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)

- Demand of bike is significantly less in spring season when compared to other seasons
- Demand of bikes significantly less during snowy weather and high during clear weather
- Bike demand is very high during Jun to Sep, however Jan is the lowest demand month.
- Bike demand was higher during year 2019
- Bike demand is less in holidays in comparison to not being holiday
- The demand of bike is almost similar throughout the weekdays, except a slight lesser on Sundays
- There is no significant change in demand irrespective of working or non-working day

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

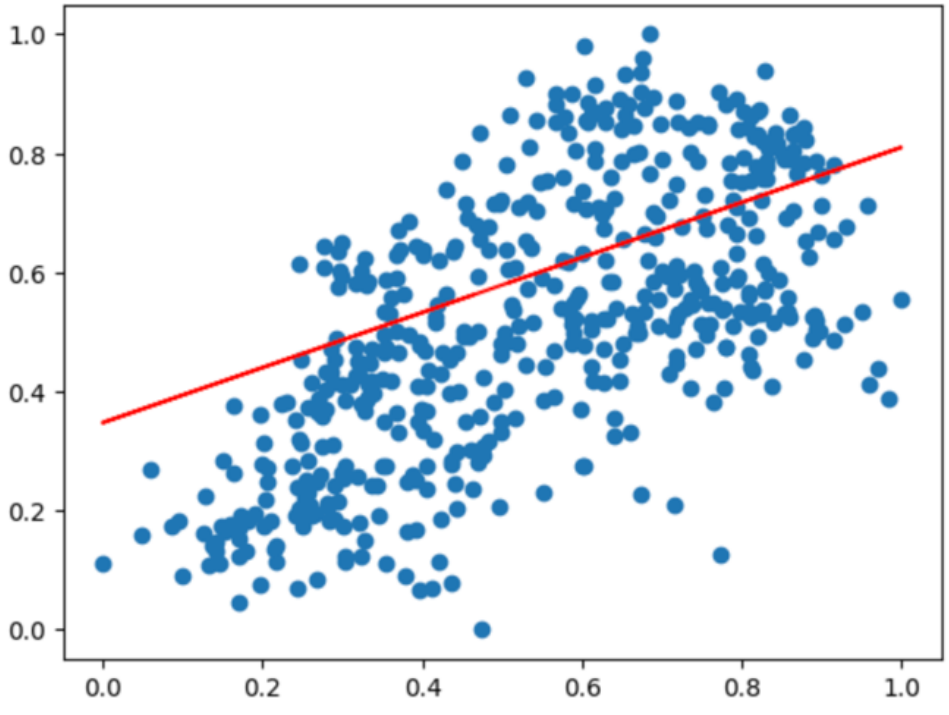
- drop_first=True drops the first column during dummy variable creation and helps to reduce the correlations created among dummy variables. So if you have N categories, it will only produce N – 1 dummy variables.

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

- Variable 'temp' has the highest correlation with 'cnt'

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

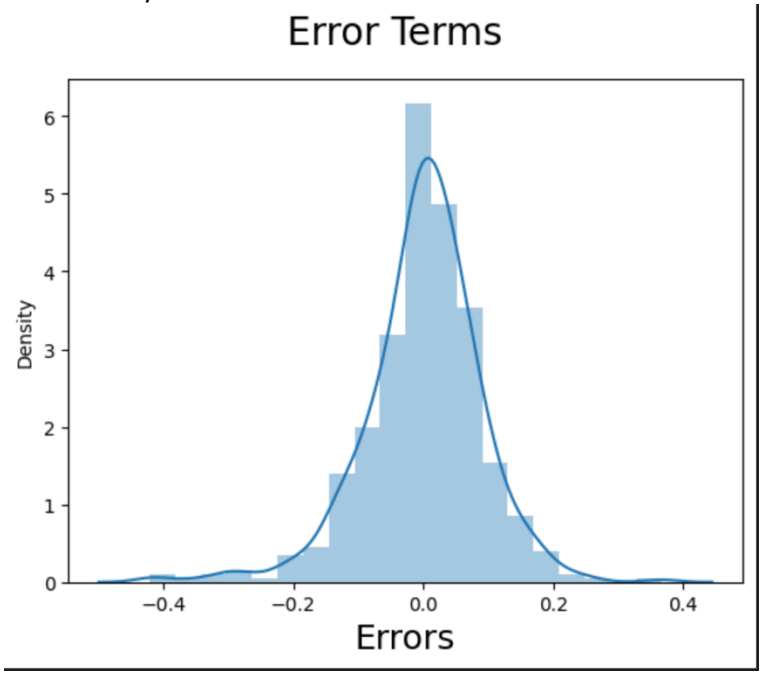
- A linear relationship is said to exist between the dependent and the independent variables. Refer the scatter plot and regression line on feature 'temp'



- No Multicollinearity. Refer the VIF values of the model features, all are less than 5.

