Assignment: Linear Regression Subjective Questions

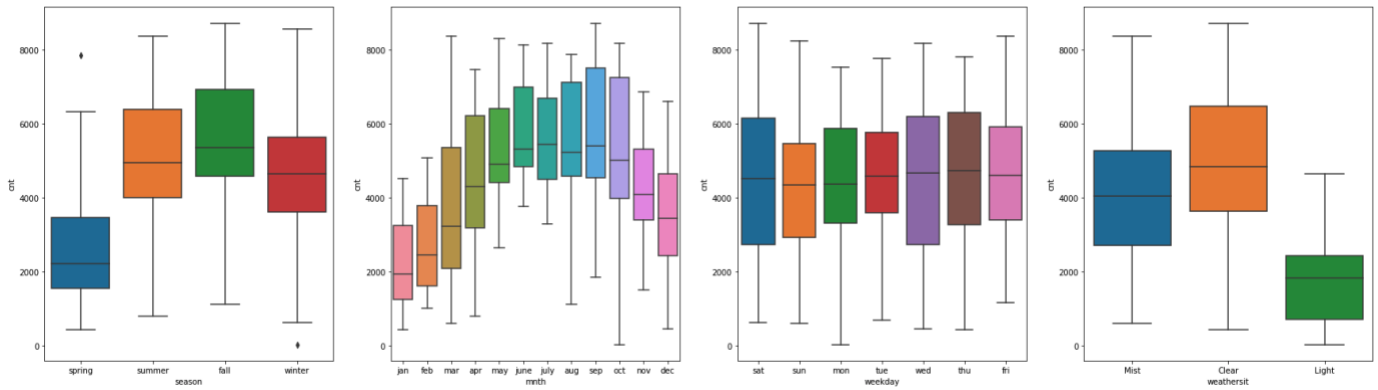
Assignment-based Subjective Question

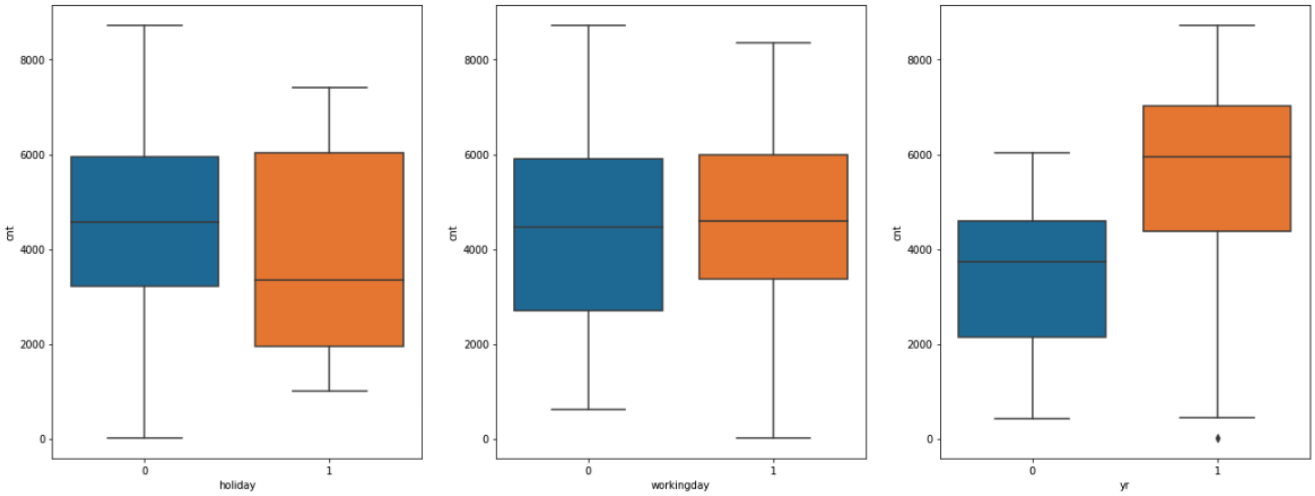
Question 1:

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

Answer:

Below is the result from boxplot from Seaborn (A box plot shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable)





- Fall season has highest demand for rental bikes
- September month has highest demand
- Demand is increasing till June and decreasing after September
- We cannot predict anything from weekday demand
- Clear weathershit has highest demand
- Booking is almost same for working and non-working days.
- Demand increased in 2019 in-comparison to 2018

Question 2:

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

Answer:

Dummy variable creation converts categorical variable into dummy/indicator variables.
drop_first=True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Creating dummy variables for categorical variable and drop the first columns as (p-1) dummies can explain p categories.

I used this for months, weekdays, season and weathersit.

