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)

- *There is more bike booking in the season of Fall .*
- *The months of May,June,July,August,Sepetember,October give higher bike bookings.*
- *Public always prefer clear weather above any other weather conditions.*
- *Thursday to Friday see higher bookings.*
- *People always book when it's not a holiday.*
- *Working day and non-working day has same booking trends.*
- *There is a year wise increase in booking*

.

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

drop_first=True

helps in reducing the extra column created during dummy variable creation.

Hence it will reduce the correlations created among dummy variables.

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

With a value of 0.63 , 'temp' variable has the highest correlation with the target variable.

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

1. *Normal distribution of error*
2. *Linear relationship among variables*
3. *Homoscedasticity- No visible pattern should be observed among residual values*
4. *Multicollinearity- no multicollinearity should be observed among the variables.*

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

Top 3 predictor variables:

- *Temperature (temp) - A coefficient value of '0.5596'.*
- *Year (yr) - A coefficient value of '0.2266'.*

- *Weather Situation 3 (weathersit_3) - A coefficient value of '-0.2678'. This causes a decrease in bike hires by 0.2678 with a unit increase.*

General Subjective Questions:

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

Linear regression is defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increases or decreases), the value of dependent variable will also change accordingly (increases or decreases).

Mathematically the relationship is represented as:−

$$Y = mX + b$$

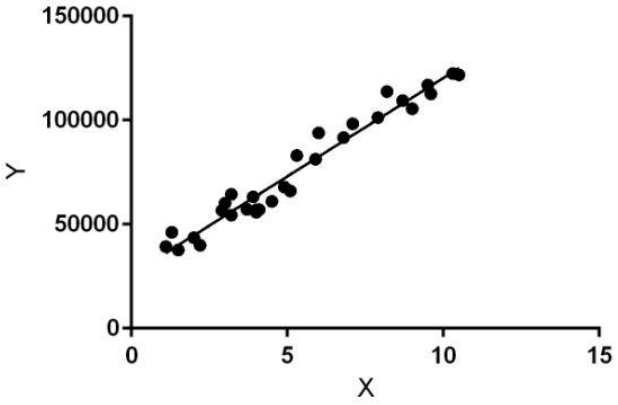
Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept.

A linear relationship will be called positive if both independent and dependent variable increases



A linear relationship will be called negative if independent increases and dependent variable decreases.

Types of Linear Regression

Linear regression is of the following two types :

- *Simple Linear Regression*
- *Multiple Linear Regression*

Assumptions

Assumptions about dataset that is made for the Linear Regression model −

Relationship between variables *– Linear regression model assumes that the relationship between response and feature variables must be linear.*

Relationship between variables *– Linear regression model assumes that the relationship between response and feature variables must be linear in neature.*

Multi-collinearity *– Linear regression model assumes that there is to be very little or no multi-collinearity in the data.*

Auto-correlation *– Linear regression model assumes is that there is very little or no auto-correlation in the data.*

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

Anscombe's Quartet is be defined as a group of four data sets which are nearly identical in simple statistics, but there are some specific factor in the dataset that fools the regression model if built.

They have very different distributions and appear differently when plotted on scatter plots.

It was created by statistician Francis Anscombe to demonstrate the importance of plotting the graphs before analysing and model building.

It tells us the importance of visualising the data before applying various algorithms to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data etc.

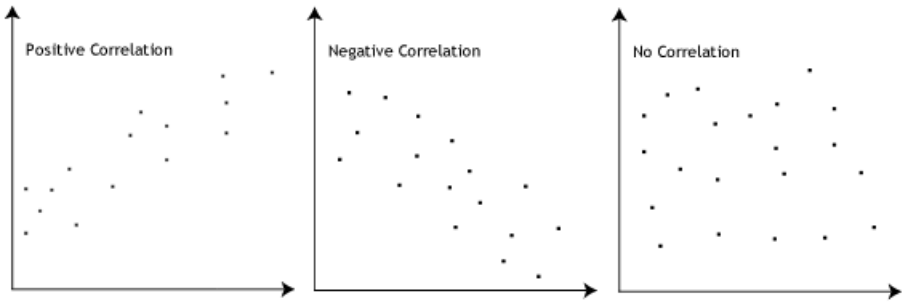
3. What is Pearson's R? (3 marks)

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r , 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 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*
- *r = -1 means the data is perfectly linear with a negative slope*
- *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*



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

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

Scaling is a method used to normalize the range of independent variables .

In data processing, it is also known as data normalization and is generally performed during the data preprocessing stage.

an example — if you have multiple independent variables like marks, age, salary, and height, with different range in values, feature scaling would scale them in same range, for example- centered around 0 or in the range (0,1) depending on the scaling technique.

Many machine learning algorithms that are using Euclidean distance as a metric to calculate the similarities will fail to give importance to the smaller feature.

Compared to age and salary, the Machine Learning algorithm will give Salary greater weightage,

Hence scaling is required.

normalized scaling	**standardized scaling**
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales	It is used when we want to ensure zero mean and unit standard deviation.
It is more affected by outliers.	It is less affected by outliers.

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

(3 marks)

When each of the features become "dependent" on each other, is that when we change an independent variable and expect a change in a dependent variable, we see that another independent variable have also changed. These two independent variables are now codependent, or collinear of each other. Add more features that are collinear of each others and we get multicollinearity. Multicollinearity occurs when independent variables in a regression model are correlated.

To solve this problem we need to drop one or more of the variables from the dataset which is causing this perfect multicollinearity.

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

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

The quantile-quantile (q-q) plot is a graphical technique to help us determine if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.

The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified.

If so, then location and scale estimators can pool both the data sets to obtain estimates of the common location and scale.

If two samples do differ, it is also useful to gain some understanding of the differences.

The q-q plot can be of help of giving more insights into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.