Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: I have observed following observation:

1. Fall season has the highest business, the best season to ride bike.

2. As the time passes, people are getting aware about the bike sharing services so business growing in 2019 as compared to 2018

3. In months, September is doing well in terms of business growth.

4. People rent bike on non-holidays, because during the holiday they spend time with families and use personal car or vehicle.

5. In days, Saturday people do not rent bike compare to other days because of weekends.

6. Clear weather conditions are most favourable for the business, people rent bike more with compare to other season days.

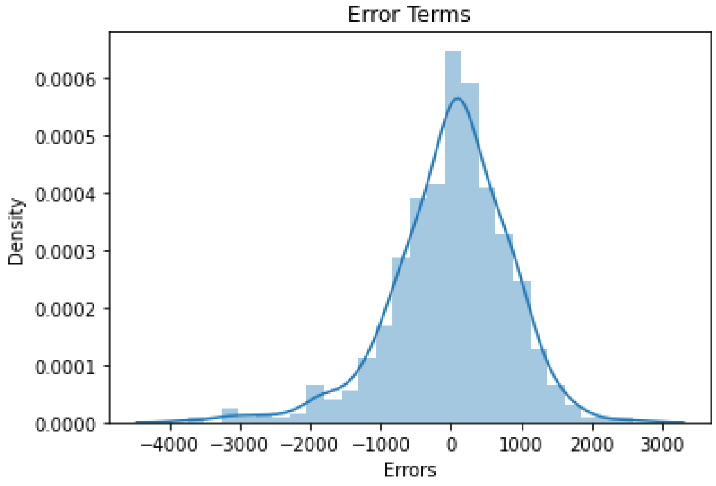
Question 2: Why is it important to use drop_first=True during dummy variable creation?

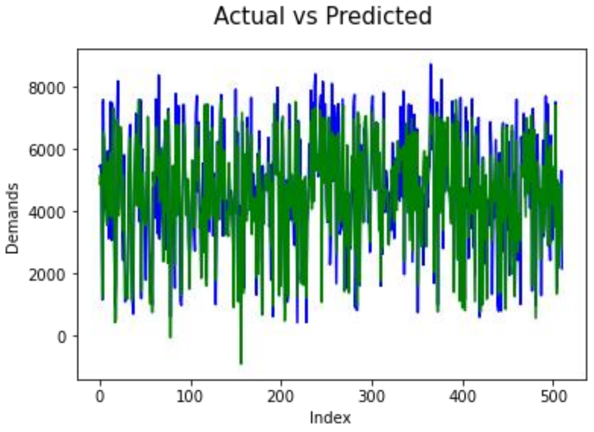
Answer: A variable with n levels is represented by n-1dummy variable. So, if we remove 1^{st} column then also we can represent the data. If the value of variable is o from 2 to n, it means the value of 1^{st} variable is 1.

Question 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: Target variable is cnt- temp atemp has the highest correlation with target variable 0.63.

Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?





By plotting the graph actual vs predicted and error terms we can justify that assumption was valid.

Question 5: Based on the final model, which are the top 3 features contributing significantly towards

Explaining the demand of the shared bikes?

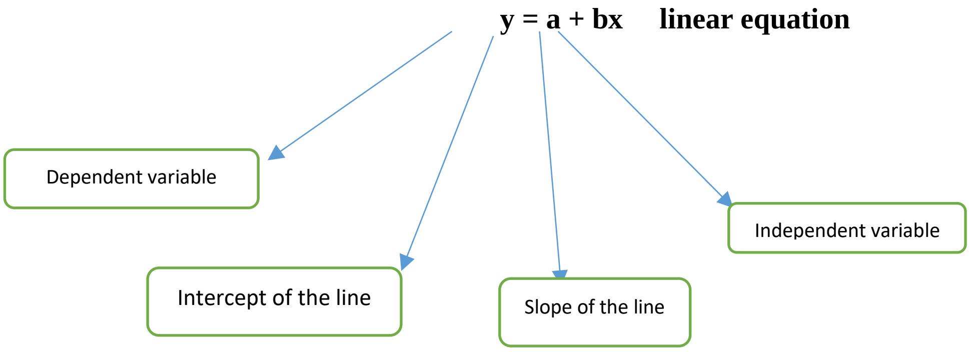
Answer: Based on final model top three features contributing significantly towards explaining the demand are:

1. Temperature
2. weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
3. year

General Question:

Question 1: Explain the linear regression algorithm in detail.

Answer: Linear regression is a very ordinary form of machine learning, where we train our data to predict behaviour of the data based on the some variables. As cleared with the name linear regression have two variables and the relationship between them is linear. One variable lies on X-axis, another on y-axis and both the variable should correlated linearly with each other. Linear regression comes under the supervised learning methods of machine learning.



$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(inter cept) = \frac{n\sum y - b(\sum x)}{n}$$

Target variable should be continuous only.

Input variable ccan be categorical or numerical.

Relationship between both the variable should be linear.

