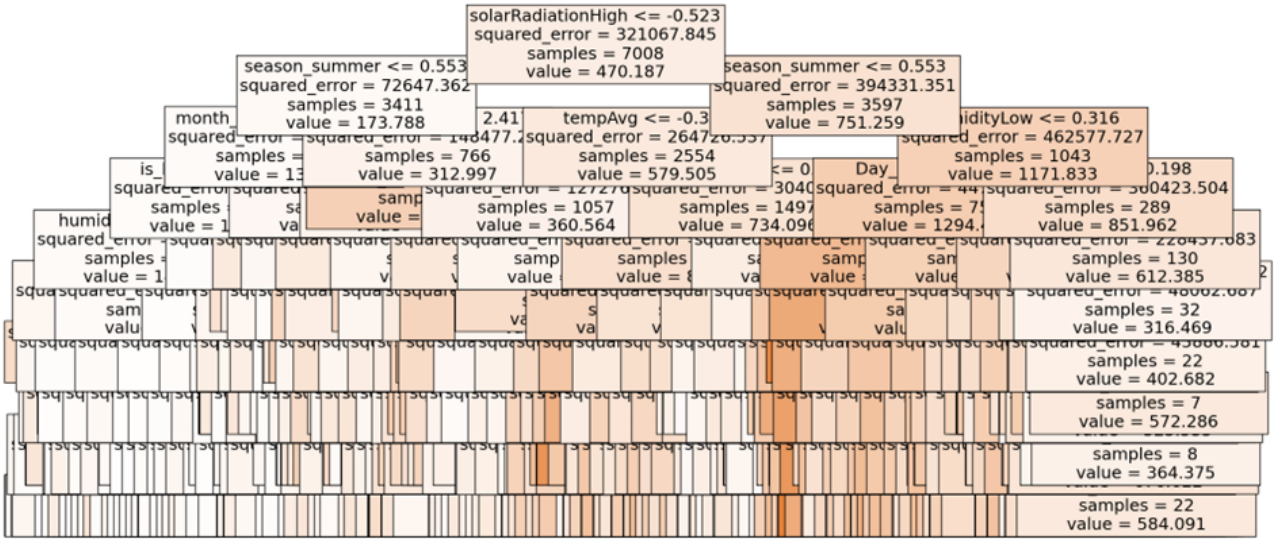
Date	Day
2020-01-01	Wednesday
2020-01-02	Thursday
2020-01-03	Friday
2020-01-04	Saturday
2020-01-05	Sunday
2020-01-06	Monday
2020-01-07	Tuesday

Pseudo-R2 (52.3%)

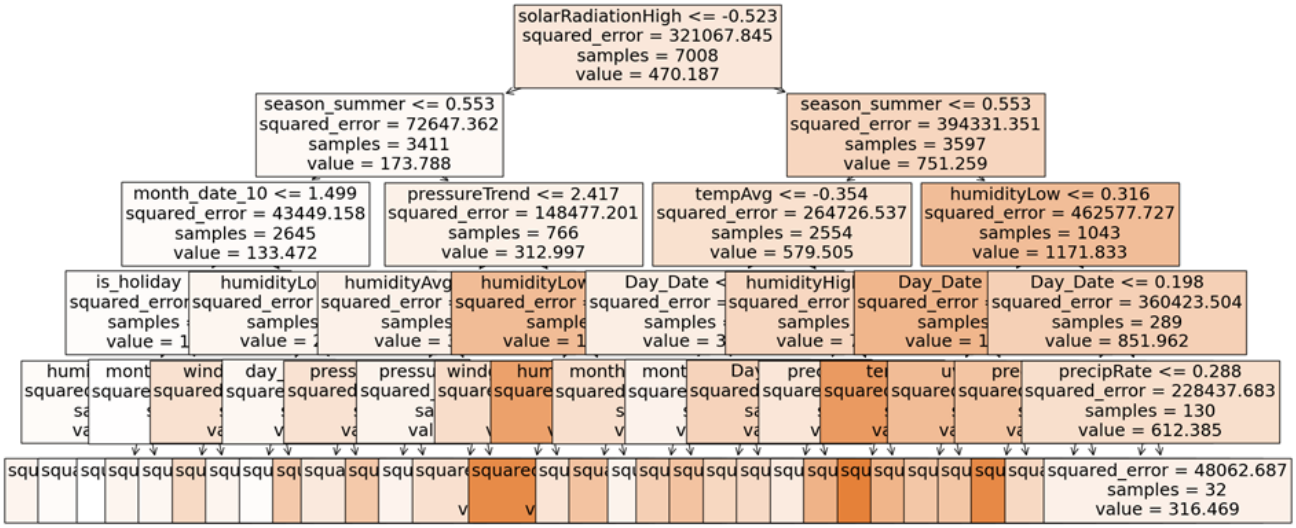
```
np.square(mf.fittedvalues.corr(dt))
```

```
0.5231932923581092
```

