# Assignment-based Subjective Questions

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

Solution:
Some of the inferences we can get from analysing the model are:

1. ***cnt*** is highly correlated with ***yr*** :
   It means that the number of bikes rented in 2019 has increased from that in 2018. This means that BoomBikes customer base has increased in the past year.

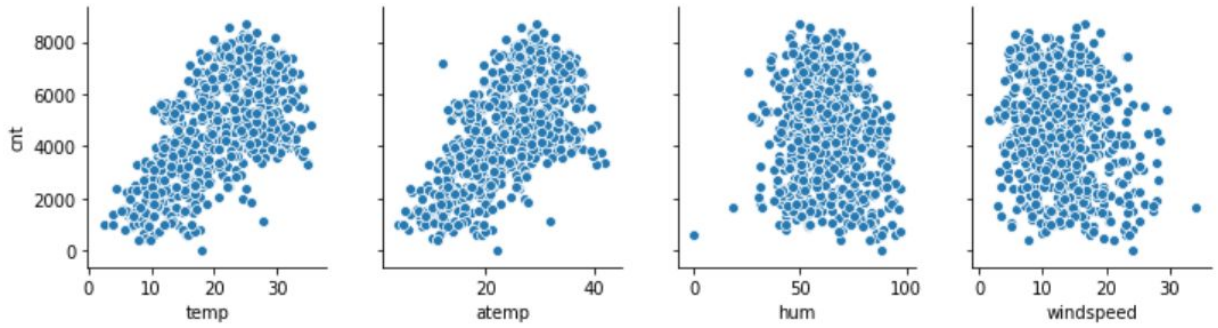
2. ***weather_2***, **weather_3** are negatively correlated:

   It signifies that when the weather conditions are misty, cloudy, snowy, or rainy the bike rental count decreases.

3. Bike rentals in Winter are more than that in Summer
4. Coefficients of categorical variables:
   a. Summer season : 0.1029
   b. Winter : 0.1361
   c. August : 0.0557
   d. September : 0.1134
   e. Weather_2 (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) : -0.0795
   f. Weather_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) : - 0.2776

**Q. 2. Why is it important to use drop_first=True during dummy variable creation?**
Solution:
If we don't use **drop_first = True**, get_dummies function will give us n dummy variables for n categorical values of any independent variable. But we require only n-1 values. This is due to the reason that when the values of all the n-1 variables is 0, that means none of the selected categories are there. This automatically left us with the value left.
So for instance, consider a categorical variable month, divided in 11 months. As we used, drop_first = True, so the column for the month of January will automatically be dropped. This means that when the value of all month related columns, i.e., from February to December are 0, then it automatically means that the respective record will be of column January, i.e., the categorical value of the month is January.

**Q. 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
Solution:
Looking at the correlation plot below, the highest correlated variable is temp.



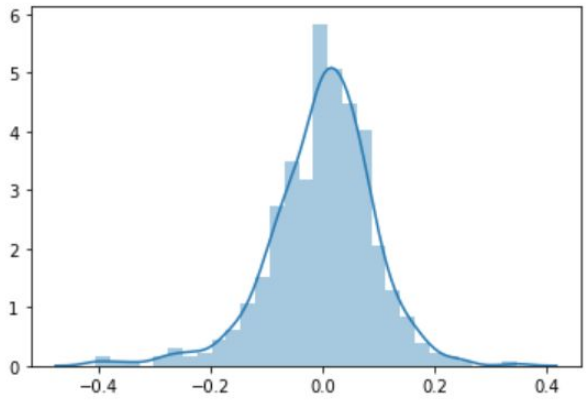

**Q. 4. How did you validate the assumptions of Linear Regression after building the model on the training set?**

Solution:

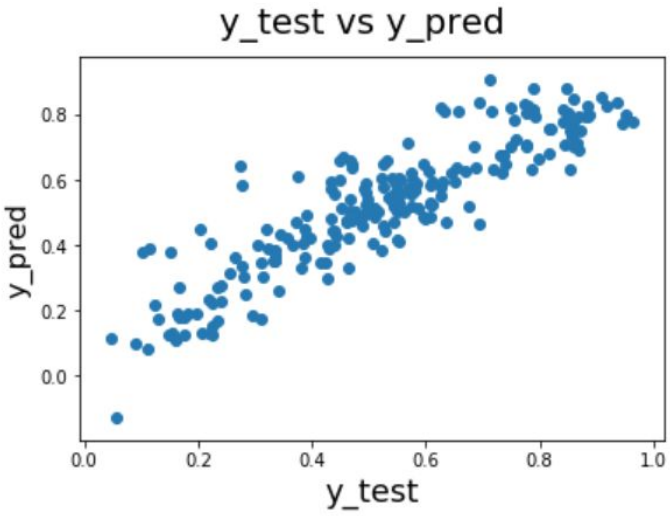
The assumptions are, 1. Linear relation between the independent variables and dependent variable. 2. Errors are normally distributed. 3. Errors are independent of each other. 4. Error terms have constant variance. For assumption 1, I have taken only those independent variables, which have a significant correlation with the dependent variable. For assumption 2,3 & 4, I have plotted them in my jupyter notebook to prove that errors are normally distributed, they are independent of each other, and Error terms have constant variance.

According to the assumptions of Linear regression,

1. Linear relation between the independent variables and dependent variable: As we have taken only those variables which are highly correlated, we have been able to depict the relationship between the dependent and independent variables using correlation values.
2. Errors are normally distributed, independent and have constant variance: As we can see in my model, the error terms distribution is a normal one.




Also the variance is constant




**Q. 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Solution: Top 3 features:

1. Temperature
2. Year
3. Weather_3 : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

# General Subjective Questions

**Q. 1. Explain the linear regression algorithm in detail.**

Solution:

Linear regression is the technique of understanding the relationship between a dependent variable and the independent variables.

There can be 2 different models in linear regression:

1. Simple Linear Regression
2. Multiple Linear Regression

**1. Simple Linear Regression**:

This is the most basic form of regression model. Here we find the relationship between 1 dependent variable and 1 independent variable. This can be explained on the basis of simple linear equation:

$$Y = \beta_0 + \beta_1 X$$

Here Y is the dependent variable, $\beta_0$ is the intercept, $\beta_1$ is the slope or correlation factor between Y and independent variable X.

Step 1: Plot a scatter plot for the above variables.
Step 2: Plot a best fit line shown by the above equation is drawn on the scatter plot between 2 variables. This line is known as Regression Line. There can be multiple regression lines in this plot. So we decide by calculating the Residual Sum of Squares (RSS).

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

For any data point on the line, the difference between the actual data point of scatter plot and the line is the residual. RSS is determined by adding the squares of these residuals for every point in the scatter plot.

