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)**

   Ans:

   1. People have used shared bikes more in season 3 (fall) followed by summer and winter. Spring season has least usage of shared bikes. So, demand is high in Fall and low in Spring season
   2. Overall, Public holidays data is showing a dip, which means people are not using much during holidays compared to non-holiday. However, on some holidays people are using more shared bike than the day before and after holiday, which we can see in the data in spreadsheet for Independence Day, Christmas Day, Columbus Day.
   3. From workday against count of bike shares, we can say that the demand is pretty much the same with working and non-working day. But the spread is more for working day. Overall, shared bikes usage shows that it increases from Tuesday to Sunday and drops on Monday.
   4. Demand is more Weather Situation 1: Clear, Few clouds, Partly cloudy, Partly cloudy followed by Weather Situation 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist, followed by 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
   5. Usage of shared bikes increase from January to June and starts to decrease a bit until September then decreases more with more fluctuations in September and October
   6. Usage of shared bikes increase from Tuesday to Friday and almost no change from Friday to Saturday but then decrease on Sunday and Monday.

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

   Ans:

   It is important to use **drop_first=True** during the creation because,

   1. We only need **n-l** categorical variables to represent in formation in **n** categories of a categorical feature.

   2. We can avoid redundancy in the variable and keep their count minimal to build model with exact same information.

3 **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   Ans:

   Numerical variable registered has the highest correlation of 0.95 among all the numerical variables with target variable ‘cnt’

4 How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

To validate the assumptions of Linear Regression,

1. Fitted a linear regression model for predicting the demand of shared biked using multiple independent features on 'hyperplane'.
2. Used Adjusted R-Square to keep the sum of squared errors minimum
3. Used P-value to understand the significance of a features and VIF (Variance Inflation Factor) to understand the multicollinearity and built a model with more important features having P-value < 0.05 and and VIF < 5
4. Validated that the errors in the model are normally distributed with a mean centered at zero by plotting a distribution of residuals
5. Used R-Square value to measure the accuracy of the Model

5 Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:

Top 3 contributing features are

1. yr (Year),
2. season (Season)
3. weathersit (Weather Situation)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans:

Linear regression algorithm is a supervised machine learning algorithm, which is used to make the prediction of a dependent continuous output/target variable by using the independent variable. For example, predicting the shared bikes using the data set with features like Temperature, humidity, season and so on.

There are two different types of Linear Regression models,

1. Simple linear regression, which can be represented as,

   $y = b0 + b1x$,

where,

b0 is constant

y is dependent variable

b1 is slope, tells the rate at which y changes wrt x

2. Multiple linear regression, which can be represented as,

y = B0 + B1x1 + B2X2+……+Bn xn

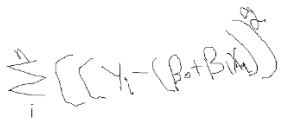
where,

b0 is constant
y is dependent variable

b1 is slope, tells the rate at which y changes wrt x1

b2 is slope, tells the rate at which y changes wrt x2 and so on

Linear regression algorithm tries to understand the relationship between independent variable and target variable, and it tries to keep the RSS (residual Sum Square) minimal, which is the total sum of errors across the sample. It is a measure of difference between expected and actual output of the target variable value

RSS = 

Where,

(b0+b1xi) represent the actual output for individual sample

yi represent the expected target value for sample i

Linear regression algorithm assumes,

1. There is linear relationship between independent and target variables
2. Error terms are normally distributed
3. Error terms are independent with each other
4. Error terms have constant variance

With multiple linear regression for feature selection the algorithm looks for couple of metrics to fit the best regression line like,

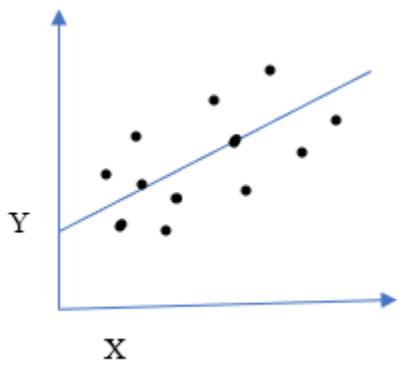
1. P-value for significance of variable,
2. VIF for checking multicollinearity and avoid variables which are multicollinear,
3. AIC and BIC.

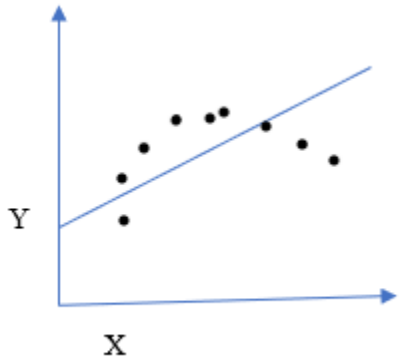
2 Explain the Anscombe's quartet in detail. (3 marks)

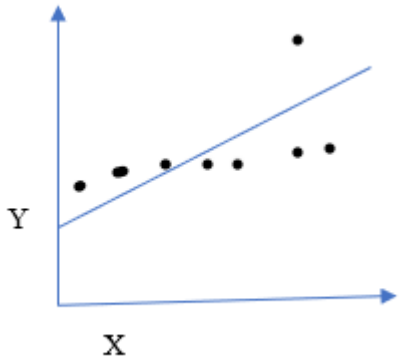
Ans:

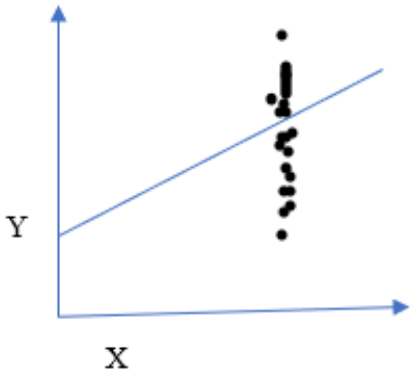
Anscombe's quartet will have four data sets that have 4 identical descriptive statistics (Mean, Sample Variance) with independent variables and identical descriptive statistics (Mean, Sample Variance). Also, the independent variable shows a high correlation with a straight regression line. But, the data looks different with more outliers when we visualize it.

It tells that we should always visualize the data than going by statistical numbers. Below are the four scatter plots with regression line showing the actual data in dots with linear regression line.









3 What is Pearson's R? (3 marks)

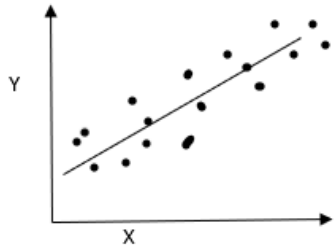
Ans:

Pearson's correlation coefficient is the measure of linear correlation between two sets of data. It is the ratio between the covariance of two variable/features and the product of their standard deviations.

Pearson's r or correlation coefficient value varies between -1 and 1

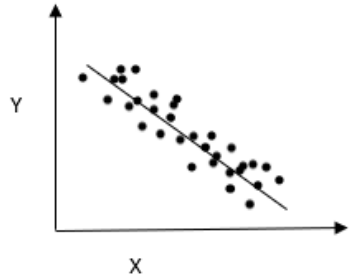
1. The value 1 mean the two variables say for example X and Y are highly positively correlated, for every positive change in X there will be a positive change in Y

   Positive Correlation



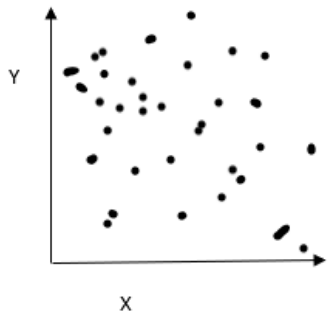
2. The value -1 mean the two variables say for example X and Y are highly negatively correlated, for every positive change in X there will be a negative change in Y

   Negative Correlation



The value 0 means the two variables are not correlated, for every increase in X there is no positive or negative increase in Y

Zero Correlation



Formula,

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

Where, x-bar and y-bar are means of x and y

4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:
Scaling is the process of bringing all variables value into one scale measurement.

It is performed to help interpret the data of multiple independent variables having their value on different scales. And, to avoid weird looking coefficients for every other variable. Overall, for ease of understanding and it helps with performance when we apply algorithms to build model.

Normalized scaling scales the values of variable between 0 and 1, whereas standardized scaling scales values of variable in such way their mean is 0 and standard deviation is 1.

5 **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

Ans:

When we see a VIF multicollinear value of infinite, it means that the variable is highly correlated with other variable/s. It happens when the change in a variable is explained 100% by other variables.

Let's take an example of two variables x1 and x2, which are highly correlated with correlation coefficient 1

VIF(x1) = 1 / (1-R-square)

R-square will be 1 when x1 is highly correlated with x2

VIF(x1) = 1 / (1 - 1)

VIF(x1) = Infinity

In the bike sharing data set, we could almost notice it between temp and atemp variables with correlation or 0.99 between them

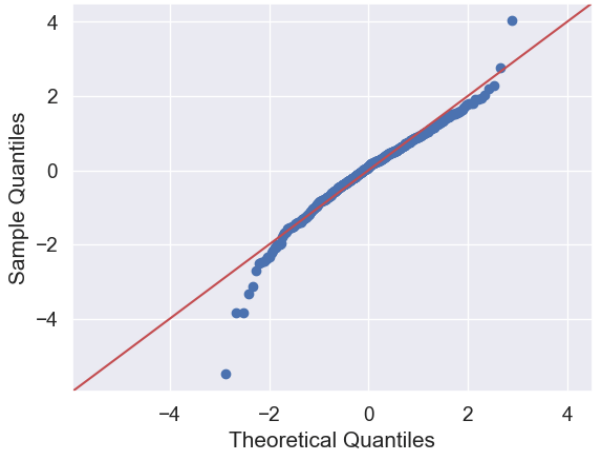
6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans:

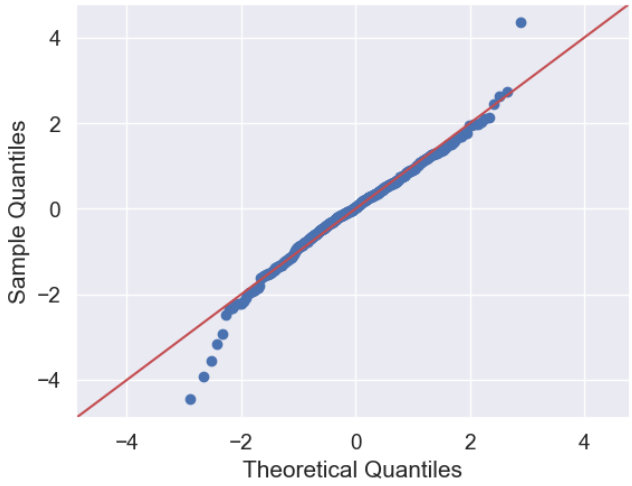
Q-Q Plot (Quantile - Quantile plot) is a probability plot, a graphical method for comparing the two distributions by plotting their quantiles against each other.

It used to graphically analyze the two probability distributions. If two distributions which we want to compare are exactly equal then the points on the Q-Q plot will lie on a straight-line y = x, 45-degree straight line

In linear regression it can be used to graphically analyze the residual distributions, to check if they are normally distributed. Which, is actually the distribution of target variable's predicted to actual distribution



Q-Q plot of the residuals of model built with manual approach to predict the demand of shared bikes



Q-Q plot of the residuals of model built with automated and manual approach to predict the demand of shared bikes

We can see from the Q-Q plot that the model built with automated and +manual approach is leaning more towards the straight line than compared to the model built through manual approach.