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?

   Each categorical variable partially affects the dependent variable. Only some of the values of the categorical variables such as ‘Spring’ in seasons, ‘2019’ in year and ‘non-working day’. They also tend to influence some of the continuous variables. For example, the season affects the temperature and also the ‘feeling temperature’.

2. Why is it important to use drop_first=True during dummy variable creation?
   If it is not dropped, the VIF value will tend to infinite.
   Secondly, a redundant variable can be reduced so that the algorithms can be optimized with lesser variables to compute.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Answer: ‘atemp’ variable.

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

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Answer: atemp, Year_1 (2019) and spring

General Subjective Questions

1. Explain the linear regression algorithm in detail.

In general, regression is a method of modelling a target value based on independent predictors. Linear regression algorithm shows a linear relationship between a dependent variable(y) and one or more independent variables (x1, x2,…xn) hence called as linear regression. A Linear Regression model provides a straight line with a slope representing the relationship between the variables.

This can be shown mathematically as

$y = mx + c$

where, y is the dependent variable

x is the predictor variable

m is the slope of the line

and c is the intercept, a constant.

Linear Regression is also deemed a statistical method for Predictive Analysis.

There are two types of Linear Regression:

a. Simple Linear Regression
    i. The numerical dependent variable is predicted using a single independent variable.
b. Multiple Linear Regression
    i. The numerical dependent variable is predicted using multiple (more than one) independent variables

In either of these cases, the line that is obtained to predict the dependent variable is called as the Linear Regression Line.

If the value of dependent variable increases with the increase in the value of the independent variable, it is termed as a Positive Linear Relationship.

If the value of the dependent variable decreases with the increase in the value of the independent variable, it is termed as a Negative Linear Relationship.

The main goal of Linear Regression is to find the Best-Fit line, the line that will have the least error i.e., the error between the predicted values and the actual values should be minimized. This means to say that we need to find the best values for m (slope) and c(intercept) in the above mathematical equation. This is achieved by calculating the Cost Function.

For Linear Regression, we use the Mean Squared Error (MSE) cost function.

The cost function is measured using the Residuals i.e., the distance between the actual value and the predicted value. Cost Function is directly proportional to the Residual.

Gradient Descent method is used to minimize the MSE by calculating the gradient of the Cost Function.

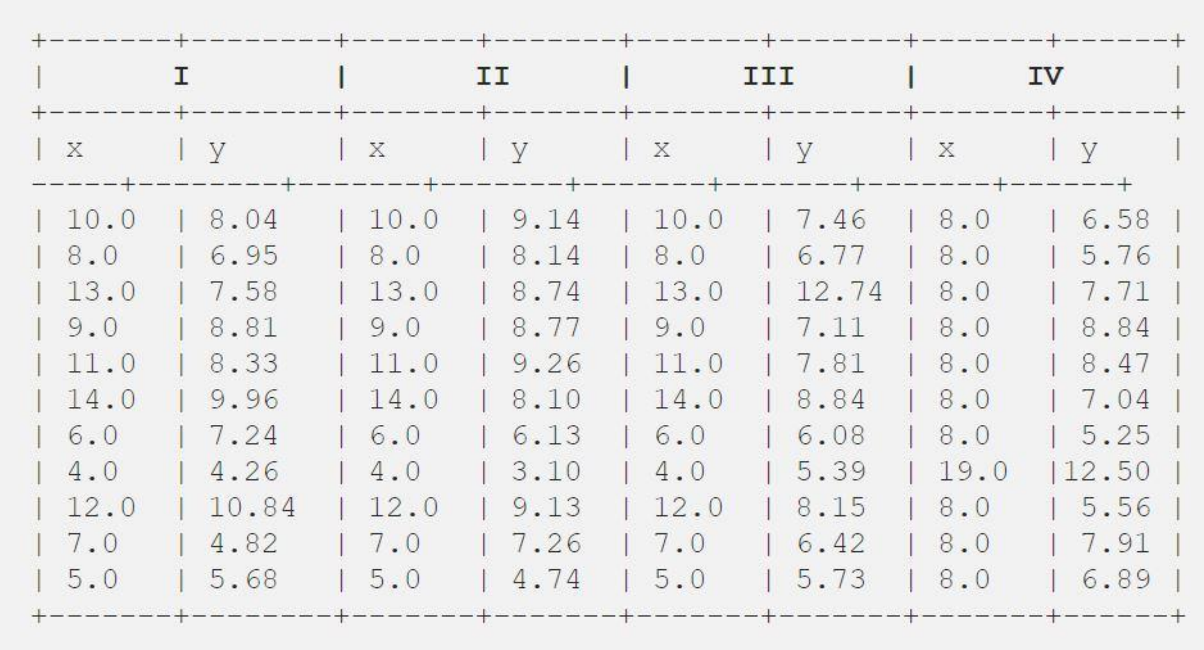
In Linear Regression, the process of finding the best model is termed as Optimization. The goodness of fit of the line of regression is determined using a statistical method called R-Squared.