Appendix H: Decision Tree Model on Overall Dataset

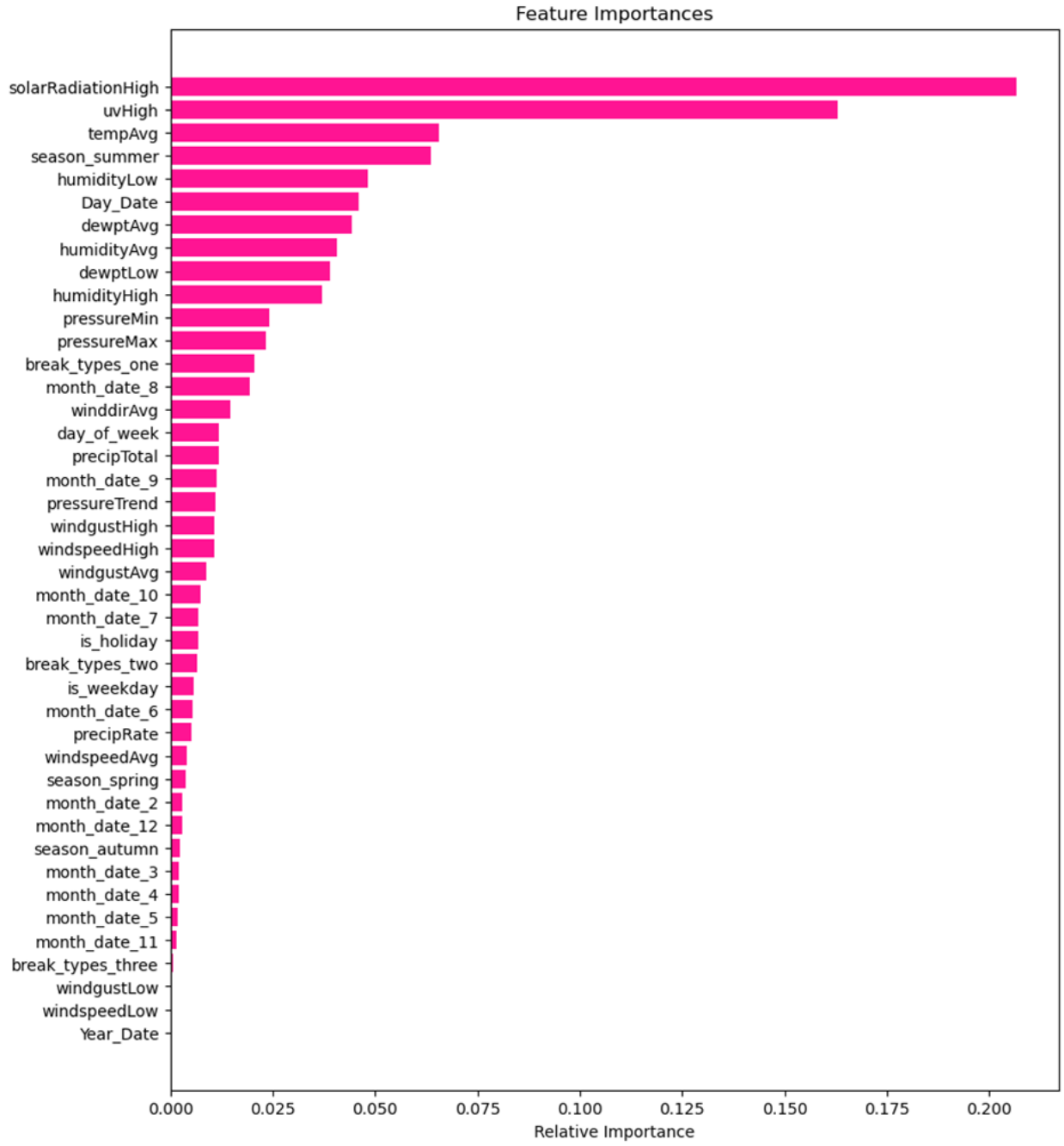


Decision Tree Model without Regularization



Decision Tree Model with Regularization

Appendix I: Random Forest Model on Overall Dataset



Appendix J: Overview of Model & Metrics on Overall Dataset

	Model	MSE	r^2
0	XGBoost	95710.328158	0.697013
1	Random Forest	114653.210122	0.637046
2	Voting Regressor	118730.451512	0.624139
3	Decision Tree	134756.654075	0.573406
4	RF Bagging	163216.176351	0.557005
5	Neural Network	139937.357591	0.483312
6	SVR	190506.061328	0.396922
7	AdaBoost	190603.710394	0.396613

Appendix K: Result Table of Negative Binomial Model on Top 20 Stations Dataset

Best model has 12 Xs (Adjusted R^2 = 0.5069413691329274 , rank deficiency = 0):

Results: Generalized linear model

Model:	GLM	AIC:	53343.9501
Link Function:	Log	BIC:	-55400.9099
Dependent Variable:	Demand	Log-Likelihood:	-26659.
Date:	2022-11-04 17:09	LL-Null:	-29779.
No. Observations:	7008	Deviance:	6538.5
Df Model:	12	Pearson chi2:	9.89e+03
Df Residuals:	6995	Scale:	1.0000
Method:	IRLS		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	2.5590	0.0815	31.4104	0.0000	2.3993	2.7187
solarRadiationHigh_sqrt	0.3642	0.0093	39.2906	0.0000	0.3460	0.3823
