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: There are 7 categorical variables

a. 'yr': as inferred from the above boxplot we can say that 2018 & 2019 year by year bookings are gradually increasing. Can be used as good predictor
b. 'holiday': bike bookings are maximum during non-holidays showing high median and holiday may not be considered as good predictor variable
c. 'workingday': 68 % percent booking done in working day for the period of two years can be good predictor
d. 'weekday': weekday variable shows very close trend for all week days having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor.
e. 'season': the bike max booking happening in season-fall with a median of over 5000 booking (for the period of 2 years), next followed by season-summer & season-winter of total booking. This indicates, season can be a good predictor for the dependent variable.
f. 'weathersit': 63% of the bike booking were happened during weathersit-clear with a median of close to 5000 booking (for the period of 2 years), next followed by weathersit-mistcloudy with 33% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the target variable.
g. 'mnth': month 5,6,7,8,9 & 10 are highest number of bookings can be good predictor for target variable.

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

Ans: Two most important by using drop_first=True argument while creating dummy variables are

Avoiding Multicollinearity: When you create dummy variables from a categorical variable with N categories, you typically end up with N-1 dummy variables. For example, if you have a "Colour" variable with categories "Red," "Blue," and "Green," creating dummy variables would result in "Red," "Blue," and "Green," with one category (usually the first one) omitted. This is done to avoid perfect multicollinearity, where one dummy variable can be perfectly predicted from the others. Perfect multicollinearity can lead to unstable coefficient estimates in regression models. when all the other columns are zero that means the first columns is 1.

Interpretability: Dropping the first category makes it easier to interpret the coefficients of the remaining dummy variables. The coefficients represent the change in the response variable associated with a one-unit change in the predictor variable while holding all other predictor variables constant. When the first category is included as a reference, the coefficients of the other dummy variables represent the change relative to the reference category.

By omitting one category, you reduce the dimensionality of your feature space, which can make your models more efficient in terms of memory and computation. It also helps avoid multicollinearity, which can lead to more stable model training.

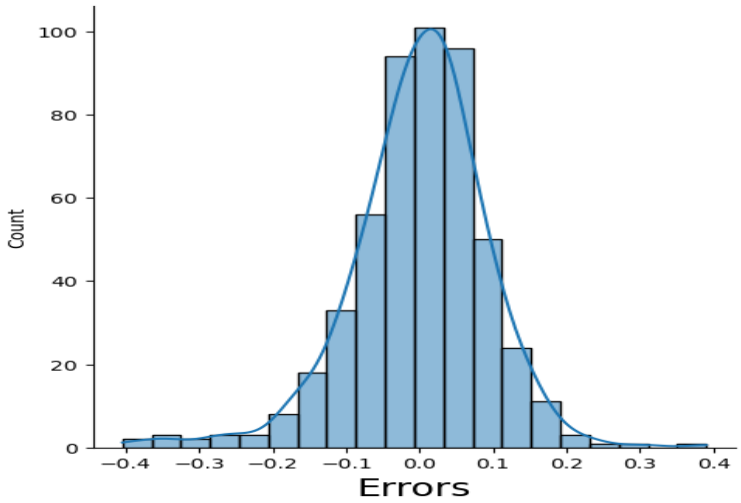
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: Variable ‘temp’ and ‘atemp’ has high correlation with target variable ‘cnt'