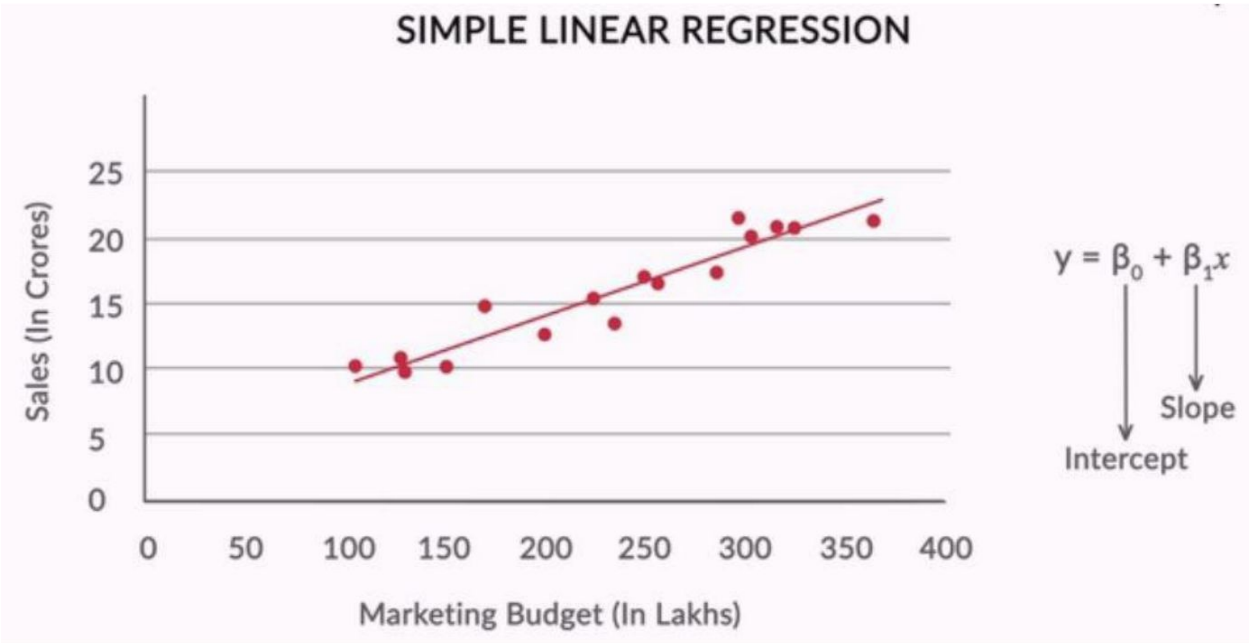



*Figure 3 - Regression Line*

Step 3: Determine the strength of the linear regression model by using $R^2$ or Coefficient of Determination:

$R^2 = (1 - RSS) / TSS$

TSS is Total Sum of Squares: It is the sum of errors of the data points from mean of response variables.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

**2.Multiple Linear Regression:**

Multiple Linear Regression is a statistical method to understand the relationship between 1 dependent and multiple independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Here *e* is error.

There are some basic ideas of this model

a. Model now fits a 'hyperplane' instead of line
b. Coefficients are still calculated using r-square
c. Assumptions of simple linear regression holds, that means that the error terms have zero mean and are normally distributed

There are some things we need to take care of in Multiple Linear Regression:

a. **Overfitting**: Adding more variables may cause overfitting. This will led the model to memorize all the data points and will led to generalisation. This will lead to decrease in accuracy.
b. **Multicollinearity**: We must check that the independent variables should not be highly correlated. It affects the interpretation of data whether the change in Y when all other variables are constant apply or not. We can use heatmaps and pairplots to find the correlation between independent variables.

**Variance Inflation Factor**(VIF): Correlation might not always depict the model accurately. There can be a situation that 2 or more variables as a group affect the dependent variable. In such a situation VIF is used.

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

So once there is multicollinearity detected, we have to deal with it:

a. Dropping variables
b. Create new variables using the interaction of other variables

**Steps of Multiple Linear Regression:**

**Step 1: Feature Scaling**

Scale all the numerical variables to a common scale. This does not affect the correlations. This is mainly done to get relatable values of coefficients. This can be done by 2 methods:

1. Standardizing
2. Min-Max Scaling

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-square, etc.

**Step 2: Handling Categorical Variables:**

To handle the categorical variable with n categories, we divide the variable into n-1 dummy variables. This method is used to determine the effect of each categorical value on the model.

