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? (3 marks)

   Ans: Inferences about the categorical variables on the dependent variable are as follows:
   1) The target variable 'cnt' representing the bike rental count is highest in fall followed by summer compared to spring, and after fall, it starts declining in winter.
   2) The 25th percentile of the bike rental count in 2019 was nearly equal to 75th percentile the count in 2018 , implying the increase in the rentals with year.
   3) The count shows gradual increase from the months from Jan-Sep; with visible intervals as Jan-Apr and then May-Sep. It declines further from Oct-Dec; which is in accordance with the first boxplot showing effect of season on the count.
   4) On a Holiday, the rental count is decreased, which may be logical as well, as people might be preferring family time over biking.
   5) Weekday or Weekend does not seem to have any significant effect on the rentals.
   6) There are no rentals in 'severe' weather, and maximum rentals in 'good' weather, which is quite logical.
   7) There are outliers in Season, which need to be treated.

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

   Ans: To convert a categorical column with *'k'* number of categories, *(k-1)* dummy variables are sufficient. With this approach, there will be less memory used. For large number of categorical features, saving 1 dummy variable's memory per feature would lead to a lot of memory efficiency.

   Additionally, it is also used to reduce the collinearity between dummy variables .

   For example to map k=3 categories, instead of using 3 dummy variables as 100, 010, 001, 2 dummy variables will be sufficient. Here, dropping the first dummy variable will lead to converted or mapped values as 00,10 and 01. Absence of any dummy variable set to 1, can still indicate the 1st value.

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

   Ans: The variable 'temp' and 'atemp' has the highest correlation of 0.63 with the target variable.

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

   Ans: Assumptions of linear regression were validated by:
   1) Residual Analysis: Error terms are normally distributed with mean 0, Error terms are independent of each other. Actual vs Predicted result follow the same pattern.
   2) R2 value for prediction on test data is very close to that for train data. That means the model performs well on unseen data as well.

3) Homoscedacity: Variance of residuals is constant across predictions i.e. error terms don't vary drastically with change in predictor variable.
4) Actual Test vs Predicted Test values plot: The two values are very close to each other.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: The top 3 features are:
1. **temp** (positive correlation coefficient: 4126.2436) ,
2. **yr** (positive correlation coefficient : 2172.3695) and
3. **windspeed** (negative correlation coefficient : -1115.4101)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It's like finding the best-fit line through a scatter plot of data points. This line is used to predict future values. If there are a set of points on a graph, each representing a pair of values, the goal of Linear Regression is to draw a straight line that best represents these points. This line can be represented by the equation $y=mx+c$, where:

y is the dependent variable to predict.
x is the independent variable we use for prediction.
m is the slope of the line.
c is the y-intercept (where the line crosses the y-axis).

The "best-fit" line minimizes the distance between the line and all data points. The difference between actual and predicted values is called error or **residual**.
Least Squares Method: Linear regression uses the least squares method to minimize the sum of the squares of these errors.
Training: Historical data is used to find the best m and b.
Prediction: Once the model is trained, y for any new x can be predicted.
Simple Linear Regression: When there is only one independent variable, it's called simple linear regression.
Multiple Linear Regression: When there are multiple independent variables, it's called multiple linear regression.
Assumptions: It assumes a linear relationship, which means changes in x are associated with proportional changes in y. The residuals in linear regression: a) show normal distribution with mean value 0, b) are independent of each other, hence do not exhibit any pattern and c) maintain a constant variance across all levels of the independent variable.
Output: The output is a line that best fits the data points, helping to make predictions.
Application: Linear regression is widely used in finance, economics, biology, and many other fields to understand relationships between variables and predict future trends.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. Each dataset consists of eleven (x, y) points. Despite their similar statistical properties—such as mean, variance, correlation, and linear regression line—their graphical representations reveal distinct patterns and outliers.
The quartet was created by the statistician Francis Anscombe to illustrate the importance of graphing data before analyzing it. His examples of the 4 graphed datasets highlight how relying solely on summary statistics can be misleading, underscoring the necessity of visualizing data to understand its underlying structure and identify potential issues.

3. What is Pearson's R? (3 marks)

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

- r=1 indicates a perfect positive linear relationship,
- r=−1 indicates a perfect negative linear relationship,
- r=0 indicates no linear relationship.

It is calculated as:

$$r = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2}\sqrt{\sum(y_i-\overline{y})^2}}$$

Where,

x_i and y_i are the individual sample points,

$\overline{x}$ and $\overline{y}$ are the means of the x and y values.

Application: Determining the strength of the relationship between variables in regression analysis, Testing hypotheses about the relationship between variables, etc.
Limitations: Pearson's r only measures linear relationships; it does not capture non-linear relationships. It is sensitive to outliers, which can distort the correlation. It assumes that the data is homoscedastic (i.e., the variance around the regression line is constant).

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

Ans: Scaling is a preprocessing technique used in data analysis and machine learning to adjust the range and distribution of feature values. The goal is to ensure that the features contribute equally to the model, especially for algorithms that are sensitive to the magnitude of feature values, such as gradient descent-based algorithms and distance-based algorithms.

Scaling is performed for:

1) Improving Model Performance: Scaling can lead to faster convergence of gradient descent algorithms by ensuring that the features are on a similar scale.
2) Ensuring Interpretability: Standardized features can make it easier to interpret the model coefficients, especially in linear models.
3) Preventing Numerical Instability: Scaling can help avoid issues with numerical stability in some algorithms, where large feature values might lead to overflow or underflow errors.

Difference between Normalized scaling and Standardized scaling:

1) Normalized Scaling (Min-Max Scaling): Transforms the data to a specific range, typically [0, 1] or [-1, 1].
   Formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

   It is suitable when you know the minimum and maximum values of your data and when you want the data to have a specific bound. It keeps all features in the same range, which is useful for algorithms that require bounded inputs. However, it is sensitive to outliers since they can significantly affect the minimum and maximum values.
2) Standardized Scaling (Z-Score Standardization): Transforms the data to have a mean of 0 and a standard deviation of 1.
   Formula:

$$X_{std} = \frac{X - \mu}{\sigma}$$

   where μ is the mean and σ is the standard deviation of the feature.
   It is suitable when the data follows a Gaussian distribution and when the algorithm assumes normally distributed data. It is less sensitive to outliers compared to normalization and ensures that the data is centered around the mean with unit variance. However, it does not bound the data within a specific range, which might not be suitable for all applications.

:

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

(3 marks)

Ans: VIF value becomes infinite when R2=1. This implies that the predictor can be perfectly predicted by a linear combination of the other predictors. In other words, there is perfect multicollinearity. To remove multicollinearity, identify and remove one of the variables involved in the perfect linear relationship; combine collinear variables into a single predictor through techniques like principal component analysis (PCA); ensure proper encoding of categorical variables and avoid including duplicate or redundant features.

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 compares the quantiles of the sample data to the quantiles of a theoretical distribution. The plot is created by: i) Sorting the sample data. ii) Calculating the corresponding quantiles of the theoretical distribution. iii) Plotting the sample quantiles against the theoretical quantiles.

Importance of Q-Q plot:

Q-Q plots are used in assessing the normality of residuals. Q-Q plots validate the assumptions underlying the linear regression model, ensuring that the results and inferences drawn from the model are reliable. Q-Q plots can help identify outliers or extreme values in the data, which may disproportionately influence the regression model. Patterns in the Q-Q plot, such as curvature or S-shapes, indicate deviations from normality, suggesting that the residuals may follow a different distribution.