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?

Ans. The variable like season, yr(year), mnth(month), holiday and weathersit(weather situation) have a direct dependent over the target variable.

Like:

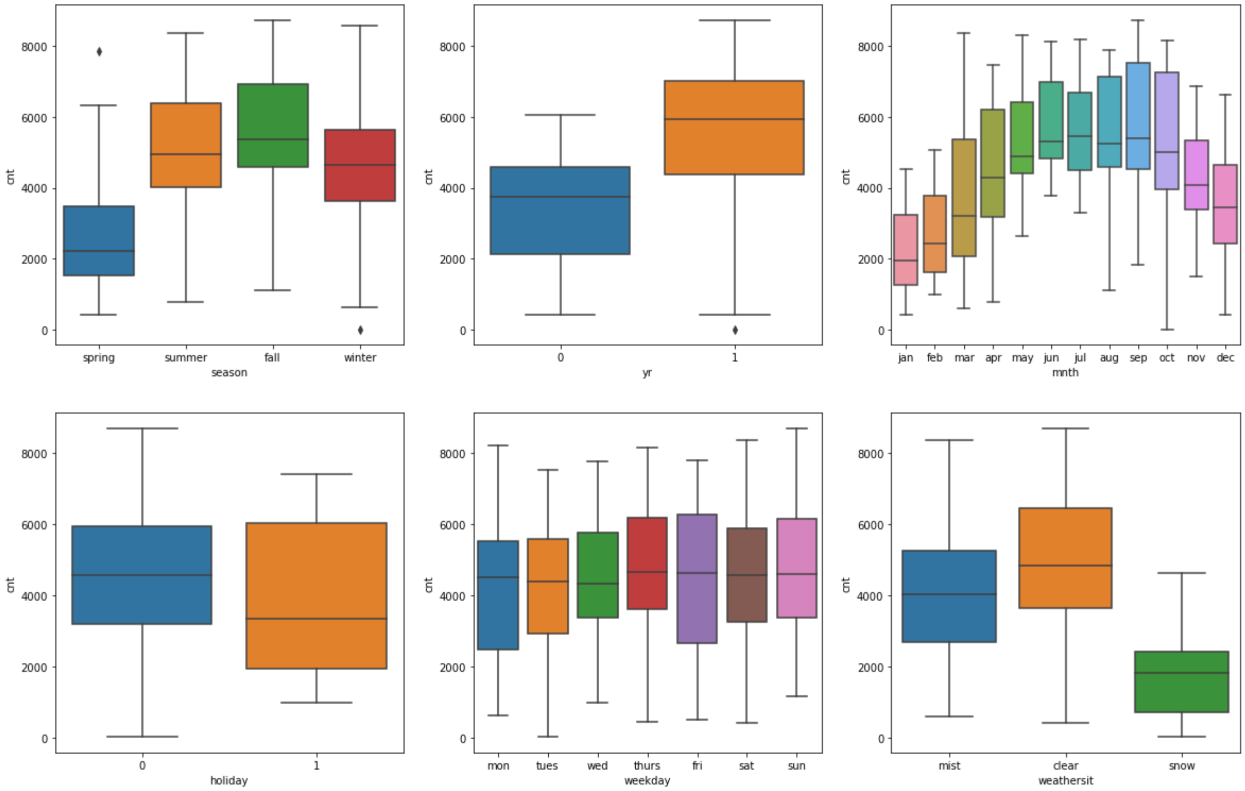
- In year 2019 the their was a hike in sale.

- There was a drop in cnt in the season of spring(whatever the reasons).

- If we observe the variable mnth we can see a hike from the month of may (starting of summer holidays) till the month of sept.

- people usually don’t prefer to take a bike ride in snow weather.

- On a holiday there was drop in the demand of bikes.



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

Ans. Drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

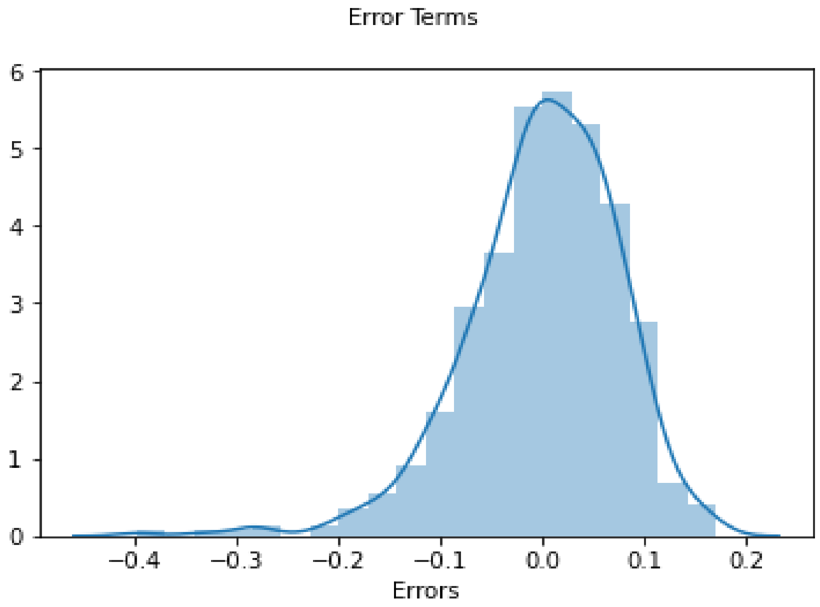
Example: We created dummy variables for seasons, there are 4 seasons so the code pd.get_dummies() created 4 dummy variables but in practice 3 dummies would be sufficient to represent seasons. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Ans. Registered variable has the highest correlation with target variable(cnt).

