

## 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)

The categorical variables shows a significant impact on cnt variable. Like on a clear and cloudy day the sales are similar and on snowy day it is less. Similarly sales during summer and winter are similar and during spring it is less.

---

**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)

Using drop\_first=True when creating dummy variables (one-hot encoding) is important to avoid the dummy variable trap, which refers to a situation where multicollinearity arises in regression models.

1. Prevents Multicollinearity
  2. Ensures Model Interpretability
  3. Reduces the Number of Features
- 

**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)

Temp, atemp, casual and registered as high correlation

---

**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)

I have dropped the columns like instant, casual and registered. When these are added, model gave the R2 as 1 which is not recommended. These variables are overfitting the model. Hence removed them.

While checking p and VIF values there are multiple variables which has high P and VIF like "workingday", "temp", "hum", "clear", "day", "mnth", "atemp". Hence removed them.

---

**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)

Year and weekday are the most contributing features

---

## 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)

<Your answer for Question 6 goes here>

Linear regression is one of the most fundamental and widely used statistical and machine learning algorithms. It models the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a straight line to the data.

Types of Linear Regression

Simple Linear Regression :Involves one independent variable.

Multiple Linear Regression: Involves multiple independent variables.

Assumptions of Linear Regression

To use linear regression effectively, the following assumptions should be met:

1. Linearity: The relationship between independent and dependent variables is linear.
  2. Independence: Observations should be independent of each other.
  3. Normality of Residuals: Residuals (errors) should be normally distributed.
  4. No Multicollinearity: Independent variables should not be highly correlated with each other.
- 

**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)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four datasets that have nearly identical statistical properties but differ significantly when graphed. It was created by the statistician **Francis Anscombe** in 1973 to illustrate the importance of **graphical analysis** in statistics.

---

**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)

<Your answer for Question 8 goes here>

Pearson's  $r$ , also known as the Pearson correlation coefficient (PCC), is a statistical measure that quantifies the linear relationship between two variables. It is denoted by  $r$  and ranges from -1 to +1.

---

**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)

<Your answer for Question 9 goes here>

Scaling is a data preprocessing technique used to adjust the range of numerical variables so that they are on a similar scale. This is especially important in machine learning algorithms that are sensitive to differences in magnitude, such as gradient-based models (e.g., linear regression, logistic regression, SVMs, neural networks).

Scaling is done for several reasons:

1. **Prevents dominance of large values** – Some features may have much larger values than others, making models biased toward those features.
2. **Improves model convergence** – Gradient descent and optimization algorithms converge faster when features are on the same scale.
3. **Enhances performance of distance-based models** – Algorithms like KNN, K-means, and SVMs rely on distances (e.g., Euclidean distance), which can be distorted by different feature magnitudes.
4. **Prepares data for regularization** – Regularized models (like Ridge or Lasso regression) work better when features have comparable scales.
5. **Improves interpretability** – Some models, like PCA, rely on variance; scaling ensures meaningful principal components.

Normalized Scaling:

- **Transforms data into a fixed range (0 to 1 or -1 to 1).**
- **Sensitive to outliers** because it depends on min and max values.
- **Best suited for cases where data distribution is not normal and has bounded values.**
- Used in deep learning models (e.g., neural networks) where values need to be scaled between 0 and 1.

Standardized Scaling:

- **Centers the data around mean 0 with a standard deviation of 1.**
- **Not affected by outliers as much as normalization** (but still sensitive).
- Works well when data follows a normal distribution.
- Used in models like SVM, PCA, logistic regression, and linear regression.

---

**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)

<Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) is a metric used to detect multicollinearity in regression models. It quantifies how much the variance of a regression coefficient is inflated due to correlation with other independent variables.

---

**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)

<Your answer for Question 11 goes here>

A Q-Q (Quantile-Quantile) Plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (usually the normal distribution). It helps in assessing whether a given dataset follows a specified distribution by plotting the quantiles of the sample data against the quantiles of the theoretical distribution.

---