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

Ans: Here are some inferences I made from my analysis of categorical variables from the dataset on the dependent variable(count).

1. Fall has the highest median, which is expected as weather condition are most optimal to ride bike followed by summer.
2. Median bike rents are increasing year on as year 2019 has higher median than 2018, it might be due to the fact that bike rentals are getting popular and people are becoming more aware about the environment.
3. Overall spread in the month plot is reflection of season plot as fall months have higher median.
4. People more rent on non-holidays compared to holidays so reason might be people like to spend time with family in holidays or they prefer to use personal vehicle rather than rental bikes.
5. Overall median across all days are same but spread for Saturdays and Wednesday is bigger may be evident that those who have plans for Saturdays might not rent bikes as it is a non-working day.
6. Working and non-working days have almost same median although spread is bigger on non-working days as people might have plans and don't want to bike rent.
7. Clear weather is most optimal for bike renting as temperature is optimal and humidity is less.

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

Ans: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

   with the target variable?

   ‘temp’ has the highest correlation coefficient of 0.63

4. How did you validate the assumptions of Linear Regression after building the model on the

   training set?

   Ans: By plotting the residual distribution. It came out to be a normal distribution with a mean value of 0.

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

explaining the demand of the shared bikes?

Ans: The following are the top three features contributing significantly explaining the demand of the shared bikes.

- Atemp(0.412)
- Yr(0.236)
- weatherSit Light Rain(-0.275)

1. Explain the linear regression algorithm in detail.

A Linear Regression algorithm tries to explain the relationship between dependent and independent variable using a straight line. It is applicable to numerical variables only. Following steps are performed while doing Linear Regression.

- The dataset is divided into training and test dataset.
- Train data is divided into features(independent) and target(dependent) datasets.

- A linear model is fitted into training dataset. Internally the API's from Python uses gradient descent algorithm to find the coefficients of best fitted line. The gradient descent algorithm works by minimizing the cost function. A typical example of cost function is residual sum of squares.
- In case of multiple features, the predicted variable is hyperplane instead of a line. The predicted variable takes the following form:

$$y_i=\beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}+\ldots+\beta_k x_{i,k}+\epsilon_i.$$

- The predicted variable is then compared with test data and assumptions are checked.

2. Explain the Anscombe's quartet in detail.

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.

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.

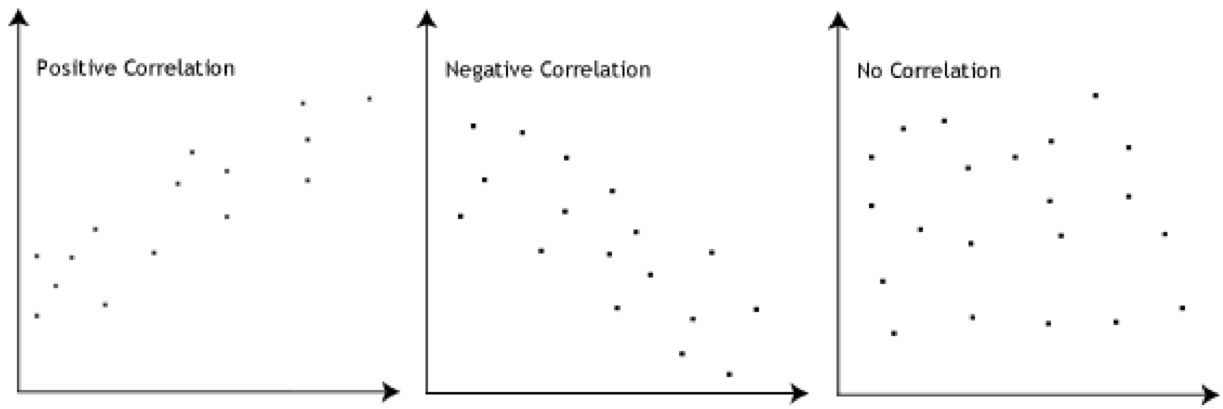
3. What is Pearson’s R?

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



Pearson r Formula

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

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling?

Ans: 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.

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

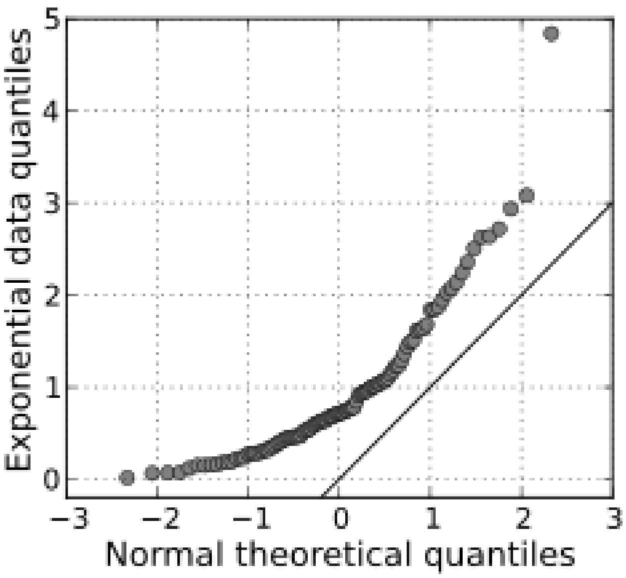
Ans: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing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

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

Ans: 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.

A Q Q plot showing the 45 degree 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.