solarRadiationHigh	-0.0207	0.0009	-23.3242	0.0000	-0.0224	-0.0189
solarRadiationHigh_sqar	0.0000	0.0000	14.4555	0.0000	0.0000	0.0000
humidityAvg	-0.0140	0.0010	-14.4357	0.0000	-0.0159	-0.0121
month_date_8	1.1267	0.0465	24.2501	0.0000	1.0356	1.2177
month_date_7	1.1527	0.0474	24.3312	0.0000	1.0599	1.2456
windgustAvg_sqrt	0.1465	0.0208	7.0525	0.0000	0.1058	0.1872
month_date_9	1.0479	0.0474	22.1024	0.0000	0.9549	1.1408
month_date_6	0.7127	0.0476	14.9840	0.0000	0.6195	0.8060
month_date_5	0.6235	0.0476	13.0939	0.0000	0.5302	0.7168
month_date_4	0.4590	0.0482	9.5214	0.0000	0.3645	0.5535
month_date_10	0.6187	0.0463	13.3634	0.0000	0.5280	0.7094

Descending order of 12 X's significance, assuming NegativeBinomial error distribution:

	Coefficient	z-stat
const	2.558987	31.410373
solarRadiationHigh_sqrt	0.364152	39.290581
month_date_7	1.152727	24.331208
month_date_8	1.126681	24.250124
solarRadiationHigh	-0.020676	-23.324185
month_date_9	1.047852	22.102390
month_date_6	0.712741	14.984024
solarRadiationHigh_sqar	0.000015	14.455525
humidityAvg	-0.013964	-14.435701
month_date_10	0.618695	13.363352
month_date_5	0.623507	13.093856
month_date_4	0.459019	9.521391
windgustAvg_sqrt	0.146506	7.052536

Rank deficiency = 0: Df Model (12) is same as number of Xs (12).

Appendix L: Time Series Forecasting Model on Top 20 Stations Dataset

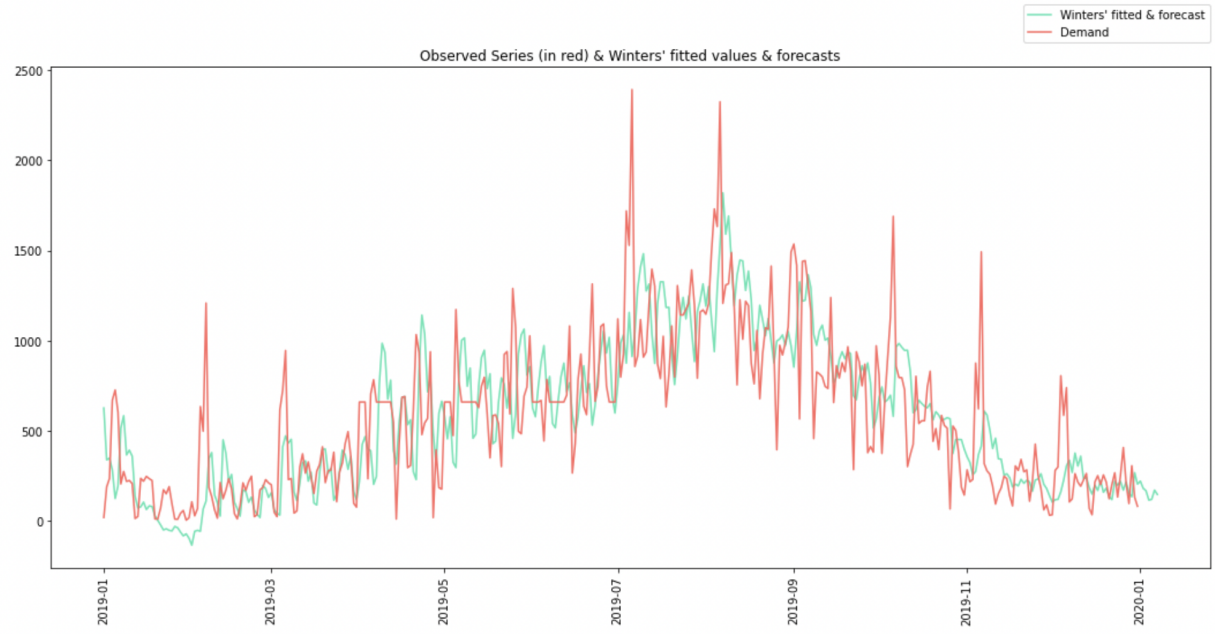
Winters' Triple Exponential Fit:

