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?

Ans: Analysis of the categorical variables from the dataset,

1. From the season vs rentals per day plot , fall has the highest average rentals followed by summer.

2. Looking at year by year rentals, 2019 has had a median 2000 increase in rentals compared to 2018.

3. From the month wise plot, September has the highest rentals, followed by the two months surrounding it.

4. It seems like the trend is explained by seasonal rentals too

Holidays show lower rental count compared to working days, with greater variability in demand on holidays.

5. There is no significant difference between rentals vs weekdays, except that Thursdays and sundays have a higher variation in rentals than others.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans: **drop_first=True** helps to prevent multicollinearity and simplifies the interpretation of the regression coefficients when dealing with categorical variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation
with the target variable? (1 mark)

Ans: atemp and temp

4. How did you validate the assumptions of Linear Regression after building the model on the
training set? (3 marks)

Ans:

1. 13 features have been selected.

2. All the features are statistically significant [low p-value].

3. The model over is a good fit with Prob (F-statistic): 2.28e-186 The model explains 83.8% variability in the training data. Adjusted R-square being 83.4%.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans:

```
cnt = 4527.31685981 + 2000.813889 yr + 685.560556
Saturday + 625.322329 winter + 586.229282 september  -
890.3115 july -561.583892 Mist_cloudy + 513.321998
workingday -1341.146803 hum -1100.976725 spring +
4099.972163 atemp -791.687695 windspeed -2051.379041
light snow/rain.

where atemp , windspeed and hum are normalized.
```

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a fundamental statistical and machine learning algorithm used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent variables (predictors) and the dependent variable (response).

Algorithm Steps:

1. Data Preparation
2. Model Initialization
3. Model Training
4. Model Evaluation
5. Model Interpretation:

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a famous example in statistics that demonstrates the importance of visualizing data before drawing conclusions. It consists of four datasets that have nearly identical statistical properties but exhibit vastly different patterns when plotted. This quartet was introduced by the statistician Francis Anscombe in 1973 to emphasize the limitations of summary statistics and the necessity of graphical exploration in data analysis.

- **Visualizing Data**: Despite having similar statistical properties (e.g., means, variances, correlations), the datasets in Anscombe's quartet exhibit vastly different patterns when plotted. This highlights the importance of visualizing data to understand its underlying structure and relationships.
- **Summary Statistics**: Relying solely on summary statistics (e.g., mean, variance, correlation coefficient) can be misleading, as datasets with different distributions can yield similar summary statistics.
- **Outliers**: Dataset IV demonstrates the influence of outliers on summary statistics and the importance of identifying and addressing outliers in data analysis.
- **Modeling Assumptions**: When performing statistical analysis or building predictive models, it's essential to validate assumptions such as linearity, homoscedasticity, and normality, as these may not hold true for all datasets.

3. What is Pearson's R?

Ans:Pearson's correlation coefficient, often denoted as *r* or Pearson's *r*, is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the association between the variables. Pearson's *r* ranges from -1 to 1, where:

- r=1: Perfect positive linear correlation (as one variable increases, the other variable increases proportionally).
- r=-1: Perfect negative linear correlation (as one variable increases, the other variable decreases proportionally).
- r=0:No linear correlation (the variables are not linearly related).

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

Ans: Scaling is a process used in data preprocessing to transform the values of variables into a specific range or distribution. It's performed to ensure that all variables have a comparable scale, which can be particularly important in machine learning algorithms that are sensitive to the magnitude of input features.

Normalized scaling and standardized scaling are two common scaling techniques, and they differ in how they transform the data:

Normalized Scaling: Also known as Min-Max scaling, it transforms the data to a specific range, typically between 0 and 1. It preserves the shape of the original distribution but compresses it to fit within the specified range.

Standardized Scaling: Also known as z-score normalization, it transforms the data so that it has a mean of 0 and a standard deviation of 1. This method standardizes the distribution of the data, making it easier to compare variables with different units and scales. It centers the data around the mean and scales it based on the standard deviation.

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

Ans: The phenomenon of having infinite values of Variance Inflation Factor (VIF) typically occurs when there is perfect multicollinearity among predictor variables in a regression model. Perfect multicollinearity means that one or more predictor variables can be expressed as a perfect linear combination of other predictor variables.

In other words, when one predictor variable can be exactly predicted by a linear combination of other predictor variables, it leads to perfect multicollinearity. This situation makes it impossible to estimate the coefficients of the regression model uniquely, resulting in infinite values of 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 Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess whether a set of data follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the data to the quantiles of a theoretical distribution, typically a normal distribution.

Here's how a Q-Q plot works:

- The data points are sorted in ascending order.
- The corresponding quantiles of the theoretical distribution are calculated.
- Each data point is plotted against its corresponding quantile from the theoretical distribution.

In linear regression, Q-Q plots are particularly useful for assessing the normality assumption of the residuals (i.e., the difference between observed and predicted values). Here's why Q-Q plots are important in linear regression:

- **Assumption Checking**: Linear regression models typically assume that the residuals are normally distributed. Q-Q plots provide a visual check to assess whether this assumption holds true. If the residuals deviate significantly from the straight line in the Q-Q plot, it suggests that the normality assumption may be violated.
- **Model Evaluation**: Q-Q plots help evaluate the adequacy of the linear regression model. If the residuals are normally distributed, it indicates that the model adequately captures the underlying relationship between the predictors and the response variable. On the other hand, non-normal residuals may indicate that the model needs improvement.
- **Detecting Outliers**: Q-Q plots can also help detect outliers in the data. Outliers may appear as points that deviate significantly from the expected linear pattern in the plot.
- **Guiding Model Improvements**: If the Q-Q plot reveals departures from normality, it suggests potential areas for model improvement. Techniques such as data transformation or using robust regression methods may be considered to address non-normality in the residuals.