**Step 3 : Model Selection:**

We take into account 2 new parameters to assess the multiple linear regression model.

a. Adjusted R-square

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

b. AIC

$$AIC = n * log(\frac{RSS}{n}) + 2p$$

**Step 4 : Feature Selection:**

There are various models of optimal feature selection:

1. Try all possible combinations (2 ^ p): Practically not feasible
2. Manual Feature elimination:
    - Build model
    - Drop features that are least helpful in prediction (high p-value)
    - Drop features that are redundant (using correlations, VIF)
    - Rebuild model and repeat
3. Automated approach:
    - Recursive Feature Elimination
    - Forward / Backward / Stepwise selection based on AIC

As a recommendation, we should follow a mixed approach.

**Q. 2. Explain the Anscombe's quartet in detail.**

This quartet was constructed in 1973 by the statistician Francis Anscmbe to demonstrate

- the importance of graphing the data before analyzing it
- Effect of outliers and other influential observations of statistical properties

This quartet comprises four data sets that have nearly identical descriptive statistics, yet they have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

For instance consider the following 4 datasets:

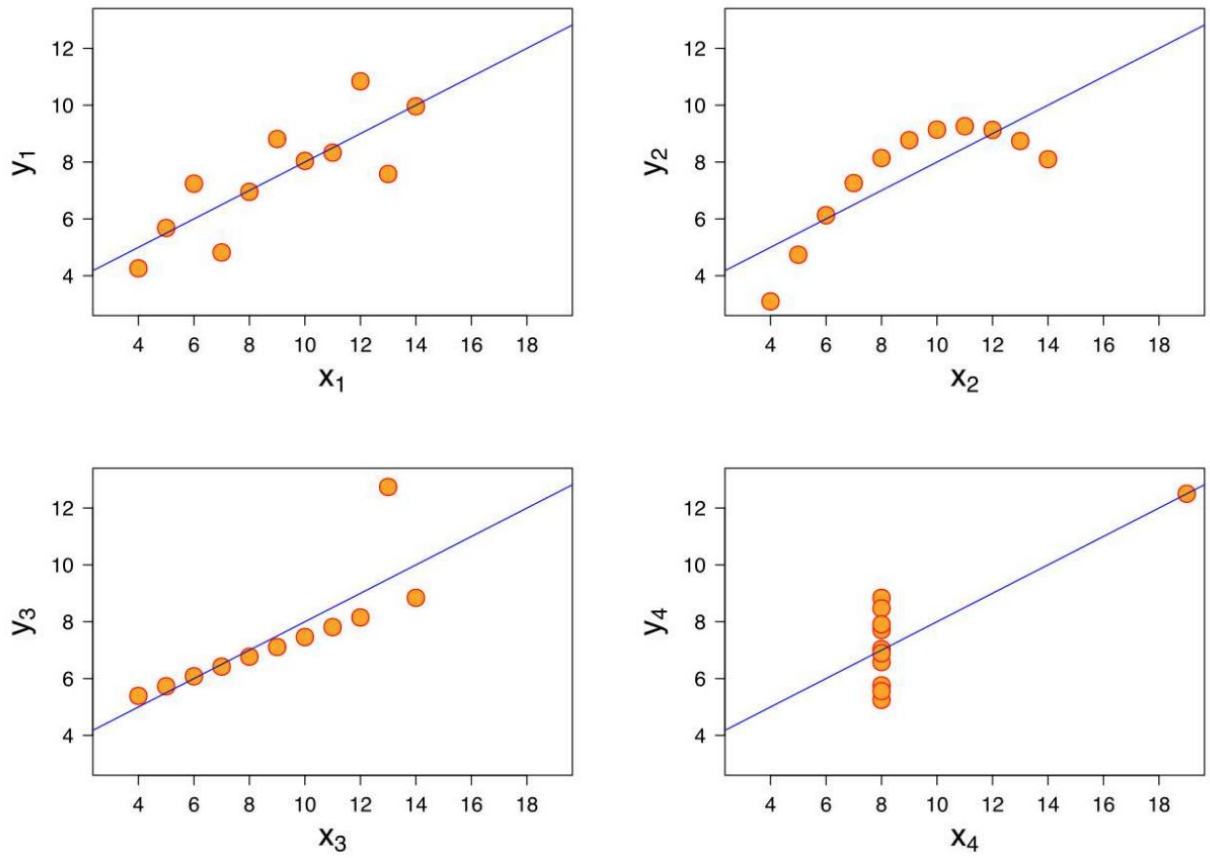
| | I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|---|
| | x | y | x | y | x | y | x | y |
| | 10 | 8,04 | 10 | 9,14 | 10 | 7,46 | 8 | 6,58 |
| | 8 | 6,95 | 8 | 8,14 | 8 | 6,77 | 8 | 5,76 |
| | 13 | 7,58 | 13 | 8,74 | 13 | 12,74 | 8 | 7,71 |
| | 9 | 8,81 | 9 | 8,77 | 9 | 7,11 | 8 | 8,84 |
| | 11 | 8,33 | 11 | 9,26 | 11 | 7,81 | 8 | 8,47 |
| | 14 | 9,96 | 14 | 8,1 | 14 | 8,84 | 8 | 7,04 |
| | 6 | 7,24 | 6 | 6,13 | 6 | 6,08 | 8 | 5,25 |
| | 4 | 4,26 | 4 | 3,1 | 4 | 5,39 | 19 | 12,5 |
| | 12 | 10,84 | 12 | 9,13 | 12 | 8,15 | 8 | 5,56 |
| | 7 | 4,82 | 7 | 7,26 | 7 | 6,42 | 8 | 7,91 |
| | 5 | 5,68 | 5 | 4,74 | 5 | 5,73 | 8 | 6,89 |
| SUM | 99,00 | 82,51 | 99,00 | 82,51 | 99,00 | 82,50 | 99,00 | 82,51 |
| AVG | 9,00 | 7,50 | 9,00 | 7,50 | 9,00 | 7,50 | 9,00 | 7,50 |
| STDEV | 3,32 | 2,03 | 3,32 | 2,03 | 3,32 | 2,03 | 3,32 | 2,03 |

