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)
Based on the analysis using box plot and bar plot, below are my findings

- Fall season has more booking, winter has less booking
- September and october has more number of booking
- Booking increased from2008 to 2019
- Booking seems to be equal in both working and non working days

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
It helps in reducing the extra column created during dummy variable creation. If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
'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)
I have validated based on below 5 assumptions
1. Normality of error terms: Error terms should be normally distributed
2. Multicollinearity check: There should be insignificant multicollinearity among variables.
3. Linear relationship validation: Linearity should be visible among variables
4. Homoscedasticity: There should be no visible pattern in residual values.
5. Independence of residuals: No auto-correlation

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

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Linear Regression is an algorithm that belongs to supervised Machine Learning. It tries to apply relations that will predict the outcome of an event based on the independent variable data points. The relation is usually a straight line that best fits the different data points as close as possible. The output is of a continuous form, i.e., numerical value. For example, the output could be revenue or sales in currency, the number of products sold, etc. In the above example, the independent variable can be single or multiple.

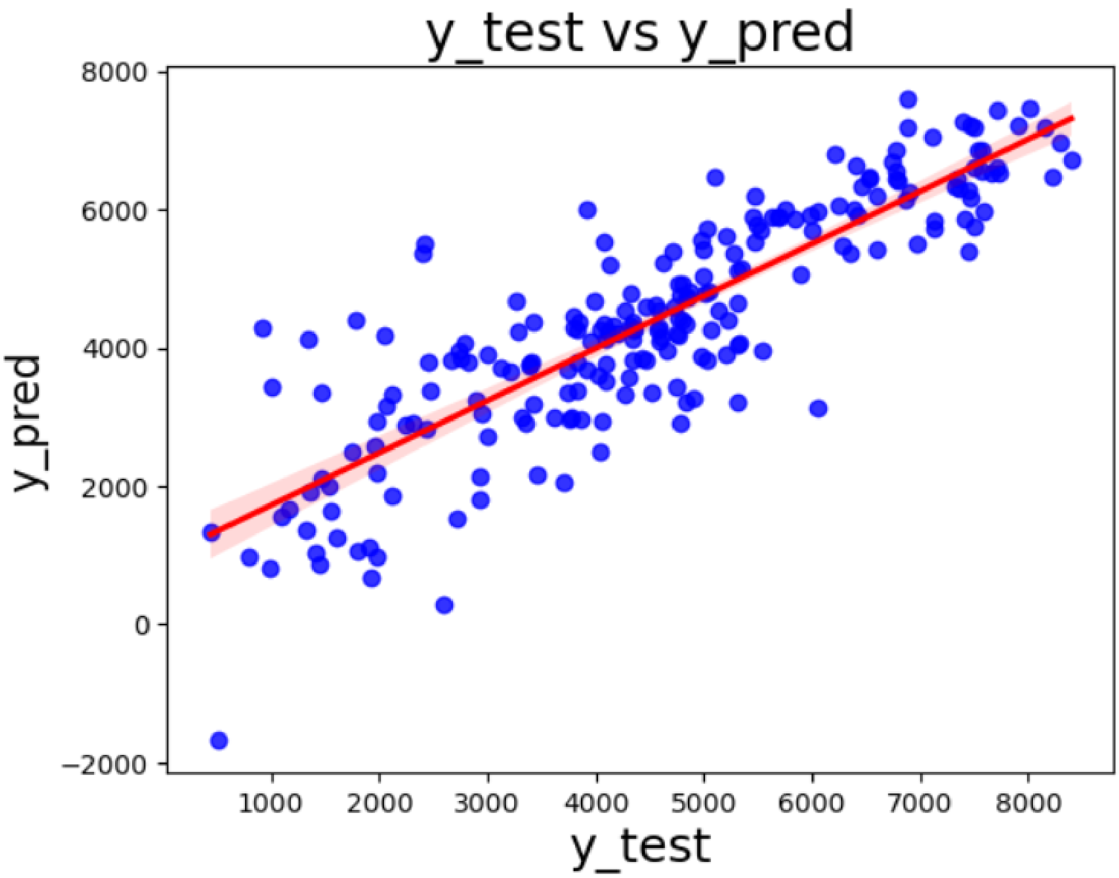
Linear regression can be expressed mathematically as:

$y = \beta 0 + \beta 1x + \varepsilon$

Where,

- Y= Dependent Variable
- X= Independent Variable
- β 0= intercept of the line
- β1 = Linear regression coefficient (slope of the line)
- ε = random error

The last parameter, random error ε, is required as the best fit line also doesn't include the data points perfectly.



2 types of Linear Regression

1. Simple Linear regression
2. Multiple Linear regression

Simple Linear Regression:

Simple linear regression has only one x and one y variable.

For instance: when we predict rent based on square feet alone that is simple linear regression

Multiple Linear regression:

Multiple linear regression has one y and two or more x variables.

For instance: when we predict rent based on square feet and age of the building that is an example of multiple linear regression

2. Explain the Anscombe’s quartet in detail. (3 marks)
Anscombe’s quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.

The four datasets that make up Anscombe’s quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson’s R? (3 marks)
The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables
The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:

- Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.
- The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It’s not a problem if the variables are a little non-normal.
- The data have no outliers: Outliers are observations that don’t follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- The relationship is linear: “Linear” means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
What is scaling?
Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Why is scaling performed?
Scaling guarantees that all features are on a comparable scale and have comparable ranges. This process is known as feature normalisation. This is significant because the magnitude of the features has an impact on many machine learning techniques.

What is the difference between normalized scaling and standardized scaling

S.NO.	Normalization Scaling	Standardization Scaling
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 MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
VIF is infinity if there is a perfect correlation. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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 for determining if two data sets come from populations with a common distribution.
Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. 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 also 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