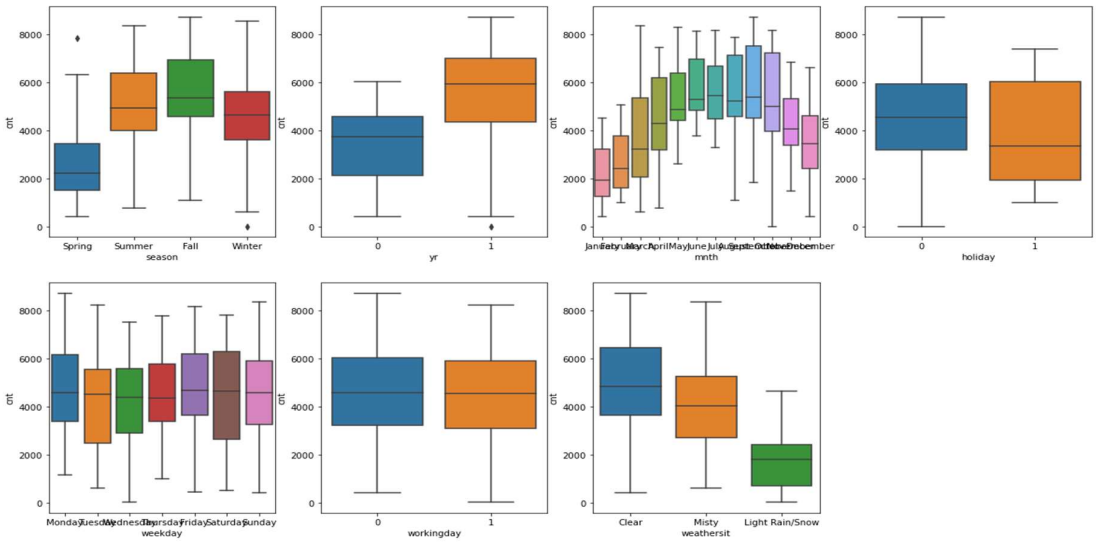
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



Answer:

From the boxplots shown above comparing various categorical variables with the dependant variable, cnt, we can clearly see the following:

- For season variable – There is higher count value for Summer & Fall followed by Winter and Spring being the last. Between Summer and Fall, Fall season gives the highest value of cnt.
- For year variable – Between 2018 & 2019, there is higher value of cnt in 2019.
- For month variable – The maximum ridership is being observed for the months from May to October with the highest being in September which also marks the start of the Fall season.
- For holiday variable – We can see that the ridership does not vary much whether it is a holiday or not.
- For weekday variable – As seen from the boxplot, there is not much variation between the days of the week with the mean being almost same for all days.
- For working day variable - During weekends and holidays the ridership is more as compared to working days.
- For weathersit variable – It can be clearly seen that the highest ridership is during clear weather followed by mist/cloudy weather and very less during light rain/snow. There is no ridership during extreme weather (heavy rain + thunderstorm)

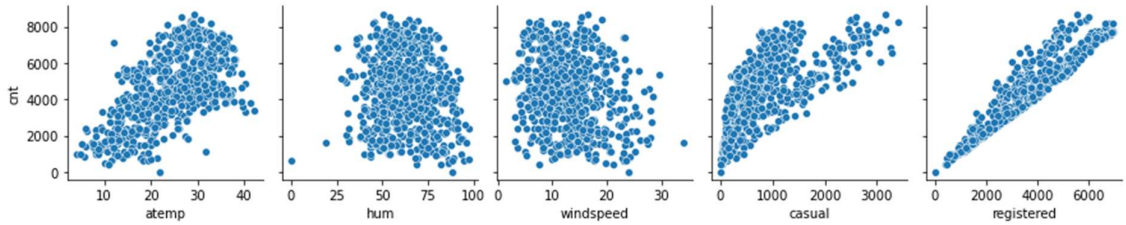
Q2. Why is it important to use **drop_first=True** during dummy variable creation?

Answer:

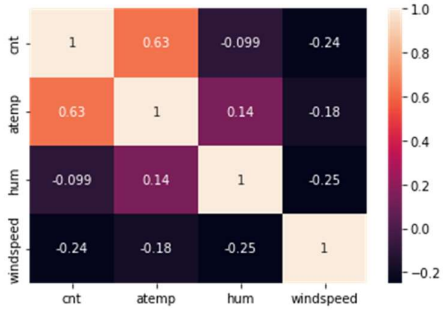
When we use pd.get_dummies() functions it creates binary vectors for the column, one each for unique values of the column represented by 1 or 0. The drop_first=True is used so we can reduce the first column which in turn helps reducing the number of columns. More dummy features make it harder for the algorithm to fit or even worse make it easier to overfit.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:



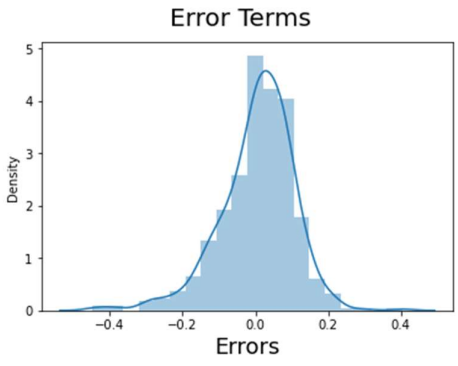
Looking at the above pairplot we can clearly observe that registered and casual have the highest correlation with the target variable but these are not taken into account while creating model since they are linealy associated with the target variable (cnt = casual + registered). Hence when we see other variables we see that atemp is having the highest correlation with the target variable (cnt). This is further proved when we plot the heatmap of the correlation matrix.

