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:-

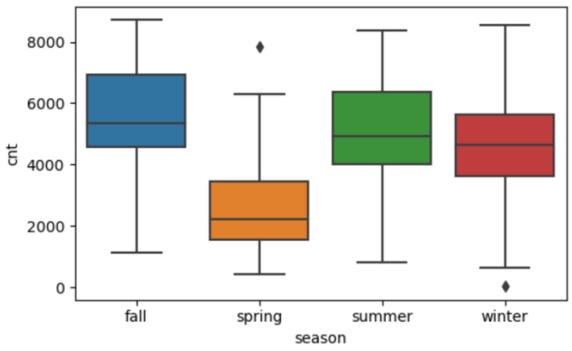
During my analysis, following variables were considered as categorical type:-

yr, season, mnth, weekday, weathersit, holiday, workingday

There effects can be summarized as below:-

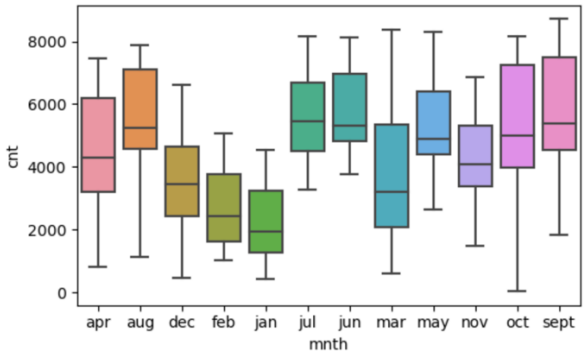
yr:- It can be seen that bike rental increased in the year 2019 compared to 2018. It seems that the bike rental service was gaining popularity among the commuters.

season:- Spring season has negative correlation with target variable. Season summer, and season fall shows positive correlation with the target variable. The same can be seen in the box plot:-

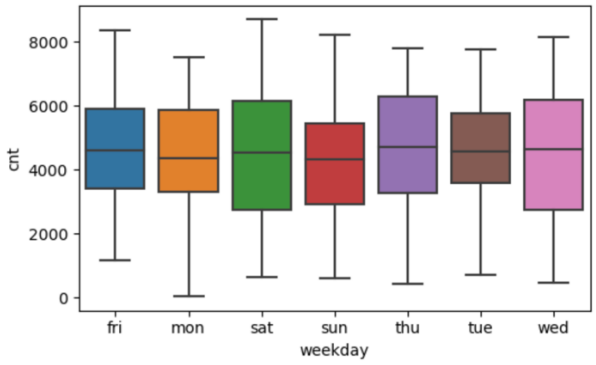


It is evident from the graph above that total rental was less in spring season, which confirms the negative correlation.

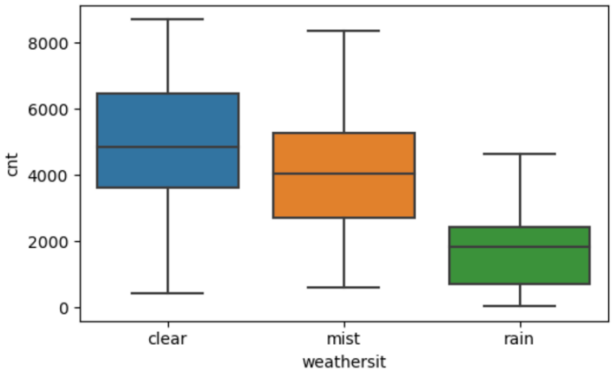
mnth:- A few months shows negative correlation and the box plot also represent the same information. Example, jun, jul, aug, sep have positive correlation and the box plot also shows higher rental events for these months. For the months where the negative correlation exist, the rental counts are also less.



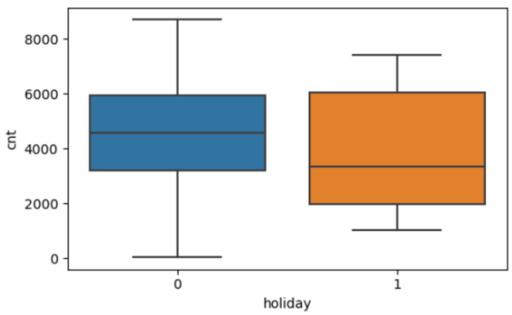
weekday:- Sunday has a small negative correlation of -0.059. Thursday has the maximum positive correlation number of 0.034. It can be seen that bike rental is more on weekdays compared to Sunday.

