Assignment-based Subjective Questions

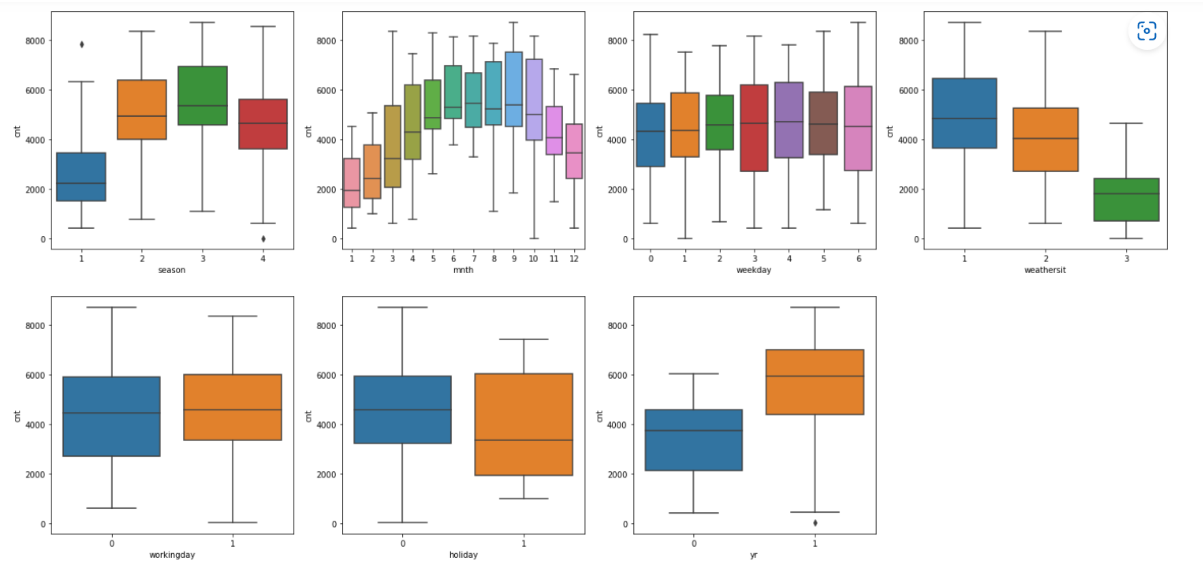
Question 1:

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

Answer:

As per our observation during the prediction of model, we have observed that independent categorial variables are highly correlated with the target dependent variable.

For better understanding we can look at the below box plots of categorial variable to see the correlations



Question 2:

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

Answer:

For categorical variable with ‘n’ levels, we can create ‘n-1’ new columns having 0 and 1 as an indicator to identify whether it exist or not.

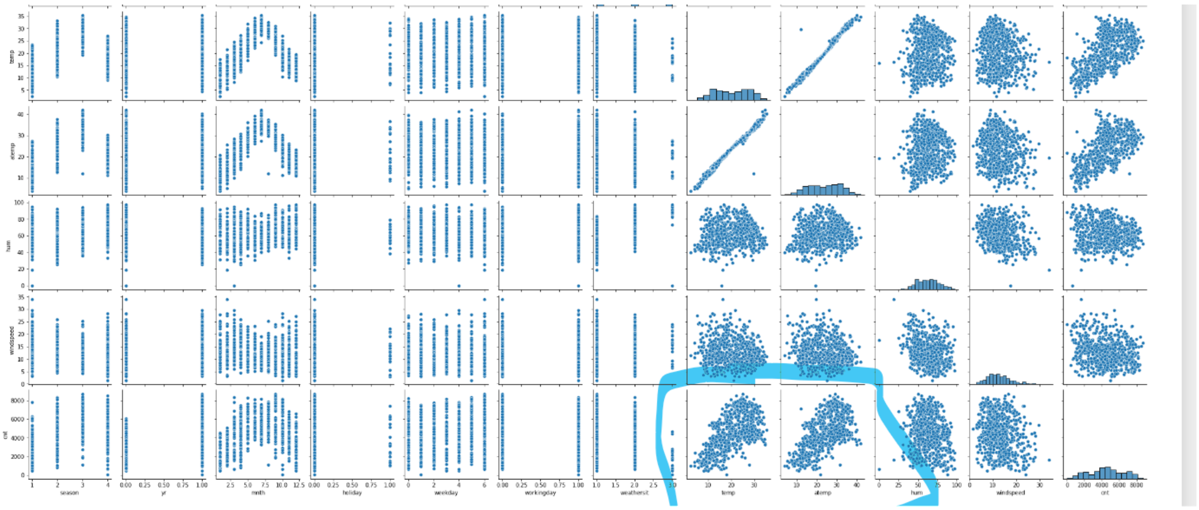
To achieve this, we are using drop_first=True, so that the resultant can match up n-1 levels and correlation can be reduced among the dummy variables

Question 3:

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

Answer:

As per the below pair-plot, we have observed that **temp** and **atemp** are highly correlated with target dependent variable **cnt**.

