To: Aaron Goldbeck & Capital Bikeshare's Senior Management Team
From: Jordan Bandy, Corinne Doty, & Rachael Hensley
Topic: Analyzing Bike Rental Demand

Background

The purpose of this report is to analyze bike rental demand for Capital Bikeshare. In order to make informed decisions about where and when to supply bikes for the bike share program, it is imperative to understand what factors affect the demand for the bikes and be able to accurately predict bike rental demand. Our analysis dives into seasonality, key factors, customer differentiation, and prediction of bike rental demand.

Previous research has provided us with some insight into month-to-month trends in demand suggesting some element of seasonality. Our research further supports this idea that there is a seasonality element to the bike rental demand. Further, our team analyzed what specific factors on any given day might affect bike rental demand. Mr. Goldbeck was aware that the demand for commuters over any given day was fairly equal, however we were able to dig deeper into this topic to determine external factors that would change the anticipated demand for a day.

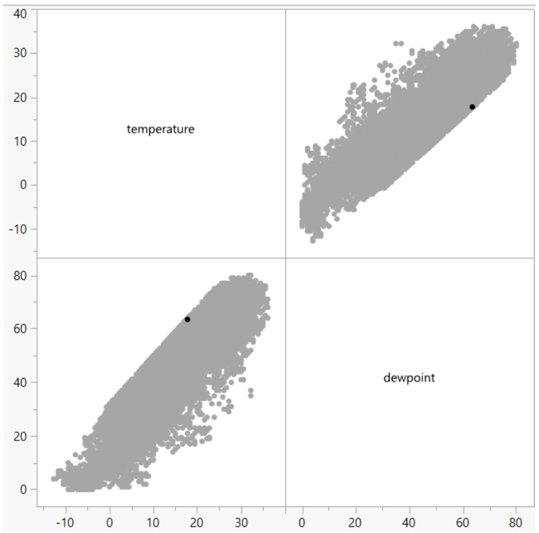
Data Preprocessing

Prior to starting the analysis, we had to perform some preprocessing steps on the data we were given. We first created dummy variables for the "seasons" variable and changed all of the time factors to categorical variables. Additionally, we created a column to account for the number of casual riders each hour. For both the member and casual riders, we created lagged columns in order to have past data available for prediction. Without creating these lagged columns, we would have factors that are partial leakers as they would provide data we would not normally have ahead of time. By lagging this column, we can use how many registered user rentals or casual rentals were initiated the previous hour.

Analysis

Factor Importance

In order to determine the factors that most affected Capital Bikeshare's bike rental demand, we set up various models to determine the feature importance. Both linear and non-linear models were chosen to model this scenario. It was important to remove any multicollinearity from our models. The identified multicollinearity can be seen below.



Using the Multivariate method in JMP, we created scatterplots and correlation matrices for every variable pair to identify multicollinearity within our data set. With a correlation of 0.8955, we found the most problematic instance of multicollinearity to be with dewpoint and temperature. Knowing this, we checked the VIF of these variables when creating our models and always removed the one with the greater VIF in order to remove this multicollinearity from our model.

Linear Model

After removing dewpoint due to introducing multicollinearity into our model (VIF = 131), we implemented back propagation to determine the significant factors. The linear model produces nine significant variables1: [lag]member, temperature, humidity, weekend, summer, precipitation accumulation, cloudy, pressure, and fall. These variables are significant with winter as the “season” base case and fair as the “weather” base case when we consider the data set as a whole.

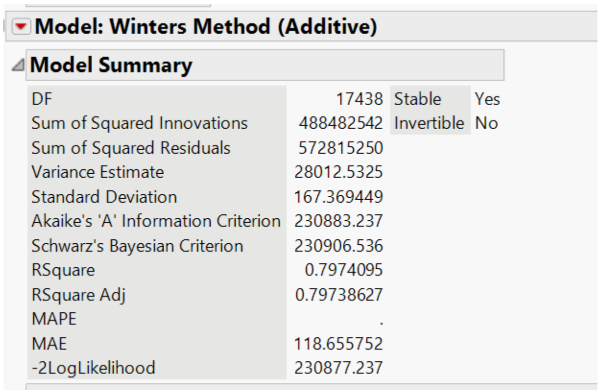
However, we know that there are two different groups of people who utilize the bike share program: members and casual riders. Members are far more regular riders, often those using Capital Bikeshare’s program for commutes to work, while casual riders utilize the bike share program less regularly. Knowing that we have these two classes of people, and that members are primarily commuters with regular schedules, we wanted to analyze the trends of the casual riders. Specifically, we wanted to gain some insight on how external factors affected the bike rentals of casual riders.