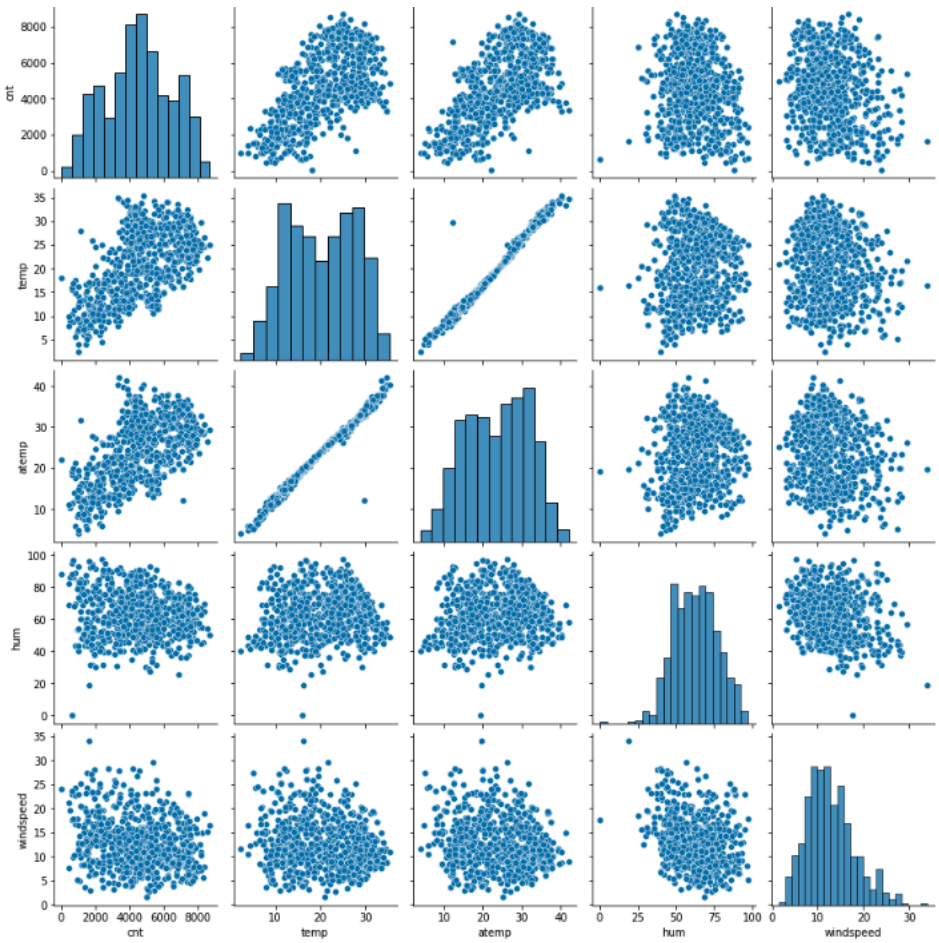
Example:

months_df=pd.get_dummies(bike_df.mnth,drop_first=True)

Question 3:

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

Answer:



We can see temp is has the highest correlation with the target variable cnt.

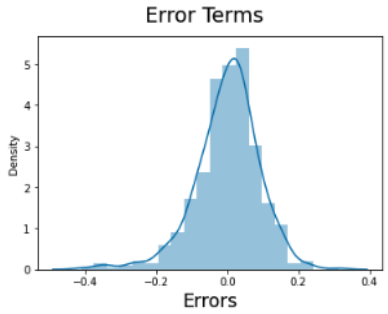
Question 4:

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

Answer:

I used below techniques to validate the assumptions of Liner Regression after building the model on the training set.

1. Absence of Multicollinearity (Multicollinearity refers to the fact that two or more independent variables are highly correlated (or even redundant in the extreme case). I used heatmap to check the same.
2. Homoscedasticity (Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable). I verified variance of the error terms is constant across the values of the dependent variable to check this.
3. Normality of Errors (Error term should be normally distributed).



4. Independence of residuals (absence of autocorrelation)
5. Linear Relationship

Question 5:

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

Answer:

Based on the final model, below are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

1. temp
2. sep
3. yr

General Subjective Questions

Question 1:

Explain the linear regression algorithm in detail.

Answer:

Linear Regression Algorithm is a machine learning algorithm based on supervised learning; it is a part of regression analysis. Regression analysis is a technique of predictive modeling that helps you to find out the relationship between Input and the target variable.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where,

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

Use Cases of Linear Regression:

- Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
- Price Prediction – Using regression to predict the change in price of stock or product.
- Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

Assumptions of simple linear regression are:

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

Question 2:

Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It illustrates the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. This tells us about the importance

of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.
The four datasets can be described as:
Dataset 1: this fits the linear regression model well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

Question 3:

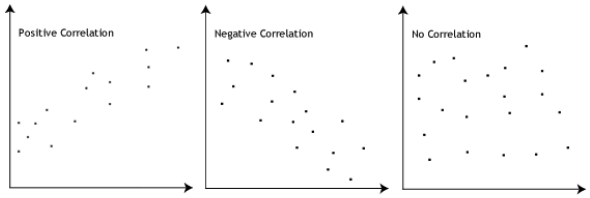
What is Pearson's R?

Answer:

Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association



Pearson r Formula

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

Here,

- r = correlation coefficient
- 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

Question 4:

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

Answer:

Scaling:
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why is scaling performed?
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Difference between normalized scaling and standardized scaling

Normalization	Standardization
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

Question 5:

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

Answer:

VIF infinite shows a perfect correlation between two independent variables

The value of VIF is calculated by the below formula:

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

Question 6:

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

Answer:

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

Usage:
QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

if we are testing if the distribution of age of employees in my team is normally distributed, we are comparing the quantiles of my team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

The Q-Q plot lets us compare how close two distributions are, and is often used to assess normality in linear regression.