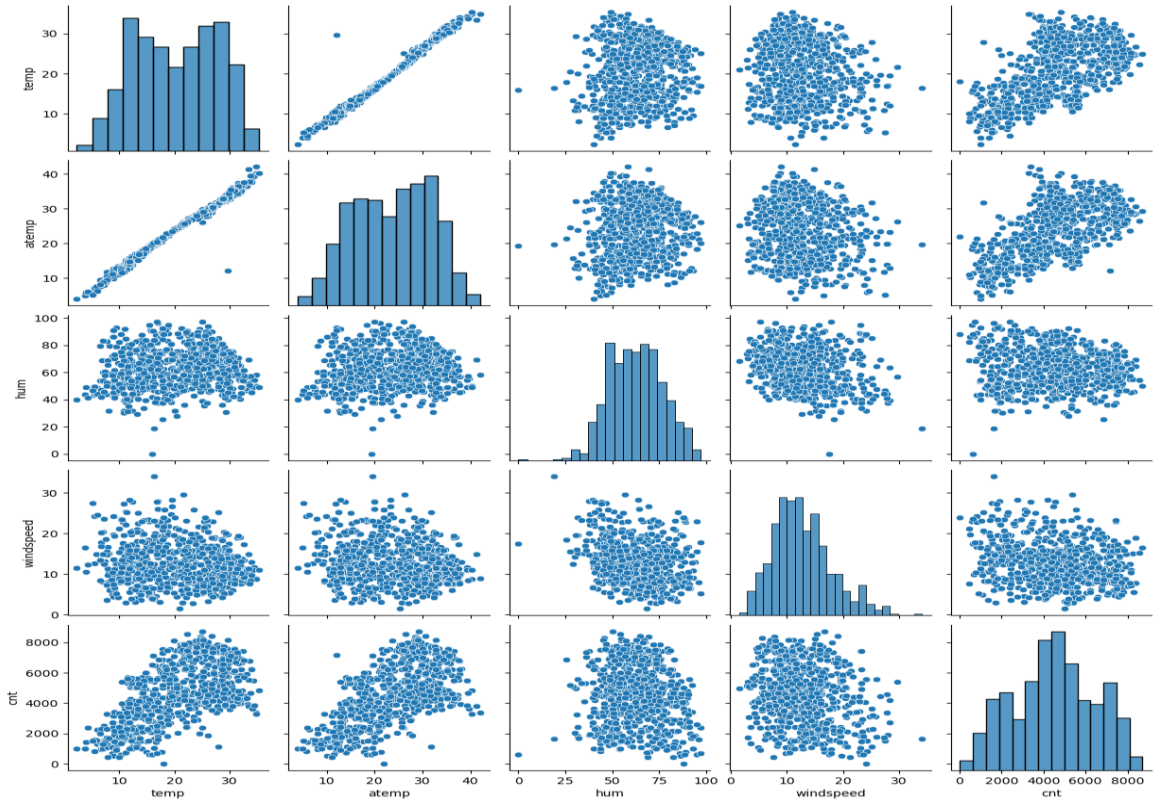
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: By calculating the Residual errors = (y_train – y_train_predict) and plot histogram or distribution plot using seaborn or matplot lib check whether they are normally distributed or not which the mean of error is ‘zero’ and S.D is ‘1’. residual measure how far away a point is from the regression line.

From bike-sharing trained model, histogram of residual errors shows that are normally distributed



b. There is linear relatioship between the vaiables



c. There is No Multicollinearity between the predictor variables¶

SrNo	Features	VIF
2	atemp	4.81
1	workingday	4.06
3	windspeed	3.41
0	yr	2.02
7	weekday_sun	1.69
4	season_summer	1.58
8	weathersit_cloudy	1.53
5	season_winter	1.40
6	mnth_sep	1.20
9	weathersit_lightrain	1.08

R2 Squared value and Adjusted R2 are within the range 82.8% & 82.5% respectively.

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

Ans: atemp: A coefficient value of '0.5779' indicated that a unit increase in temp variable, increases the bike hire numbers by 0.5779 units.

weathersit_lightrain: A coefficient value of '-0.3070' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable, decreases the bike hire numbers by 0.3070 units.

yr: A coefficient value of '0.2342' indicated that a unit increase in yr variable, increases the bike hire numbers by 0.2342 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable (also known as the target or outcome variable) and one or more independent variables (predictors or features). It assumes a linear relationship between the predictors and the target variable. Here's a detailed explanation of the linear regression algorithm:

Objective:

Linear regression aims to model the relationship between a dependent variable (Y) and one or more independent variables (X) such that it can predict the value of Y based on the values of X.

Assumptions:

Linearity: It assumes that the relationship between the predictors and the target variable is linear.

Independence: It assumes that the errors (residuals) are independent of each other.

Homoscedasticity: It assumes that the variance of the residuals is constant across all levels of the predictors.

Normality: It assumes that the residuals are normally distributed.

Simple Linear Regression (Single Predictor):

In simple linear regression, you have one predictor variable (X) and one target variable (Y).

The relationship is modelled as: $Y = \beta_o + \beta_1 X + \varepsilon$, where β_o and β_1 are the coefficients to be estimated, and ε represents the error term.

Multiple Linear Regression (Multiple Predictors):

In multiple linear regression, you have more than one predictor variable (X_1, X_2, X_3, ...) and one target variable (Y).

The relationship is modelled as: $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \varepsilon$, where β_o, β_1, β_2, β_3, ... are the coefficients to be estimated, and ε represents the error term.

Model Training:

The goal is to find the values of the coefficients (β_o, β_1, β_2, ...) that minimize the sum of squared differences between the observed values (Y) and the predicted values ($\hat{Y}$) made by the model.

This is typically done using optimization techniques like Ordinary Least Squares (OLS).

Prediction:

Once the model is trained and coefficients are estimated, you can use it to make predictions.

Given new values of the predictor variables, you can calculate the predicted value of the target variable.

Evaluation:

Linear regression models are evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) to assess how well the model fits the data.