[1] Insignificant, and therefore removed variables, included drizzle, spring, precipitation, wind gust, and heavy rain.

When controlling for season and whether it was a weekend, we found multiple significant external factors that affected these riders. When it is drizzling or there is heavy rain, there are approximately 10 fewer bike rentals by casual riders each hour than there would be if the weather was fair. Windspeed and wind gusts also have a negative effect on casual riders' bike rentals, but they cause less than 1 rental fewer each hour. The final external factor we found to significantly affect casual rider's bike rentals was temperature. Contrary to those variables describing poorer weather, for every three-degree increase in the temperature, Capital Bikeshare can expect one more casual bike rider rental every hour. As far as Goldbeck's concern for timing throughout the day, the highest demand for bikes from casual riders happens from 9:00-11:00am.

Predictive Modeling

In order to determine the number of bike rentals, we built a square root model, a linear regression, and compared multiple times series methods. The time-series methods we conducted were winters method, linear exponential smoothing, and simple exponential smoothing which were all unconstrainted. We are looking for the model with the smallest mean absolute error (MAE) as MAE represents the average absolute errors between the predicted and observed values when comparing the time series models. Thus, measuring the average magnitude of the errors forecasted. Comparing all three methods, winter's method had the smallest MAE number of 118.656. This means that on average, the model's predictions differ from the true values by about 118.655 rentals. When producing this model, we found that having the observations per period set to 24 produced the lowest MAE number. Following winter's method was simple exponential smoothing with a MAE of 124.309 and closely behind was linear exponential smoothing with a MAE value of 124.334.

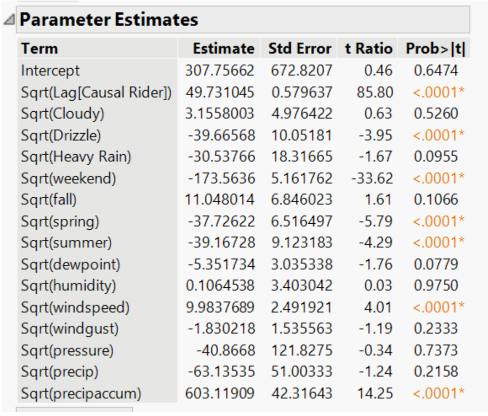


Model: Winters Method (Additive)

Model Summary

DF	17438	Stable	Yes
Sum of Squared Innovations	488482542	Invertible	No
Sum of Squared Residuals	572815250		
Variance Estimate	28012.5325		
Standard Deviation	167.369449		
Akaike's 'A' Information Criterion	230883.237		
Schwarz's Bayesian Criterion	230906.536		
RSquare	0.7974095		
RSquare Adj	0.79738627		
MAPE	.		
MAE	118.655752		
-2LogLikelihood	230877.237		

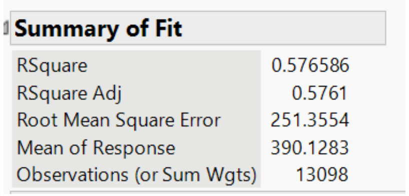
To start building the linear regression and square root model, we created a validation column that had 75% as training and 25% as testing. This means that out of the 17,466 rows, 13,099 are a part of the training dataset and the rest dedicated to the testing. When building the square root model, we included the following variables:



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	307.75662	672.8207	0.46	0.6474
Sqrt(Lag[Causal Rider])	49.731045	0.579637	85.80	<.0001*
Sqrt(Cloudy)	3.1558003	4.976422	0.63	0.5260
Sqrt(Drizzle)	-39.66568	10.05181	-3.95	<.0001*
Sqrt(Heavy Rain)	-30.53766	18.31665	-1.67	0.0955
Sqrt(weekend)	-173.5636	5.161762	-33.62	<.0001*
Sqrt(fall)	11.048014	6.846023	1.61	0.1066
Sqrt(spring)	-37.72622	6.516497	-5.79	<.0001*
Sqrt(summer)	-39.16728	9.123183	-4.29	<.0001*
Sqrt(dewpoint)	-5.351734	3.035338	-1.76	0.0779
Sqrt(humidity)	0.1064538	3.403042	0.03	0.9750
Sqrt(windspeed)	9.9837689	2.491921	4.01	<.0001*
Sqrt(windgust)	-1.830218	1.535563	-1.19	0.2333
Sqrt(pressure)	-40.8668	121.8275	-0.34	0.7373
Sqrt(precip)	-63.13535	51.00333	-1.24	0.2158
Sqrt(precipaccum)	603.11909	42.31643	14.25	<.0001*

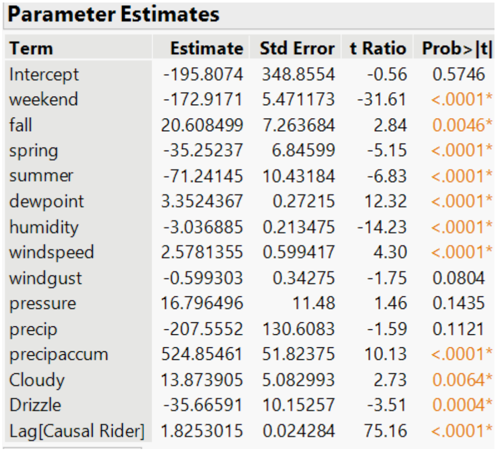
The only variable we did not keep was temperature as we had dewpoint already and they are showing very similar data points and introducing multicollinearity into our model. We included lag[Casual Rider] as we are more focused on investigating the number of bikes sold for the causal riders as they are going to be the group that is more so targeted for the promotion. When looking more closely at this model, we can see that the root mean square value (RMSE) is 251.355. This represents the differences between observed and predicted values. Thus, the lower the RMSE the better the fit. This suggests that on average, the square root model is 251.355 rentals away from the actual observed value.



Summary of Fit

RSquare	0.576586
RSquare Adj	0.5761
Root Mean Square Error	251.3554
Mean of Response	390.1283
Observations (or Sum Wgts)	13098

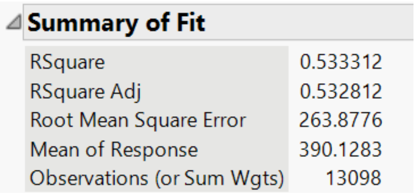
Following this model, we produced a linear regression model with the following features (same as the square root model):



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-195.8074	348.8554	-0.56	0.5746
weekend	-172.9171	5.471173	-31.61	<.0001*
fall	20.608499	7.263684	2.84	0.0046*
spring	-35.25237	6.84599	-5.15	<.0001*
summer	-71.24145	10.43184	-6.83	<.0001*
dewpoint	3.3524367	0.27215	12.32	<.0001*
humidity	-3.036885	0.213475	-14.23	<.0001*
windspeed	2.5781355	0.599417	4.30	<.0001*
windgust	-0.599303	0.34275	-1.75	0.0804
pressure	16.796496	11.48	1.46	0.1435
precip	-207.5552	130.6083	-1.59	0.1121
precipaccum	524.85461	51.82375	10.13	<.0001*
Cloudy	13.873905	5.082993	2.73	0.0064*
Drizzle	-35.66591	10.15257	-3.51	0.0004*
Lag[Causal Rider]	1.8253015	0.024284	75.16	<.0001*

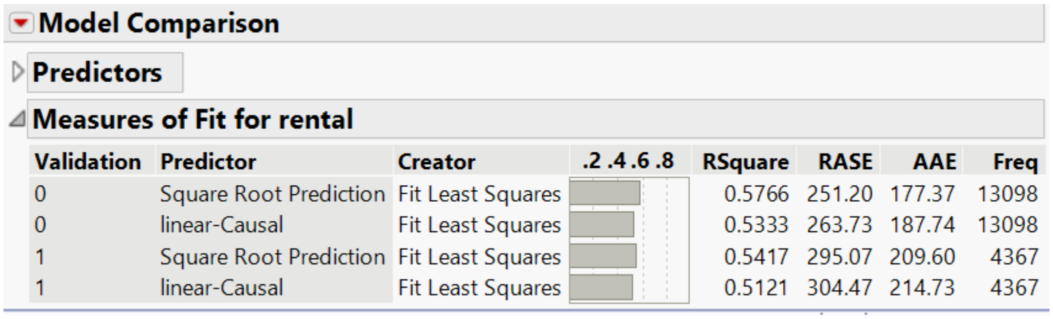
Similar to the square root model, we are looking to see what RMSE this model produces. The RSME equated to 263.878. This suggests that on average, the linear regression's predictions are approximately 283.878 rentals away from the actual observed values.