As it can be observed here, that summary statistics of each data set are the same,

- mean of x is 9
- mean of y is 7.50,
- sum and variance are also the same

Now if we plot them, we will infer following observations:

a. Dataset 1 is a linear model.
b. Data set 2 is not normally distributed.
c. The distribution of the data set 3 is linear, but it has an outlier.
d. The dataset 4 shows that one outlier is enough to change the mean and variance, and correlation and coefficient.



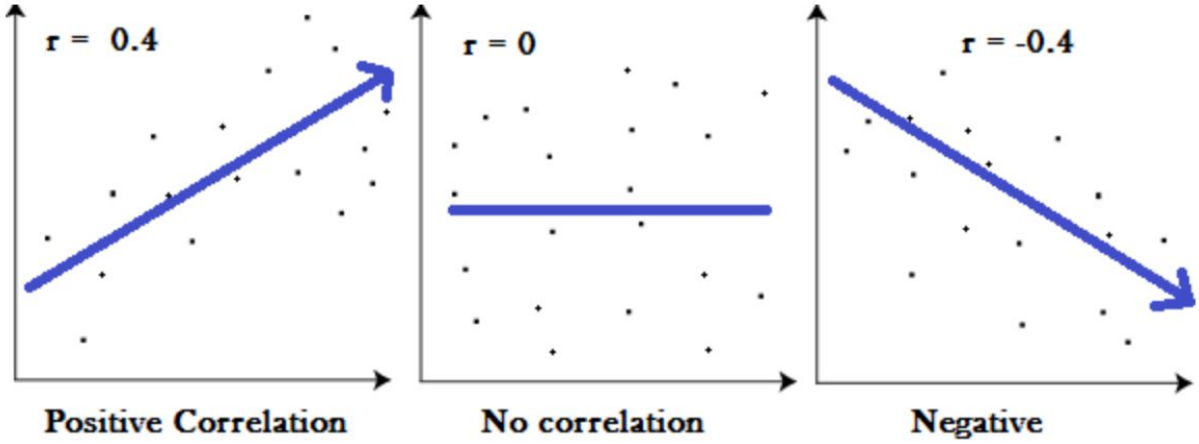

Q. 3. What is Pearson's R?

The Pearson's R or simply Pearson Correlation Coefficient is a statistical formula that measures the strength between variables and relationships. It gives an understanding of how well 2 variables are correlated with each other.

