Bike sharing case study

Naren Reddy palupunoori

Table of Contents

[Problem Statement 2](#_Toc100706985)

[Analysis 2](#_Toc100706986)

[Relation analysis 2](#_Toc100706987)

[Season and weather impact 3](#_Toc100706988)

[Correlation 4](#_Toc100706989)

[Model training 7](#_Toc100706990)

[Final Model analysis 8](#_Toc100706991)

[Conclusions 8](#_Toc100706992)

[Suggestions 9](#_Toc100706993)

[Confidence level 9](#_Toc100706994)

[Final equations of model 10](#_Toc100706995)

[Subjective 10](#_Toc100706996)

[Assignment-based Subjective Questions 10](#_Toc100706997)

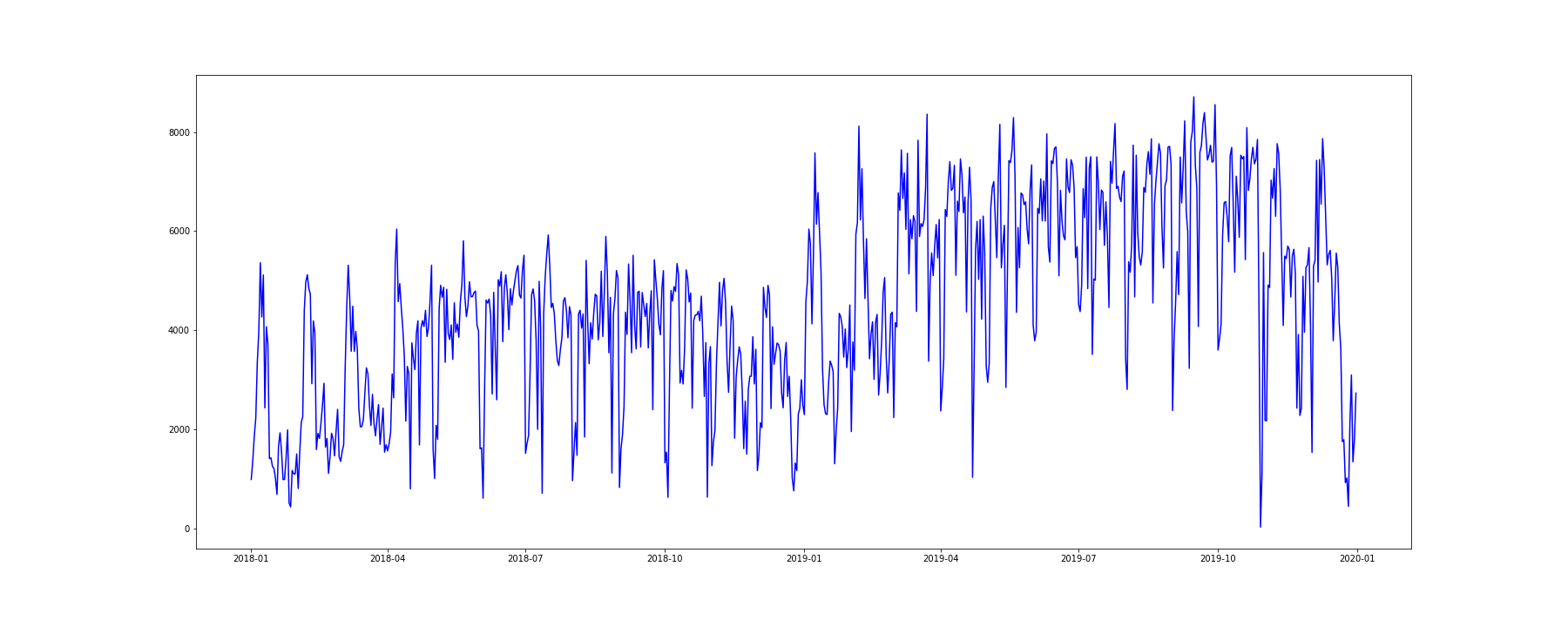
[General Subjective Questions 15](#_Toc100706998)

# Problem Statement

Understand the demand for bike sharing for given customers in the year of 2018 and 2019, to suggest marketing and predict demand for upcoming years post pandemic.