Summary of Fit

RSquare	0.533312
RSquare Adj	0.532812
Root Mean Square Error	263.8776
Mean of Response	390.1283
Observations (or Sum Wgts)	13098

Model comparison: When looking to choose a predictive model between the linear regression and square root model, a model comparison must be done. In order to produce a model comparison, we need to save the prediction formula columns for both models. This is imperative when building the model comparison as you are comparing the two prediction columns and having them grouped by the validation column. After running this comparison, we compared RASE numbers for when validation equals one. The RASE represents the square root of the average of the squared differences between the observed and predicted values thus lower values will indicate a better model. The square root prediction has a RASE of 295.07 compared to the linear regression prediction which has a RASE of 304.47. This leads us to the conclusion that the square root model prediction has a better model accuracy than linear regression. This is given that the RASE and RSME are smaller in each transformation.



Model Comparison

Predictors

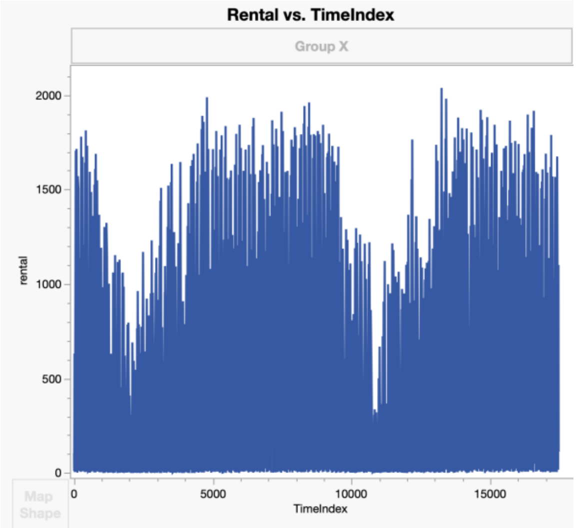
Measures of Fit for rental

Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
0	Square Root Prediction	Fit Least Squares		0.5766	251.20	177.37	13098
0	linear-Causal	Fit Least Squares		0.5333	263.73	187.74	13098
1	Square Root Prediction	Fit Least Squares		0.5417	295.07	209.60	4367
1	linear-Causal	Fit Least Squares		0.5121	304.47	214.73	4367

Seasonality

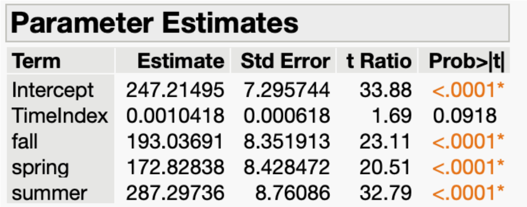
As a bike business, it is important for Capital Bikeshare to consider seasonality to determine the best time to launch promotion strategies. First, we investigated riders' patterns within different seasons for the past two years for trend analysis. Following this, we created a seasonality report to determine how different seasons affect riders.

In order to look at the patterns within the different seasons for the past two years, we utilized the graph builder to see any trend lines. The dataset starts in October and continues throughout the months into the next years. So, after October you see a decline in rentals into the end of the year and beginning of the next. This trendline repeats itself going into the following years with an incline of rentals and again another decline. Based on this graph we can see seasonality occurring through the time index, it is just difficult to see where exactly the different months are located which is why we then ran seasonality reports for more in-depth analysis.



For the seasonality report, we wanted to cover our basis, so we analyzed the different member types such as casual riders, members, and overall riders as well as different seasonality types to consider such as seasons and months. After doing so, we found that all member types had similar seasonality reports and therefore chose to focus on the overall rental numbers since the final recommendation would have been the same from casual to member riders.

