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)
Ans: From the visulization of categorial data we found out:
- Most of the bookings are done in fall season
- There is sharp increase in number of bookings in 2019 as compared to 2018
- In the month of May, June, July, August, September and October there were most number of bookings
- During non-holiday we have more number of bookings
- Clear weather condition results in most number of bookings
- Apart from workingday column all the variables seems to be good predictor of dependent variables we will select variables according to their p-vales and VIF values

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans: It is important to use drop_first=True during dummy variable creation because by removing extra column we can decrease the number of dependent columns and even if remove that extra columns other columns can easily explain the result of removed column too.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Ans: atemp 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)
Ans: 1. We checked the p values for significance of the variables any variable with more than 0.05 p-value is removed.
2. We checked at each step if removing a variable is drasticaaly decreasing the r^2 and adjusted r^2 values.
3. During residual analysis the curve was normally distributed which indicate that it is homogeneouslydistributed.
4. When used test data with predicted values the difference between r^2 and adjusted r^2 was in 5%-10%

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

General Subjective Questions

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

Ans: Linear regression is used in supervised machine learning to predict continuous target variable using one or more independent variable. We predict dependent variable by drawing a best fit line using those independ variable.

- Generally the expression for Linear regression line is: **Y=mX+c**
  Where **m** is the slope of line
  Y is the variable we are trying to predict
  X is the independent variable and
  c is the Y intercept which remains constant
- Now according to the value of m that linear regression can be positive as well as negative.
  Positive Linear Relationship: If value of m(slope) is positive in nature that means value of Y will increase with the increase in value of value of X
  Negative Linear Relationship: If value of m(slope) is negative in nature that means value of Y will increase with the increase in value of value of X
- Now we have two 2 categories of linear regression too:
  Simple Linear Regression: In simple linear regression, we have one independent variable (predictor) and one dependent variable (the variable we want to predict). The relationship between the independent and dependent variables is assumed to be linear. The equation for simple linear regression is typically expressed as: **Y=mX+c**
  Multiple Linear Regression: In multiple linear regression, there are two or more independent variables (predictors) and one dependent variable. The relationship is still assumed to be linear, but it's more complex because it involves multiple predictors. The equation for multiple linear regression is expressed as: **Y=m1X1 + m2X2 + m3X3…..+c**
  Where m1,m2,m3 are coefficients for the independent variables
- Things to take care of while creating linear regression model
  a. All the variables should be significant.
  b. There should be no multicollinearity between the independent variables.
  c. Error terms should be normally distributed.
  d. Select appropriate evaluation metrics like Mean Squared Error (MSE), R-squared, or others depending on your specific goals.
  e. Interpret coefficients to understand the impact of each independent variable on the dependent variable.

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

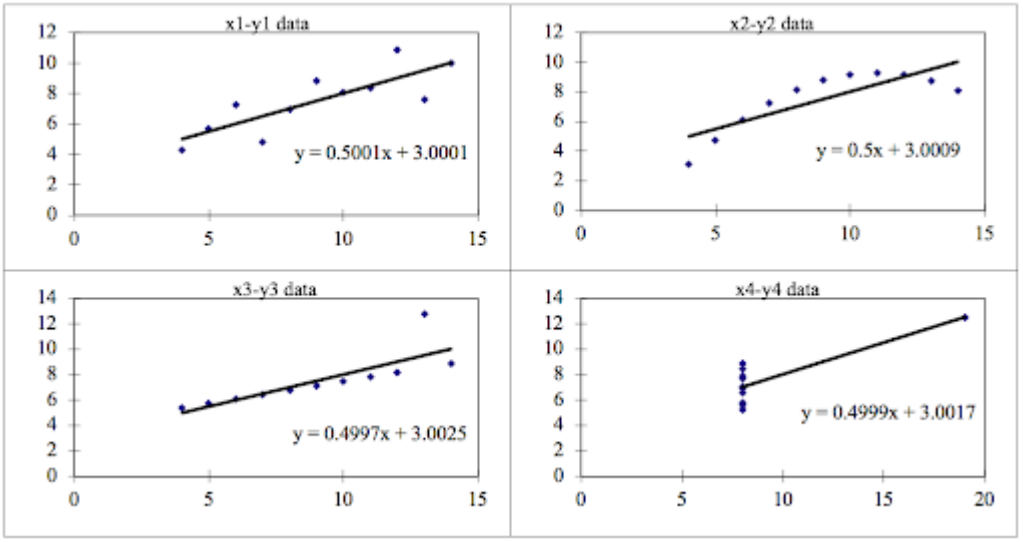
Ans: Anscombe’s quartet is a powerful reminder of the importance of visualizing data before applying various modeling algorithms. This quartet consists of four different datasets, each with its unique characteristics. It serves as a cautionary tale, highlighting the need to examine your data graphically to gain insights.