There can be a positive, negative or no correlation.

1. Positive Correlation: Variable 1 is directly proportional to Variable 2. It means if one variable is increasing the other will also increase and vice versa.
2. Negative Correlation: Variable 1 is inversely proportional to Variable 2. It means if one variable is increasing the other will decrease and vice versa.

3. No Correlation: This means that with every increase or decrease in one variable, there is no effect on another variable. The variables are not correlated.




Pearson's Formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,]\,[\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Here r is the Pearson's r which is calculated for variables x and y. This denotes the relation between 2 variables.

**Q. 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Solution:

In datasets, some numeric variables lie in very high range while the others lie in very low range. If we build the model using the same variables, then we will find the weird values of correlation coefficients. So we can say we need scaling for 2 reasons:

1. Ease of interpretation
2. Faster convergence for Gradient Descent Methods

We have to scale all the numerical variables to a common scale. This does not affect the correlations. This is mainly done to get relatable values of coefficients.

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-square, etc.

This can be done by 2 methods:

1. Standardizing: The variables are scaled so that the mean is zero and standard deviation is 1.

$$x = \frac{x - mean(x)}{sd(x)}$$

2. Min-Max Scaling (Normalized scaling): Using the maximum and minimum values of data, we scale the variable in a way that its value lies between 0 and 1.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

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

Solution:

VIF can be calculated as

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

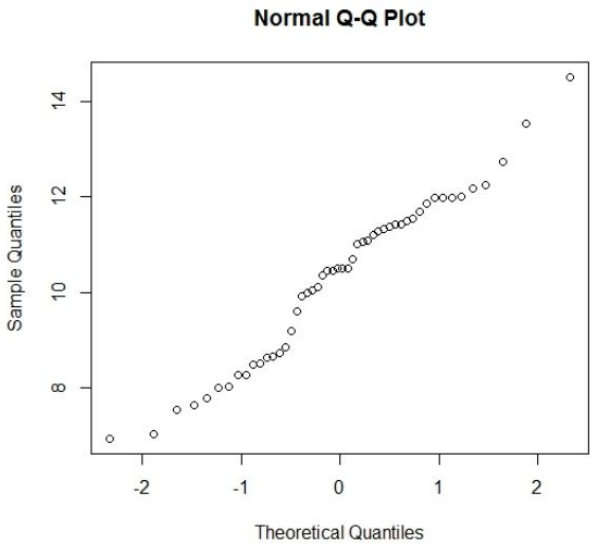
So if the denominator is 0, that means R-square equals to 1 will cause the VIF to be infinite. This will happen only when there is a perfect correlation between the dependent variable and the independent variable. This will come only when there are situations such as over-fitting and high multicollinearity. In other words the model has memorized the data and hence the accuracy of prediction is not guaranteed.

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

Solution:

Q-Q Plot or quantile-quantile plot is a scattered plot created by plotting 2 sets of quantiles against one another. It is a graphical tool which determines if data can be approximated by a statistical distribution such as Normal, exponential or Uniform distribution.

If both the sets of quantiles which are plotted came from same distribution, then we will get a line which is roughly straight.




A Q-Q plot is a probability plot. First, the set of intervals for the quantiles is chosen. A point $(x, y)$ on the plot corresponds to one of the quantiles of the second distribution ($y$-coordinate) plotted against the same quantile of the first distribution ($x$-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.
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 $y = x$.