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)

- Bike rentals are increased in summer and fall season as compared to spring and winter
- Bike rentals are increased significantly in year 2019 than year 2018
- Bike rentals increased from Jan to May, peaked between May to Oct and decreased from Nov to Dec in each year
- On holidays rental count has slightly dropped
- There is no significant change as per weekday or weekends on rental counts
- There is no significant difference in rental counts for working and non-working days
- Clear weather attracted most bikers to rent bike. During heavy rain not a single bike got rented, which is obvious

---

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)

- drop_first=True is important because it helps reducing the number of columns created during dummy variable creation. It helps to reduce possibility of multi-collinearity while creating multiple linear regression model.
- If a categorical variable has 3 possible values X, Y, Z then while crating dummy variable columns, we can create only two columns X and Y. If the variable's value is not X and not Y then it is obviously Z. So we don't need 3rd column to identify Z

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)

- 'temp' and 'atemp', both variables have highest correlation with target variable

---

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)

- Linear relationship
  - Linear relationship is observed between temp and cnt variable during bivariate analysis
- Normality of error terms
  - Error terms are observed normally distributed with mean equal to zero during residual analysis
- Independence of error terms
  - There is no visible pattern among error terms (like time series data where next value depends on previous one) during bivariate analysis

- Homoscedasticity
    - Variance of error terms is constant, it is not increasing or decreasing as error values change
- Multicollinearity check
    - There is no significant multicollinearity observed. This is done by checking VIF value among independent variables

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)

According to magnitude of the coefficient following 3 variables are most significant

- weathersit_LightRain
- season_spring
- yr

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)

Concept

Linear regression is a statistical method used for modeling the relationship between a dependent variable Y and one or more independent variables X

The general formula for simple linear regression (when there is one independent variable) is:

$$Y = \beta 0 + \beta 1X + \epsilon$$

Where:

Y is the dependent variable (output we want to predict)

X is the independent variable (input used to make predictions)

β0 is the intercept term (constant)

β1 is the slope (coefficient) of the line, showing how Y changes with X

ϵ is the error term (residual), which accounts for the difference between the observed value and the predicted value

In multiple linear regression, when there are multiple independent variables, the formula is generalized to:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + \epsilon$$

Objective

The goal of linear regression is to find the values of β0 and β1 that minimize the sum of squared residuals. The residual (ϵ) is the difference between the observed value and the predicted value.

Finding Regression Coefficients

These coefficients, β0 and β1 are typically determined using calculus and optimization techniques. After the coefficients β0 and β1 are calculated (i.e. model is trained), the model can make predictions for new data points.

Prediction

The predicted value Y^ for any given X is:

$$Y^ = \beta 0 + \beta 1X$$

Evaluation of model performance

Once the model is trained, we need to evaluate its performance. The key metric used for evaluation include:

R-squared(R2): A statistical measure of how well the regression line approximates the real data points. It is the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R2 = 1 - (\sum(Yi - Yi^)2 / \sum(Yi - Y^-)2)$$

Where:

Yi are the actual values.

Yi^ are the predicted values.

Y^- is the mean of the actual values.

Assumptions of Linear Regression

For linear regression to work well, the following assumptions must be satisfied:

Linearity:

The relationship between the independent variable(s) and the dependent variable is linear.

Independence:

The residuals (errors) are independent of each other.

Homoscedasticity:

The variance of the residuals is constant across all levels of the independent variable.

Normality:

The residuals are normally distributed.

Multi-collinearity:

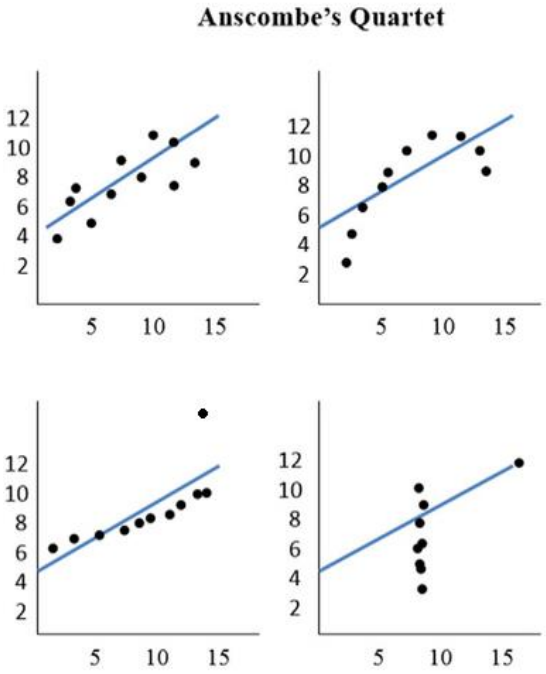
There is very little or no multi-collinearity in the data. Multi-collinearity occurs when the independent variables or features have dependency in them.

If any of these assumptions are violated, the model may not perform well, and adjustments might be needed.

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)

Anscombe's Quartet is a set of four data sets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but very different distributions and visual patterns when plotted.

Anscombe's Quartet of Different XY Plots of Four Data Sets Having Identical Averages, Variances, and Correlations



Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	5.76
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	8.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	7.26	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Source: Adapted from Anscombe (1973, pp. 19-20)

1

The four datasets I, II, III, IV each result into same statistical properties, i.e. Mean, Sample Variance, Correlation and Linear Regression. But when they are plotted, visually they tell a different story.

- Dataset I points lies around a straight line, suggesting linear relationship between X and Y
- Dataset II points forms curved shape, indicating linear model will not fit well here
- Dataset III points follow a linear trend but one extreme outlier present here pulls regression line in a direction that does not accurately represent majority of data points
- Dataset IV points are vertically aligned but one extreme outlier present here changes direction of regression line and correlation is misleading

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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)
Peason's correlation coefficient (R), is a measure of strength and direction of the linear relationship between two variables. It quantifies how closely two variables are related to each other on the scale of -1 to +1.

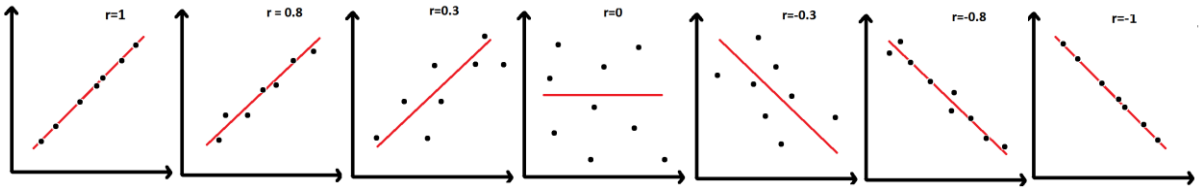
Interpretation

- R = +1: As X increases, Y increases in a perfectly linear manner (perfect positive linear relationship)
- R = -1: As X increases, Y decreases in a perfectly linear manner (perfect negative linear relationship)
- R = 0: There is no linear relationship between X and Y
- 0 < R < 1: As X increases, Y tends to increase, but the relationship is not perfectly linear
- -1 < R < 0: As X increases, Y tends to decrease, but the relationship is not perfectly linear

Strength Of Relationship

- R = 0.1 to 0.3: Weak positive correlation
- R = 0.3 to 0.5: Moderate positive correlation
- R = 0.5 to 1.0: Strong positive correlation
- R = -0.1 to -0.3: Weak negative correlation
- R = -0.3 to -0.5: Moderate negative correlation
- R = -0.5 to -1.0: Strong negative correlation

Visualizing R



Diagrams above show relationship between data points at various values of R

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)
Scaling refers to the process of transforming the values of features in a dataset into a specific range or distribution. This is typically done to adjust features so that they are on a similar scale.

Scaling is performed for:

- Improving algorithm Performance
  - If features have very different ranges, (0-1 vs 0-1000), then algorithms which are gradient descend based (linear regression) might perform poorly or inefficiently
- Improving convergence speed in optimization
  - the optimization in gradient descend can be slow, as the algorithm might take very small steps in some directions (for large-valued features) and large steps in others (for small-valued features), leading to slow convergence.

- Equal importance of features
    - In algorithms like kNN, features with larger ranges may dominate the learning process if scaling is not applied
- Distance-based Models
    - For models that rely on distance (e.g., k-NN, k-means clustering), unscaled data can lead to skewed results, as larger values can dominate the distance calculation

Difference between normalized and standardized scaling is as follows,

Aspect	Normalized Scaling	Standardized Scaling
Range of data	Transforms data to a fixed range. Usually, 0 to 1 or -1 to 1	Transforms data to have a mean of 0 and standard deviation of 1
Effect of outliers	It is affected by outliers	It is much less affected by outliers
Data Distribution	No assumption about the distribution of data. Best used when you want to scale to a fixed range.	Assumes the data is approximately normally distributed. Best used when algorithms that require standardization (e.g. linear regression) are used.

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)

A VIF value of infinity typically occurs when there is perfect multicollinearity between two or more independent variables in the model. This means one variable can be perfectly predicted by a linear combination of other variables, causing the model's estimates to become unstable.

Reasons

- If one predictor variable is a perfect linear combination of others, the R-squared value for that predictor, when regressed against all other predictors, becomes 1. The formula for VIF is $1/(1-R2)$, so when $R2=1$, the denominator becomes zero, resulting in an infinite VIF.
- If redundant or duplicate variable is present in a dataset, it can cause a perfect multicollinearity, leading in an infinite VIF

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)

The quantile-quantile (Q-Q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not.

Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.

By quantile we mean percent of points below the given value. For example, by value at 0.3 quantile (or 30 percentile) we mean, 30% of values fall below this value. A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

For reference purposes, a 45-degree line is also plotted. If the two samples are from the same population, then the points should approximately lie along this line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Interpretation

Straight Line: If the points lie close to the straight line, this indicates that the observed data closely follows the reference distribution (e.g., normal distribution).

S-shaped curve: If the points form an S-shape, the data may have heavy tails or lighter tails compared to the reference distribution.

Curved pattern: If the points show a curve that deviates significantly from the line, it indicates skewness (data may be right-skewed or left-skewed).

Outliers: Points far from the line represent potential outliers in the data.

Importance of Q-Q Plot in Linear Regression

Check Normality of Residuals

The Q-Q plot is a simple way to visually check if the residuals of your linear regression model are approximately normal. If the residuals deviate significantly from a straight line in the Q-Q plot, it suggests that the normality assumption may be violated.

Detect Skewness and Kurtosis

A right-skewed distribution (where the tail is longer on the right side) will show a bend in the Q-Q plot above the line at the right tail.

A left-skewed distribution (where the tail is longer on the left side) will show a bend below the line at the left tail.

A heavy-tailed distribution will show a significant deviation from the straight line at both ends of the plot.

Assumption Validation

If the Q-Q plot of residuals indicates a significant deviation from normality, this suggests that the results of statistical tests like t-test or F-test might not be valid. These tests assume normal distribution of residuals.

Identify Outliers

The Q-Q plot can also help identify outliers in the residuals. Outliers will appear as points far away from the straight line on the plot.