Cases:

Linear regression is widely used in various fields, including economics, finance, social sciences, and engineering, for tasks such as predicting bike sharing booking as we built on model using liner regression library other examples sales, housing prices, stock prices, and more.

linear regression is a fundamental algorithm used for modelling the relationship between variables in a linear fashion. It's relatively simple to understand and implement, making it a valuable tool for both predictive modelling and statistical analysis. However, its effectiveness depends on the linearity assumption and the suitability of the data for the chosen model.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) but exhibit very different characteristics when graphed. This quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and not relying solely on summary statistics.

Anscombe's quartet consists of four datasets, each containing 11 (x, y) data points.

These datasets were designed to have nearly identical summary statistics, such as mean, variance, and correlation coefficients, to demonstrate that summary statistics alone may not reveal the true nature of the data.

The Four Datasets:

Dataset I: Linear Relationship

This dataset forms a clear linear relationship, where the relationship between x and y can be closely approximated by a straight line.

The summary statistics for this dataset closely resemble those of a simple linear regression.

Dataset II: Non-Linear Relationship

Unlike Dataset I, Dataset II exhibits a non-linear relationship between x and y.

A linear regression model would not accurately capture the underlying pattern in this dataset.

Dataset III: Outlier

This dataset appears to have a linear relationship similar to Dataset I, but it contains an outlier that significantly affects the regression line's fit.

Dataset IV: No Relationship

Dataset IV has no apparent relationship between x and y.

The data points are scattered randomly, and there is no meaningful linear relationship.

Anscombe's quartet serves as a compelling reminder of the limitations of summary statistics. Despite having nearly identical statistical properties, these datasets have vastly different patterns when visualized.

The quartet highlights the importance of data visualization in data analysis, as it can reveal trends, outliers, and patterns that might go unnoticed when relying solely on numerical summaries.

It emphasizes the need to choose appropriate data analysis techniques based on the characteristics of the data, as blindly applying methods like linear regression without visual inspection can lead to erroneous conclusions.

In summary, Anscombe's quartet consists of four distinct datasets with similar summary statistics but different underlying patterns. It underscores the significance of data visualization and the potential pitfalls of relying solely on summary statistics in data analysis.

3. What is Pearson’s R?

Pearson's R is a statistical measure that quantifies the linear correlation between two variables. It is a number between -1 and 1, where:

- r = 1 indicating a perfect positive correlation,
- r = -1 indicating a perfect negative correlation, and
- r = 0 indicating no correlation.

Pearson's R is a parametric test, which means that it assumes that the data is normally distributed. It is also sensitive to outliers, so it is important to remove outliers from the data before calculating Pearson's R.

Formulae to calculate Pearson's R value

$R = (n * \Sigma xy) - (\Sigma x * \Sigma y) / \sqrt{((n * \Sigma x^2) - (\Sigma x)^2) * \sqrt{((n * \Sigma y^2) - (\Sigma y)^2)}}$

- n is the number of data points
- Σxy is the sum of the products of the corresponding values of x and y
- Σx is the sum of the values of x
- Σy is the sum of the values of y
- Σx^2 is the sum of the squared values of x
- Σy^2 is the sum of the squared values of y

Interpretation: The sign of r (+ or -) indicates the direction of the relationship (positive or negative), and the magnitude of r indicates the strength of the relationship. Values closer to -1 or 1 represent stronger linear relationships, while values closer to 0 suggest weaker or no linear relationships between the variables.

example, Pearson's R could be used to:

- Identify whether there is a relationship between student grades and the amount of time they spend studying.
- Determine whether there is a correlation between advertising spending and

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

   Scaling is a preprocessing technique in data analysis and machine learning that is performed on numerical features or variables to bring them to a similar scale or range. It's done to ensure that different variables contribute equally to model training and to prevent certain variables from dominating the learning process. Scaling is particularly important for algorithms that are sensitive to the scale of input features, such as many distance-based algorithms and gradient-based optimization algorithms. There are two common types of scaling: normalized scaling and standardized scaling.

   ### Normalized Scaling:

   Normalization, also known as min-max scaling, scales the features to a specific range, typically between 0 and 1.

The formula for normalization is:

X_normalized = (X - X_min) / (X_max - X_min)

Here, X is the original value, X_min is the minimum value of the feature, and X_max is the maximum value of the feature.

Normalization is useful when you want to preserve the original distribution of the data but ensure that all features have the same scale.

Standardized Scaling (Standardization):

Standardization, also known as z-score scaling, scales the features to have a mean of 0 and a standard deviation of 1.

The formula for standardization is:

X_standardized = (X - X_mean) / X_std

Here, X is the original value, X_mean is the mean of the feature, and X_std is the standard deviation of the feature.

Standardization is useful when you want to transform the data into a standard normal distribution with a mean of 0 and a standard deviation of 1.

It centres the data around zero and adjusts the spread of data, making it suitable for algorithms that assume normally distributed data.

Key Differences:

Range:

Normalization scales the data to a specific range, often [0, 1].

Standardization scales the data to have a mean of 0 and a standard deviation of 1.