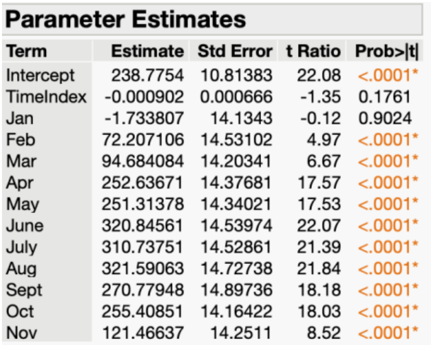
For the first seasonality report, we analyzed the number of rentals and the four seasons of Fall, Spring, Summer, and Winter for a broader beginning range. To do so, we created a time series with rentals being the dependent variable, and TimeIndex, Fall, Spring, and Summer being the parameters with Winter being the base case. Based on the chart below, there is seasonality within the four seasons. Fall, Spring, and Summer are considered significant meaning the number of riders in these seasons were significantly different than those in Winter. It is fair to assume that with Winter comes less bike riders, but it is best to observe that through analysis just in case. If we stopped here, we would advise promotions to be in the Winter months for all riders, but we chose to dig a little deeper to be more precise.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	247.21495	7.295744	33.88	<.0001*
TimeIndex	0.0010418	0.000618	1.69	0.0918
fall	193.03691	8.351913	23.11	<.0001*
spring	172.82838	8.428472	20.51	<.0001*
summer	287.29736	8.76086	32.79	<.0001*

The four seasons can be strong with seasonality, but so can other forms of timing such as months if we are looking to be more specific. Below is a seasonality report for the total number of rentals for the months of January through December with December being the base case. Based on the results, we can conclude that there is seasonality, and that February through March are significant against the base case. This means that the number of rentals for these months are all significantly different than December. We are also able to determine that these months are larger than December due to the positive coefficients with August being the most significantly different and largest. January, however, is not significantly different from December, having a large p-value of 0.9024. January also has a coefficient of –1.7338 meaning it has lower rentals than December. Although the time index is not significant, the time series still shows a strong form of seasonality with the months.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	238.7754	10.81383	22.08	<.0001*
TimeIndex	-0.000902	0.000666	-1.35	0.1761
Jan	-1.733807	14.1343	-0.12	0.9024
Feb	72.207106	14.53102	4.97	<.0001*
Mar	94.684084	14.20341	6.67	<.0001*
Apr	252.63671	14.37681	17.57	<.0001*
May	251.31378	14.34021	17.53	<.0001*
June	320.84561	14.53974	22.07	<.0001*
July	310.73751	14.52861	21.39	<.0001*
Aug	321.59063	14.72738	21.84	<.0001*
Sept	270.77948	14.89736	18.18	<.0001*
Oct	255.40851	14.16422	18.03	<.0001*
Nov	121.46637	14.2511	8.52	<.0001*

Conclusion & Recommendations

After analyzing Capital Bikeshare's bike rental demand from the given dataset, we were able to produce strong insights and recommendations for the company.

We analyzed member and casual riders separately, as we know member rides have a more consistent and predictable rental pattern. With member riders often being commuters who use bikes before and after work, casual riders ride the most frequently in the late morning and are much less likely to rent in any type of rainy weather. Knowing that between 7:00pm-12:00am is the slowest casual rental time of the day, Capital Bikeshare can use this time if they ever need to perform maintenance on the bikes or have fewer rentals available for any reason, in order to minimize loss.

Based on our trend analysis, ridership patterns show there is a similar trend within the different seasons. Our findings show that towards the end of the year and into the beginning of the next year, there is a decline in bike rentals. Then, following that decline, the bike rentals begin to have higher demand and therefore have an incline. This trend continues to repeat itself within the two years of data that the dataset provided. Breaking it down further, we were also able to discover seasonality within the bike rentals. We pin-pointed that bike rentals are lowest in January, second lowest in December, and third lowest in February. Based on these findings, we recommend to Capital Bikeshare that you offer a promotion to all bike renters within the winter months, specifically December leading into January. With these months holding the lowest numbers, a strong promotion could potentially help boost demand and encourage riders to use the bikes.

We hope that with the analysis and recommendations provided in this report, the Capital Bikeshare company can further refine their strategic expansion plan and maintain profitability while continuing to promote sustainability in Washington DC.