

## 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)**

- The demand of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Lightsnow and light rainfall. We do not have any data for Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog, so we can not derive any conclusion. Maybe the company is not operating on those days or there is no demand of bike.

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

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 4 types of values in Categorical column of season. And the values are spring, summer, fall and winter. and we want to create dummy variable for that column. If one variable is not spring, summer and semi\_furnished, then it is obvious 'fall'. So we do not need 4th variable to identify the 'fall'.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

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

Validated the assumptions of Linear regression by checking the VIF, error distribution of the residuals and linear relationship between the dependent variables and a feature variable.

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

The Top 3 features contributing significantly towards the demands of share bikes are:weathersit\_Light\_Snow(negative correlation).

yr\_2019(Positive correlation).

atemp(Positive correlation).

## **General Subjective Questions**

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

**Solution -**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

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

While training the model we are given :

x: input training data (univariate – one input variable(parameter))

y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best  $\beta_0$  and  $\beta_1$  values.

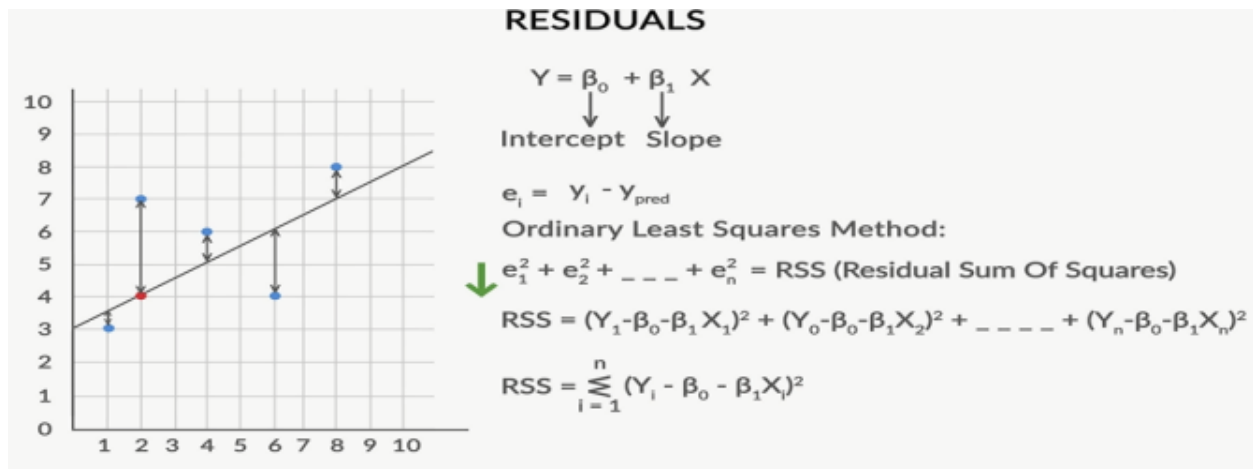
$\beta_0$  : intercept

$\beta_1$  : coefficient of x

Once we find the best  $\beta_0$  and  $\beta_1$  values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

### Best fit line:

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable



### Strengths of Linear Regression:

The strength of the linear regression model can be assessed using 2 metrics: 1. R<sup>2</sup> or Coefficient of Determination 2. Residual Standard Error (RSE). You also learnt an alternative way of checking the accuracy of your model, which is R<sup>2</sup> statistics. R<sup>2</sup> is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual

outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data. Mathematically, it is represented as:  $R^2 = 1 - (RSS / TSS)$

### R2 Formula

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where

RSS= Residual sum of square

TSS= Sum of errors of the data  
from mean

RSS (Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

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

TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

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

## 2. Explain the Anscombe's quartet in detail.

**Solution :**

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. 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.

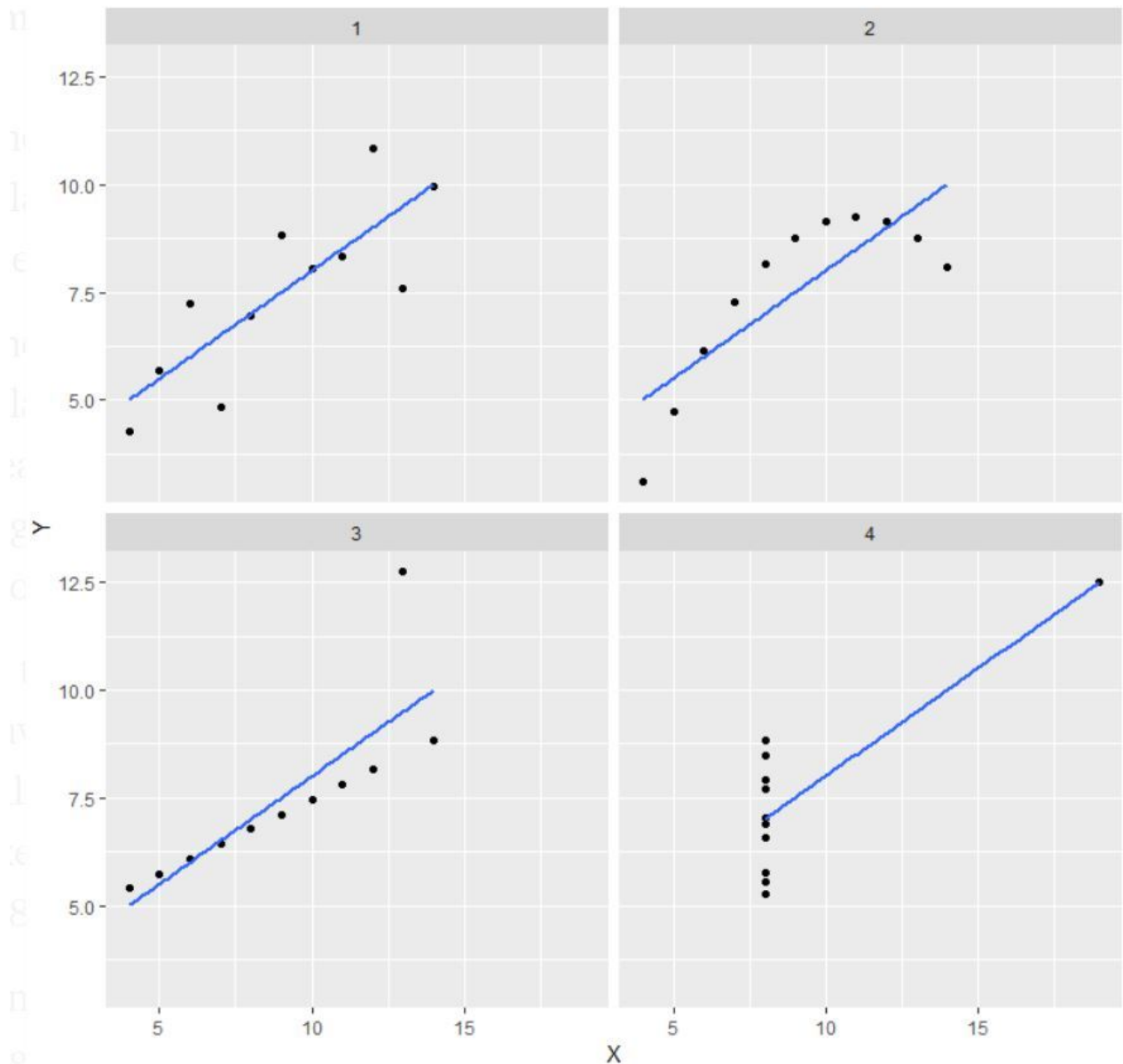
Simple understanding:

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

Summary						
Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)	
1	9	3.32	7.5	2.03	0.816	
2	9	3.32	7.5	2.03	0.816	
3	9	3.32	7.5	2.03	0.816	
4	9	3.32	7.5	2.03	0.817	

Output:



### Explanation of this output:

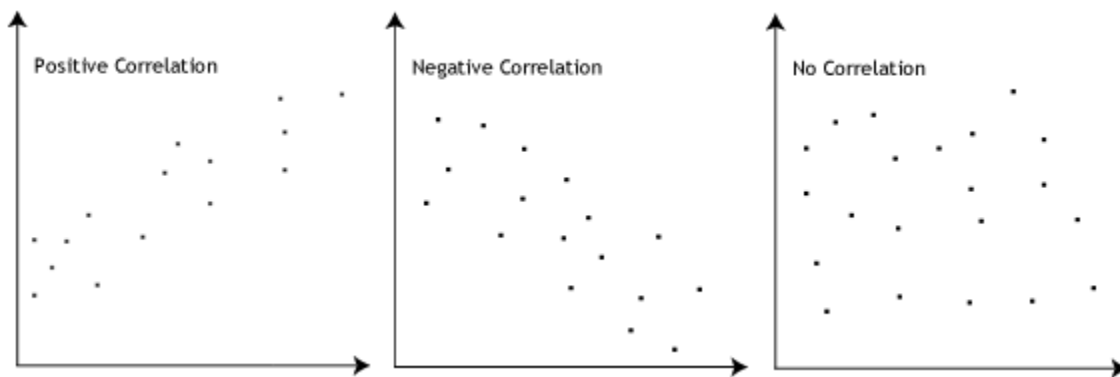
- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

### 3. What is Pearson's R?

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,

- $r$  =correlation coefficient
- $x_i$  =values of the x-variable in a sample
- $\bar{x}$  ,=mean of the values of the x-variable
- $y_i$  =values of the y-variable in a sample
- $\bar{y}$  :=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?

##### Solution:

Scaling : 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.

Need of the scaling: 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.

Difference between normalized and standardized scaling:

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO.	Normalisation	Standardisation
-------	---------------	-----------------

1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called <b>MinMaxScaler</b> for Normalization.	Scikit-Learn provides a transformer called <b>StandardScaler</b> for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

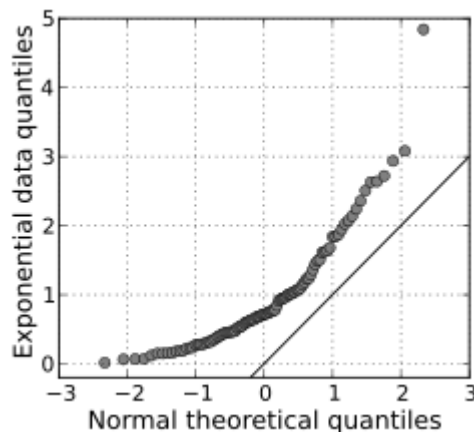
If there is perfect correlation, then  $VIF = \infty$ . This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get  $R^2 = 1$ , which leads to  $1/(1-R^2)$  infinity. To solve this problem we need to drop one of the variables from the dataset which 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.

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.

