Assignment-based Subjective Questions

Q-From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A -Categorical variables does have impact on dependent variables. Categorical variables were initially converted to dummy variables and then were encoded to binary values. When we look into the coefficients these variables had both negative and positive values indicating the impact on the target variable. Some of the categorical variables presented high collinearity like – weekday/working day. One of them were removed in building model. categorical variables, when properly encoded and included in a linear regression model, can provide valuable insights into the relationships between different categories and the target variable, leading to a more accurate and interpretable model.

Q-Why is it important to use drop_first=True during dummy variable creation?

A-Creating dummy variable is a way to convert the categorical variables into a binary encoded format. But if we have K categories, K-1 fields/variables are sufficient. By setting drop_first=True, we drop one of the dummy variables, typically the first category. This resolves the multicollinearity issue because it removes the redundancy. For a categorical variable with kkk categories, this will leave us with k−1k-1k−1 dummy variables, which are independent and can be used without causing multicollinearity.

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

A- temp has highest correlation of 0.63 with target variable.

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

A- R2 and Adjusted R2 was checked first. Once the R2 was higher than 0.6, p value of features selected was validated. P-value should be less than 0.05, then VIF was checked between the predictor variables to check variance inflation factor. It should be less than 10. Also F-statistic value and Prob (F-statistic) was checked. Prob-Statistic should be as low as possible. Three important factors R2/AdjustedR2, p-value and VIF to start with.

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

A- Temperature, year and Windspeed were top3 features contributing significantly. Year could slightly be dependent on weather conditions (if that particular year had more clearer days with pleasant weather)

General Subjective Questions

Q- Explain Linear regression algorithm in detail.

A- Linear regression is a fundamental and widely used statistical method for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes a linear relationship between the dependent and independent variables. Here is a detailed explanation of the linear regression algorithm, from its formulation to the solution and evaluation.

Two types of linear regression

1 – Simple Linear regression

$y=\beta0+\beta1x$

where X is independent variable, β0 and β1 are co-efficient and y is target variable.

2- Multiple linear regression

$y=\beta0+\beta1x1+\beta2x2+...+\beta pxp$

here the target variable depends on multiple features. βx are all the coefficients and x1, x2 are predictor variables.

To determine the beta coefficients, multiple methods are chosen.

1- Derivative method and equate the cost function to 0
2- Gradient Descent. Iterative learning and identifying best possible Beta coefficients

Linear regression relies on several key assumptions:

- Linearity: The relationship between the dependent and independent variables is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The residuals have constant variance.
- Normality: The residuals of the model are normally distributed (particularly important for inference).
- No multicollinearity: Independent variables are not highly correlated with each other.

To identify the model fit, R^2 is used:

$R2=1-RSS/TSS$

RSS : Residual Sum Squares

TSS: Total Sum Squares.

Q-Explain the Anscombe's quartet in detail.

A- Anscombe's quartet is a collection of four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. This was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphical analysis of data before relying solely on summary statistics. Each dataset in the quartet includes 11 (x, y) points. Here are the detailed properties of each dataset:

Summary Statistics

1. **Mean of x**: Each dataset has the same mean of x, approximately 9.
2. **Mean of y**: Each dataset has the same mean of y, approximately 7.5.
3. **Variance of x**: Each dataset has the same variance of x, approximately 11.
4. **Variance of y**: Each dataset has the same variance of y, approximately 4.12.
5. **Correlation between x and y**: Each dataset has a correlation of approximately 0.816.
6. **Linear regression line**: Each dataset has a linear regression line of y=3+0.5xy = 3 + 0.5xy=3+0.5x.
7. **Coefficient of determination (R^2)**: Each dataset has an R^2 value of approximately 0.67.

Despite these similarities in summary statistics, the visualizations of these datasets reveal very different patterns and highlight the significance of graphing data:

Dataset1

This dataset forms a roughly linear pattern with some scatter. It closely follows the linear regression line y=3+0.5x

x	y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Dataset 2

This dataset is also linear but has one outlier that affects the overall analysis. Most points align perfectly on a line, but one point significantly deviates.

x	y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74

Dataset 3

This dataset forms a perfect quadratic relationship. All points lie on a parabola, deviating significantly from the linear regression line.