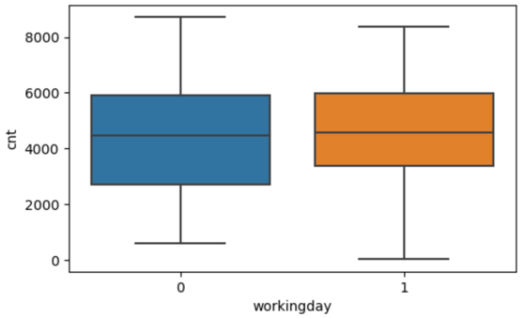


weathersit:- From the box plot below, it is clear that weather situation has impact on the target variable. Good weather situation promotes bike rental. As the weather situation deteriorates, the bike rental numbers also slip. Even the correlation values for mist and rain weather situation are -0.17 & -0.24 respectively, which confirm the pattern.



holiday & workingday :- Bike rental is more on workingday as compared to holiday. Positive correlation value for workingday also confirms the same. There is a negative correlation value for holiday. The values are 0.063 (for workingday) & -0.069 (for holiday).





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

Answer:-

When we create dummy variables, n dummy variables get created. These n dummy variables will have correlation among themselves which is known as multicollinearity. However, n states can be represented by n-1 variables. Hence, drop_first =True is used to drop the extra variable.

Example:- Suppose there is a dataset which has a column to store the following states of a food item:- Hot, Warm, Cold. If we don't use drop_first = True, 3 dummy variables will be created for each type of the values. However, we can easily represent the three values with just 2 variables. The three states can be thus represented using two variables as shown in table below:-

State	Hot	Warm
Hot	1	0
Warm	0	1
Cold	0	0

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

Answer:-

temp & atemp have the highest positive correlation with the target variable.

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

Answer:-

To validate the linear regression model's assumptions, I checked below things:-

1. Performed Residual analysis and verified that the distribution is a normal distribution, centered around 0.
2. R-squared and Adjusted R-squared values were found to be high nearly 80%.
3. p-values of predictors were less than a threshold (in our case 0.05), and the VIF values were less than 5.
4. Besides, F-statistics was high and Prob (F-statistic) was 0 or near zero.
5. Performed predictions using the model and drew scatter plot between the predicted values and actual values (on test data). For the model to be good, the values were expected to be in a close range (which also was the case).
6. Calculated R-squared and Adjusted R-squared values on test data and those values were similar to the values obtained on training data.

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

Answer:-

1. Temperature (temp):- temp had high positive coefficient.
2. Good weather(weathersit_clear):- good weather had a positive impact on bike demand.
3. Windspeed (windspeed):- windspeed had negative impact on bike's demand.

Fall and summer had higher bike demand, and that may be due to the good weather condition in those seasons.

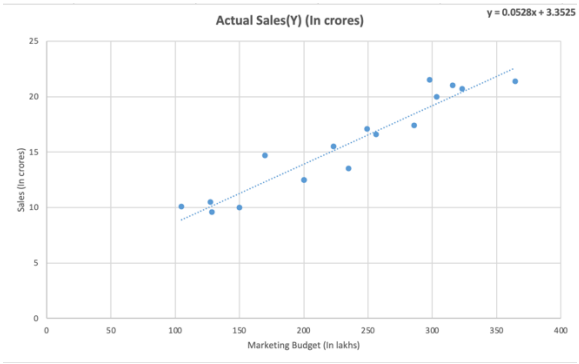
Bike rental has seen growth in 2019 compared to 2018 and it can only be assumed that after the normalcy gets restored, the bike rental will improve as the trend shows that people are increasingly liking the idea of renting/sharing bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

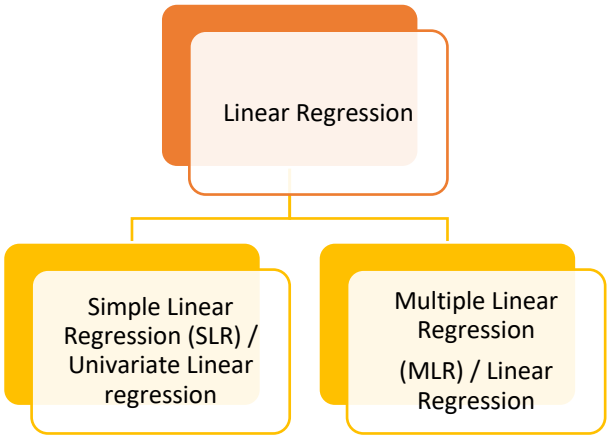
Answer:-

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of continuous numeric values. This is based on the popular equation "y = mx + c" that is also called equation of straight line. Linear regression is useful when there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).



