Table Of Contents

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?

By analysing the categorical variables we can infer that:

- The demad of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non-working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Lightsnow and light rainfall. We do not have any dat for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we can not derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

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

> When creating dummy variables from categorical variables with multiple categories (levels), it's common practice to use the drop_first=True parameter. This parameter controls whether to drop the first category when creating dummy variables. This might seem counterintuitive, but it's important for several reasons:

1. **Avoiding Multicollinearity:**
   Including all dummy variables (one for each category) without dropping one can lead to multicollinearity, which is a situation where two or more independent variables are highly correlated. This can cause issues in linear regression models, as it becomes difficult to distinguish the individual effects of the variables. Dropping one category helps mitigate this by ensuring that the dummy variables are linearly independent.
2. **Redundant Information:**
   When you include dummy variables for all categories, the information about the omitted category can be derived from the others. This means that the dropped category's information is redundant and can be captured by the intercept term in the model.
3. **Interpretation of Coefficients:**
   Dropping one category sets it as the reference category, and the coefficients of the remaining dummy variables represent the difference between their respective categories and the reference category. This makes the interpretation of the coefficients more intuitive and straightforward.
4. **Simpler Models:**
   Including fewer dummy variables (by dropping one category) simplifies the model and reduces the number of parameters that need to be estimated. This can lead to a more stable and less complex model.

---

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

Looking at the pair-plot we can infer a few things:

1. **'registered'** - Seems to have the highest correlation with the target variable i.e. **'cnt'** but as we know '*cnt = registered + casual*' this behaviour is expected.
2. **'temp'** - Seems to have a pretty good correlation with the target variable i.e. **'cnt'**. After **'registered'**, **'temp'** is the variable which seems to have highest correlation with the **'cnt'**.

---

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

> Assumptions of Linear Regression:-
>
> 1. There exists some kind of linear relationship between X and y.
> 2. Error terms are normally distributed.
> 3. Error terms are independant of each other.
> 4. Error terms have constant variance(Homosedacity).

- **Validating Linear Relation between X and y:**
  - We drew a pair plot and found that there exists a linear relation between **'cnt'** and **'temp'**.

- **Validating that error terms are independant of each other**
    - Drew a heatmap on the train set correlation.
    - Calculated VIF for all the predictors and removed predictors with high VIF.
    - VIF(Variance Inflation Factor) - High VIF indicates high correlation between the predictors.
- **Validating that Error terms are normally distributed:**
  we drew a distplot for the residuals(error terms) from that we can infer that:
    - There is no pattern in the error terms.
    - Mean is almost 0.
    - The plot appears to be a normal distribution.
    - There doesn't seems to be pattern in the error terms.
- **Validating Homosedacity:**
    - Drew a scatter plot between residuals and y_train and found out that there exists a constant variance in the error terms.

---

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

The top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows:

1. **Temperature** - β1 value is 0.4914(High temperature indicates high count of bikes sharing count).
2. **Year** - β2 value is 0.2334(In 2019 Bike sharing count was comparatively very high than in 2018).
3. **Season(Winter)** - β3 value is 0.0970(In winters bike sharing count is comparatively high).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

> Linear regression is a fundamental statistical algorithm used for modeling the relationship between a dependent variable (also called the target) and one or more independent variables (also called predictors or features). It's a supervised learning algorithm commonly used for tasks such as prediction and inference. Linear regression assumes a linear relationship between the variables, which means that changes in the independent variables are associated with proportional changes in the dependent variable.

Here's a detailed explanation of the linear regression algorithm:

1. **Problem Statement:**
   Linear regression addresses problems where you have a dataset with pairs of observations: input features (independent variables) and corresponding target values (dependent variable). The goal is to find the best-fitting linear relationship between the inputs and the target.
2. **Model Representation:**
   In simple linear regression, there's only one independent variable, and the relationship is represented

by the equation:
y = mx + b

where:
'y' is the predicted target value.
'x' is the input feature.
'm' is the slope of the line, representing the change in y for a unit change in x.
'b' is the y-intercept, the value of y when x is 0.