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

Ans. By performing Residual Analysis on the Training Set, i.e. I drew a histogram of error terms, they were normally distributed with mean=0. Therefore Validating our assumption.



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

Casual: People usually use bikes for renting casual.

Snow: Snowy weather is a top featuring variable affecting our model as in snowy weather people doesn't prefer using bikes.

Year (yr) : In the year 2019 the demand for boom bikes have increased significantly, which makes it third most important variable.

General Subjective Questions.

1. Explain the linear regression algorithm in detail.

Ans. Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

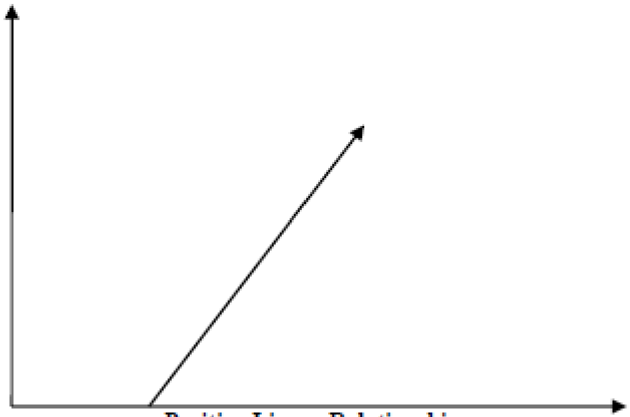
Y=mX+bY=mX+b

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the YY-intercept. If X = 0,Y would be equal to bb.



Positive Linear Relationship

Furthermore, the linear relationship can be positive or negative in nature as explained below −

Positive Linear Relationship

A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −

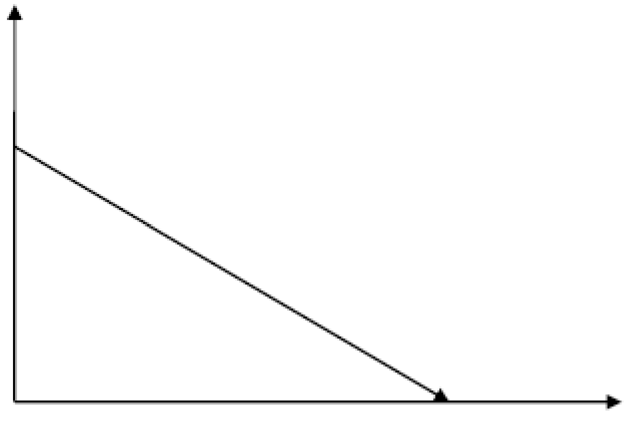
Negative Linear relationship

A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –

Types of Linear Regression

Linear regression is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression



Negative Linear Relationship

Assumptions

The following are some assumptions about dataset that is made by Linear Regression model –

Multi-collinearity – Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.

2. Explain the Anscombe's quartet in detail.

Ans . Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."[1]

3. What is Pearson’s R?

Ans. Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatter plot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.

Pearson’s r measures degree of correlation or correlation coefficient between 2 numerical variables.

Its value varies between -1 and 1.

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

0<r<5 means there is a weak association

5< r< 8 means there is a moderate association

8 < r means there is a strong association

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

Ans.

<u>What?</u>

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.

<u>Why?</u>

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. …

Normalization vs. standardization

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.
However, at the end of the day, the choice of using normalization or standardization will depend on your problem and the machine learning algorithm you are using. There is no hard and fast rule to tell you when to normalize or standardize your data. You can always start by fitting your model to raw, normalized and standardized data and compare the performance for best results.

It is a good practice to fit the scaler on the training data and then use it to transform the testing data. This would avoid any data leakage during the model testing process. Also, the scaling of target values is generally not required.

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

Ans. An infinite VIF value indicates that the dependent variable may be expressed exactly by a linear combination of other variables. VIF = 1/ (1-R2), when R2=1 then VIF = Infinity Example: In our Assignment, Registered Users + Casual Users = Total no. of Users If we fit the model including these 2 variables then VIF will be infinity because of this.

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

Ans. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. this helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Uses of Q-Q plot

a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior