## 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)

From Univariate and bivariate analysis performed we received the following information:

- Most of the bikes are rented on working days.
- Least number of bikes are rented on public holidays.
- Fall season had highest bike rentals.
- 2019 had more bike rentals than 2018.
- Peak bike rentals occurred during September.
- The highest bike rentals were observed during clear weather conditions.

---

**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)

When creating dummy variables, using drop_first=True ensures that one category is dropped from each feature. This prevents the model from having redundant information, which could lead to incorrect results due to 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 displays the highest correlation with the target variable.

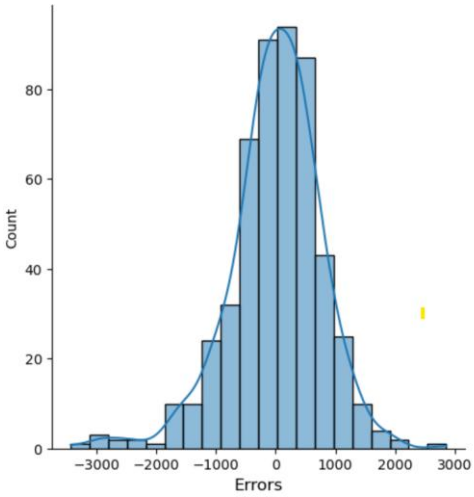
---

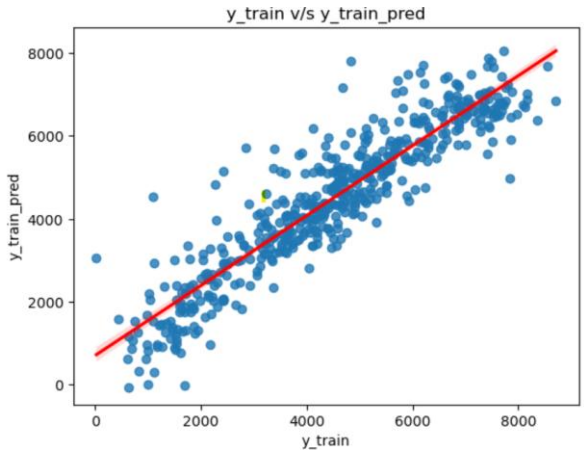
**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

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

To validate the assumptions of Linear Regression after building the model, I performed residual analysis. This involved plotting a graph of the actual values (y_train) versus the predicted values (y_train_pred). In the graph, the points were scattered around the fitted line, which indicated that the residuals were randomly distributed.




---

**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)
From the model params we can determine that the top 3 contributing features are:

- Temperature (temp)
- Winter
- Year(yr)

[172]:
```
# Model Params
print(model.params)
```

```
const       4491.303327
yr           997.283116
holiday     -144.317309
temp        1114.637217
hum         -210.089508
windspeed   -272.332177
Summer       338.007812
Winter       530.943515
Sep          247.104324
Sun         -142.228636
Light_snow  -357.515816
Misty       -224.310729
```

---

## 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 a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes that there is a linear relationship between the variables, meaning the change in the dependent variable can be explained by the linear combination of the independent variables.

- Simple linear regression: We have one independent variable.

  Equation: $y=\beta0+\beta1\cdot x+\epsilon$

- Multiple Linear Regression: we extend the simple linear regression to more than one independent variable.

  Equation: $y=\beta0+\beta1\cdot x1+\beta2\cdot x2+\cdots+\beta n\cdot xn+\epsilon$

To find the best-fitting line, linear regression uses the Ordinary Least Squares (OLS) method. OLS minimizes the sum of the squared differences (residuals) between the observed values and the predicted values.

Following are the methods to evaluate performance:

- R-squared ($R^2$): Indicates how well the model explains the variance in the target variable. Higher $R^2$ means better fit.
- Adjusted R-squared: Adjusts $R^2$ for the number of predictors, useful when comparing models with different numbers of features.
- Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values.
- Root Mean Squared Error (RMSE): The square root of MSE, providing error in the same unit as the dependent variable.

**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 consists of four datasets with identical summary statistics (mean, variance, and correlation), but vastly different data patterns. It was created to highlight the importance of data visualization and how summary statistics alone can be misleading.

Here are the four datasets:

Despite these datasets having identical basic statistical properties, the visual patterns in each dataset are vastly different

| X | Y1 | Y2 | Y3 | Y4 |
|---|---|---|---|---|
| 10 | 8.04 | 9.14 | 7.46 | 7.74 |
| 8 | 6.58 | 8.14 | 6.77 | 7.78 |
| 13 | 7.58 | 8.74 | 12.74 | 7.82 |
| 9 | 8.81 | 8.77 | 7.11 | 8.84 |
| 11 | 8.33 | 9.26 | 7.81 | 8.47 |
| 14 | 9.96 | 8.10 | 8.84 | 7.04 |
| 6 | 7.24 | 6.13 | 6.08 | 5.94 |
| 4 | 4.26 | 3.10 | 5.39 | 5.26 |
| 12 | 10.84 | 9.13 | 8.15 | 12.74 |
| 7 | 4.82 | 7.26 | 6.42 | 5.49 |
| 5 | 5.68 | 4.74 | 6.34 | 5.19 |

- Dataset 1: Shows a clear linear relationship.
- Dataset 2: Displays a curved relationship.
- Dataset 3: Has a linear relationship but is influenced by an outlier.
- Dataset 4: Appears linear but has a vertical alignment with one outlier.

---

**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 is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is a value between -1 and 1, where:

- 1 indicates a perfect positive linear relationship (as one variable increases, the other also increases in a perfectly linear manner).
- -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases in a perfectly linear manner).
- 0 indicates no linear relationship between the variables.

---

**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 adjusting the range of data to ensure features are on a similar scale, which improves the performance and accuracy of many machine learning models.

Scaling is performed:

- Ensures features are treated equally.
- Speeds up convergence in algorithms.
- Prevents bias towards features with larger values.

Normalized Scaling vs Standardized Scaling:

- Normalized Scaling: Rescales data to a fixed range (0 to 1) => used when data need to be in a bounded range.
- Standardized Scaling: Centers data to have a mean of 0 and a standard deviation of 1 => this is used when we have normally distributed data.
- One major difference is Normalization is sensitive to outliers; Standardization is more robust.

---

**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 formula for calculating VIF is 1/(1-R^2).

- In case of perfect correlation value of R=1, so in this case the value of VIF is infinite.
- If a variable is a direct linear combination of other variables, the model cannot distinguish between them, and the VIF for these variables may become infinite.
- If there are duplicate columns in the dataset, VIF will be infinite for those columns as they provide no additional information.

---

**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 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 for the conclusion that the two data setshave come from populations with different distributions.