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?

Answer:

We can infer following-

>> Year has good dependency on count of bookings, we only have 2 years of data but based on trend we can say that it has grown to a good extend in 2019 compared to 2018.
>> Bookings are impacted negatively in weather conditions of Mist+Cloudy, Light rain or Light snowfall.
We can say people mostly use this in clear weather, in other cases they might prefer taxi or stay at home.
>> Summer and winter season improves the count of bookings.
>> Months like January and July negatively impacts the overall bookings but this reverses in case of September which leads to more bookings (Highest in all months.)
>> Holiday impacts the bookings negatively, maybe because customer mostly uses these bikes as a commute for their work.

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

Answer:

We uses drop_first= True as in case of dummy variables if number of categories are n, they can be defined by n-1 dummy variables.
We do not need one of the variables as it is already defined by its other dummy variables.
This will increase co-linearity in independent variables and will un-necessarily make computation more complex.

Eg :	Green	Red	Blue
	1	0	0
	0	1	0
	1	0	0

Here 1 represent the colour of the ball, if we remove the first column “Green”, we can still interpret that Red - 0 and Blue – 0 will represent green colour.

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

Answer:

Temperature has highest correlation with booking count.

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

Answer:

To calculate validate the assumptions of linear regression on training set we did following:

We predicted the results for training set using the linear model we created, with that we calculated the error terms in our prediction.

Error = y_train – y_train_predicted

Once we got the error terms, we plotted following graphs.

1) Distribution plot/ histogram of error terms and found that they are creating a normal distribution with mean at zero, which helped us with the assumption that error terms are normally distributed with mean at zero.
2) We plotted the error terms against the y-train set and found that errors do not form a pattern and are evenly distributed across vertical axis, which confirmed They have constant variance.

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

Answer:

Temperature, Light Snow or light Rain and year arranged in descending order of their dependency.

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a supervised machine learning technique which used principle of fitting a line or a hyperplane based on given data to predict the output within the range of data points.

Linear regression assumes that dependent variable must be in linear relation with independent variable or predictor variables.

We provide the data and output to linear model to train it and it provides us the best possible constant and coefficient values.

y = b0 + b1X1 + b2x2 … bnXn

y = variable needed to be predicted
b0 = constant(y - intercept)
X1 to Xn = independent variables
b1 to bn = respective coefficients of X1 to Xn

linear regression achieves best possible values by minimizing RSS(Residual sum of squares).

Res = y – y_pred

RSS = $\sum_{i=1}$ (*Res*)^2

Res = residuals
y = actual value of output
y_pred = predicted value

Linear regression uses gradient descend algorithm to minimize the cost function (RSS), which uses a recursive technique that keeps on optimizing the values of coefficients until best possible RSS is achieved.

Further we can evaluate how well our model is able to define the variance of predicted variable using parameters like r square, adjusted, r square.

We can check P values and VIF to remove more features based on their significance and correlation with other independent variables.

2) Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a statistical example that helps us understand importance of visualization of data.

It consists of 4 dataset with identical statistical properties like mean, variance, regression line, correlation, but when visualised they revealed distinct patters.

1. **Linear relationship**: A straightforward linear relationship between x and y.
2. Non-linear relationship: A non-linear, curved relationship between x and y.
3. **Outlier**: A linear relationship with an outlier point that significantly affects the regression line.
4. **Non-linear, dependent y**: A non-linear relationship where y is dependent on x, but not vice versa.

It depicts that statistical measures are alone not sufficient to understand structure of data hence visualization is important.

3) What is Pearson's R?

Answer:

Pearson's R is a statistical measure that calculates the strength and direction of a linear relationship between two continuous variables. It's a value between -1 and 1, where 1 and -1 represents perfect positive and negative correlation respectively, 0 represents no correlation and value near to 0 weakens the correlation while near to 1 increases it.

Formula : $R = \Sigma[(xi - \bar{x})(yi - \bar{y})] / (\sqrt{\Sigma}(xi - \bar{x})^2 * \sqrt{\Sigma}(yi - \bar{y})^2)$

xi and yi are individual data points
$\bar{x}$ and $\bar{y}$ are the means of X and Y
Σ denotes the sum of the values

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

Answer:

Scaling can be defined as changing the scale or range of values of features in a way that they have common or similar range as others.

Scaling is an important step in data processing part because.

1. As different features might have different ranges, here just because a feature will have higher values it will become dominant in prediction.
   Eg: we have a dataset having columns income with range (1000 to 1000000) and age with range (10 - 100), here income variable will have more influence just because its range is greater.
2. Model have to go through more complex computation with unscaled features.
3. Interpretation of coefficients will be much better.

Standardize scaling:

It scales the data in a way that mean of the data shift to zero and data is distributed normally. It divides data by standard deviation to have unit variance.
Data can still have large range and it is sensitive to outliers.

Formula:-

$Z = (X - \mu) / \sigma$

Z = standardized value
X = original value
μ = mean of the dataset
σ = standard deviation of the dataset.

Minmaxscaling scaling:

It scales data between a range of 0-1, data is not shifted to zero mean, it preserves the distribution of data and it is not much affected by outliers.

Formula:-

Formula:

$X' = (X - min) / (max - min)$

X' = normalized value
X = original value
min = minimum value in the dataset
max = maximum value in the dataset

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

Answer:

Infinite VIF value can indicates:

1. **Perfect multicollinearity**: Two or more predictor variables in a regression model are perfectly correlated, making it impossible to estimate the regression coefficients.
2. **Singularity**: The correlation matrix is singular, meaning its determinant is zero, and the matrix is not invertible.

In practical terms, an infinite VIF value means that, Features are highly correlated, making it difficult to distinguish their individual effects.

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

Answer:

A Q-Q plot also known as Quantile-Quantile plot is a graphical tool used to compare the distribution of two datasets or a dataset and a theoretical distribution. In linear regression, Q-Q plots are essential for:

1. Verifying if the residuals (errors) follow a normal distribution, which is a crucial assumption in linear regression.
2. Detecting skewness, heavy tails, or other non-normal patterns in the residuals.

Here's how to read a Q-Q plot:

X-axis: Theoretical quantiles (expected values) from a normal distribution.
Y-axis: Observed quantiles (actual values) from the residuals.
Points: Each point represents a quantile (a percentage of data) from the residuals.
Line: A reference line (usually red) indicating perfect normality.

Points close to the line depicts that Residuals follow a normal distribution while points deviating from the line: Non-normality, indicating potential issues with the regression model.