Understand variables effecting the bike demanding and how much these variables are affecting the demand.

# Analysis



* Demand for bike sharing has been steadily increasing from 2018 to 2019.

## Relation analysis

* Season 1, spring has negative effect on bike sharing usage.
* Bike sharing has drastic uptick in 2019.
* Weekday and working day also have minimal effect.
* Snow has very drastic effect on bike sharing usage.

A picture containing text, room

Description automatically generated

## Season and weather impact

There are two major observations from above graphs

1. Season 1, spring has negative effect on bike sharing usage, irrespective of the weather.
2. Snow has a negative effect on bike sharing irrespective of the season.

Conclusion here irrespective of seasons and weather, Snow and Spring has negative effect.

A screenshot of a computer

Description automatically generated with low confidence

## Correlation

* `cnt` is highly correlated to `registered` and `casual` which makes sense, `cnt` = `registered` + `casual`
* `atemp` and `temp` are highly dependent on the bike sharing. Understandable, based on temperature customers are deciding whether to use bike sharing or not.
* Also, there is some linear relation between `atemp`, `temp` and `cnt` so linear regression can be used here
* `season` and `weathersit` are also correlated to `cnt` as we saw `spring` and `snow` have high effect on customer use of bike sharing.

A picture containing text

Description automatically generated

Qr code

Description automatically generated with low confidence

A picture containing table

Description automatically generated

# Model training

Table

Description automatically generated

Doing RFE model training as giving us a starting set of features. Analyzing the feature dependency,

* Holiday has very low impact on register customers but has a big effect on casual customers. As a result, there is some impact on overall demand.
* Even though working day does not have impact on overall demand, there is impact on registered and casual customers independently.
* During summer and winter there is no effect on demand, mostly due to casual customers.
* January and July there is demand change from registered customers no demand change from casual customers
* In March and October there is demand change for casual customers but no effect on overall demand.
* In November there is demand change form registered customers which is affecting overall demand.
* There is change in demand on Saturday from causal customers that is affecting overall demand.
* There is change in demand on Wednesday from registered customers, this is not affecting overall demand.

At this point of we can only analyze that there is change in demand. What the change is and how much is changed will be determined after doing the full model training.

## Final Model analysis

Table

Description automatically generated

After final analysis there can some conclusions made and some suggestions to the marketing team

### Conclusions

* With no other elements effecting the demand, there is more demand in casual customers then the registered customers.
* There is a positive demand increase in bikes for both registered and casual customers when it gets warmer.
  + For every 2 units increase in temperature there is 1 unit increase in demand.
* Demand for bike from casual customers tend to increase in summer. Increase in demand is very significant.
* Winter has more increase in demand then summer, all the demand increase is from registered customers.
* There is good increase in demand form registered customers in 2019, there is almost 25% increase in demand
* On windy days there is obvious drop in demand from all customers.
* There is loss of customers in snowy and cloudy days.
  + There is drop in demand from registered customers, Casual customers don’t have much effect on demand during this season.
* Working day has opposite effect on demand from registered and casual customers, drop in demand from casual customers is compensated by demand from registered customers.
* Humidity as huge effect on casual customer’s demand. Casual customers like to use bike’s when the weather is better.
* During Sunday’s there is uptick in demand from casual customers.

### Suggestions

* Customer service team and supply team need be ready to handle demand increase in warmer days.
* Marketing team need to focus on getting more casual customers in summer and converting them to register customers, so they continue to become more regular users of bike and keep the demand up in winter also.
* In general, people are getting more health concussions this might contribute to increase in demand year on year.
* On working days looks like register customers are using bike sharing for commute to work. So, demand for bikes in residential area might increase during daytime and increase in demand in industrial area during evenings.
  + Marketing team need to focus on having some deals to get more casual customer to use during the evenings, like happy hour.
  + Loss of demand from casual customers is negating the demand increase from registered customers.
  + It is important to get more casual customers on working days.
* To handle the drop in demand in windy days, company need to think to upgrade to aerodynamic bikes and advertise it accordingly.

