Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

yr: Dependent variable appears to be affected by year. 2019 has higher demand than 2018.

mnth: The dependent variable is relatively more sensitive to month with higher demand from Mar-Oct.

season: The dependent variable is relatively more sensitive to season. Spring has lower demand.

weekday: The dependent variable is less sensitive to weekday

holiday: The variable should be excluded as it can result in skew in analysis as there are very few datapoint with holiday.

workingday: The dependent variable is less sensitive to working day

weathersit: The dependent variable is relatively more sensitive to weather situation.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

While creating dummy variables drop_first=True helps to eliminate variable which can be inferred from other dummy variables thus taking care of multicollinearity.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

"atemp" has the highest correlation with the target variable

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

We can reaffirm the assumption of Linear Regression after building the model on the training set using:

Residual Analysis: We can plot a distribution plot for the residual errors and understand it is centered around zero

VIF: Low variance inflation factor (VIF) of less than 5 between the features indicates that there is no multicollinearity amongst the features

---

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The top 3 features contributing significantly are:

Year

Feels like temperature

Winter

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a supervised machine learning algorithm that provides a linear relationship between one or many independent variables and dependent variable by fitting a linear equation to observed data.
There are 2 main types of Linear Regression:
Simple Linear Regression: It is the simplest form of linear regression and involves only one independent variable and one dependent variable. The equation for simple linear regression is:
$y = \beta_0 + \beta_1.X$ where,
y is the dependent variable
X is the independent variable
β_0 is the intercept
β_1 is the slope
Multiple Linear Regression: It involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:
$y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \ldots\ldots + \beta_n.X_n$ where,
y is the dependent variable
$X_1, X_2, .. X_n$ are the independent variable
β_0 is the intercept
$\beta_1, \ldots \beta_n$ are the slopes

The goal of the linear regression algorithm is to find the best fit line which minimizes the error between the predicted and actual values.

The cost function for a linear regression model is the Mean Squared Error (MSE) which is the mean of the squared errors between the predicted values and the actual values.

Cost Function J = $\frac{1}{n}\sum_{n=1}^{n}\left[\hat{y}_i - y_i\right]^2$

Simple Linear Regression makes the following assumptions:
1. **Linearity**: The independent and dependent variables have linear relationships
2. **Independence**: The observations in the dataset are independent of each other.
3. **Homoscedasticity**: The variance of the errors is constant across all the levels of the independent variables which indicates that the value of independent variables does not affect the model.
4. **Normality**: The residuals should be normally distributed.

Multiple Linear Regression makes the following assumptions:
1. **No multicollinearity:** There is no high correlation between the independent variables. Multicollinearity makes it difficult to understand the affect of each independent variable on the

dependent variable.

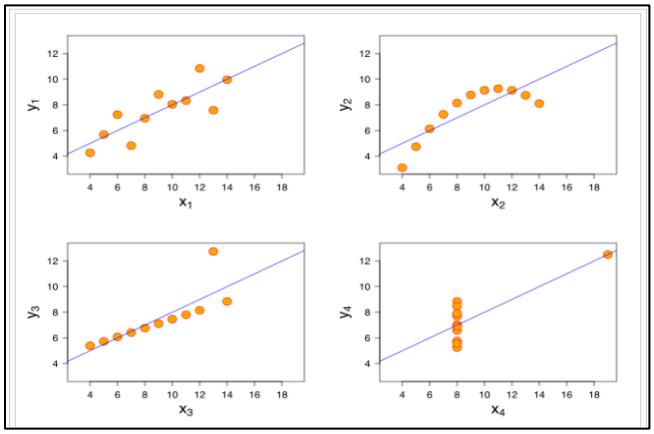
2. **Feature Selection:** Including irrelevant or redundant variables would lead to overfitting and complicate the interpretation of the model.
3. **Overfitting:** Overfitting occurs when the model fits the training data too closely, capturing noise or random fluctuations that do not represent the true underlying relationship between variables. This can lead to poor generalization performance on new, unseen data

---

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet proposed by statistician Francis Anscombe comprises of four datasets each with 11 data points and have nearly identical descriptive statistics yet have very different distributions and appears very different when graphed. This dataset emphasizes the importance of data visualization when analyzing it and the effect of outliers and other influential observations on statistical properties.

Please see below the graph for the 4 datasets designed by Anscombe:



For all the four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places

For the first scatter plot, the relation appears to be a simple linear relationship between two variables.

For the second scatter plot, the relations between two variables is not linear and Pearsons correlation coefficient is not relevant.
For the third scatter plot, the modelled relationship though linear should have been a better line since it is offset by one of the outliers.
For the fourth scatter plot, when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R proposed by Karl Pearson is a statistical measure that quantifies the degree of linear correlation between two variables. It is a score between -1 to +1.
A value of +1 is the result of a perfect positive relationship between two or more variables. Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship. Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A value of zero indicates no correlation.

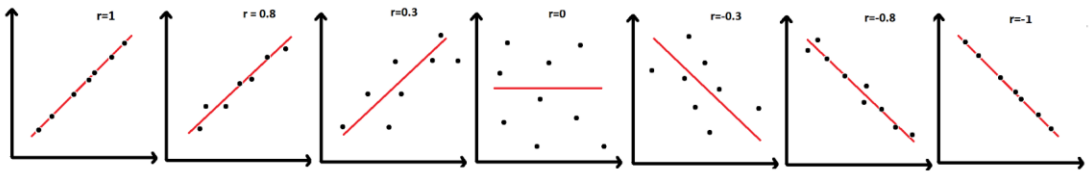
The formula for Pearsons's R is as under:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y.

