Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Effect on the Dependent Variable (cnt):

Season: Visualized using a boxplot (sns.boxplot), which shows that bike demand varies across seasons. Summer and fall have higher demand, while winter has relatively lower demand.
Weather Situation: Likely analyzed using dummy variables for correlation or regression modeling. Clear weather might correlate positively with bike demand, while mist or adverse weather conditions could have a negative effect.
Holiday: Correlation analysis suggests that bike demand is generally lower on holidays compared to working days.
Working Day: A positive association with bike demand, as weekdays (working days) typically see higher bike usage due to commutes.

The correlation heatmap (sns.heatmap) includes encoded variables to analyze their relationship with cnt.

This analysis demonstrates that categorical variables significantly influence bike demand and are essential for predictive modeling.

- Further we can see the exact effect by best fit equation (i.e. cnt = 0.2272 + 0.2338 × yr + 0.4663 × temp − 0.1532 × windspeed − 0.0235 × updt_holiday − 0.0834 × season_spring + 0.0387 × season_summer + 0.0744 × season_winter − 0.2777 × weathersit_Light_Snow_Rain − 0.0748 × weathersit_Mist)

❖ from that 'yr', 'updt_holiday',' season_summer', 'season_winter', 'season_spring', 'weathersit_Light_Snow_Rain', 'weathersit_Mist', are the categorical variables.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using drop_first=True during dummy variable creation prevents the dummy variable trap, which occurs when all dummy variables for a categorical variable are included, causing multicollinearity. drop_first=True take place by dropping one dummy variable in the model.

- It will avoid redundancy and instability in regression.
- like if we not use drop_first=True then it Including all k dummy variables creates a linear dependency because the k^{th} variable can be derived from the other k−1.
- For instance: $Category_k = 1-(Category_1+Category_2+\ldots+Category_{k-1})$
- This redundancy leads to instability in parameter estimation.

- Uses the dropped category as the reference for comparison.
- Ensures coefficients are interpretable and computationally efficient.
- This keeps the model stable and meaningful.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

By looking at the pair-plot among the numerical variables (i.e. 'temp', 'hum', 'windspeed') the 'temp' (i.e. temperature) has the highest correlation with the target variable 'cnt'.

Which can further conform by looking at sns.heatmap from that we can know that 'temp' var has the highest correlation with 0.63 value.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)
I have done the validation of following assumptions of Linear Regression-
1. Normality of Error Terms – by using a histogram Plot of residuals. (Should be normal distribution curve plot)
2. Homoscedasticity – by a scatterplot - Plotting the residuals against the predicted values. (The residuals should scatter randomly around the zero line without any visible patterns.)
3. Independence of Residuals – by checking the Durbin-Watson statistic from the model summary. (Value should close to 2)
4. Verification of Linearity – by using the Component and Component Plus Residual (CCPR) plots.
5. Multicollinearity Verification – by checking the VIF values of predictors. (Values should - VIF<5)

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Based on the final model, here are the top 3 features contributing significantly towards explaining the demand of the shared bikes

- highest positive features -

1.' temp' with β2 value of 0. 4663.
2. 'yr' with β2 value of 0.2338.
3.'season_winter' with β2 value of 0.0744.

- highest negative features –

1. 'weathersit_Light_Snow_Rain' with β2 value of (−0.2777).

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is a supervised learning algorithm used for modeling the relationship between a dependent variable (y) and one or more independent variables (X_1, X_2, …, X_n). It predicts the value

of y based on a linear combination of the input features.

1. Assumptions of Linear Regression
Linearity: The relationship between the dependent and independent variables is linear.
Independence: Observations are independent of each other.
Homoscedasticity: The variance of residuals is constant across all levels of independent variables.
Normality of Residuals: Residuals (errors) are normally distributed.
No Multicollinearity: Independent variables are not highly correlated.

2. Mathematical Representation
For a simple linear regression model: $y = (\beta_0 + \beta_1 * X + \epsilon)$

y: Dependent variable (target).
X: Independent variable (feature).
β_0: Intercept (value of y when X=0).
β_1: Slope (rate of change of y with X).
ϵ: Residual error (difference between actual and predicted values).

For multiple linear regression:
$(y=\beta_0 + \beta_1 \times 1 + \beta_2 \times 2 + \cdots + \beta_n \times n + \epsilon)$

3. Steps in Linear Regression
a. Feature section
by using Automated approach like Recursive Feature Elimination (RFE)
or Manual approach like Backward Elimination
(I have used both)

b. Fit the Model
by using statsmodel.api or by using sklearn library.
The algorithm calculates the values of coefficients (β_0 , β_1, …, β_n) by minimizing the sum of squared errors (SSE):
Ordinary Least Squares (OLS) is the method used to find the best-fit line:

c. Evaluate the Model
Calculate metrics to measure model performance:
R-squared (R2): Proportion of variance in y explained by X.
Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

c. Validate Assumptions
Check residuals for normality, linearity, and homoscedasticity.
Use VIF to check for multicollinearity.

By following this linear regression algorithm one can build model that good and reliable for consistent and accurate predictions.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet is a collection of four datasets created by the statistician Francis Anscombe in 1973. These datasets show the importance of visualizing data before analyzing it. Despite having nearly identical statistical properties, the datasets show different patterns when plotted, emphasizing how reliance only on summary statistics can be misleading.

1.Properties of Anscombe's Quartet

Each dataset in the quartet has:

The same mean for the X and y variables.

The same variance for X and y.

The same correlation coefficient (r=0.816) between x and y.

The same linear regression line: y=3+0.5x.

The same coefficient of determination (R2).

