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? (3 marks)

Answer:

Below is the analysis obtained after reviewing the Uni and Bivariate plots for the categorical variables:

- Fall season has the greatest number of bookings and the booking count has significantly increased from 2018 to 2019.
- As evident from the remaining columns, 2019 attracted a greater number of bookings as compared to the previous year. This shows a good overall progress in business
- The highest number of bookings are made during may, june, july, aug, sep and oct, post that we can observe a dip which help us interpret that a smaller number of bookings are made at the year-end (being a festive season). Further to this the number of booking also increased for each month in 2019 as compared to 2018
- People don't prefer booking bikes on a holiday.
- Thursday, Friday, Saturday and Sunday record highest number of booking as compared to the start of week, in both the years. With higher booking counts recorded in 2019.
- The bookings don't seem to be impacted by the working days as the number are fairly close.
- Clear weather attracts the greatest number of bookings. Booking numbers also increased in 2019.

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

Answer:

When you have categorical variables with let's assume n levels, the idea is to create n-1 (indicating the levels) dummy variables. With each unique categorical feature represented in a separate column in binary form (i.e 0 and 1).

Without drop_first=True, we will end up using all the columns in a regression model which can lead to multicollinearity. This introduces redundancy in the model and can cause issues with the regression coefficient estimation, making them unreliable and unstable.

Furthermore, by reducing the number of columns, make the model more efficient in terms of computation and simplifying model interpretation.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

Upon analysing the pairplot we can say that the target variable ('cnt') has highest correlation with **'temp'** (independent numerical variable).

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

In order to validate the assumption of Linear Regression model to make the inferences, I have authenticated the below pointers:

- The error terms are normally distributed with zero mean, by plotting the histogram of residual/error terms.
- The Error terms are independent of each other, for which I have attached Residuals vs. Order of Observations plot, ACF and PACF Plots and calculation of Durbin-Watson statistic value.
- Error terms have a constant variance i.e. Homoscedasticity for which there was no pattern and the spread of pattern found constant across range of fitted values.
- Linear relationship validation for the dependent and independent variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

The top 3 features contributing significantly towards explaining the demand of the shared bikes are mentioned in the order of their coefficient's absolute values:

1. 'temp' with 0.4786
2. 'Light Snow' with -0.2913
3. yr with a coefficient of 0.2339

These features have a largest absolute value of their coefficients, indicating that they are have the most significant impact on the target variable 'cnt'.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear regression is a fundamental statistical method used to model the relationship between a dependent/target variable (which is continuous in nature) and one or more independent/feature variable. The goal is to find the best-fitting line that predicts the dependent variable based on the independent variables.

Following are the two important types of linear regression models:

1. Simple linear regression
2. Multiple linear regression

Simple Linear Regression:

Simple linear regression models the relationship between two variables by fitting a linear equation to the observed data. The equation of a line is:

$$y = B0 + B1x + E$$

Where, y is the dependent variable

x is the independent variable

B0 is the y-intercept

B1 id the slope of line

E is the error term

Multiple Linear Regression:

When there are multiple independent variables, the model is extended to:

$$y = B0 + B1x1 + B2x2 + \ldots\ldots\ldots\ldots + Bpxp + E$$

Where, x1, x2, x3, …….. xp are the coefficients.

Assumptions:

Following are some assumptions about dataset that is made by Linear Regression model

- Linearity: The relationship between the independent and dependent variables is linear.
- Normality: The Error terms or residuals (errors) are normally distributed with zero mean.
- Independence: Error terms are independent of each other.
- Homoscedasticity: Error terms have a Constant variance.

By careful data understanding and data preparation methods, we can build an effective model. Further, evaluating these models using coefficients and plots, we can ensure model's reliability and accuracy in predicting the target variable.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, yet appear very different when graphed. It was developed by statistician Francis Anscombe in 1973. The quartet demonstrates importance of graphing data before analysing it and also the limitations of using summary statistics alone to understand the data.

Key Properties:

Mean value and variance of x and y variables is same across all four datasets

The correlation coefficient, which measures linear relationship between x and y, is same across all four sets.

The linear regression line has same slope and intercept for all four datasets.