For multiple linear regression (more than one independent variable), the equation extends to:
*y = b0 + b1x1 + b2x2 + ... + bn*xn*

where:
'x1', 'x2', ..., 'xn' are the independent variables.
'b0', 'b1', 'b2', ..., 'bn' are the coefficients (weights) associated with each independent variable.

3. **Objective Function:**

> The goal is to find the best-fitting line that minimizes the difference between the predicted values and the actual target values. This is achieved by defining a cost or loss function that measures the difference between the predicted values and the actual values. One common loss function is the Mean Squared Error (MSE):
> *MSE = (1/n) * Σ(y_actual - y_predicted)^2*

where n is the number of data points.

4. **Finding Coefficients (Weights):**

> The coefficients (b values) are adjusted to minimize the cost function. This process is often done using optimization techniques like gradient descent. Gradient descent iteratively updates the coefficients in the direction that reduces the loss until convergence.

5. **Gradient Descent:**
   Gradient descent involves the following steps:
   - Initialize the coefficients with some values.
   - Calculate the predicted values using the current coefficients.
   - Calculate the gradient of the loss function with respect to each coefficient.
   - Update each coefficient by subtracting a small fraction (learning rate) of the gradient.

6. **Training the Model:**

> The algorithm iteratively updates the coefficients using gradient descent until the loss converges to a minimum or reaches a predefined number of iterations. This process optimizes the model to fit the data better.

7. **Inference and Prediction:**

> Once the model is trained, you can use it to make predictions on new, unseen data by plugging in the values of the independent variables into the regression equation.

8. **Evaluation:**

> The model's performance is evaluated using metrics such as R-squared, which measures the proportion of the variance in the target variable that is predictable from the independent variables.

> Linear regression can be extended and adapted in various ways, such as adding polynomial terms to capture non-linear relationships, regularization techniques to prevent overfitting, and handling multicollinearity among independent variables.

---

2. Explain the Anscombe’s quartet in detail.

> Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, yet they exhibit vastly different patterns when graphed and analyzed. This quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics. It serves as a powerful reminder that looking at data graphically can reveal patterns that might be missed by simply looking at numerical summaries.

The four datasets in Anscombe's quartet share the following properties:

- They each consist of 11 data points.
- They have the same mean and variance for both the x and y variables.
- They have the same linear regression line (y = 3 + 0.5x) for each dataset.

However, when plotted, these datasets reveal dramatically different patterns, highlighting the importance of visual representation in data analysis.

Here's a description of each dataset in Anscombe's quartet:

1. **Dataset I:**
    - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
    - y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68
    - When plotted, this dataset roughly follows a linear relationship, and the linear regression line fits reasonably well.
2. **Dataset II:**
    - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
    - y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74
    - This dataset also shows a linear relationship, but with a slight upward curvature. The linear regression line is still a decent fit.
3. **Dataset III:**
    - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
    - y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
    - This dataset is nonlinear and has a clear quadratic pattern. A linear regression line would be a poor fit for this data.
4. **Dataset IV:**
    - x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
    - y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89

- This dataset doesn't follow any apparent linear or nonlinear pattern. It has an outlier that significantly affects the linear regression line.

The key takeaway from Anscombe's quartet is that summary statistics like mean, variance, and even the equation of a linear regression line can't fully capture the complexity of relationships in data. Visualizations, such as scatter plots, help to reveal patterns, trends, and potential outliers that are critical for making accurate interpretations and decisions in data analysis. It serves as a reminder that exploratory data analysis should involve both numerical summaries and graphical representations to gain a comprehensive understanding of the dataset.

---

3. What is Pearson's R?

> Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It's a common method to assess the degree to which changes in one variable are associated with changes in another variable. Pearson's correlation coefficient ranges between -1 and 1:

- Positive Correlation (r > 0): When one variable increases, the other tends to increase as well.
- Negative Correlation (r < 0): When one variable increases, the other tends to decrease.
- No Correlation (r ≈ 0): There's little to no linear relationship between the variables.

The formula to calculate Pearson's correlation coefficient between two variables, X and Y, is as follows:

$r = \Sigma((X - \bar{X}) * (Y - \bar{Y})) / \sqrt{(\Sigma(X - \bar{X})^2 * \Sigma(Y - \bar{Y})^2)}$

Where:

'$\bar{X}$' *is the mean of variable X.*

'$\bar{Y}$' *is the mean of variable Y.*

Pearson's correlation coefficient is sensitive to the linear relationship between variables, meaning that it measures how well the data can be approximated by a straight line. It doesn't capture nonlinear relationships, outliers, or other complex associations.

Interpretation of the magnitude of Pearson's r:

-**'r ≈ 1'**: Strong positive linear correlation.

-**'r ≈ -1'**: Strong negative linear correlation.

-**'r ≈ 0'**: Little to no linear correlation.

However, it's important to note that correlation does not imply causation. A high correlation between two variables does not necessarily mean that changes in one variable cause changes in the other. Correlation only indicates the strength and direction of the linear relationship between variables.

Pearson's correlation coefficient is widely used in various fields such as statistics, social sciences, finance, and more, to explore and quantify relationships between continuous variables.

---

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

> Scaling is a preprocessing step in data preparation for machine learning and statistical analysis. It involves transforming the features (variables) of a dataset so that they are on a similar scale, usually within a specific range. Scaling is performed to ensure that the features have comparable magnitudes and to improve the performance and stability of various algorithms that are sensitive to the scale of the data.

Why Scaling is Performed:

The reasons for scaling are as follows:

Scaling is a preprocessing step in data preparation for machine learning and statistical analysis. It involves transforming the features (variables) of a dataset so that they are on a similar scale, usually within a specific range. Scaling is performed to ensure that the features have comparable magnitudes and to improve the performance and stability of various algorithms that are sensitive to the scale of the data.

Why Scaling is Performed: The reasons for scaling are as follows:

1. **Algorithm Sensitivity:** Many machine learning algorithms, such as k-nearest neighbors, support vector machines, and gradient descent-based optimization algorithms, are sensitive to the scale of the input features. Features with larger scales can dominate the optimization process, causing the algorithm to perform poorly.

2. **Distance Metrics:** Algorithms that rely on distance metrics, like k-nearest neighbors, calculate distances based on the differences between feature values. If features are on different scales, features with larger values could disproportionately influence distance calculations.

3. **Convergence Speed:** Gradient-based optimization algorithms (e.g., in neural networks) converge faster on data where the features are similarly scaled. This speeds up the training process.

Normalized Scaling vs. Standardized Scaling:

1. **Normalized Scaling (Min-Max Scaling):** Normalization scales the features to a specific range, often between 0 and 1. The formula to normalize a feature is:

x_normalized = (x - min) / (max - min)

where **'x'** is the original feature value, **'min'** is the minimum value of the feature, and max is the maximum value of the feature.
Normalization is a good choice when you know that the distribution of your data does not follow a Gaussian (normal) distribution.

2. **Standardized Scaling (Z-Score Scaling):** Standardization transforms features to have a mean of 0 and a standard deviation of 1. The formula for standardizing a feature is:

x_standardized = (x - mean) / standard_deviation

where **'x'** is the original feature value, **'mean'** is the mean of the feature, and **'standard_deviation'** is the standard deviation of the feature.

Standardization is suitable when features have different units or when the distribution of the data is close to Gaussian.

- **Key Differences:**
  Range:
  - Normalization scales features to a specific range (e.g., 0 to 1).
  - Standardization centers features around 0 with a standard deviation of 1.
- **Interpretation:**
  - Normalization preserves the relative relationships between values, but it doesn't handle outliers well.
  - Standardization handles outliers better because it uses the mean and standard deviation, but it can change the interpretation of feature importance.
- **Application:**
  - Normalization is commonly used when you have prior knowledge that your data should fall within a certain range.
  - Standardization is generally a good choice when the distribution of data is not known or when features have varying units.

In summary, scaling is important to ensure that features are on a similar scale, helping machine learning algorithms to perform effectively. Normalization and standardization are two common scaling techniques, chosen based on the distribution of the data and the requirements of the specific problem.

---

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

> The situation where the Variance Inflation Factor (VIF) becomes infinite is known as "perfect multicollinearity." This occurs when one or more of the independent variables in a multiple linear regression model can be perfectly predicted by a linear combination of the other independent variables. In other words, one of the independent variables is a perfect linear function of the others, making its coefficient unidentifiable in the regression equation.

> When perfect multicollinearity exists, the mathematical calculations for calculating VIF break down, resulting in an infinite VIF value. This is because the formula for VIF involves dividing by the residual sum of squares, and when one of the variables is perfectly correlated with others, the residual sum of squares becomes zero, leading to division by zero.

Here's why perfect multicollinearity can cause infinite VIF values:

1. **Mathematical Issue:**

   The VIF formula includes the term $1 - R^2$, where R^2 is the coefficient of determination of the linear regression model that regresses the variable of interest against all the other independent variables. When perfect multicollinearity exists, the R^2 value becomes 1, making the term $1 - R^2$ equal to 0. As a result, the denominator becomes zero, leading to division by zero and thus an infinite VIF.

2. **Identifiability Issue:**

   In the presence of perfect multicollinearity, the affected variable's coefficient cannot be uniquely determined because it's entirely predictable from the other variables. This lack of identifiability creates

numerical instability and issues in estimation.

Perfect multicollinearity is a problem in linear regression because it complicates the interpretation of the coefficients and the overall model. It makes it challenging to understand the individual impact of each variable on the target variable, as well as to generalize the model to new data.

To address the issue of perfect multicollinearity, one or more variables involved in the multicollinearity should be removed from the model. Additionally, the underlying reasons for multicollinearity, such as redundant or highly correlated variables, should be carefully examined and addressed during the data preprocessing and feature selection stages to ensure a stable and interpretable model.

---

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

> A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess whether a given dataset follows a specific theoretical distribution, such as the normal distribution. It's particularly useful for comparing the quantiles of the observed data against the quantiles of a theoretical distribution. The Q-Q plot visually represents how well the data fits a particular distribution by plotting the quantiles of the observed data against the quantiles of the theoretical distribution.

Here's how a Q-Q plot works:

1. **Theoretical Quantiles:** Start with a set of theoretical quantiles from the chosen distribution (e.g., normal distribution) based on the probabilities (percentiles). These quantiles represent what the data should look like if it perfectly follows the theoretical distribution.

2. **Observed Quantiles:** Sort the data in ascending order and calculate the quantiles for the same probabilities used in step 1. These are the observed quantiles from your dataset.

3. **Plotting:** Plot the observed quantiles on the y-axis against the theoretical quantiles on the x-axis. Each point on the plot represents a pair of values—one from the theoretical distribution and one from the observed data.

If the observed data closely follows the theoretical distribution, the points on the Q-Q plot will fall roughly along a straight line with a slope of 1 (45-degree angle). Deviations from this diagonal line indicate departures from the assumed distribution.

Use and Importance of Q-Q Plot in Linear Regression:

Q-Q plots are valuable tools in linear regression and other statistical analyses for the following reasons:

1. **Normality Assumption:** In linear regression, it's often assumed that the errors (residuals) of the model follow a normal distribution. A Q-Q plot can help you assess whether the residuals of your regression model approximate a normal distribution. If the points on the Q-Q plot deviate significantly from the diagonal line, it suggests that the normality assumption might not hold.

2. **Model Validation:** Checking the residuals' normality is a crucial aspect of model validation. If the residuals do not approximate a normal distribution, it could impact the validity of statistical tests and confidence intervals associated with the model.

3. **Outlier Detection:** Q-Q plots can also help identify outliers in the data. Outliers can cause deviations from the theoretical distribution, leading to points that deviate from the diagonal line in the Q-Q plot.

4. **Improvement of Model Performance:** If non-normality is identified in the residuals, it might indicate issues with the model. Addressing non-normality by transforming the dependent variable or introducing appropriate transformations to the predictors can sometimes improve the model's performance.

In summary, Q-Q plots are useful tools to assess the fit of a dataset to a theoretical distribution, especially the normal distribution in the context of linear regression. They help verify assumptions and guide necessary adjustments to ensure the accuracy and validity of statistical analyses and model interpretations.