	Features	VIF
0	const	59.65
2	temp	2.99
5	season_spring	2.54
3	hum	1.89
6	season_winter	1.77
7	weathersit_Mist	1.56
9	mnth_Jul	1.30
8	weathersit_Snow	1.24
4	windspeed	1.17
10	mnth_Sep	1.10
1	yr	1.03
11	weekday_Sun	1.01

- Normal distribution of error terms. Refer the graph below, we can see that residual error terms are normally distributed.



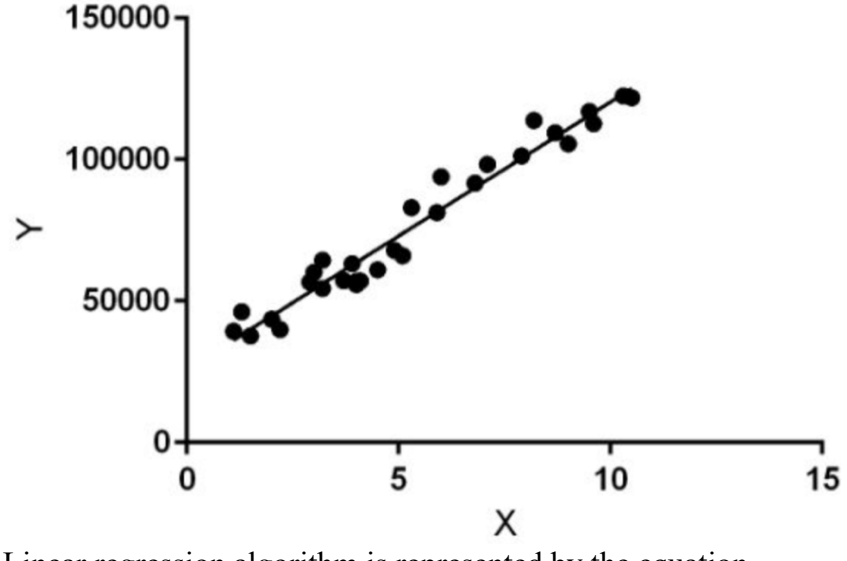
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- **Temperature** (0.479)
- **Year** (0.230)
- **Weathersit** - **Light Snow**, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds(-0.245)

General Subjective Questions

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

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis We build a predictive model through this algorithm which predicts the behaviour of data based on certain variables.
The variables which are on the x-axis and y-axis should be linearly correlated.
Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables
Below fig represents a linear relationship between dependent and independent variable



Linear regression algorithm is represented by the equation

$Y = \beta_0 + \beta_1 X$
X - independent variable
Y – dependent variable
β_0 - Intercept(where y intercepts x=0)
β_1 - slope of the line
Let us assume that we have drawn the regression line using the following set of x and y values:
(x1,y1),(x2,y2),(x3,y3).....(xn,yn)

The following formulas give the y-intercept and the slope of the equation.

$$\beta_1 = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$\beta_1 = \frac{\sum y - m\sum x}{n}$$

There are 2 types of linear regressions,

1. **Simple linear regression**: Here target variable is dependent on only 1 independent variable.
   It is represented by $Y = \beta_0 + \beta_1 X$
2. **Multilinear regression**: Here target variable depends on multiple independent variable.
   It is represented by $Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta pXp+\epsilon$

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

Anscombe's quartet is a group of four datasets which are nearly identical simple statistical properties yet appear different when plotted on a graph. Each dataset consists of eleven (x,y) points.

It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analysing and model building, and the effect of other **observations on statistical properties**.

This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

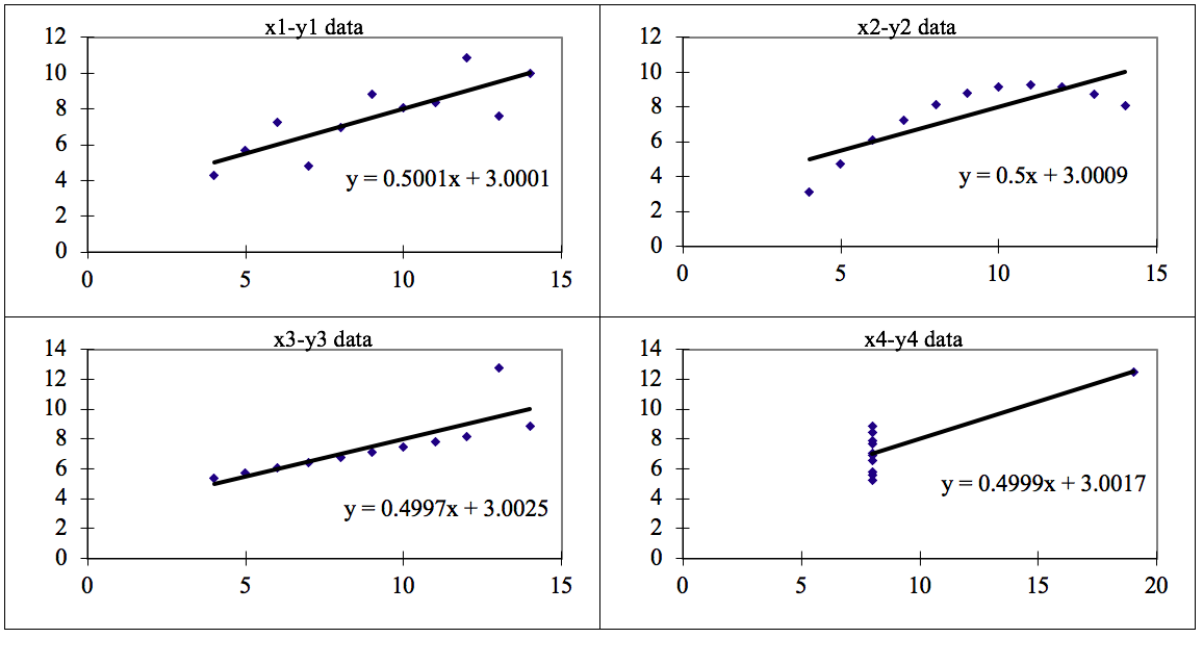
We can define these four plots as follows:

Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for these four data sets are approximately similar.
We can compute them as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

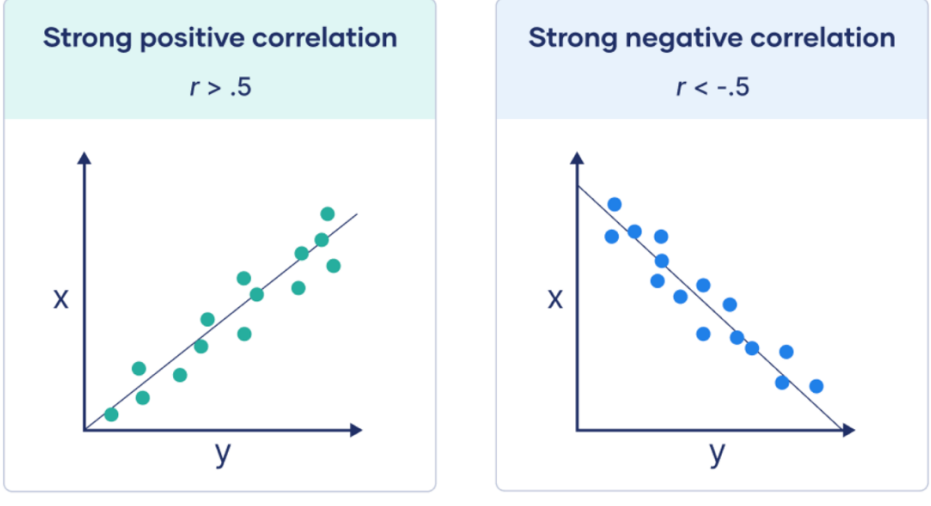
3.What is Pearson's R? (3 marks)

Pearson R is the bivariate correlation, measure of linear correlation between 2 sets of data. It is the covariance of two variables, divided by the product of their standard deviations. The Pearson's R always lies between -1 and 1

