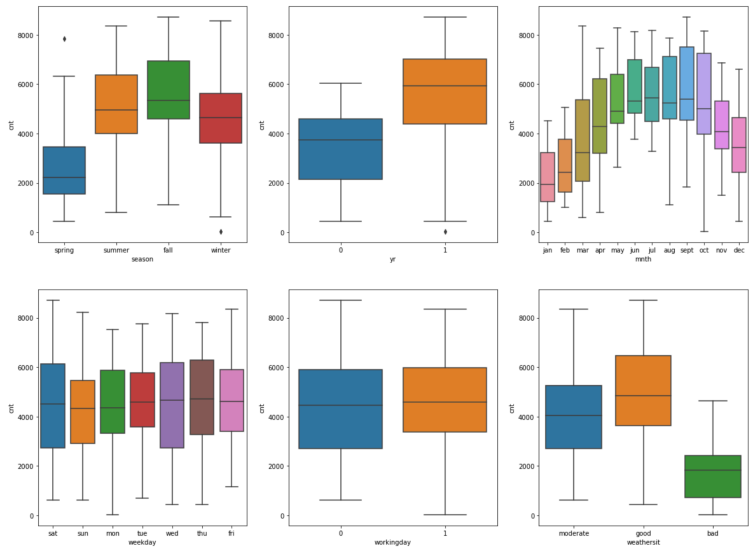
Assignment-based Subjective Questions

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

Answer : Following are the 6 categorical variables provided in the dataset that have a major effect on dependent variable “cnt”:

a) season
b) mnth
c) yr
d) weekday
e) working day
f) weathersit

Here is an inference derived using the box plot :



Mostly season, weather situation and month shows a good influence on cat with median of over 5000 bookings which indicates they are good predictors. However, weekday may not be a good predictor based on the plot which we will have to let the model decide.

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

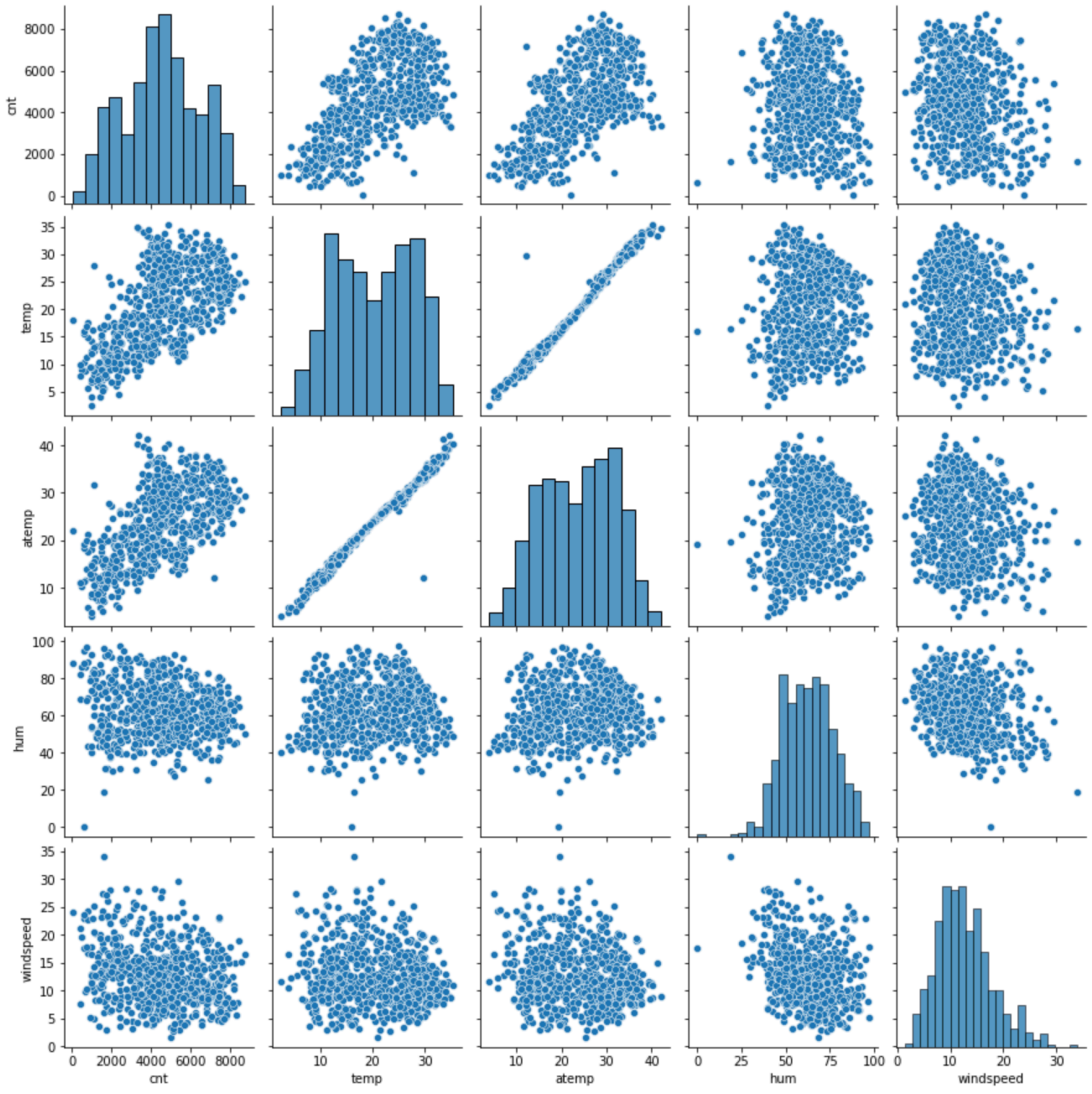
Answer : It is important to use drop_first=True because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let’s assume we have three columns in a category i.e. Female, Male and Others. Here, we do not need 3rd variable to identify the Others to avoid redundant correlation.