ExponentialSmoothing Model Results

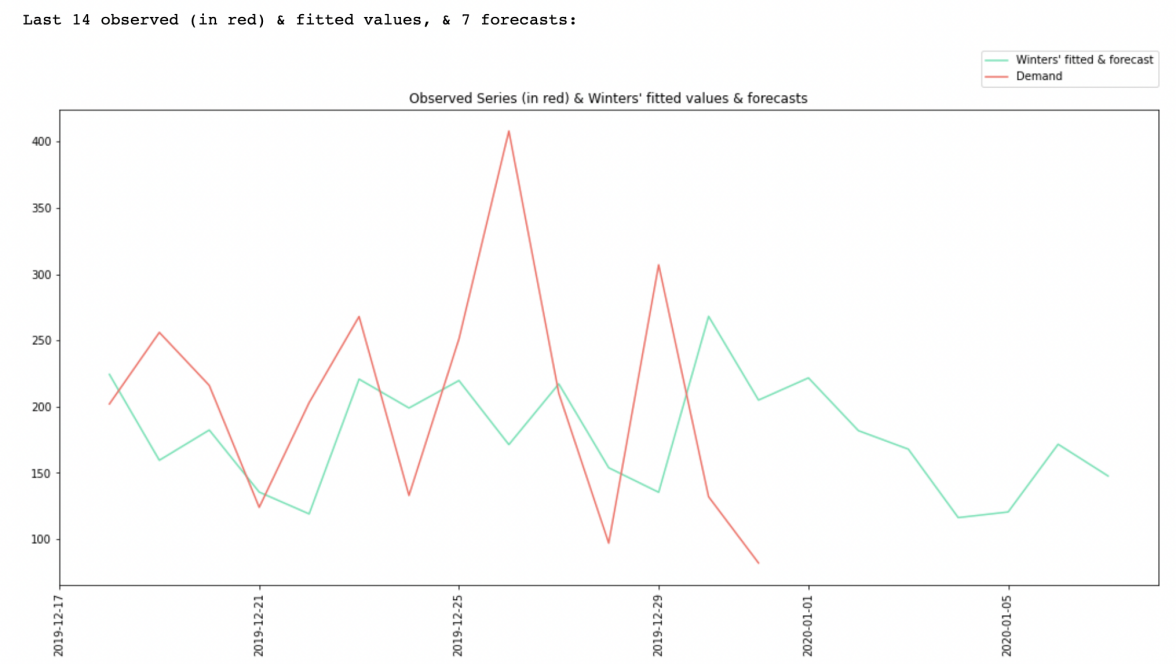
Dep. Variable:	Demand	No. Observations:	365
Model:	ExponentialSmoothing	SSE	32791577.145
Optimized:	True	AIC	4185.111
Trend:	Additive	BIC	4228.010
Seasonal:	Multiplicative	AICC	4186.149
Seasonal Periods:	7	Date:	Fri, 04 Nov 2022
Box-Cox:	False	Time:	17:09:49
Box-Cox Coeff.:	None		

	coeff	code	optimized
smoothing_level	0.1817857	alpha	True
smoothing_trend	0.0151488	beta	True
smoothing_seasonal	0.0629396	gamma	True
initial_level	482.13333	l.0	True
initial_trend	-34.813853	b.0	True
initial_seasons.0	1.4015704	s.0	True
initial_seasons.1	1.0210314	s.1	True
initial_seasons.2	1.2941408	s.2	True
initial_seasons.3	1.3079371	s.3	True
initial_seasons.4	0.5324615	s.4	True
initial_seasons.5	0.4767999	s.5	True
initial_seasons.6	0.9660589	s.6	True

Last 365 observed (in red) & fitted values, & 7 forecasts:



Last 14 observed (in red) & fitted values, & 7 forecasts:




```
Last observation & 7 Winters' forecasts:

2019-12-31     82.000000
2020-01-01    221.680400
2020-01-02    181.868362
2020-01-03    167.900806
2020-01-04    116.225739
2020-01-05    120.544653
2020-01-06    171.622374
2020-01-07    147.669736
Freq: D, dtype: float64
```

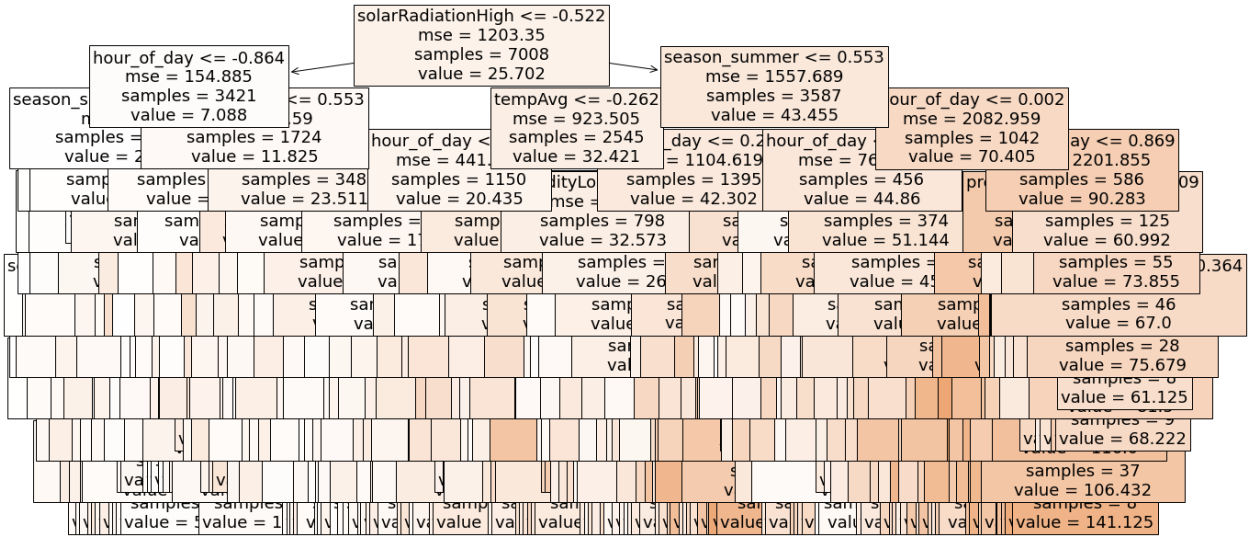
Date	Day
2020-01-01	Wednesday
2020-01-02	Thursday
2020-01-03	Friday
2020-01-04	Saturday
2020-01-05	Sunday
2020-01-06	Monday
2020-01-07	Tuesday

Pseudo-R2 (55.0%)

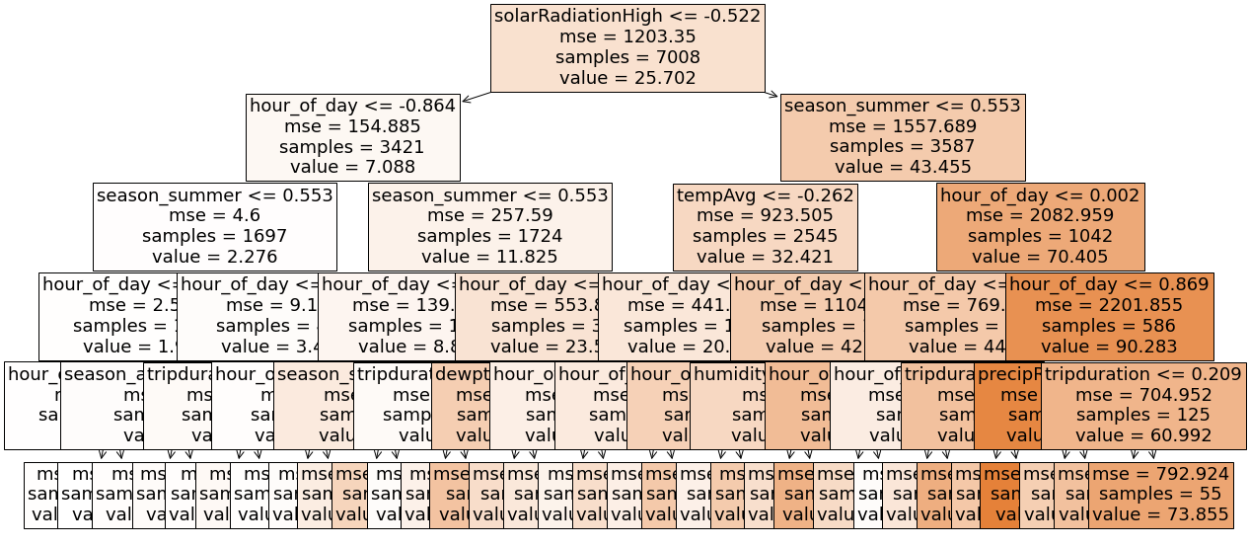
```
np.square(mf.fittedvalues.corr(dt))
```

```
0.5497959639717654
```