Linear regression are basically of 2 types:

1. Simple Linear Regression: SLR examines the relationship between the dependent variable and a single independent variable

   $y_i = \beta_0 + \beta_1 x_i$

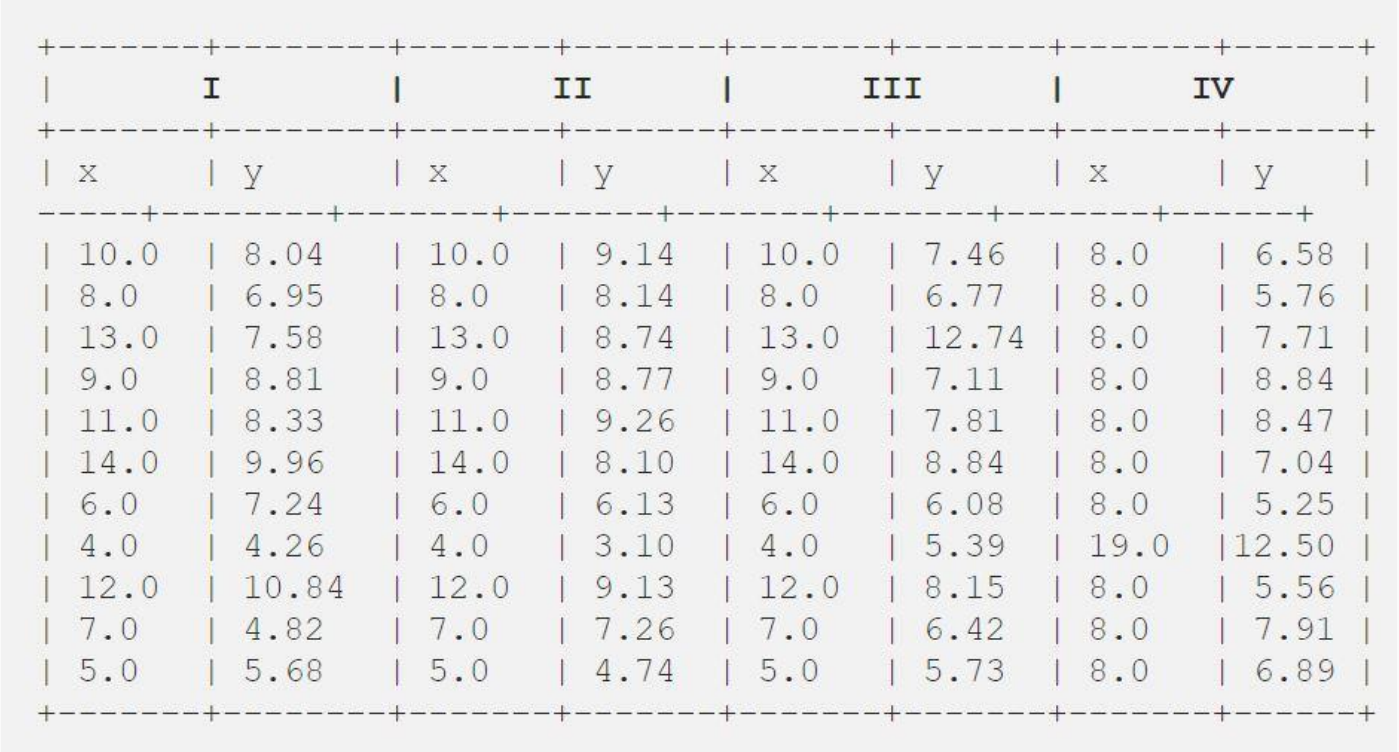
2. Multiple Linear Regression: MLR examines the relationship between the dependent variable and multiple independent variables.

   $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ \beta_p x_{ip} + \varepsilon_i$ **for** $i = 1,2, \ldots\ n.$

Question 2: Explain the Anscombe’s quartet in detail.

Answer: Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Simple understanding:Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.



```
+-------+--------+-------+-------+-------+-------+-------+-------+------+
|       I        |       II      |       III     |       IV             |
+-------+--------+-------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y            |
-----+--------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+-------+------+
```

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

Question3: What is Pearson’s R?

Answer: In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}}$$

Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Question 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

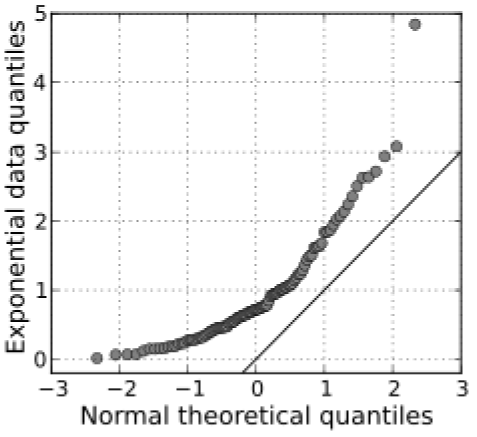
Answer: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Question6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Answer: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.


Exponential data quantiles
Normal theoretical quantiles