2. Datasets in the Quartet

Each dataset consists of 11 points, but their distributions differ:

for example-

a. Dataset 1:

The data forms a nearly perfect linear relationship.

The linear regression line accurately represents the data.

b. Dataset 2:

The data is non-linear, following a quadratic-like curve.

The regression line poorly fits the data.

c. Dataset 3:

All points except one fall on a perfect linear relationship.

The outlier significantly affects the regression line, demonstrating the impact of outliers.

d. Dataset 4:

Most points have the same X value, forming a vertical cluster, with one point far away.

The regression line is heavily influenced by the single distant point.

3. Understandings from the Anscombe's Quartet:

Visualization is important: Summary statistics (mean, variance, correlation) can be identical for different datasets, but the patterns in the data might differ significantly.

Understand the Context: Always analyze data in the context of the problem, not just through numbers.

Detect Outliers and Patterns: Visualizing data can reveal anomalies or non-linear relationships that might not be evident from statistics.

4. Conclusion

Anscombe's quartet highlights the limitations of relying solely on numerical summaries. Data visualization should be a fundamental step in data analysis to revel hidden insights and to make sure accurate interpretation.

Question 8. What is Pearson's R? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R, also known as the Pearson Correlation Coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two variables. It is widely used in data analysis to determine how two variables are closely linearly related.

Pearson's R is a powerful metric for identifying and quantifying linear relationships between variables, but it must be used carefully, considering its assumptions and limitations.

The formula for calculating Pearson's R is:

$r = \sum(Xi - \bar{X}) \cdot (yi - \bar{y}) / \sqrt{(\sum(Xi - \bar{X})2 \cdot (yi - \bar{y})2)}$

Where:

Xi and yi: Individual data points of variables X and Y.

$\bar{X}$ and $\bar{y}$: Means of X and y.

r: Pearson Correlation Coefficient.

Pearson's R ranges between -1 and +1:

r = +1: Perfect positive linear relationship.

r = −1: Perfect negative linear relationship.

r = 0: No linear relationship.

Interpretation

Strength

for example-

0.0 < |r| <= 0.3: Weak linear relationship.

0.3 < |r| <= 0.7: Moderate linear relationship.

0.7 < |r| <= 1.0: Strong linear relationship.

- Direction:

r > 0: Positive relationship (as X increases, Y increases).

r < 0: Negative relationship (as X increases, Y decreases).

Usage-

Feature selection in machine learning.

Identifying potential predictors in regression analysis.

Limitations

Pearson's R only measures linear relationships and may not capture non-linear patterns.

It is sensitive to outliers, which can distort the coefficient.

Assumes the data is continuous and normally distributed.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is a preprocessing technique used in machine learning to adjust the range of features in a dataset. It ensures that numerical values of features are transformed to a specific scale without distorting differences in their distribution.

Scaling is Performed -

To Handle Feature Variability:

Features with different units (e.g., age in years vs. income in dollars) can have widely varying scales, leading to biased results in algorithms like gradient descent-based models or distance-based models.

Improving Model Performance:

Proper scaling speeds up convergence and ensures optimal model performance.

Ensuring Fair Feature Weightage:

Without scaling, features with large values can dominate the learning process, overshadowing features with smaller ranges.

The difference between normalized scaling and standardized scaling-

Normalized scaling like (Min-Max scaler) transforms data into a specific range, typically [0, 1]. This is achieved by subtracting the minimum value and dividing by the range (max - min). It is particularly useful when features have varying units or scales.

It is most widely used method for scaling.

However, normalization is highly sensitive to outliers, as extreme values can distort the scaling process.

Standardized scaling, on the other hand, adjusts the data to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean and dividing by the standard deviation of the feature. Standardization is ideal for data that is approximately normally distributed and for algorithms assuming Gaussian distribution, such as linear regression. Unlike normalization, it is less sensitive to outliers, though extreme values can still have some influence. The choice between normalization and standardization depends on the dataset and the algorithm's requirements.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

Variance Inflation Factor (VIF) quantifies multicollinearity among independent variables in regression.

A VIF value becomes infinite when there is perfect multicollinearity between two or more independent variables.

This Happens because When one independent variable is an exact linear combination of other variables, the denominator in the VIF formula becomes zero:

VIF= $1/(1-R^2_j)$

Here:

R^2_j is the coefficient of determination from regressing the j^{th} variable on all other independent variables.

If R^2_j =1 (perfect linear dependency), then 1−R j2 = 0, making VIF infinite.

Causes of Infinite VIF-
Duplicate Variables:
Including the same feature more than once, such as X1 and its exact duplicate X1 * 1.
Linear Dependencies:
Perfectly correlated features (e.g., height in meters and height in centimeters).
Feature Engineering Issues:
Creating features that are deterministic functions of others (e.g., a column for total (cnt) that is the sum of other feature columns (Casual, Registered).

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>
A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset against a theoretical distribution, often the normal distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution. If the points lie approximately along a 45-degree line, it indicates that the observed data closely follows the theoretical distribution.

Use of Q-Q Plot in Linear Regression-
In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps to assess this assumption by comparing the residuals against a normal distribution.

If the residuals form a straight line on the Q-Q plot, it suggests they are normally distributed.
Deviations from the straight line indicate departures from normality, such as skewness or the presence of heavy tails.

Importance of Q-Q Plot in Linear Regression-
Validates Model Assumptions:
Normality of residuals is crucial for making valid inferences, such as confidence intervals and hypothesis tests.
A Q-Q plot identifies violations of this assumption.
Identifies Outliers:
Extreme deviations in the Q-Q plot highlight potential outliers in the data.
Improves Model Diagnostics:
It helps decide if transformations or alternative models are needed to better fit the data.