[pic:- example linear regression graph]

Linear Regression is broadly divided into :



Simple / Univariate Linear Regression:-

This is used when the dependent variable is predicted using only one independent variable. In this regression technique, we try to find out a relationship between a dependent variable (Y) and one independent (X).

Its mathematical equation is given as:

$$Y = \beta_0 + \beta_1 * x$$

where

- Y is the response or the target variable
- x is the independent feature
- β_1 is the coefficient of x
- β_1 is the intercept

Multiple / Multivariate Linear Regression: -

This is used when the dependent variable is predicted using multiple independent variables. Linear regression algorithm uses independent variables to model a goal prediction value. The equation for multiple linear regression is similar to the equation for a simple linear equation, i.e., $Y = \beta_0 + \beta_1 * x$ plus the additional weights and inputs for the different features which are represented by $\beta_n * n$. The formula for multiple linear regression would look like,

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta p * Xp + \epsilon$$

2. Explain the Anscombe's quartet in detail.

Answer:-

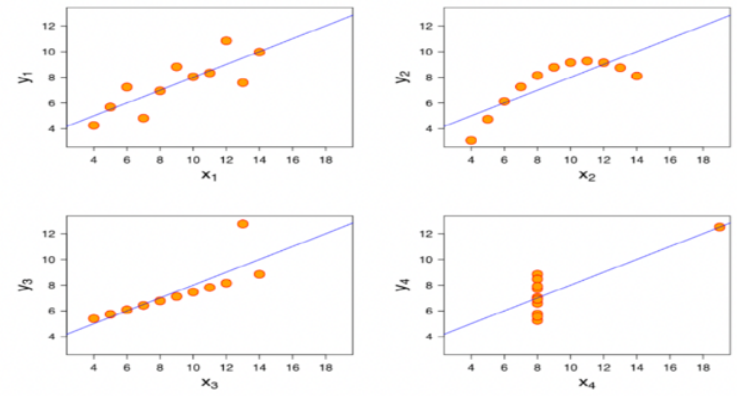
Anscombe's quartet emphasizes on the need of visualizing the data, critically examine any and all the assumptions, and deploy various analytical tools to reach a fair conclusion. It states that we shall not put our weight behind summary statistics and look beyond what is visible.

Anscombe's quartet's dataset (source: Wikipedia) :-

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The 4 data sets created by Anscombe would produce nearly identical mean, standard deviation, and regression line. But when they were plotted, they appeared as below:-



Observation

- The first plot (top left) shows that there is a linear relationship between x & y.
- Top right data set does not show any linear relationship between X and Y.
- The bottom left graph shows some outliers which could not be explained by the linear regression model.
- The last data set has a high leverage point which is enough to produce a high correlation coefficient.

3. What is Pearson's R?

Answer:-

Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's.
Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all

The latter indicating a perfectly positive and linear correlation and the former indicating a perfectly linear negative regression. The values in between denotes the relative collinearity of two variables.

Interpretation:

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.

- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

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

Answer:-

Feature scaling is the process of normalising the range of features in a dataset. It is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step. We need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

For example, if we have multiple independent variables like age, salary, and height; with their range as (18–60 Years), (25,000–75,000 Rupees), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range,

Scaling is necessary for a model to be functional with the appropriate range of coefficients.

There are two types of scaling:

1. Normalized scaling:

Normalization is good to use when the distribution of data does not follow a Gaussian distribution. The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. Normalization is affected by the presence of outliers, so they must be removed from the data before it can be applied.

2. Standardized scaling:

Standardization can be helpful in cases where the data follows a Gaussian distribution. Though this does not have to be necessarily true. Since standardization does not have a bounding range, so, even if there are outliers in the data, they will not be affected by standardization.

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

Answer:-

In the limit, when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity. This can be easily understood, if we examine the formula of VIF:-

$$VIF_i = 1 / 1 - R^2$$

This equation can have infinite value if and only if $1 - R^2 = 0$. R^2 of 1 is possible only when the predictions are identical to the observed values, that is, the two values are perfectly correlated.

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. It is a graphical tool to assess if sets of data come from the same statistical distribution.

Q-Q plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behaviour.

Importance and benefits of Q-Q plot:

- Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, so we don't need to care about the dimensions of values.