Therefore when there is a categorical variable with n-levels, we can 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?

Answer : I observed that there is the highest correlation between the variables ‘temp’ and ‘atemp’ variables as compared to the rest with target variable as ‘cnt’ as per the following observed pair plot :



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

Answer : We could validate the Linear Regression models on the basis of following :

- Linearity
- No auto-correlation
- Normality of error
- Multicollinearity

The F-Statistics value > 1 i.e. F-statistic = 180.2 and the p-value ~ 0 indicates that the model is significant.

Also the VIF calculation indicates that there is no multicollinearity existing in predictor variables as its below 5 almost.

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 final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are temperature, year and season.

General Subjective Questions

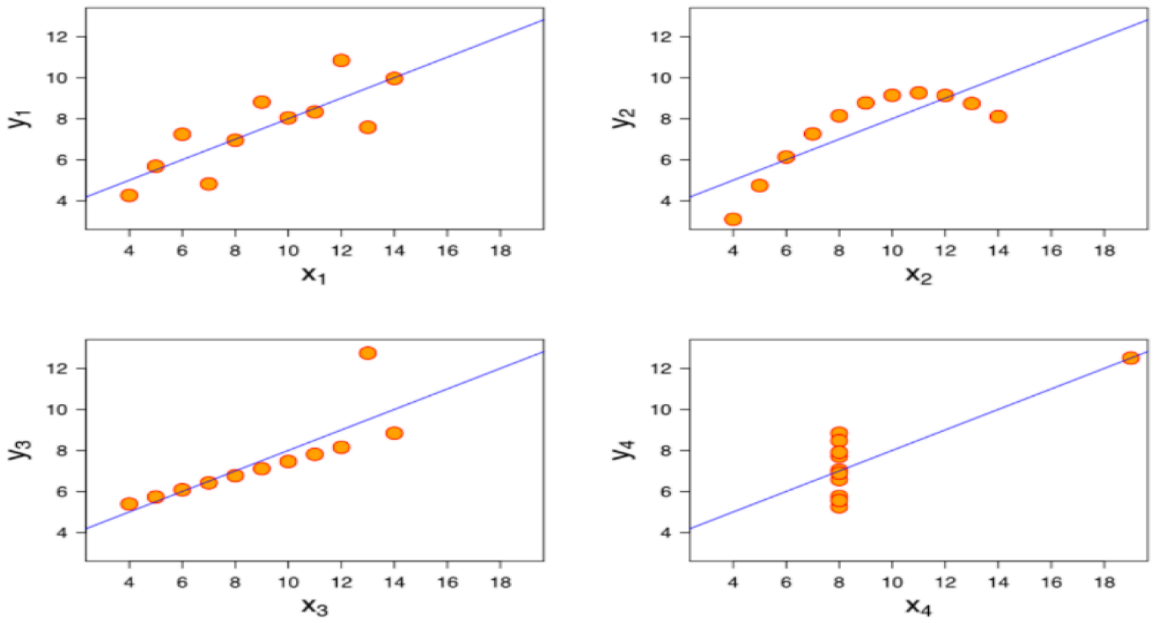
1. Explain the linear regression algorithm in detail.

Answer : Linear regression is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds

how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.
A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error.
In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

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 was constructed to illustrate 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.



- 1 st data set fits linear regression model as it seems to be linear relationship between X and y
- 2 nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.
- 3 rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
- 4 th data set has a high leverage point means it produces a high correlation coeff.

Its conclusion is that regression algorithms can be fooled so, it's important to data visualisation before build machine learning model.

3. What is Pearson's R?

Answer : In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a

statistic that measures the linear correlation between two variables.

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
The figure below shows some data sets and their correlation coefficients. The first data set has an r=0.996, the second has an r = -0.999 and the third has an r= -0.233

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

Answer : Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.
Difference between Normalizing Scaling and Standardize Scaling:
1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don’t know about the distribution whereas standardized scaling is used when distribution is normal.
6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer : VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:
A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.
A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity

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

Answer : Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression :

In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also
- 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

Q-Q plot use on two datasets to check

- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape
- If both datasets have tail behavior