Please see the below graph for the different values of Pearson's R:



r value	Interpretation
$r = 1$	Perfect positive linear correlation
$1 > r \geq 0.8$	Strong positive linear correlation
$0.8 > r \geq 0.4$	Moderate positive linear correlation
$0.4 > r > 0$	Weak positive linear correlation
$r = 0$	No correlation
$0 > r \geq -0.4$	Weak negative linear correlation
$-0.4 > r \geq -0.8$	Moderate negative linear correlation
$-0.8 > r > -1$	Strong negative linear correlation
$r = -1$	Perfect negative linear correlation

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling also known as feature scaling is a technique to standardize the independent variables in the data to a fixed range.
Scaling data provides the below benefits:

1. Scaling features makes ensuring that each characteristic is given the same consideration during the learning process. Without scaling, bigger scale features could dominate the learning, producing skewed outcomes. This bias is removed through scaling, which also guarantees that each feature contributes fairly to model predictions
2. When the features are scaled, several machine-learning methods, including gradient descent-based algorithms, distance-based algorithms (such k-nearest neighbours), and support vector machines, perform better or converge more quickly. The algorithm's performance can be enhanced by scaling the features, which can hasten the convergence of the algorithm to the ideal outcome.
3. Numerical instability can be prevented by avoiding significant scale disparities between features. Examples include distance calculations or matrix operations, where having features with radically differing scales can result in numerical overflow or underflow problems. Stable computations are ensured, and these issues are mitigated by scaling the features.

There are several scaling techniques:

- Absolute Maximum Scaling
- Min-Max Scaling
- Normalization
- Standardized Scaling

Normalized Scaling
The scaled variable is calculated by subtracting each entry by the mean value of the whole data and then dividing the result by the difference between the minimum and the maximum value.

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{X_{\text{max}} - X_{\text{min}}}$$

Standardized Scaling
This method of scaling uses the central tendencies and variance of the data to scale the variables.
The scaled variable is defined as:

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{\sigma}$$

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

In the Variance Inflation Factor (VIF) method, we assess the degree of multicollinearity by selecting each feature and regressing it against all other features in the model. This process calculates how

much the variance of a regression coefficient is inflated due to the correlations between independent variables.

The VIF for each feature is given by the following formula:

$$:VIF = \frac{1}{1-R^2}$$

Where, R-squared is the coefficient of determination in linear regression. Its value lies between 0 and 1 from the linear regression of one feature against the others. R-squared measures the proportion of variance in the dependent variable that is predictable from the independent variables, and it ranges between 0 and 1. A higher R-squared value suggests a stronger relationship between the feature and other predictors, which results in a higher VIF. If R-squared is close to 1, this indicates a high degree of multicollinearity, as the feature can be largely explained by the other variables in the model.

If R2 = 1 due to complete predictability of the independent variable by other independent variables then VIF would be infinite.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The quantile-quantile (q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

To draw a Quantile-Quantile (Q-Q) plot, you can follow these steps:

Collect the Data: Gather the dataset for which you want to create the Q-Q plot. Ensure that the data are numerical and represent a random sample from the population of interest.
Sort the Data: Arrange the data in either ascending or descending order. This step is essential for computing quantiles accurately.
Choose a Theoretical Distribution: Determine the theoretical distribution against which you want to compare your dataset. Common choices include the normal distribution, exponential distribution, or any other distribution that fits your data well.
Calculate Theoretical Quantiles: Compute the quantiles for the chosen theoretical distribution. For example, if you're comparing against a normal distribution, you would use the inverse cumulative distribution function (CDF) of the normal distribution to find the expected quantiles.
Plotting:
Plot the sorted dataset values on the x-axis.
Plot the corresponding theoretical quantiles on the y-axis.
Each data point (x, y) represents a pair of observed and expected values.

Connect the data points to visually inspect the relationship between the dataset and the theoretical distribution

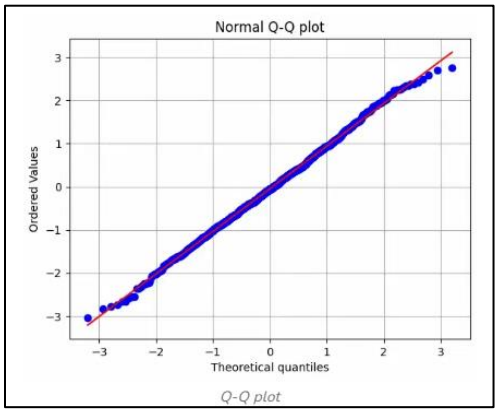
Interpretation of Q-Q plot

If the points on the plot fall approximately along a straight line, it suggests that your dataset follows the assumed distribution.

Deviations from the straight line indicate departures from the assumed distribution, requiring further investigation.

Exploring Distribution Similarity with Q-Q Plots

Exploring distribution similarity using Q-Q plots is a fundamental task in statistics. Comparing two datasets to determine if they originate from the same distribution is vital for various analytical purposes. When the assumption of a common distribution holds, merging datasets can improve parameter estimation accuracy, such as for location and scale. Q-Q plots, short for quantile-quantile plots, offer a visual method for assessing distribution similarity. In these plots, quantiles from one dataset are plotted against quantiles from another. If the points closely align along a diagonal line, it suggests similarity between the distributions. Deviations from this diagonal line indicate differences in distribution characteristics.



Advantages of Q-Q plot

- **Flexible Comparison**: Q-Q plots can compare datasets of different sizes without requiring equal sample sizes.
- **Dimensionless Analysis**: They are dimensionless, making them suitable for comparing datasets with different units or scales.
- **Visual Interpretation**: Provides a clear visual representation of data distribution compared to a theoretical distribution.
- **Sensitive to Deviations**: Easily detects departures from assumed distributions, aiding in identifying data discrepancies.
- **Diagnostic Tool**: Helps in assessing distributional assumptions, identifying outliers, and understanding data patterns