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:
Based on the the analysis of the categorical variables, the following can be inferred

- The Fall season had the highest booking.
- The year 2019 saw more bookings than the year 2018 which showed an increase in business.
- The months of June, July, August and September had the highest booking.
- The bookings were mostly evenly distributed across the days, except Sundays where the bookings were a little lesser than other days.
- The bookings were usually high on a ‘clear’ weather day and decreased substantially on a rainy day.
- The bookings were a little lesser during a holiday compared to the non-holiday days.

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

Answer:
The main purpose of setting the parameter *drop_first* = *True* is to avoid the extra column created during dummy variable creation. This further helps in reducing the correlations created among dummy variables

e.g..If a variable has 3 levels say A, B, C. While generating dummy variables for this categorical column, if the drop_first parameter is set, we just get 2 columns and if the values of both these columns are 0, we already know that the category is set to 3rd level.

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

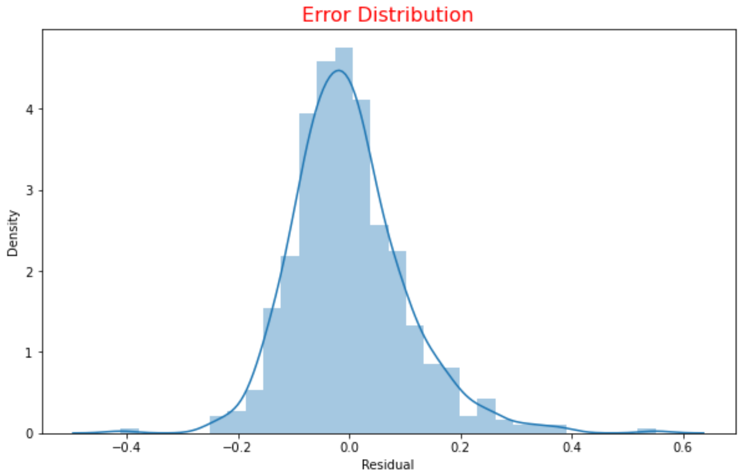
Answer:
The *‘temp’* and *‘atemp’* variables had the highest correlation with the target variable as ‘cnt’.

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

Answer:

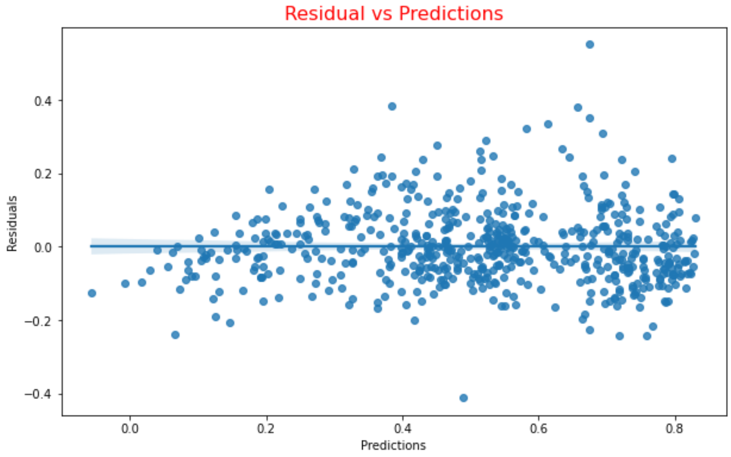
I have validated the assumption of Linear Regression Model based on the following assumptions

- Error terms should be normally distributed



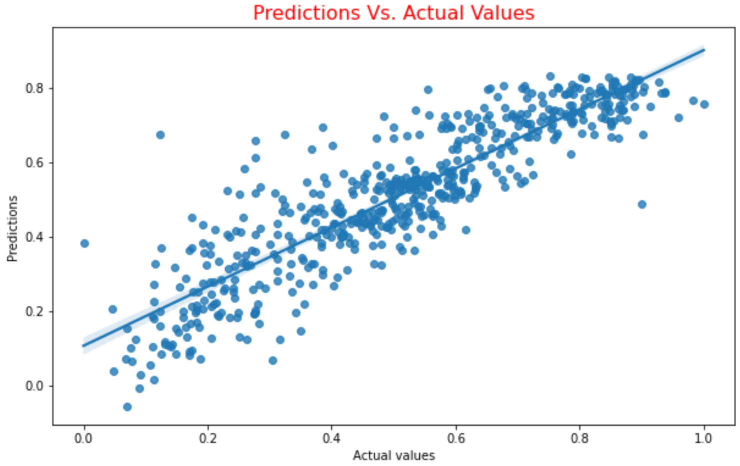
Error Distribution

- Error tems should be independent and there should be no specific pattern observed.



