# 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)

List of categorical variables

- Season
  - Spring
  - Summer
  - Fall
  - Winter
- Year
  - 2018
  - 2019
- weathersit
  - Clear
  - Mist
  - Light Rain
  - Heavy Rain
- holiday
  - Yes
  - No
- month
  - Jan
  - Feb
  - Mar
  - Apr
  - May
  - Jun
  - July
  - Aug
  - Sep
  - Oct
  - Nov
  - Dec
- weekday
  - Yes
  - No

1. Seasons:
   - Winter (0.1859, $p < 0.001$) and Summer (0.1062, $p < 0.001$) have a significant positive impact.
   - Fall (0.0352, $p = 0.235$) is not statistically significant.
2. Year:
   - 2019 (0.2297, $p < 0.001$) shows a significant increase compared to the reference year.
3. Holiday:

- Negative effect (-0.0726, p = 0.008) on the dependent variable, indicating lower values on holidays.

4. Working Day:
   - Positive effect (0.0178, p = 0.052), but not statistically significant at the 5% level.
5. Dropping the month as it's not having much significance.

Key Takeaways:

- Winter and Summer increase the dependent variable, while Fall is not significantly different from Spring (reference category).
- 2019 had a significantly higher effect compared to the base year.
- Holidays lead to a drop in the dependent variable, while working days have a small positive impact but are not strongly significant.

**Season**

| | Coef | p-value |
|---|---|---|
| seasson_fall | 0.0352 | 0.235 |
| seasson_winter | 0.1859 | 0.000 |
| seasson_summer | 0.1062 | 0.000 |

**Year**

| | Coef | p-value |
|---|---|---|
| yr_2019 | 0.2297 | 0.000 |

**Holiday**

| | Coef | p-value |
|---|---|---|
| holiday | -0.0726 | 0.008 |

**Working day**

| | Coef | p-value |
|---|---|---|
| working-day | 0.0178 | 0.052 |

**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** in dummy variable creation helps **avoid multicollinearity** by removing one category from the dummy variables.

**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)

Temperature is having strong correlation with the Bike rental count

---

**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)

By checking whether the residual errors are normally distributed.
By checking the VIF value.
By plotting the residuals vs. fitted values; the points should be evenly spread.

---

**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)

- temp
- seasson
- Yr

---

## 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>

Explanation of the Linear Regression Algorithm

Linear Regression is a statistical method used to find the relationship between one or more independent variables (X) and a dependent variable (Y). It helps predict future values based on this relationship. Linear Regression is a simple but powerful technique for making predictions based on relationships between variables.

The algorithm finds the best-fitting line for the data using the Least Squares Method.

Steps in Linear Regression:

1. Define the equation : The model assumes a linear relationship:
   y=mx+c + error
   m - slope
   c - intercept
   x - independent variable
1. Use the Least Squares Method : The algorithm calculates coefficients.
2. Evaluate Model Performance : Several parameters help assess if the model is reliable:
   - p-value → Checks if a variable is statistically significant (lower p-value means it's important).
   - R-squared ($R^2$) → Measures how well the model explains the variation in data (closer to 1 means a better fit).
   - Coefficient (Coef) → Shows how much Y changes with a one-unit increase in X.
   - Standard Error (Std Err) → Measures how accurate the coefficient estimates are (lower is better).
1. Check Assumptions : Linear Regression assumes:
   - A linear relationship between X and Y.
   - Residuals (errors) are normally distributed.
   - No multicollinearity (independent variables should not be highly correlated).
   - Homoscedasticity (constant variance of residuals).

---

**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 different datasets that have nearly the same statistical properties (mean, variance, correlation, and regression line) but look very different when plotted.

**Why is it Important?**

It shows that relying only on summary statistics (like mean, variance, and correlation) can be misleading. Visualizing data is crucial to understanding patterns, outliers, and relationships.

**The Four Datasets Include:**

1. A normal linear relationship.
2. A curved relationship (not suitable for linear regression).
3. A strong linear relationship with an outlier affecting the trend.
4. A dataset where one extreme outlier dominates the regression.

**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 (or Pearson correlation coefficient) measures the strength and direction of the linear relationship between two variables.

- Range: -1 to 1
- 1 → Perfect positive correlation
- -1 → Perfect negative correlation
- 0 → No correlation

It helps determine how strongly two variables are related.

**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>

**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>

**What is Scaling?**
Scaling is the process of transforming numerical values so they fit within a specific range. It ensures that all features in a dataset have a similar scale, preventing certain features from dominating others due to their larger values.
**Why is Scaling Performed?**
**Improves model performance** – Many machine learning models work better when features have a similar range.
**Faster convergence** – Algorithms like gradient descent work more efficiently with scaled data.
**Difference Between Normalization and Standardization**
Normalization : Scales between 0 and 1 (or -1 and 1)
Standardization : Mean = 0, Std = 1

**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>

The Variance Inflation Factor (VIF) becomes infinite when there is perfect **multicollinearity** between two or more 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 graph used to check if a dataset follows a normal distribution

**Use & Importance in Linear Regression**

In linear regression, one key assumption is that the residuals (errors) should be normally distributed. A Q-Q plot helps check this assumption:

1. If residuals follow a **normal distribution**, the regression model is likely reliable.
2. If residuals deviate from normality, the model may violate assumptions, affecting prediction accuracy.