The final model can be accepted if it passes or fulfils the Linear Regression Assumptions. They are:

- The dependent variable should be linearly related to the independent variable(s).
- The residuals should be independent of each other.
- The residual errors should be Homoscedastic i.e., they should have constant variance.
- The residual errors should be normally distributed.

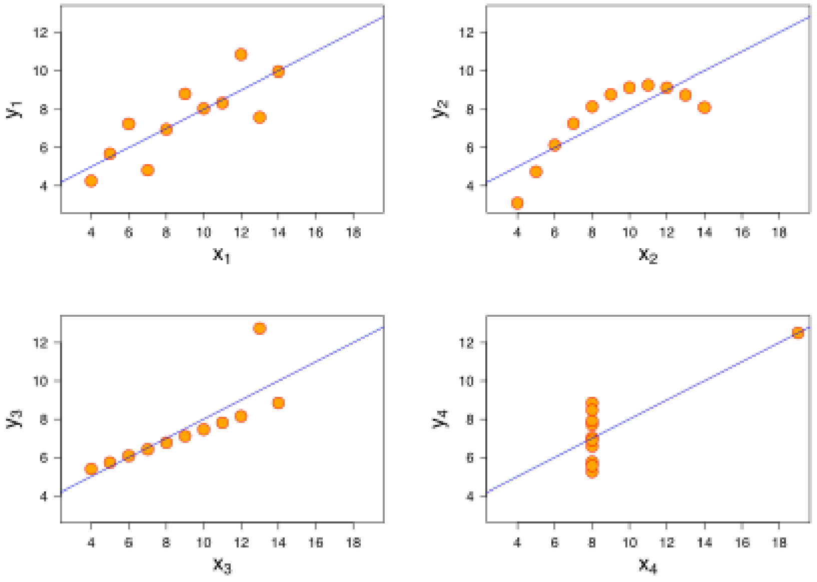
2. Explain the Anscombe's quartet in detail.

Simply put, Anscombe's quartet shows that multiple data sets with similar statistical properties can still be very different from one another when they are represented graphically.
This comprises four datasets that have nearly identical statistical properties. Each dataset consists of eleven (x,y) points.



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

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.



The main purpose of this is to tell us about the importance of visualizing the data before applying various algorithms. It is very important for Linear regression since a Regression line can be considered fit only for data with linear relationships and not for other types of relationships.

3. What is Pearson's R?

Pearson's R measures the strength of the linear relationship between two variables. Its value is always between -1 and 1.

It is also known as **Pearson correlation coefficient or correlation coefficient.** It is a measure of

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

It is a step taken during the pre-processing of data. It basically helps to normalise the data within a particular range. It can also help in speeding up the calculations in an algorithm. This step is very crucial during multiple linear regression.

The independent variables might be in different scales owing to the nature of data. One of the challenges with data being in different scales will be the interpretation of the coefficients once the model is created.

Consider two variables one which has values in tens, basically less than 100 and the other that has values in thousands or ten thousands. Clearly these are in different ranges and thus scale differently. Now when the coefficients are checked at the later part of the model building, the values in tens will have larger coefficients and the larger values of data will have lesser coefficients. This can lead to the confusion or wrong interpretation of the coefficients. Hence to uniformly identify the better coefficients, the data needs to be uniformly scaled.

For example:

Coefficient of variable A (whose data value is less than hundred) can be 0.004

Coefficient of variable B (whose data is in thousands) can be 356

So here if we interpret that the variable B is a much better predictor than variable A due to its high coefficient, it would be wrong.

If the data is rescaled, the algorithms such as Gradient Descent would work in a more optimized way owing to the scale or range of the data being between o and 1 uniformly. The algorithm could run faster with rescaling.

Normalized Scaling: Also known as min-max scaling, this type of scaling converts/compresses the data between 0 and 1. So the max value of the data will be 1 and the minimum value of the data will be 0.

The formula used is (xi – xmin) / (xmax – xmin)

Where xi represents each value of x.

So, if xi equals xmin, the value will be zero.

If xi equals xmax, the value will be 1.

While using this method of scaling, the outliers can be handled well since any value is brought into the range of 0 and 1. This is an advantage of using normalization against standardization in linear regression.

Standardized Scaling: This type of scaling converts the data such that the mean value will be 0 and standard deviation of the values will be 1.

The formula used is (xi – mu) / sigma,

Where xi represents each value of x

mu is the mean

and sigma is the standard deviation.

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

The value of VIF basically indicates the extent of collinearity between the independent variables. Hence if this value is inf, it indicates that the variable is very highly correlated with one or more of the variables in the data set.

It could also happen when the dummy variables are used but the first value is not dropped.

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data came from some theoretical distribution such as a Normal or exponential. If we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It helps us visualize of the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a Scatter plot created by plotting two sets of quantiles against each other. If both sets of the quantiles come from the same distribution, the resultant plot will be roughly straight.

In all the above, the quantiles referred to are the data, below which, a certain proportion of the complete data fall.