Residual vs Predictions

- Homoscedasticity - Error terms should be equally distributed and there should be a homogeneity of variance



Predictions Vs. Actual Values

Question 5

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

Answer:

The top 3 features that explained the demand of the shared bikes the most are in respective order

- Year (positive)
- Light rain (negative)
- Spring (negative)

General Subjective Questions

Question 1

Explain the linear regression algorithm in detail

Answer:

Linear regression can be defined as the statistical model that analyses the linear relationship between a dependent variable with a given set of independent variable/variables. Linear relationship between variables means a change in the independent variables affects the dependent variables which can be explained by a linear equation.

Thus, the relationship can also be represented with the help of the following equation *Y = mX + c.*

Here, *Y* is the dependent variable whereas *X* is the independent variable that is used to predict. *m* is the slope of the line, also called the *coefficient*. *c* is a constant, also known as the *Y-intercept*.

This linear relationship can also be positive or negative in nature as explained below

- Positive Linear Relationship: A linear relationship is called positive if an increase in a dependent variable also increases the independent variable. The slope in such a relationship is positive.

- Negative Linear relationship: A linear relationship is called negative if an increase in dependent variable causes the independent variable to decrease. The slope in such a relationship is negative.

Further, Linear regression can be divided into 2 types based on the dependent variables

- Simple Linear Regression : One dependent and one independent variable.
  e.g.. *Y = mX + c*

- Multiple Linear Regression: Multiple dependent and one independent variable.
  e.g.. *Y = aX1 + bX2 + cX3 + k*
  Note: In this case, the regression line is infact a hyperplane

However, in order to come up with a linear regression model, the following are some assumptions made about the datasets

- Normality of error terms i.e.. the error terms should be normally distributed

- Independence of error terms i.e.. the error terms should be independent of each other

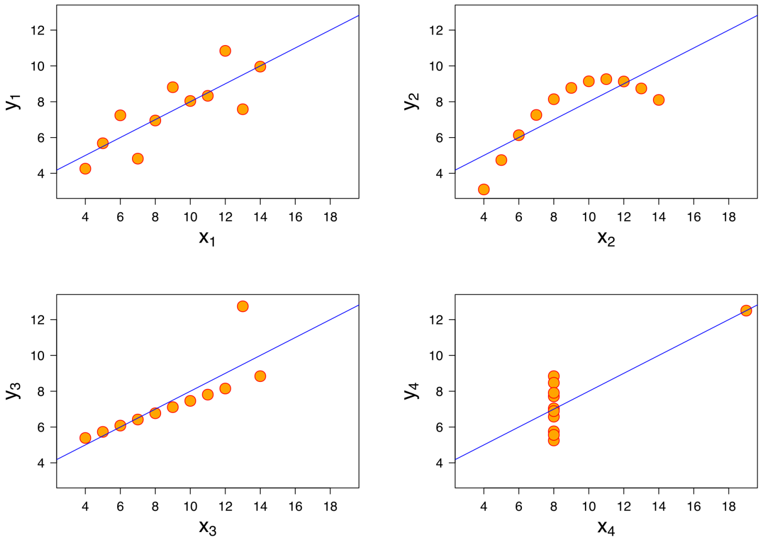
- Homoscedasticity i.e.. there is a homogeneity of the variance of error terms.

Question 2

Explain the Anscombe's quartet in detail.

Answer

Anscombe's Quartet may be defined as a group of four data sets which are nearly identical in simple descriptive statistics but can be very different when plotted.Each of these datasets consists of eleven (x,y) points. The importance of these datasets is to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.



