Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: -

- In `Summer` and `Fall` the number of count is higher than other season and less in `Spring` season.
- In `2019` the number of count is high and less in `2018`.
- The number of count is more on `Clear` wheather and less in `Light_Snow_Rain`.
- Number of bikes are higher when there is no `holiday`.
- As we can see the count of bikes are high on `workingday`.
- Number of Bikes are higher on the day when the weather is `clear`.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans: - After creating dummy variables we need to remove the redundant variables so that we use drop_first = True, we know that if we have k categorical label in a categorical column we need k -1 dummy columns.

Here the below example explain briefly

Spring	Summer	Winter
1	0	0
0	1	0
0	0	1
0	0	0

Total 4 categorical label (fall, spring, summer, winter) was present and after applying drop_first = True, The fall categorical was dropped and the above table shows that,

1 0 0 – spring,

0 1 0 – summer,

0 0 1- winter,

0 0 0 – fall

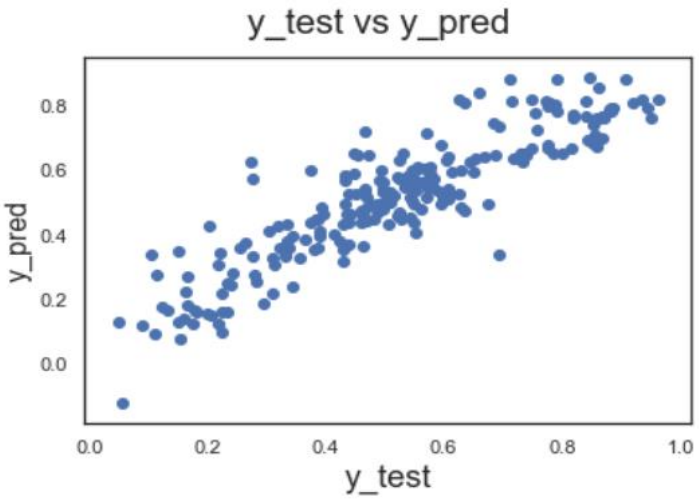
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: - By looking at the pair plot we found that 'temp' variable have highest correlation with target variable.

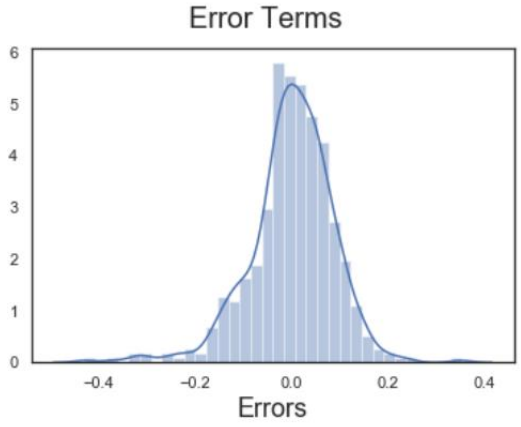
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans –

- The relationship between dependent and independent variable should linear.
  To find the nonlinearity we can see the observed vs. predicted values or residuals vs. predicted values. The desired outcome is that points are symmetrically distributed around a diagonal line in the plot.



- The error between actual values and predicted values should be normally distributed and centered at zero.



- Then look at the Variance Inflation Factors (VIF). It is calculated by regressing each independent variable on all the others. And all the values are below 5 and most of the values are below 2 which show good value.

	Features	VIF
2	temp	3.93
0	yr	1.95
5	summer	1.79
7	Aug	1.56
6	winter	1.47
4	Mist_Cloudy	1.44
8	Sep	1.29
3	Light_Snow_Rain	1.06
1	holiday	1.03

- Homoscedasticity. A scatter plot of residuals versus predicted values is good way to check for Homoscedasticity. There should be no clear pattern in the distribution; if there is a cone-shaped pattern, the data is heteroscedastic

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: - temp, yr, summer and winter are the top features which are contributing significantly towards explaining the demand of the shared bikes. As per the final Model.

General Subjective Questions

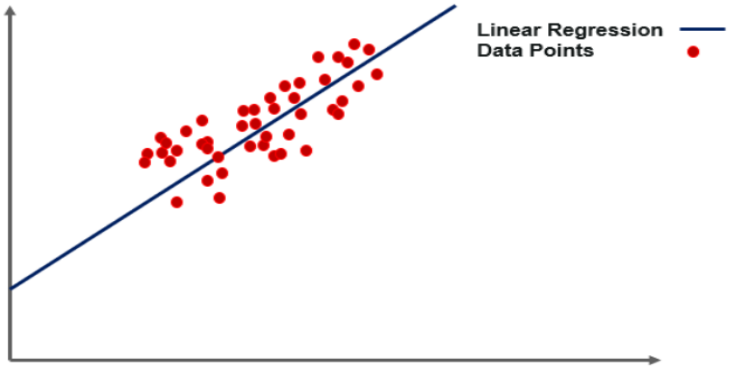
Q1. Explain the linear regression algorithm in detail.

Ans: - In linear regression is a regression technique in which the independent variable has a linear relationship with the dependent variable. The straight line in the diagram is the best fit line. The main goal of the simple linear regression is to consider the given data points and plot the best fit line to fit the model in the best way possible.

Types of Regression

The following are types of regression.

1. **Simple Linear Regression**
2. **Polynomial Regression**
3. **Support Vector Regression**
4. **Decision Tree Regression**
5. **Random Forest Regression**



- Linear Regression Algorithm is a machine learning algorithm based on **supervised learning.**
- In linear regression we train our model to predict the behavior Of data based on some variable, as name suggest he two variables which are on the x-axis and y-axis should be linearly correlated.
- **Mathematically,** we can write a linear regression equation as: Simple linear regression: This is used when the number of independent variables is 1.

$$Y = \beta o + \beta 1X$$

Where:
y is dependent variable
x is independent variable
βo is intercept
β1 is the coefficient for x

- **Multiple linear regressions:** This is used when the number of independent variables is more than 1.

$$F(X, Y, Z) = \beta o + \beta 1X + \beta 2Y + \beta 3Z$$

Linear Regression Terminologies:-

The following terminologies are important to before moving on to the linear regression algorithm.

- Cost Function
    - By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum .
    - So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y predicted value and y actual value.
- **Gradient Descent :-**
    - When there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS/TSS)$.

• **RSS**: Residual sum of squares

• **TSS**: Total sum of squares

High R^2 means model is good, but in multiple linear regression we need to see adjusted R^2.

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model, the difference between adjusted R^2 should be minimum.

Assumption of Linear regression:

- **Linear relationship:-** The relationship between independent variable and dependent variable should be linear Residual
- **Normality:-** error between actual values and predicted values should be normally distributed No or little
- **Multicolinearity:-** Multicolinearity refers to a situation when 2 or more variable are highly correlated.

- **Homoscedasticity:-** Homoscedasticity describes a situation in which the error term is the same across all values of the independent variables.

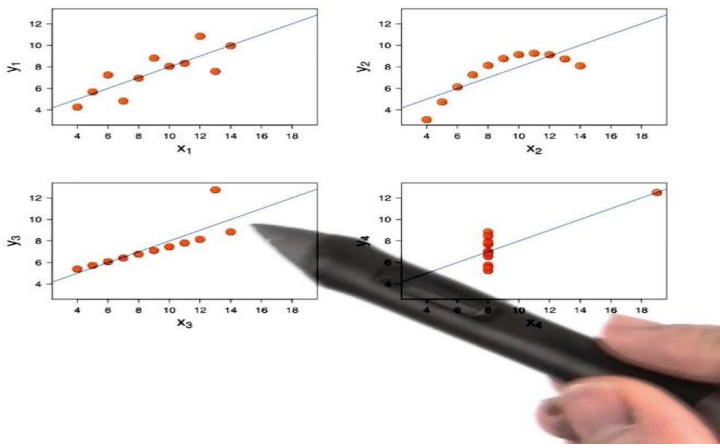
Advantages And Disadvantages :-

Advantages	Disadvantages
Linear regression performs exceptionally well for linearly separable data	The assumption of linearity between dependent and independent variables
Easier to implement, interpret and efficient to train	It is often quite prone to noise and overfitting
It handles overfitting pretty well using dimensionally reduction techniques	Linear regression is quite sensitive to outliers
One more advantage is the extrapolation beyond a specific data set	It is prone to Multicolinearity

Q2. Explain the Anscombe's quartet in detail.

Ans: - It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. When they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

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



The average x value is 9 for each dataset

• The average y value is 7.50 for each dataset

• The variance for x is 11 and the variance for y is 4.12
• The correlation between x and y is 0.816 for each dataset
• A linear regression (line of best fit) for each dataset follows the equation
$y = 0.5x + 3$

- Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance.
- Dataset II fits a neat curve but doesn't follow a linear relationship (maybe its quadratic?).
- Dataset III looks like a tight linear relationship between x and y, except for one large outlier.
- Dataset IV looks like x remains constant, except for one outlier as well.

This isn't to say that summary statistics are useless. They're just misleading on their own. It's important to use these as just one tool in a larger data analysis process.

Visualizing our data allows us to revisit our summary statistics and recontextualize them as needed. For example, Dataset II from Anscombe's Quartet demonstrates a strong relationship between x and y, it just doesn't appear to be linear. So a linear regression was the wrong tool to use there, and we can try other regressions. Eventually, we'll be able to revise this into a model that does a great job of describing our data, and has a high degree of predictive power for future observations.

Q3. What is Pearson's R?

Ans: -

- In statistics, the Pearson correlation coefficient referred to as Pearson's r,
- It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.
- Pearson's r is a numerical summary of the strength of the linear association between the variables.

- If the variables tend to go up and down together, the correlation coefficient will be positive.
- If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
- For the Pearson r correlation, both variables should be normally distributed, There should be no significant outliers,
- Each variable should be continuous, The two variables have a linear relationship and Homoscedascity(equal variance)

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: -

Scaling is applied to independent variables or features of data. It basically helps to normalize the data within a particular range.
Scaling is performed because ML algorithm works better when features are relatively on a similar scale and close to Normal Distribution.

Normalized Scaling	Standardized Scaling
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.	Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation.
Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution.	Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution.
Formula for normalization: $X = \frac{X - Xmin}{Xmax - Xmin}$ Xmax and Xmin are the maximum and the minimum values of the feature respectively	Formula for standardization: $X = \frac{X - \mu}{\sigma}$ μ is the mean of the feature values, σ is the standard deviation of the feature values.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: -

In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated with other features.

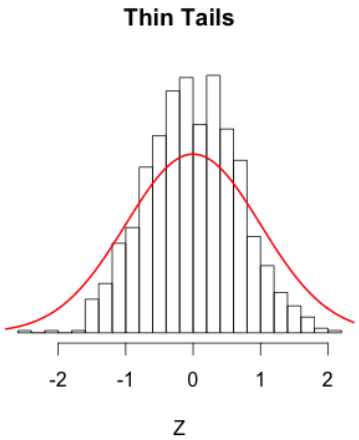
$$VIF = \frac{1}{1- R^2}$$, So when R^2 reaches 1 then VIF= inf

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.
If we have VIF infinity for feature we should drop that feature.

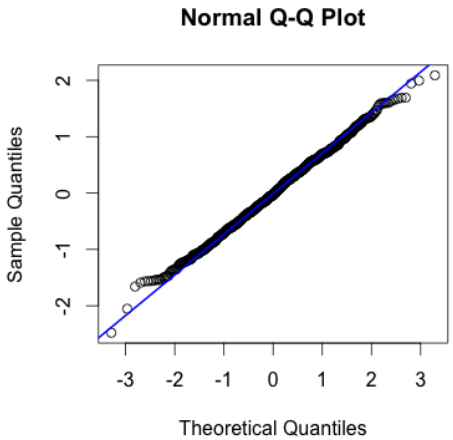
Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: - A Q–Q plot (quantile-quantile plot) is a plot of the quantiles of two distributions against each other, Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian distribution, Uniform Distribution, and Exponential Distribution.

If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is normally distribution because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot.



Normally distributed data



Normal Q- Q Plot

Hence, we can use q-q plot also in residual analysis,

The q-q plot is used to answer the following questions:

- Do two data sets come FROM populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior