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)

Going through the categorical variables, one by one.

1. *Year: The year 2019 (represented by 1) had more bike rentals than the year 2018 (represented by 0)*
2. *Season : Fall season (represented by 3) showed the highest number in bike sharing than any other season. The lowest was the spring time.*
3. *Weekday : All weekdays had equal bike sharing numbers with very little difference.*
4. *Holiday : On holidays, people rented bikes more than on working days.*
5. *Weathersit (condition of weather) : On a clear or a partly cloudy cloudy day(represented by the code 1), bike rentals shot up.*

Method of inference : Box plot and median comparison.

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

Suppose, we have a variable 'furnishing status' that has 3 values (or levels) - furnished, semi-furnished and unfurnished. Only 2 dummy variables will effectively convey the level. There is no need of a third.
If furnished is 11 and semifurnished is 01, then unfurnished can be 00.
Having an extra variable creates several problems such as

1. *Excessive number of variables in the model*
2. *High correlation among different variables which can reduce the efficiency of the model.*
3. *High correlation also leads to the problem of interpreting the coefficients. We cannot predict which variable is affecting the target variable more.*

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

The variable 'temp' (containing) the record of temperature of the day has the highest correlation with the target variable.

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

1. Residual Analysis: After building the model, a histogram of the the residuals was plotted. The mean of the residuals was centred around zero. In other words, the residuals followed a normal distribution. This is one of the assumptions of a linear regression model.

2. Homoscedasticity (Constant Variance): The residuals were plotted against the predicted values. We found a more or less equal distribution along various values of x, with an insignificant amount of outliers.

3. Linearity: We plotted a scatter plot of the dependent and independent variables and tried to fit a straight line through. It fitted perfectly. This confirms another of the assumptions of *linear regression, i.e there exists a linear relationship between the dependent and independent variables.*

4. Multicollinearity: Checked the VIF of various variables to check their colinearity with other variables. Variables with a VIF of 5 or above were removed.

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

1. *Temperature (has a coefficient of 5172)*
2. *The dummy variable light rain/snow (has a negative coefficient of 2038)*
3. *Year (has a coefficient of 1982)*

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used to predict a continuous outcome variable (dependent) based on one or more input variables (independent). It assumes a linear relationship, finds coefficients to minimize errors, and makes predictions using a linear equation. There are two types of regression models:

1. *Simple Linear Regression : We have one independent variable and one dependent variable. The equation followed is that of a straight line :* $Y = \beta_0 + \beta_1 X$
2. *Multiple Linear Regression : We have several independent variables that predict the dependent variable. The equation is as follows :* $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ *….*

β_0 *is the intercept. This means even if the value of X is 0, Y would still increase or decrease by a factor of* β_0 *.* β_1 *is coefficients or slope. This means that if X changes by 1 unit, Y would change by* β_1 *unit. In case of multiple variables, we assume that all the other independent variables are constant while we state this statement.*

Steps of doing a linear regression model:

1. *Cleaning and preparing the data: This might include basic data analysis steps, removal of unnecessary variables or null values, changing data types and creating dummy variables from categorical variables.*
2. *Rescaling the features : We rescale the numerical values to fit within a certain range (0 and 1 in case of normalization). This is done so as to not have unusually high or low coefficients, which become incomparable to each other.*
3. *Selecting the variables (in case of multiple linear regression) : It is not best practice to put all the variables into the models due to several factors like multicollinearity and overfitting. So we have a few automated and manual methods to select the best and most relevant variables.*
4. *Building the model : We make the model learn the coefficients of the training data set*
5. *Residual analysis: We analyse the error terms (the other possible values of Y for a a given X). If their mean is centred around zero, we proceed ahead with the following steps.*
6. *Evaluating the model: The common methods for evaluating the performance of the model are R-squared, MSE, RMSE. Adjusrted R-squared in case of multiple linear regression penalises the model for the number of features.*
7. *Testing the model: Testing the performance of the model on a test data set.*

Assumptions of a linear regression model:

1. *Linearity: Linear regression assumes that the relationship between the independent variables and the dependent variable is linear. This means that changes in the independent variables result in proportional changes in the dependent variable.*
2. *Independence: It assumes that the observations are independent of each other. This means that the value of the dependent variable for one observation does not depend on the values of the dependent variable for other observations.*
3. *Homoscedasticity: This assumption implies that the variance of the errors (residuals) is constant across all levels of the independent variables. In other words, the spread of the residuals should be roughly the same for all values of the predictors.*
4. *Normality: Linear regression also assumes that the residuals follow a normal distribution.*

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

Anscombe's Quartet is a classic and thought-provoking example in statistics that emphasizes the importance of data visualization and challenges our reliance on summary statistics.

1. *Anscombe's Quartet comprises four distinct datasets, labeled I, II, III, and IV.*
2. *Each dataset consists of 11 (x, y) data points, resulting in 11 pairs of values.*
3. *The summary statistics of the four data sets comes out to be identical.*
4. *For example, the mean of x and y, the variance of x and y, and the correlation coefficient between x and y are very close or identical across the four datasets.*
5. *Despite having nearly identical summary statistics, the datasets have very different underlying patterns.*
6. *However, scatterplots painted a very different picture. When plotted, Dataset I forms a clear linear pattern, Dataset II shows a curved pattern, Dataset III has a linear pattern with an outlier, Dataset IV reveals two distinct clusters.*

Anscombe's Quartet serves as a cautionary example that statistics alone can be deceptive and highlights the importance of data visualization.

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

Pearson's Correlation Coefficient (r) is a statistic used to measure the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1:
- 1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- Close to 0 suggests a weak or no linear relationship.

The sign of r (+ or -) shows the direction:
- Positive (r > 0) means as one variable increases, the other tends to increase.

- Negative (r < 0) means as one variable increases, the other tends to decrease.

It's calculated using a formula based on the means of the variables. Pearson's r is widely used in fields like statistics and data analysis to assess relationships between variables. Remember, it specifically measures linear relationships and does not imply causation.

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

Scaling is the process of adjusting the range or distribution of feature values. It's done for several reasons:

1. Algorithm Sensitivity: Many machine learning algorithms are sensitive to feature scales, so scaling ensures fair treatment of all features.

2. Faster Convergence: Scaling can speed up algorithm convergence during training.

3. Interpretability: Scaled data is easier to interpret and compare.

4. Regularization: Some techniques require standardized features.

5. Distance-Based Algorithms: Scaling is crucial for distance-based algorithms.

Two common scaling methods are:

1. Normalized Scaling (Min-Max Scaling):
- Scales data to a specified range, usually 0 to 1.
- Preserves original units and can be sensitive to outliers.

2. Standardized Scaling (Z-score Standardization):
- Centers data around mean 0 and standard deviation 1.
- Removes original units, less sensitive to outliers, and suitable for distance-based algorithms.

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

A VIF (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity in the dataset. This means that one or more independent variables can be exactly predicted from others, leading to an issue where coefficients cannot be estimated uniquely. VIF measures the degree of multicollinearity, and when perfect multicollinearity exists for a variable, its VIF becomes infinite. To address this, identify the problematic variables causing multicollinearity and either remove or combine them in the model to avoid unreliable coefficient estimates and misinterpretations.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

A Quantile-Quantile plot, often abbreviated as a Q-Q plot, is a graphical tool used in statistics and data analysis to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. It compares the quantiles (ordered values) of the dataset against the quantiles expected from the chosen theoretical distribution.

Q-Q plots are valuable in linear regression for several reasons:

1. Checking Residual Normality: One of the key assumptions in linear regression is that the residuals (the differences between observed and predicted values) follow a normal distribution. Deviations from normality can affect the validity of statistical tests and confidence intervals associated with regression coefficients. Q-Q plots are used to visually assess the normality of residuals. In a Q-Q plot for residuals, data points should roughly follow a straight line if the residuals are normally distributed.

2. Detecting Outliers: Outliers can distort regression models. Q-Q plots can help identify outliers by highlighting extreme data points that deviate from the theoretical distribution line. Outliers may indicate data entry errors, influential observations, or issues with model assumptions.

3. Assessing Model Fit: Q-Q plots can provide insights into the adequacy of the chosen model. If the residuals follow the theoretical distribution closely, it suggests that the model is a good fit for the data. Deviations from the theoretical line might suggest that the model does not capture the data distribution adequately.

4. Quantifying Skewness and Kurtosis: Besides normality, Q-Q plots can reveal deviations in skewness (asymmetry) and kurtosis (tailedness) of the data distribution. Departures from normality in these aspects can impact the model's performance and interpretation.

5. Identifying Data Transformations: If the Q-Q plot indicates that the data deviates significantly from a normal distribution, it may be necessary to consider data transformations (e.g., log transformation) to make the data more suitable for linear regression modeling.