- The first data set(top-left) fits linear regression model as it seems to be linear relationship between x and y
- The second data set(top-right) does not show a linear relationship between x and y, which means it does not fit the linear regression model.
- The third data set(bottom left) shows some outliers present in the dataset which can't be handled by a linear regression model.
- The fourth data set(bottom right) has a high leverage point meaning it produces a high correlation coeff.

Finally, it can be said that regression algorithms can be confused, so it's important to do data visualisation before building a machine learning model.

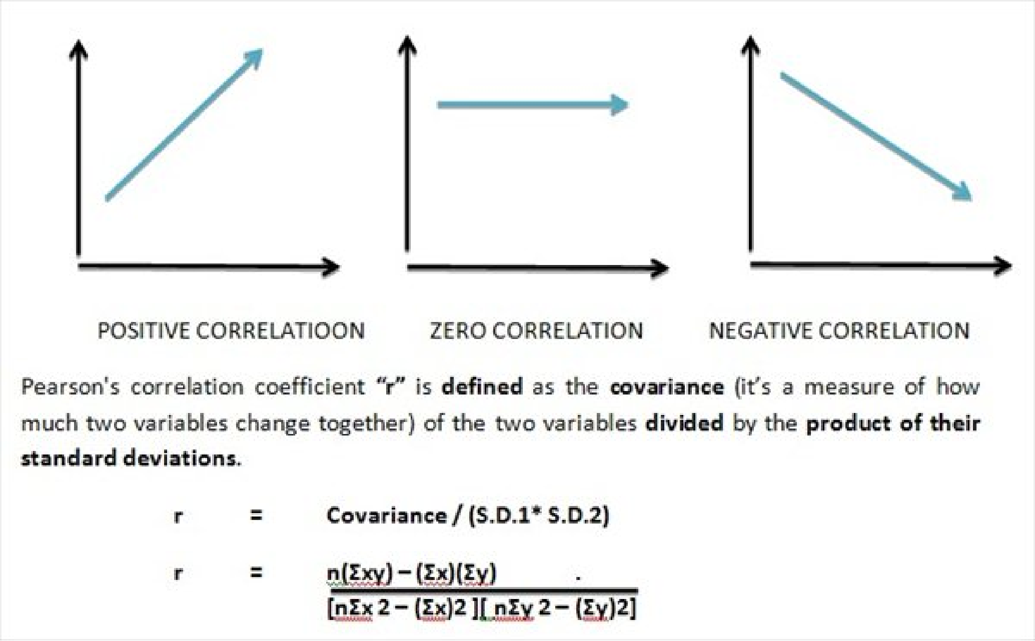
Question 3

What is Pearson’s R?

Answer

The Pearson correlation coefficient or Pearson’s R is a descriptive statistic, meaning that it summarises the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.



Pearson's correlation coefficient **“r”** is **defined** as the **covariance** (it’s a measure of how much two variables change together) of the two variables **divided** by the **product of their standard deviations.**

$$r = \text{Covariance} / (\text{S.D.1} * \text{S.D.2})$$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y) \quad .}{[n\Sigma x\,2 - (\Sigma x)2\,][\,n\Sigma y\,2 - (\Sigma y)2]}$$

Question 4

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

Answer:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed on the datasets to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.Thus, scaling can make a difference between a weak machine learning model and a better one.

The most common techniques of feature scaling are Normalization and Standardization.

A normalized scaling uses minimum and maximum values of the features to come up with a scaling formula. The values of a normalized scaling usually lie between [0,1] or [-1,1]. It can be affected by outliers.

Standardized scaling uses Mean and standard deviation of the features to come up with a scaling formula. The values are generally not bound to a certain range and this type of scaling is usually less affected by outliers.

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

Answer:
Infinite VIF means that there is a case of perfect correlation. For a general statement, any VIF more than 10 can be considered really high.

VIF = 1 / (1-R2)

Mathematically, the value of VIF is infinite for a R-Squared value 1. This again confirms a perfect correlation. In order to solve this, the feature with this high correlation must be dropped.

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

Answer:
A Q–Q(Quantile-Quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Thus, it helps to assess if a set of data possibly came from a theoretical distribution like Normal or Uniform distribution. For a similar distribution, the Q-Q tends to be linear.

Importance of QQ Plot in Linear Regression:
For two different samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If the two samples differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than regular analytical methods.