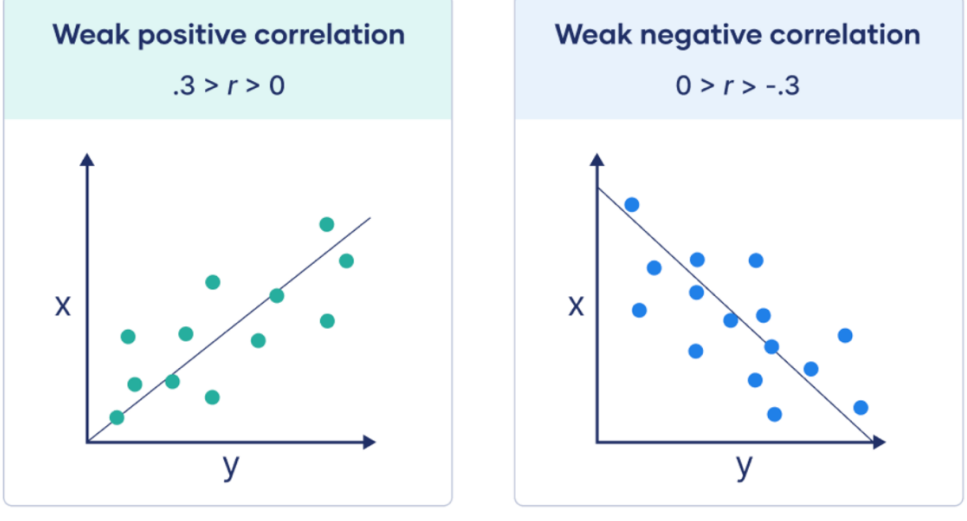
- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 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

We can also visualize the same as follows

When *r* is greater than .5 or less than –.5, the points are close to the line of best fit:



When *r* is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:



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

Scaling 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.

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.

Normalized /Min-Max Scaling:

- It brings all of the data in the range of 0 and 1
- sklearn.preprocessing.MinMaxScaler helps to implement normalization in python
- Formula to do min-max scaling is
    - MinMax Scaling : $x = \frac{x - \min(x)}{\max(x) - \min(x)}$

Standardized Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- sklearn.preprocessing.scale helps to implement standardization in python.
- Formula to perform standardized scaling is
    - Standardized Scaling : $x = \frac{x - \text{mean}(x)}{\text{sd}(x)}$
    -
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

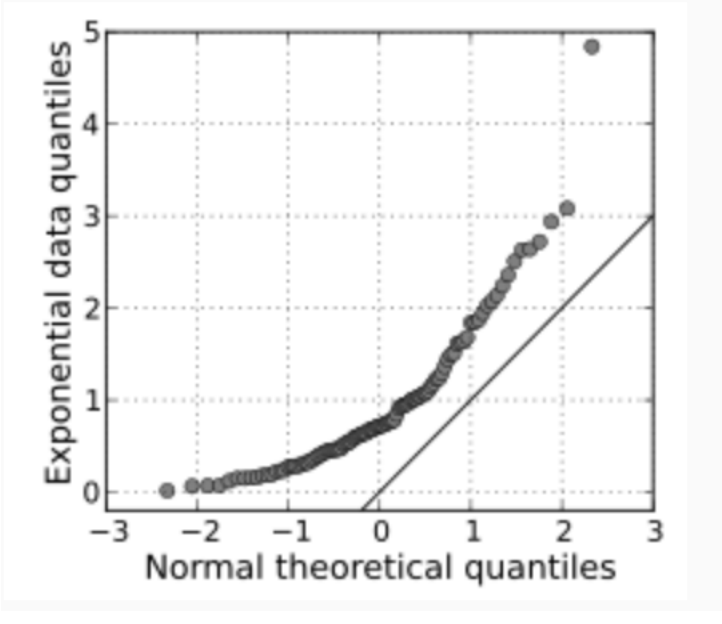
If 2 independent variables in a dataset are correlated 100 percent then we get the VIF as infinity. The value of R2 will be 1 if 2 variables are 100 percent correlated.
VIF = (1/(1-R2)) hence 1/0 leads to infinity. This high correlation between 2 independent variables is known as Multicollinearity .
To solve this problem we will need to drop one of the variables when 2 variables have high Multi collinearity.

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

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

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.