Preservation of Distribution:

Normalization preserves the original distribution of the data within the specified range.

Standardization transforms the data to follow a standard normal distribution.

Use Cases:

Normalization is useful when you have features with varying ranges and you want to bound them within a consistent range.

Standardization is suitable when you want to ensure that features have similar means and variances, which can be beneficial for algorithms that assume standardized data.

Sensitivity to Outliers:

Normalization can be sensitive to outliers, as it depends on the minimum and maximum values.

Standardization is less sensitive to outliers because it relies on the mean and standard deviation, which are less affected by extreme values.

In summary, both normalization and standardization are scaling techniques used to bring numerical features to a common scale, but they have different objectives and outcomes. The choice between them depends on the specific requirements of your data and the algorithms you plan to use.

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

Ans: The Variance Inflation Factor (VIF) measures the degree of multicollinearity among predictor variables in a regression analysis. A high VIF for a particular variable indicates that this variable can be strongly predicted from the other predictor variables, which implies multicollinearity. VIF values greater than 1 are typically a cause for concern, but VIF can become infinite under certain conditions. Here's why this can happen:

Perfect Multicollinearity:

Infinite VIF occurs when there is perfect multicollinearity between predictor variables. Perfect multicollinearity means that one or more predictor variables can be expressed exactly as a linear combination of other predictor variables.
In such cases, the coefficient estimates in regression models become undefined because there is no unique solution to estimate the individual effects of these variables.
Mathematical Dependency:

When two or more variables are perfectly correlated or linearly dependent, the determinant of the matrix used in the VIF calculation becomes zero.
The VIF is calculated as the reciprocal of the determinant of the correlation matrix of the predictor variables.
When the determinant is zero (perfect multicollinearity), the reciprocal becomes infinity, resulting in infinite VIF values.

Examples:
A common example of perfect multicollinearity is when you have a binary categorical variable with only two levels and you include both dummy variables in the regression model. For example, if you have "Male" and "Female" as binary variables, including both "Male" and "Female" as predictors will result in perfect multicollinearity.
Another example is when you include the same variable twice in the model without any transformation or interaction term.

Consequences:
Infinite VIF values indicate that the predictor variables involved are not suitable for inclusion in a multiple regression model.
In practice, you need to identify and address multicollinearity issues by either removing one of the correlated variables, transforming the variables, or using dimensionality reduction techniques such as Principal Component Analysis (PCA).

In summary, infinite VIF values occur when there is perfect multicollinearity among predictor variables, which means that some variables can be exactly predicted from others. It is a mathematical issue that results in undefined coefficient estimates in regression models, and it should be addressed by removing or transforming the problematic variables to ensure a well-posed regression problem.

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

Ans: A Q-Q plot, short for "Quantile-Quantile plot," is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as a normal distribution. It is a valuable technique in statistics and data analysis for checking the assumption of normality and can also be used to compare the distribution of two datasets.

Application:

A Q-Q plot is used to visually assess whether the observed data follows a specific theoretical distribution.

It compares the quantiles of the observed data (empirical quantiles) to the quantiles expected from the theoretical distribution (theoretical quantiles).

If the points in the Q-Q plot closely align along a diagonal line, it suggests that the data closely follows the theoretical distribution.

Importance in Linear Regression:

Q-Q plots are particularly important in linear regression for several reasons.

Assumption of Normality: Linear regression models often assume that the residuals (the differences between observed and predicted values) are normally distributed. Checking this assumption is crucial because if it's violated, it can affect the validity of statistical tests and confidence intervals associated with the regression coefficients.

Residual Analysis: After fitting a linear regression model, you can create a Q-Q plot of the residuals to check whether they follow a normal distribution. If the points in the Q-Q plot deviate significantly from the diagonal line, it may indicate departures from normality in the residuals.

Outlier Detection: Q-Q plots can help detect outliers in the data. Outliers are data points that deviate substantially from the expected distribution. In a Q-Q plot, outliers may appear as points that deviate from the diagonal line, indicating that they do not conform to the expected distribution.

Model Validation: In linear regression, it's essential to validate the assumptions of the model to ensure its reliability. Q-Q plots provide a simple and effective way to check the normality assumption, which is one of the key assumptions of linear regression.

Interpretation:

In a Q-Q plot, if the points closely follow the diagonal line, it suggests that the data follows the expected distribution (e.g., normal distribution). Deviations from the diagonal line indicate departures from the assumed distribution.

Decision-Making:

Based on the Q-Q plot, you can make informed decisions about whether to proceed with the linear regression analysis as is or whether data transformations or alternative modelling techniques are needed to address non-normality or outliers.

In summary, a Q-Q plot is a useful graphical tool for assessing the normality assumption and detecting outliers in linear regression. It allows you to visually compare the distribution of the observed data to a theoretical distribution and make informed decisions about the adequacy of your regression model.