Appendix M: Decision Tree Model on Top 20 Stations Dataset

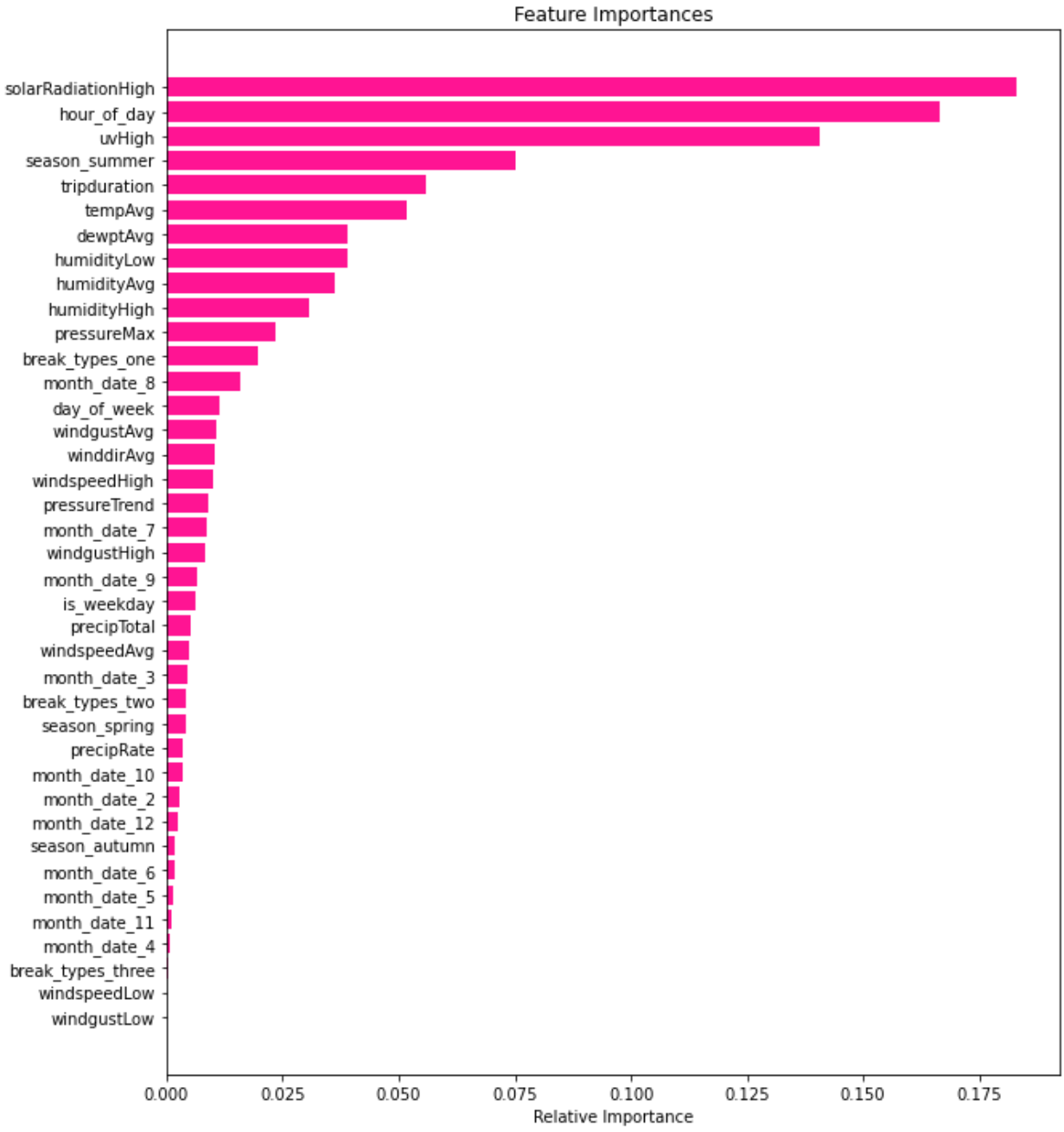


Decision Tree Model without Regularization



Decision Tree Model with Regularization

Appendix N: Random Forest Model on Top 20 Stations Dataset



Appendix O: Overview of Model & Metrics on Top 20 Stations Dataset

	Model	MSE	r^2
0	Voting Regressor	285.019202	0.737086
1	Decision Tree	316.245190	0.708282
2	Random Forest	317.764436	0.706881
3	RF Bagging	481.156458	0.685662
4	SVR	362.934318	0.665214
5	XGBoost	411.658267	0.620269
6	Neural Network	340.767143	0.556161
7	AdaBoost	539.819708	0.502048