## Confidence level

* From the model R2 value we can be approximately 80% confidence on the suggestions and conclusions made for all customers and registered customers.
* For Casual customers we can be 70% confidence on the suggestions and conclusions.

Table

Description automatically generated

## Final equations of model

# Subjective

## Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
   1. Season 1, spring has negative effect on bike sharing usage.
   2. Bike sharing has drastic uptick in 2019.
   3. Weekday and working day also have minimal effect.
   4. Snow has very drastic effect on bike sharing usage.
   5. Conclusion here irrespective of seasons and weather, Snow and Spring has negative effect.

A picture containing text, room

Description automatically generated

1. Why is it important to use **drop\_first=True** during dummy variable creation? (2 mark)
   1. Drop first helps us to reduce the number of columns without losing actual content of the data.
2. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   1. Registered has the highest correlation with the target variable cnt.
   2. Here since registered and casual columns are eliminated in the model, Next highest correlations ins with temp/atemp

A picture containing text

Description automatically generated

1. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   1. By doing residual analysis.
      1. Mean of residual analysis is zero

Table

Description automatically generated

Logo

Description automatically generated

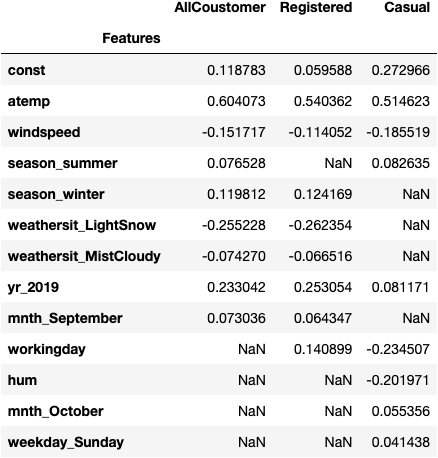
* 1. By checking for Homoscedasticity

Fitted line on residuals pass is almost flat with centered around 0, We can say residuals have equal variance. Another assumption of linear regression is satisfied

A picture containing text

Description automatically generated

1. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   1. Temperature and Temperature feel
   2. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
   3. Year 2019
   4. Wind speed



## General Subjective Questions

1. Explain the linear regression algorithm in detail.
   1. Linear regression algorithm is supervised algorithm explaining a liner relation between independent variable and dependent variable.
   2. Linear regression is defined by linear eq
      1. Where c is the constant, intercept, value of dependent variable when there is no independent variable.
      2. Y is the dependent variable.
      3. X is the independent variable.
      4. A is the slope of the equation explaining the rate of change of dependent variable when one unit of independent variable is changed.
2. Explain the Anscombe’s quartet in detail.
   1. Anscombe’s quartet is where the data distribution is not linear but have the same statical properties like,
      1. Same mean
      2. Same variance
      3. Same Min, Max
      4. Same Mode
   2. But when a line is drawn or residual analysis of linear model is done on this data, they fail to meet the linear regression assumption.
3. What is Pearson’s R?
   1. Pearson correlation coefficient is a measure of linear correlation between two sets of data.
   2. Pearson coefficient is calculated as
      1. is the covariance
      2. is standard deviation of X
      3. us standard deviation of Y
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   1. Scaling is process of scaling all the values of a data set to same scale.
   2. There might be data with large value but very low correlation with dependent variable, misguiding the beta values, to avoid this we want all the features on same scale.
   3. In normalized scaling we map the feature values between 0 and 1.
      1. In normalized data there is a worry of the we might lose data spread since all the data is getting compressed between 0 and 1.
      2. Outliers might be difficult to identify.
   4. In standardized scaling the data to have mean of 0 and standard deviation of 1.
      1. Here data still maintains the same spread of data.
      2. Easy to identify outliers here.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   1. VIF infinite between 2 variable means, these 2 variables can be perfectly explained with each other.
   2. 2 variables can be explained exactly by a linear combination.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   1. Plot to graphically determine if 2 data sets came from the population with common distribution.
   2. Advantages:
      1. Sample size do not need to be equal
      2. Distributions of 2 data sets can be handled at the same time as outliers.
   3. Need in linear regression:
      1. Helps to know if the test data or the data samples come from the population with same distribution.
      2. If they are not same helps in understanding how they are different.