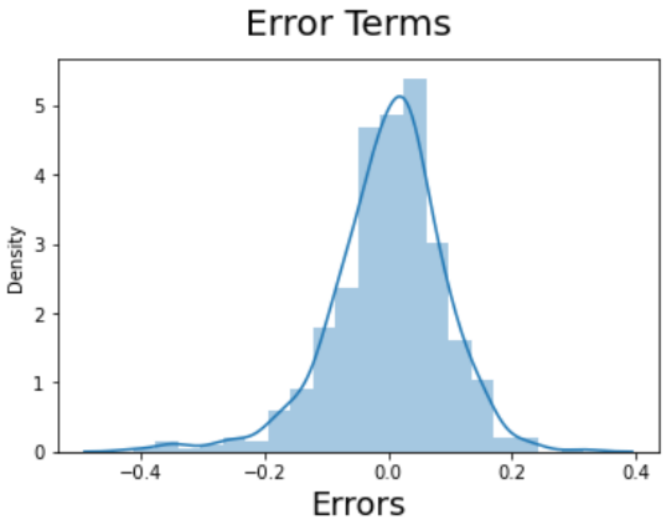


Question 4:

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

Answer:

1. We have found out the linear relationship between the dependent target variable and independent variable using the pair-plot and took the best one to predict the model using linear regression
2. Multicollinearity occurs when two independent variables are highly correlated with each other so we have calculated the VIF to drop the columns to reduce the correlations
3. Residual should follow the normal distribution and centred around mean = 0, as seen in below Error Term histogram.



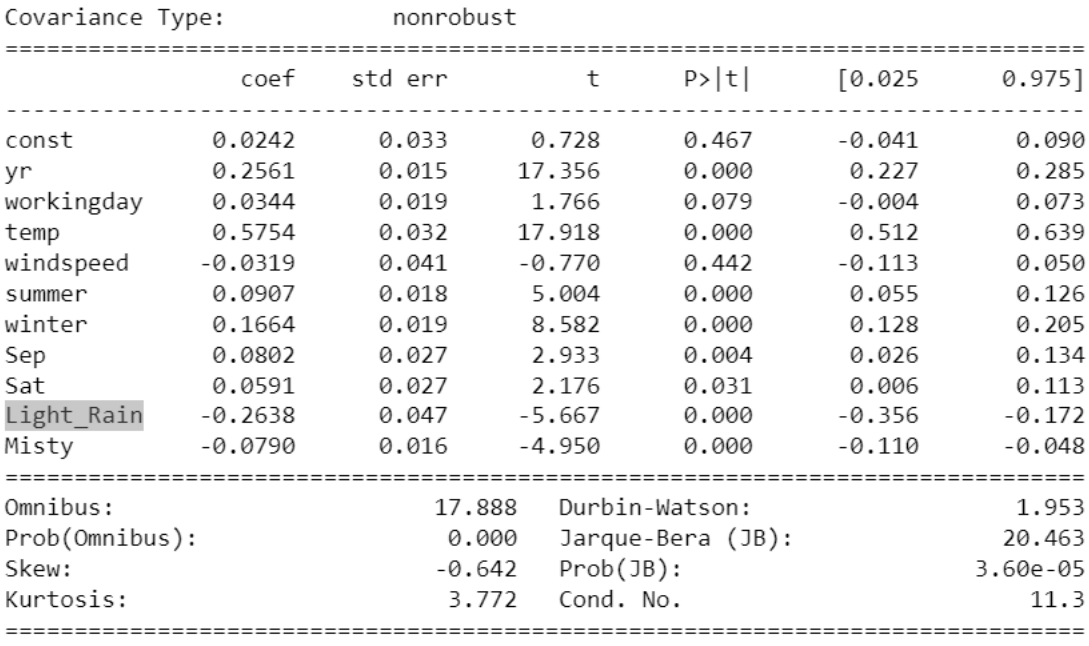
Question 5:

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

Answer:

As per below screenshot, top 3 features which are contributing significantly towards explaining the demand of the shared bikes are:

1. temp
2. yr
3. light_Rain (season)



```
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0242      0.033      0.728      0.467      -0.041       0.090
yr             0.2561      0.015     17.356      0.000       0.227       0.285
workingday     0.0344      0.019      1.766      0.079      -0.004       0.073
temp           0.5754      0.032     17.918      0.000       0.512       0.639
windspeed     -0.0319      0.041     -0.770      0.442      -0.113       0.050
summer         0.0907      0.018      5.004      0.000       0.055       0.126
winter         0.1664      0.019      8.582      0.000       0.128       0.205
Sep            0.0802      0.027      2.933      0.004       0.026       0.134
Sat            0.0591      0.027      2.176      0.031       0.006       0.113
Light_Rain    -0.2638      0.047     -5.667      0.000      -0.356      -0.172
Misty         -0.0790      0.016     -4.950      0.000      -0.110      -0.048
==============================================================================
Omnibus:                       17.888   Durbin-Watson:                   1.953
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               20.463
Skew:                          -0.642   Prob(JB):                     3.60e-05
Kurtosis:                       3.772   Cond. No.                         11.3
==============================================================================
```

General Subjective Questions

Question 1:

Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting

Below is the formula for Linear Regression Equation:

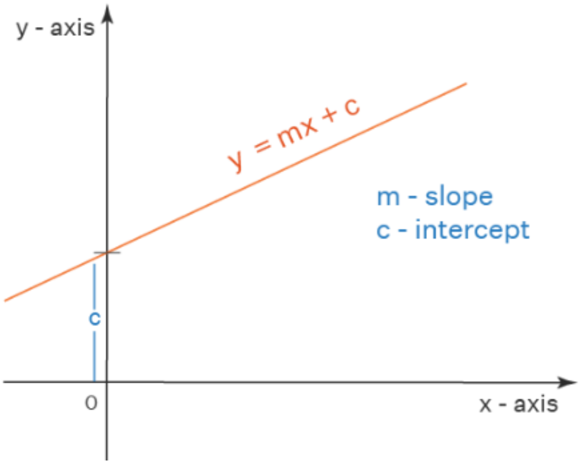
$$y = mx + c$$

Where, y is target variable

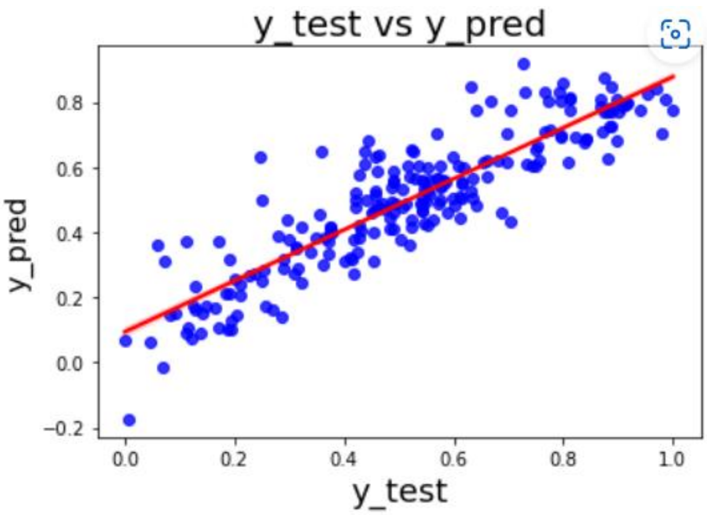
x is independent variable

m is slope of the line

c is intercept



We have also predicted below model in our assignment



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 was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing 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 visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order 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 be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

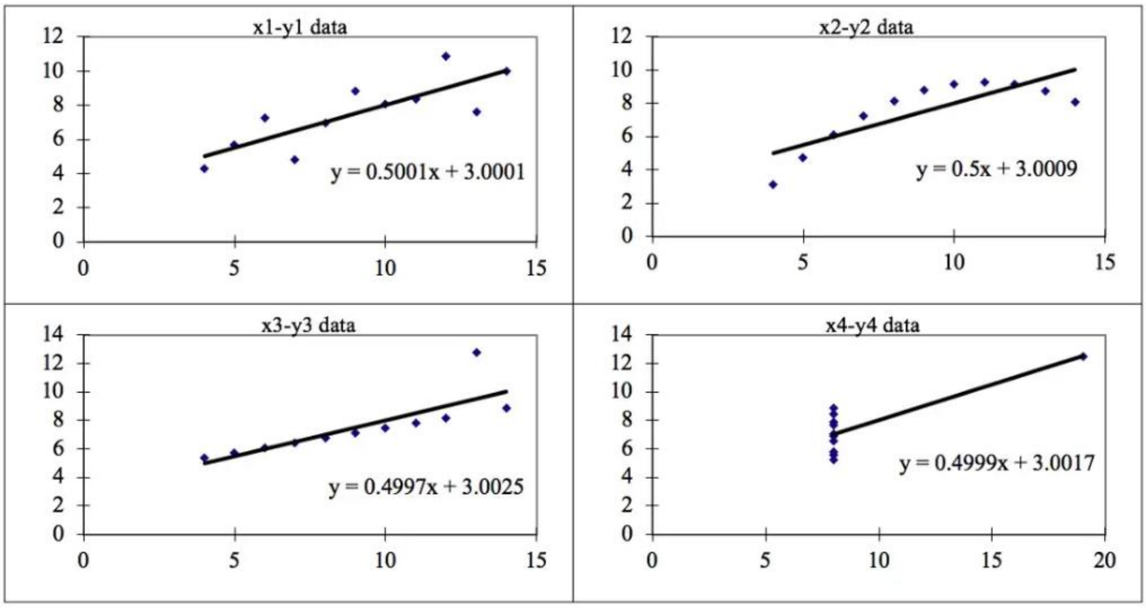


Image by Author

The four datasets can be described as:

1. Dataset 1: this fits the linear regression model pretty well.
2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

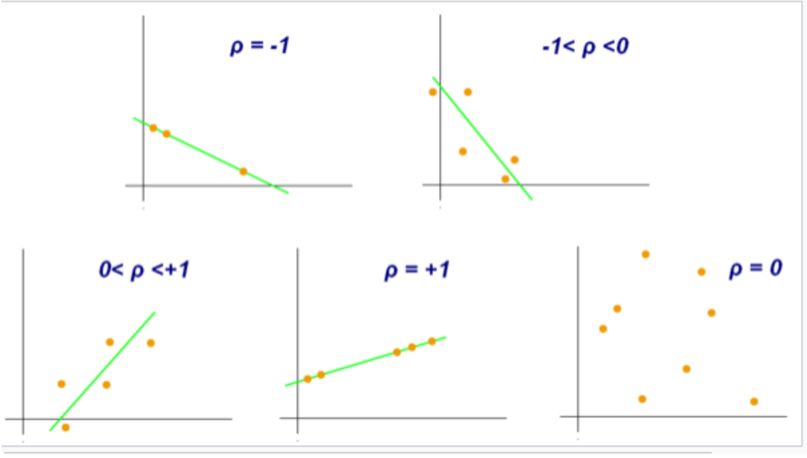
Question 3:

What is Pearson's R?

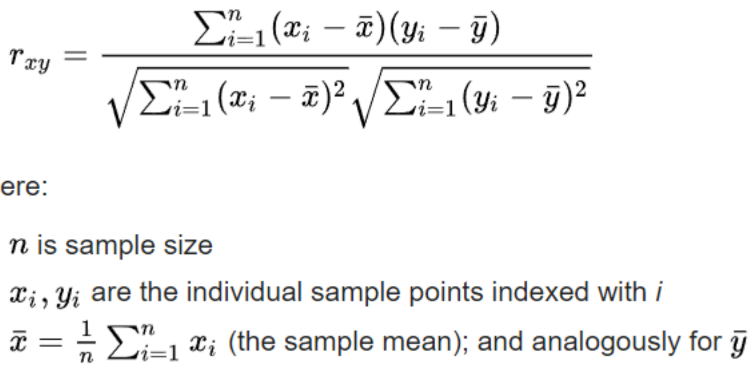
Answer:

In statistics, the Pearson correlation coefficient also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and 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. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation)

Examples of scatter diagrams with different values of correlation coefficient



Formula for calculating the coefficient:



$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Question 4:

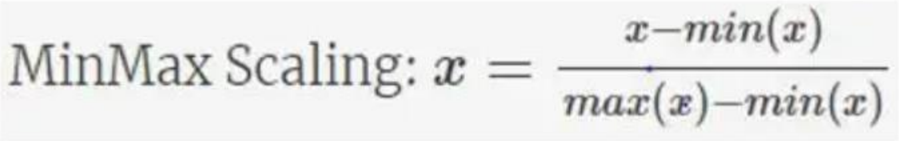
What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling

Answer:

Scaling 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.

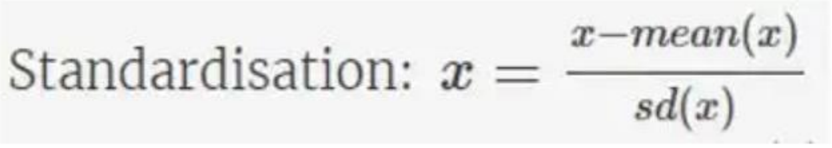
Scaling is performed because most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the 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/Min-Max Scaling brings all of the data in the range of 0 and 1



MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

whereas, 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). One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.



Standardisation: $x = \frac{x - mean(x)}{sd(x)}$

Question 5:

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

Answer:

If there is perfect correlation, then VIF = infinity. This 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.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Question 6:

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

Answer:

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. Before we dive into the Q-Q plot, let’s discuss some of the probability distributions.

The power of Q-Q plots lies in their ability to summarize any distribution visually.

QQ plots is very useful to determine

1. If two populations are of the same distribution
2. 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.
3. Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line