x	y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

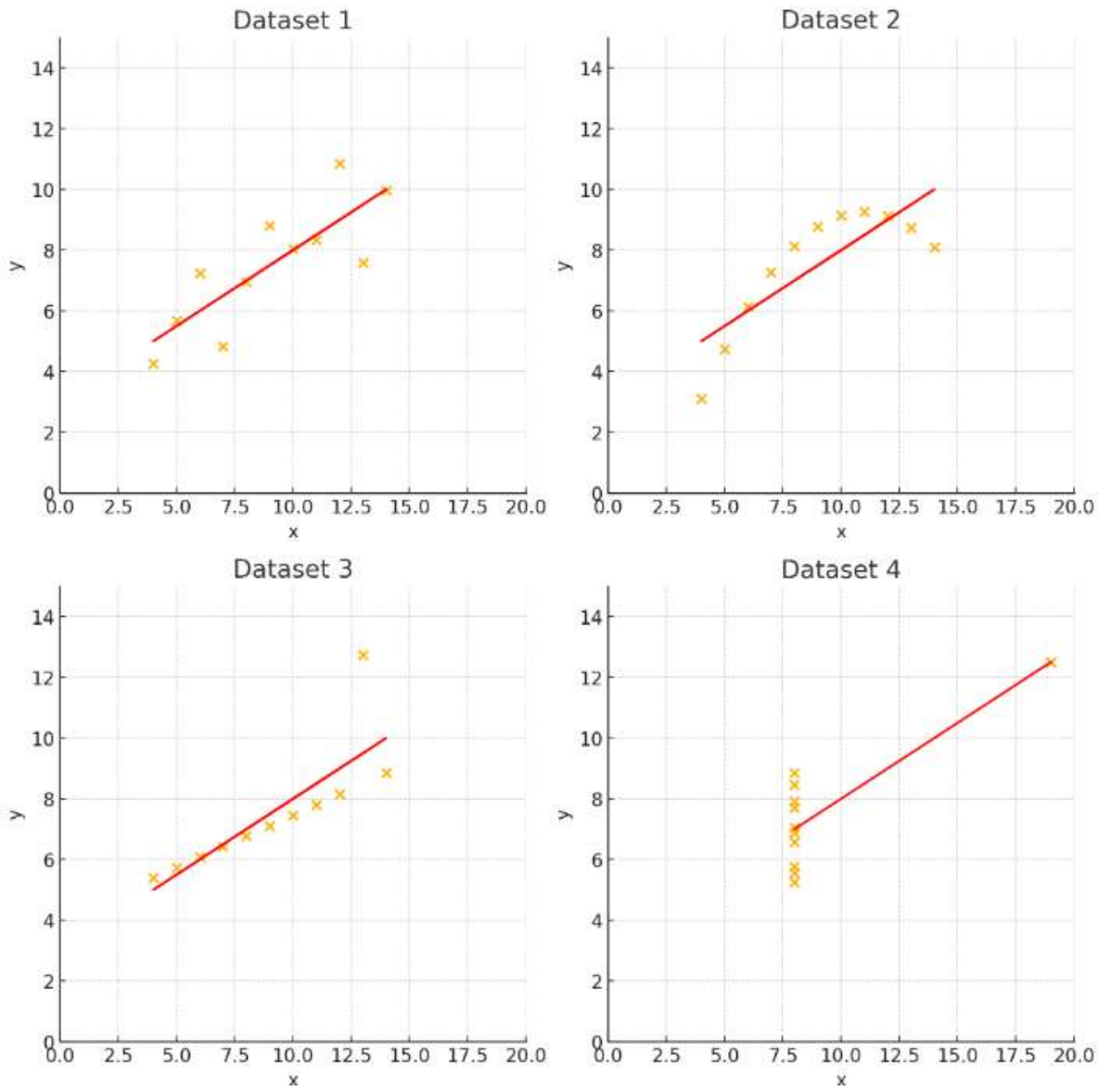
Dataset4

This dataset has a vertical line structure with one outlier that influences the linear regression results significantly. Most points have the same x value but different y values.

x	y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47

8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

Graphical representation: The visual differences will be apparent and emphasize the importance of graphical analysis in statistics.



The plots for Anscombe's quartet visually highlight the key differences between the datasets:

1. **Dataset 1**: Shows a roughly linear relationship with some scatter around the regression line y=3+0.5xy = 3 + 0.5xy=3+0.5x.
2. **Dataset 2**: Also roughly linear but with an outlier that affects the linear regression.
3. **Dataset 3**: Demonstrates a clear quadratic pattern, deviating significantly from the linear regression line.
4. **Dataset 4**: Consists of a vertical line structure with one significant outlier, which drastically influences the regression line.

These visual differences underline the crucial lesson of Anscombe's quartet: relying solely on summary statistics can be misleading, and graphical analysis is essential to fully understand the data.

Q -What is Pearson's R?

A- The Pearson **r** value, also known as the Pearson correlation coefficient, measures the linear relationship between two variables. It quantifies the degree to which two variables are related.

Characters of Pearson r

1. **Range**: The value of rrr ranges from -1 to 1.
   - r=1r = 1r=1: Indicates a perfect positive linear relationship between the two variables.
   - r=−1r = -1r=−1: Indicates a perfect negative linear relationship between the two variables.
   - r=0r = 0r=0: Indicates no linear relationship between the variables.
2. **Direction**:
   - A positive r value indicates that as one variable increases, the other variable also increases.
   - A negative r value indicates that as one variable increases, the other variable decreases.
3. **Magnitude**:
   - The closer the value of r is to 1 or -1, the stronger the linear relationship between the two variables.
   - The closer the value of r is to 0, the weaker the linear relationship.

Formula for Pearson r

The Pearson correlation coefficient between two variables XXX and YYY is calculated as:

$$r = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

where:

- X_i and Y_i are the individual sample points.
- $\bar{X}$ and $\bar{Y}$ are the means of the X and Y variables, respectively.

Q- What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A- Scaling is the process of adjusting the range and distribution of values in a dataset so that they can be compared on a similar scale. This is particularly important in machine learning and data preprocessing to ensure that different features contribute equally to the model and improve its performance.

Scaling is performed for several reasons:

1. **Improving Model Performance**: Many machine learning algorithms, such as gradient descent-based methods, perform better when features are on a similar scale. This can lead to faster convergence and improved accuracy.
2. **Avoiding Dominance**: Features with larger ranges can dominate the learning process if scaling is not applied, leading to biased model results.
3. **Distance-Based Algorithms**: Algorithms that rely on distance metrics (e.g., K-nearest neighbors, SVM, and clustering algorithms) are sensitive to the scale of data. Scaling ensures that each feature contributes equally to the distance computation.

Types of Scaling:

Normalized Scaling (Min-Max Scaling)

Normalization (or min-max scaling) adjusts the values of a feature to a fixed range, typically [0, 1] or [-1, 1]. The formula for normalization is:

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

Where :

- X is the original value.
- X' is the normalized value.
- $X_{\min}$ is the minimum value of the feature.
- $X_{\max}$ is the maximum value of the feature.

Standardized Scaling (Normalization)

Standardization (or Z-score normalization) adjusts the values of a feature so that they have a mean of 0 and a standard deviation of 1. The formula for standardization is:

$$X' = \frac{X - \mu}{\sigma}$$

where:

- X is the original value.
- X' is the standardized value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

Key Differences

1. **Range**:
   - **Normalization**: Transforms data to a fixed range (e.g., [0, 1]).
   - **Standardization**: Transforms data to have a mean of 0 and standard deviation of 1.
2. **Impact on Distribution**:
   - **Normalization**: Affects the range but does not necessarily change the underlying distribution of the data.
   - **Standardization**: Centers the data and scales it, making it easier to compare features with different units and distributions.
3. **Usage**:
   - **Normalization**: Preferred when the data needs to be bounded, such as in neural networks.

- **Standardization**: Preferred when dealing with features of different units and distributions, particularly in algorithms sensitive to the scale of data.

In summary, the choice between normalization and standardization depends on the specific requirements of the machine learning algorithm and the characteristics of the dataset.

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

A-The Variance Inflation Factor (VIF) measures the amount of multicollinearity in a set of multiple regression variables. A high VIF indicates that a predictor variable has a high degree of correlation with other predictor variables, making it difficult to isolate its individual effect on the response variable.

A VIF value becomes infinite when there is perfect multicollinearity, meaning that one predictor variable can be expressed as an exact linear combination of one or more of the other predictor variables. This situation occurs when

Formula for VIF:

VIF(Xi) = 1/(1-R^2i)

Where R^2 is the coefficient of determination of the regression of Xi on all the predictors. When there is Perfect correlation, R^2 will become 1 and thus VIF will be infinity.

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

A- A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, typically a normal distribution. The plot displays the quantiles of the data against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points will roughly form a straight line.

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

In the context of linear regression, a Q-Q plot is particularly useful for assessing the assumption that the residuals (errors) of the model are normally distributed. This is important because many inferential statistics, such as hypothesis tests and confidence intervals, rely on this assumption. Here's how the Q-Q plot is used and why it's important:

1. **Checking Normality of Residuals**:

    - **Assumption**: One of the key assumptions of linear regression is that the residuals (differences between observed and predicted values) are normally distributed.
    - **Q-Q Plot**: By plotting the residuals against a normal distribution, the Q-Q plot can visually show if the residuals deviate from normality. If the residuals are normally distributed, the points will lie along a 45-degree reference line.
2. **Identifying Deviations from Normality**:
    - **Heavy Tails**: If the points in the Q-Q plot curve away from the line at the ends, it indicates heavy tails (more extreme values than expected).
    - **Light Tails**: If the points curve towards the line at the ends, it indicates light tails (fewer extreme values than expected).
    - **Skewness**: If the points form an S-shape, it suggests skewness in the data (asymmetry in the distribution).
3. **Model Diagnostics and Validation**:
    - **Influence on Model Fit**: Non-normal residuals can indicate problems with the model, such as omitted variables, incorrect functional form, or the presence of outliers.
    - **Improving Model**: Identifying non-normality through the Q-Q plot can prompt further investigation and model improvement, such as transforming variables or adding missing predictors.
4. **Comparing Multiple Models**:
    - **Model Selection**: When comparing multiple regression models, Q-Q plots can help determine which model better meets the normality assumption of residuals, aiding in model selection.

Interpreting the Q-Q Plot

1. **Straight Line**: If the points lie close to the 45-degree line, the residuals are approximately normally distributed.
2. **Deviations from Line**: Any systematic deviations from the line indicate departures from normality, such as skewness or heavy/light tails.

Conclusion

A Q-Q plot is a vital diagnostic tool in linear regression for assessing the normality of residuals. Ensuring that the residuals are normally distributed is crucial for the validity of many inferential statistics and helps in identifying potential issues with the model.