Proportion of variance (Coefficient of determination – R2) in the dependent variable that is predictable from the independent variable is same across all four datasets.

	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

Properties for each dataset:

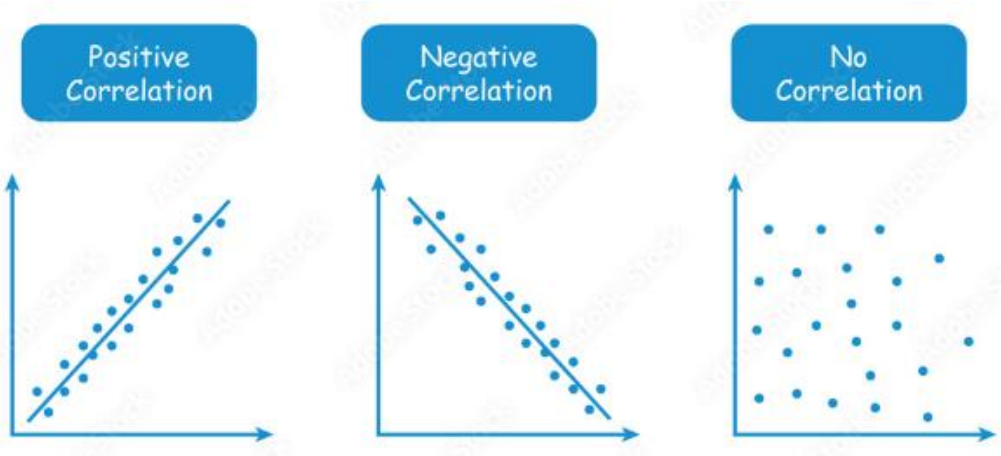
- Mean of x: 9 and Mean of y: 7.5
- Variance of x: 11 and Variance of y: 4.12
- Correlation between x and y: 0.816
- Regression Line:
    - y=3.00+0.500x
- Coefficient of Determination (R^2): 0.67

Therefore, Anscombe's quartet emphasizes the importance of using both, numerical and graphical methods in data analysis to gain a comprehensive understanding of the data.

3. What is Pearson's R? (3 marks)

Answer:

Pearson's R or the Pearson correlation coefficient (r), is a measure of the strength and direction of the linear relationship between two variables. It was developed by Karl Pearson and is one of the most widely used statistics in various fields such as biology, economics and psychology.



Importance characteristics of Pearson's R

Range: Value of Pearson's r ranges from -1 to 1

- r = 1: Indicates a positive linear relationship between variables.
- r = -1: Indicates a negative linear relationship between variables.
- r = 0: Indicates no linear relationship between variables.

Strength:

- Closer the value of r to 1 or -1, the stronger the linear relationship is between the variables.
- On the other hand, closer the value of r is to 0, the weaker the linear relationship.

Direction:

- Positive r: If one variable increases, the other variable also increases.
- Negative r: If one variable increases, the other variable decreases.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is the process of adjusting the range and distribution of data features so that they fit within a specific scale. This is an essential preprocessing step, especially for algorithms that are sensitive to the magnitudes of the data, such as gradient descent-based algorithms, k-nearest neighbors, and support vector machines.

Improving Algorithm Performance:

- Many machine learning algorithms perform better when their features are on a similar scale.
- It helps in achieving faster convergence for gradient-based optimization algorithms.

Eliminating Bias:

- Features with larger ranges can dominate the calculations and skew the results in algorithms.

Ensuring Stability:

- It improves the numerical stability in calculations, avoiding potential issues with large numbers.

Comparability:

- It ensures that each feature contributes equally to the analysis and interpretation.

Normalized scaling	**Standardized scaling**
Rescales data to a fixed range, typically [0, 1] or [-1, 1]	Rescales data to have a mean of 0 and a standard deviation of 1
$X_{normalized}=(X-X_{min})/(X_{max}-X_{min})$	$X_{standardized}=(X-\mu)/\sigma$
Mean not necessarily 0	Mean is 0
Standard Deviation not necessarily 1	Standard Deviation is 1
Highly sensitivity to Outliers	Lower sensitivity to outliers (outliers affect mean and standard deviation, but less drastically)
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

Both Normalized and Standardized scaling techniques aim to improve the performance and stability of machine learning models. However, the choice between these two depends on specific characteristics and requirements of dataset and the algorithm used.

