Assignment

Assignment-based Subjective Questions

Q1) . From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans) 1. The box plots for weather situations show that bike rentals are more frequent during periods of clear skies, few clouds, and partly cloudy conditions.

2. the year box plots shows that 2019 experienced a greater number of bike rentals compared to other years.

3. The majority of bookings occur between May and October, with a notable upward trend from the beginning of the year until the middle of the year, followed by a gradual decline towards the year's end.

4. Thursday, Friday, Sunday, and Saturday exhibit higher bookings compared to the earlier part of the week.

5. According to the month box plots, September stands out as the month with the highest number of bike rentals.

Q2) Why is it important to use drop_first=True during dummy variable creation?

Ans) Setting `drop_first = True` is crucial as it helps mitigate multicollinearity issues among dummy variables by removing one redundant column created during dummy variable creation. Also gives more accurate and efficient representation of the data in the model.

Q3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans) we observe a substantial correlation of 0.63 between the 'temp' and 'atemp' variables with the target variable 'cnt'. Although 'casual' and 'registered' exhibit stronger correlations with 'cnt', it's crucial to recognize that these variables essentially contribute to the 'cnt' value itself. we prioritize examining 'temp' due to its significant correlation with 'cnt', while respecting the requirement to exclude 'casual' and 'registered' as stipulated by the data dictionary.

Q4) How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans) Linearity: Plot observed vs. predicted values or residuals vs. predicted values to check for a linear relationship.

Normal Distribution of Residuals: Use a Q-Q plot or a histogram to ensure residuals are normally distributed, similar to the histogram in the image you provide.

Independence of Residuals: Perform a Durbin-Watson test to test for independence.

Multicollinearity: Use Variance Inflation Factor (VIF) analysis to check for multicollinearity among predictors.

Q5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans) the top 3 features contributing significantly towards explaining the demand of the shared bikes are summer, temp, sept.

General Subjective Questions

Q1) Explain the linear regression algorithm in detail.

Ans) Linear regression is an algorithm which helps like drawing a straight line through points on a graph. It gives linear relationship between independent and dependent variable. It is a tool used in data science and machine learning to guess what might happen next based on what we already know. The regression model helps us figure out how changes in the independent variable affect the dependent variable.

Mathematically it is shown as follows,

$Y = m*X + b$

Where X= dependent variable

Y = independent variable

m = slope of the line

There are 2 types of regressions:

1) Simple linear regression
2) Multiple linear regression:

A regression line is like a ruler on a graph that helps us see how the dependent and independent variables relate to each other in a straight-line pattern. It is of 2 types

1)positive linear relationship
2) negative linear relationship.

Cost function: It shows how well the model is doing by measuring the difference between the predicted values and the actual values.

The best fit line : Aims to minimize the gap between predicted and actual values, making the error as small as possible.

Gradient descent : It is an optimization technique employed to minimize the Mean Squared Error (MSE) by iteratively adjusting parameters based on the gradients of the cost function.

Assumptions of Linear Regression

1)Autocorrelations: that the errors are independent of each other across observations.

2)Multicollinearity **:** The linear regression model assumes minimal or negligible multicollinearity among the independent variables, so that that each predictor contributes unique information to the model without being overly redundant with others.

3)Normal distribution of error terms: In linear regression, it's assumed that the errors, also known as residuals, conform to a normal distribution pattern.

4)Linear relationship between the features : there exists a linear association between the dependent and independent variables, allowing for the construction of a straight-line model to describe their relationship.

Q2) Explain the Anscombe’s quartet in detail.

Ans) Anscombe's quartet consists of four distinct datasets that share identical descriptive statistical properties, including means, variances, correlations, and linear regression lines. each dataset exhibits unique patterns when visualized through scatter plots, illustrating the importance of graphical exploration in data analysis. Anscombe's quartet underscores the critical role of data visualization in uncovering trends, outliers, and other significant insights that may not be readily apparent through summary statistics alone.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Dataset I: there is a formation of a simple linear relationship with a slight outlier.

Dataset II: It also forms a linear relationship but with a significant outlier, which influences the regression line.

set III: there can be seen a non-linear relationship with an outlier, demonstrating the impact of a single outlier on correlation and regression.

set IV: It shows relationship except for a single point, which causes the correlation coefficient to be close to zero.

Anscombe's quartet is a valuable resource for teaching statistical concepts effectively. By incorporating the quartet into educational activities, educators

can actively engage learners in hands-on exploration of fundamental statistical principles such as mean, variance, correlation, and regression analysis.

Q3) What is Pearson's R?

Ans) Correlation, as implied by its name, denotes the association between two variables, often referred to as their co-relation. The correlation coefficient serves as a quantitative measure of this relationship. It's calculated using a formula aimed at assessing the connection between two datasets. Specifically, the Pearson correlation coefficient, also known as the Pearson product-moment correlation coefficient, is commonly used to evaluate the linear dependency between datasets. This coefficient ranges from -1 to +1, where a value of 0 suggests no relationship between the datasets. A coefficient of +1 indicates a perfect positive correlation, while -1 signifies a perfect negative correlation.

The Pearson correlation coefficient is denoted by the letter "r". The formula is as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,
r = Pearson correlation coeffient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Correlation strength can be categorized as follows:

- Perfect: Values close to ±1 signify a perfect correlation.
- Moderate Degree: Values between ±0.30 and ±0.49 suggest a moderate correlation.
- Low Degree: Values below +0.29 denote a weak correlation.
- High Degree: there is a high correlation Value between ±0.50 and ±1
- No Correlation: A correlation coefficient of zero indicates no discernible relationship between the variables.

Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans) Scaling involves geometric modifications that uniformly enlarge or diminish objects. Objects or rules possessing the property of scale invariance remain unaltered when factors such as length, energy, or other variables are multiplied by a common factor. The dataset contains features with significant variations in terms of their magnitudes, units, and ranges.

Scaling is performed as it ensures that all features are represented on a comparable scale with similar ranges, a process commonly referred to as feature normalization. This practice holds significance because the magnitude of features can influence the performance of various machine learning techniques. To prevent numerical instability, it's essential to mitigate significant differences in scales between features.
In the absence of scaling, features with larger scales may disproportionately influence the learning process, potentially leading to biased or skewed outcomes. Scaling ensures that each feature contributes proportionally to model predictions, thus promotes fair and balanced learning outcomes.

Difference between normalized scaling and standardized scaling:

1)Values on the scale are confined within the range of [0, 1] or [-1, 1] whereas values on a scale can vary across a continuum without being restricted to a specific range.

2)normalized scaling is called scaling normalization and standardized scaling is called Z-score normalization.

3) Having features on various scales can be advantageous in certain contexts and standardized scales the model by utilizing the mean and standard deviation.

4) When the distribution of features is ambiguous normalized method proves beneficial and the standardized method remains beneficial when the distribution of features is uniform.

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans) When VIF reaches infinity, it indicates perfect correlation between two independent variables. When VIF exceeds 10, it indicates substantial multicollinearity that requires remediation. For instance, a VIF of 4 implies that variance is inflated by a factor of 4 because of multicollinearity A high VIF value suggests increased variance in model coefficients due to multicollinearity. Perfect correlation results in an R-squared value of 1, leading to an infinite VIF value. To address this issue, it's necessary to eliminate one of the variables causing perfect multicollinearity from the dataset.

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans) Q-Q plots, short for Quantile-Quantile plots, compare the quantiles of a sample distribution with those of a theoretical distribution. By visually assessing the alignment of these quantiles, analysts can discern whether the dataset conforms to a specific probability distribution, such as normal, uniform, or exponential.

QQ plots are useful as follows:

1. Assessing whether two populations share the same distribution by comparing their quantiles.
2. Verifying if residuals from a regression model adhere to a normal distribution, validating the assumption of normality in regression analysis.

3. Evaluating the skewness of a distribution by examining the deviation of quantiles from a diagonal line on the plot.

Q-Q plots are commonly employed to visually examine whether a dataset conforms to a particular probability distribution, such as the normal distribution. It holds particular importance in numerous statistical analyses, as the accuracy of statistical inferences relies heavily on the validity of distributional assumptions.Outliers are data points that exhibit considerable deviation from the majority of the dataset. Q-Q plots serve as a particularly useful tool for evaluating the normality of a dataset.
Deviations from anticipated patterns in the plot may indicate alterations in the underlying processes, prompting the need for deeper investigation. Q-Q plots are utilized in the validation process of predictive models.