Linear Regression

Assignment based Subjective Questions.

1. From your analysis of categorical variables from the dataset what could you infer about their effect on the dependent variable?
   - 2019 increase in number of bookings cnt compared for 2018.
   - Less booking seen with Weathersit-3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.
   - Highest bookings seen in Weathersit_1 ie. Clear, Few clouds, Partly cloudy, Partly cloudy
   - Gradual increase in bookings seen in 2nd and 3rd quarter for both years.
   - Highest bookings seen when there is Holiday(0) with median lying 4800.
   - The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
   - More bikes are rent during september month.
   - More bikes are rent during saturday

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

It helps in reducing the extra column created during dummy variable creation. Also to prevent multicollinearity, which occurs when two or more predictor variables are highly correlated

3. Looking at the pairplot among the numerical variables, which one has the highest correlation with the target variables.
   Temp is highest corelated with Cnt variable. We have removed atemp since its corelated with independent variable temp.

4. How did you validate the assumptions of linear regression after building the model on the training set?
   Linearity
   Mean of residuals was almost equal to zero ie. 4.504673758192974e-12
   R-square = Model: OLS Adj. R-squared: 0.837, which is good coverage.
   Prob (F-statistic): 2.01e-188 is almost zero implies that overall, the regression is meaningful.

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

   Here are the top three features:

   - Temp (positive corelation)
   - Year (positive corelation)
   - Weathersit_3 (negative corelation)

General Subjective Questions

1. Explain linear regression algorithm in detail.

Linear regression uses a linear equation to predict the value of a dependent variable based on one or more independent variables.
There are different types of linear regression.

1. Simple linear regression - Simple linear regression uses one independent variable to predict value of dependent variable.
2. multiple linear regression- it uses more than one independent variable to predict value of dependent variable.

Linear regression at each X finds the best estimate of Y.

The equation of best fit regression line $Y= \beta 0 + \beta\ 1X + E$ can be found by minimizing cost function (RSS In this case using ordinary least squares method) which is done using following two methods.

1. Differentiation.
2. Gradient Descent method.

$\beta\ 0$ and $\beta 1$ are two unknown constants representing the regression slope, whereas E(epsilon) is the error term.

In multiple linear regression analysis, the dataset contains one dependent variable and multiple independent variables. The linear regression line function changes to include more factors as follows:

$Y= \beta 0{*}X0 + \beta 1X1 + \beta\ 2X2+\ldots\ldots \beta\ nXn + E$

As the number of predictor variables increases, the β constants also increase correspondingly.

The strength of Linear regression model is mainly explained by R square.

R-square = 1 – (RSS/TSS)

Where R-square is how much variability in y we can explain through model.

RSS = Residual Sum of squares

TSS = Total Sum of squares

Linear regression divides the data into training set and test set. It then trains the data using algorithm which is then refined until model is built. Which is later applied on test dataset to evaluate the model.

Assumptions of linear regression model:

1. Linear relationship b/w X and y.
2. Error terms are normally distributed.
3. Error terms are independent of each other.
4. Error terms have constant variance.

Hypothesis testing in linear regression:

1. To determine the significance of beta coefficients.
2. H0: $\beta1 = 0$; HA : $\beta1$ not equal 0.
3. T test on the beta coefficient.
4. T score = $\beta cap/SE(Bcap1)$
5. Building Linear model.

   OLS: method in stats model to fit the line.

   Summary statistics:

      F(Statistic, R squared coefficients and their p values.)

Residual analysis: Residuals are used to measure prediction accuracy. A residual is the difference between the observed data and the predicted value.

Histogram of error terms to check normality.

Plot of the error terms with X or y to check independence.

Predictions:

Making predictions on the test set using predict().

2.Explain the Anscombe’s quartet in detail

Anscombe's quartet is a set of four datasets that have similar summary statistics, but appear very different when graphed. A set of four datasets with similar means, standard deviations, and correlation coefficients.

- Why it's important: Anscombe's quartet is used to demonstrate the importance of graphing data before analyzing it, and the effect of outliers on statistical properties.
- Who created it: Statistician Francis Anscombe created Anscombe's quartet in 1973.
- What it shows: Anscombe's quartet shows that summary metrics can be misleading, and that important trends can be missed if data isn't visualized and broken down.
- Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit.

3. What is pearson -R?

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**	**Example**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Baby length & weight: The longer the baby, the heavier their weight.

0	No corelation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

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

Feature scaling : Scaling or not to scale variables/features.

- Easy of interpretation.
- It is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Faster convergence for gradient decent method.

Co-efficients change when you scale variables.

Scaling methods:

1. Min Max Scaler

   sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

2. Standardization

   Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

Both methods don't change distribution of original variables. It just scales them.

Standardization = x= x- mean(x)/sd(x)

Min-max scaling = x = x- min(x)/max(x)-min(x)

We use Min Max Scaler in the bike dataset since it takes care of Outliers as well.

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

The Variance Inflation Factor (VIF) can be infinite when there is perfect correlation between independent variables. This can happen when one independent variable is strongly correlated with many other independent variables.

- A large VIF indicates a high degree of multicollinearity, which is when an independent variable is highly correlated with other independent variables in a regression equation. Multicollinearity can make statistical inferences less reliable.
- VIF and threshold

A VIF of 5 is often used as a threshold, meaning that any independent variable with a VIF greater than 5 should be removed from the dataset.

- VIF and orthogonal variables

When all independent variables are orthogonal to each other, the VIF is 1.0

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

A Q-Q plot, or quantile-quantile plot, is a graphical tool that compares two probability distributions by plotting their quantiles against each other. It can be used to assess if a data set is normally distributed, or to compare the shapes of distributions.

Q-Q plots show how the quantiles of two distributions line up. The x-axis usually shows the theoretical distribution, while the y-axis shows the model residuals.

If the two distributions are similar, the points on the plot will be close to the line y = x. If the distributions are linearly related, the points will be close to a line, but not necessarily on the line y = x.

Q-Q plots can detect shifts in location, scale, and symmetry, as well as the presence of outliers.

Q-Q plots can be used to assess normality in regression models by plotting the standardized residuals from the model.