Visualizing data allows you to detect anomalies, like outliers, understand the data's diversity, and assess whether it exhibits linear relationships. Linear regression, a widely used

modeling technique, is suitable primarily for data that follows linear patterns. It struggles when faced with datasets that have non-linear relationships, making it crucial to explore your data visually to determine its suitability for linear regression.
Let's say we have 4 datasets:

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	

When we plot best fit line in all these datasets we get:



We can see the difference that in dataset1 line it fit's perfectly, in 2nd it doesn't it because relationship is not linear, in 3rd there are some outliers and in 4th outliers cannot be handled with linear regression
This demonstrates how easily a regression algorithm can be misled if we don't first understand our data visually. Therefore, it's essential to start with data visualization to gain insights, identify patterns, and ensure that our subsequent modeling efforts are well-informed and appropriate for the dataset in question.

3. What is Pearson's R? (3 marks)
Ans: Pearson's R, also known as the Pearson correlation coefficient, is a statistical metric that plays a crucial role in analyzing the relationships between two continuous variables. It

was introduced by the renowned statistician Karl Pearson and serves as a fundamental tool in data analysis.

This coefficient, which can range from -1 to 1, provides insights into how variables are related:

1. **Positive Correlation (R > 0):** When the Pearson's R is positive, it signifies a positive linear relationship between the two variables. In simple terms, as one variable increases, the other tends to increase as well.

2. **No Correlation (R = 0):** If the coefficient is approximately zero, it suggests the absence of a linear relationship between the variables. Changes in one variable do not predict or influence changes in the other.

3. **Negative Correlation (R < 0):** A negative Pearson's R indicates a negative linear relationship. In this scenario, as one variable increases, the other tends to decrease.

Pearson's R helps in understanding and quantifying linear relationships between variables, facilitating predictions based on these relationships. However, it's essential to note that this coefficient is designed for linear relationships and may not effectively capture more complex, non-linear associations between variables.

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

Ans: Scaling refers to the process of standardizing or normalizing the numerical values of features within a dataset. The primary objective of scaling is to bring all features to a similar scale or range, ensuring that no single feature dominates the others during the modeling process. With the help of scaling, we transforms these features so that they both have similar scales, often between 0 and 1 or within a standardized range like -1 to 1. This ensures that each feature contributes equally to the model's learning process.

There are two type of scaling:

MinMax scaling is typically done via the following equation:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization (Z-score Scaling):

$x(new)=(x - \mu) / \sigma$

where: x: Original value. μ: Mean of data. σ: Standard deviation of data.

MinMax scaling	Standardization (Z-score Scaling)
Range: Scales data to a range, typically 0 to 1.	Centers data to have a mean of 0 and scales by the standard deviation.
Formula: $X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$	X(new)=(X – μ) / σ
Sensitive to outliers, which can distort scaling.	More robust to outliers, as it relies on the mean and standard deviation.

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

Ans: The phenomenon of VIF (Variance Inflation Factor) becoming infinite occurs due to perfect multicollinearity in the regression model. Perfect multicollinearity means that one or more independent variables in the regression model can be perfectly predicted by a linear combination of the other independent variables. This leads to the following consequences:

1. Infinite VIF: When VIF is calculated using the formula VIF = 1 / (1 - R^2), where R^2 is the coefficient of determination in a linear regression of one independent variable against all the others, R^2 becomes 1 (perfect correlation), causing VIF to become infinite.

2. Practical Implications: Infinite VIF indicates that one or more independent variables are redundant, and their effects cannot be separated from others. In such cases, it's necessary to identify and remove one of the correlated variables from the model to resolve the issue of infinite VIF.

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

Ans: A Quantile-Quantile (Q-Q) plot is a graphical tool used in linear regression to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. Its importance lies in helping analysts and data scientists check the assumption of normality, which is crucial in many statistical techniques, including linear regression.

Use: A Q-Q plot visually compares the quantiles of the data against those expected from a theoretical distribution. If the points on the plot closely follow a straight line, it suggests that the data fits the assumed distribution. Deviations from a straight line indicate deviation from that distribution.

Importance in Linear Regression: In linear regression, it's often assumed that the residuals (the differences between observed and predicted values) follow a normal distribution. Q-Q plots of residuals help verify this assumption. If the plot deviates

significantly from a straight line, it indicates that the residuals may not be normally distributed, which can affect the validity of regression results.

In summary, Q-Q plots are a valuable diagnostic tool in linear regression, aiding in the assessment of the normality assumption, which, if violated, can impact the reliability of regression analysis.