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

(3 marks)

Answer:

The Variance Inflation Factor (VIF) measures how much of the variance of a regression coefficient is inflated due to multicollinearity with other predictors. If the VIF value is infinite, it indicates that there is perfect multicollinearity.

Perfect multicollinearity occurs when one predictor variable in a multiple regression model can be perfectly predicted from the others. In other words, there is an exact linear relationship between some of the predictor variables.

VIF for a predictor variable Xi is calculated as:

$VIF(Xi) = 1/(1-Ri2)$

Where Ri2 is the coefficient of determination of regression of Xi on all the other predictor variables.

If Ri2 = 1, which means Xi can be perfectly predicted by the other predictors, then:

$VIF(Xi) = 1/(1-1) = 1/0 = \infty$

An infinite VIF value signals perfect multicollinearity, meaning one predictor variable is a perfect linear function of the others. This can lead to an unstable coefficient estimates and inflated standard errors, by further making the model unreliable. Identifying and resolving multicollinearity is crucial for a robust regression analysis.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

A Q-Q (Quantile-Quantile) plot is a graphical tool to assess if a dataset follows a specified theoretical distribution, such as the normal distribution. It plots the quantiles of the data against the quantiles of the specified distribution. In order to drive the plot, Components of a Q-Q Plot has Theoretical quantiles (from the specified distribution) on the X-Axis and Sample quantiles (from the observed data) on the Y-Axis.

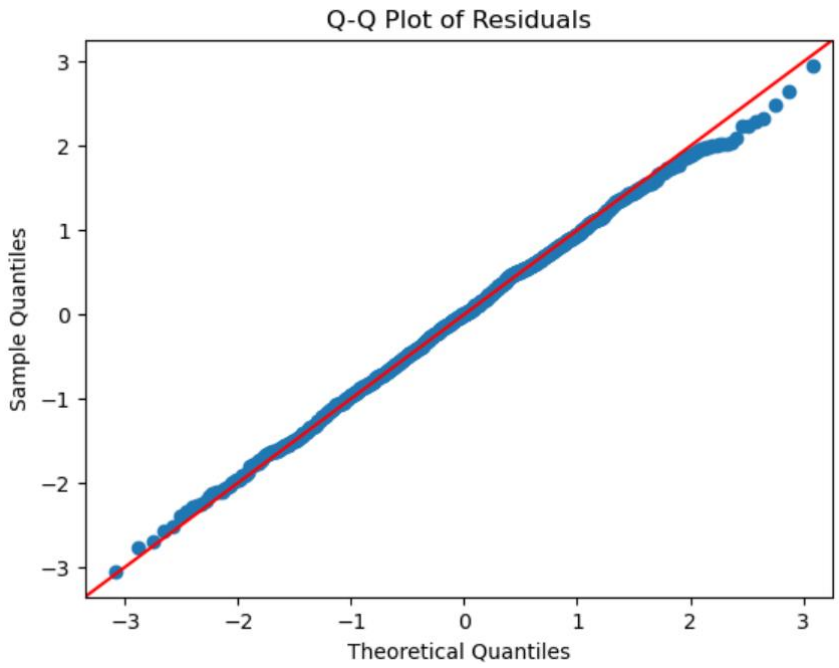
In linear regression, the Q-Q plot is primarily used to assess the assumption that the residuals (errors) are normally distributed. This assumption is crucial because many inferential statistics, such as t-tests and F-tests, rely on the normality of residuals.

Key Uses in Linear Regression

- **Assess Normality of Residuals**:
    - The residuals should be normally distributed for the validity of various statistical tests.
    - A Q-Q plot of residuals helps in diagnosing deviations from normality.
- **Detect Outliers**:

- Outliers can be identified as points that deviate significantly from the reference line in the Q-Q plot.

- **Model Validation**:
  - Confirming the normality of residuals supports the assumption of linear regression and the reliability of confidence intervals and hypothesis tests.



- If the points lie on or near the 45-degree reference line, the residuals are approximately normally distributed.
- If the points show a systematic pattern or curvature, the residuals deviate from normality.
- Points far from the reference line can be considered outliers.