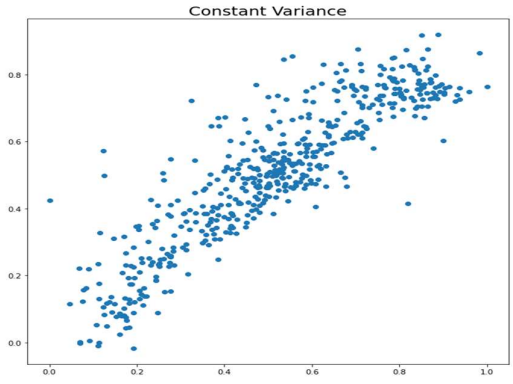


Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

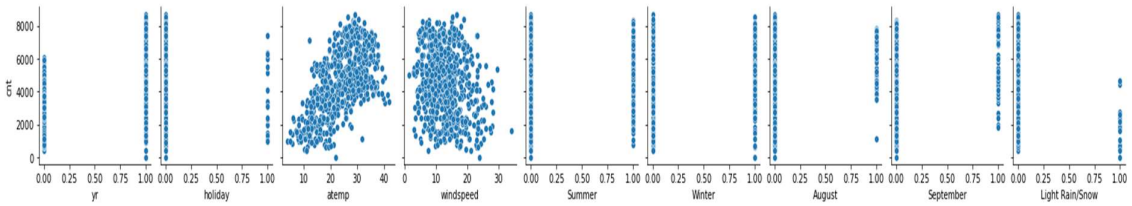
The most significant assumption of Linear Regression is that the error terms are normally distributed and another assumption is that the error terms have constant variance (homoscedasticity). This is clearly proved after looking at the following graphs





Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:



Looking at the above pairplot and also the VIF values, atemp, windspeed and summer are the top three features contributing to the demand of shared bikes.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Answer:

A linear regression model attempts to explain the relationship between a dependant variable and independent variable using a straight line. The independent variable is also known as predictor variable and the dependant variable is known as output variable. From a statistical point of view, in linear regression at each 'x' there is a distribution on the values for 'y'. Model predicts a single value therefore there is distribution of error terms.

For linear regression, cost function also known as Mean Squared Error (MSE), is used as an algorithm to reach an optimal solution. The physical significance of R-squared is higher the value of R-squared is very good for the model. The datapoints vary along a straight line, as the fit becomes poorer R-squared approaches zero. As R-squared approaches zero, the linear model's quality becomes poorer. R-squared value tells you how much variance in the data has been explained.

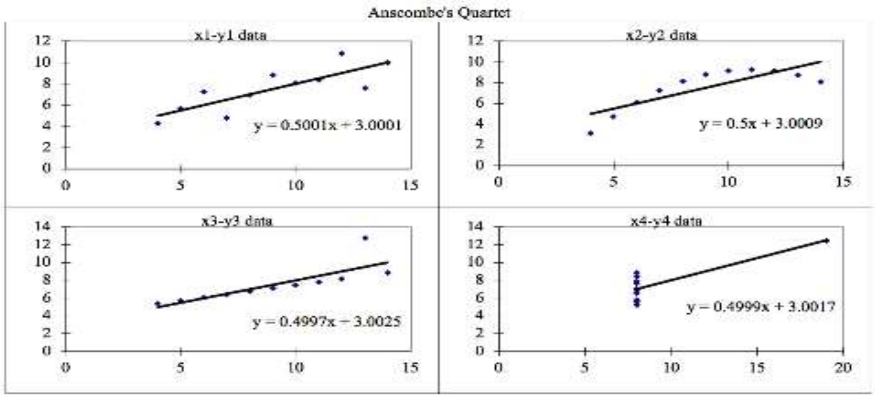
Assumptions of linear regression are:

- There is a linear relation between x and y → x & y should display some sort of linear relationship, otherwise there is no use of fitting a linear model between them.
- Error terms are normally distributed → If the error terms are not normally distributed then the p-values obtained during the hypothesis test become unreliable.
- Error terms are independent of each other → Error terms have to be independent of each other unlike a time-series data where each value is dependant on the previous one.
- Error terms have a constant variance (homoscedasticity) → The variance should not increase or decrease as the error values change. Also it should not follow any pattern as the error terms change.

Q2. Explain the Anscombe's quartet in detail

Answer:

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. It highlights the importance of plotting data to confirm the validity of the model fit. In each panel, the Pearson correlation between the x and y values is the same, r = . 816. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values. 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. These datasets were intentionally created so to prove the importance of plotting the data since regression models can be fooled, before implementation of any machine learning program data visualization is important so as to obtain a good fit.



Q3. What is Pearson's R ?

Answer:

Pearson's Correlation coefficient, R, is the measurement of the strength of the relationship between two variables & their association with each other, in other words it calculates the effect of change in one variable when the other variable changes (the strength and the direction of the relationship). It seeks to draw a line through the data of two variables to show their relationship. This linear relationship can be positive or negative. Pearson coefficient correlation r, determines the strength of the linear relationship between two variables. The stronger the association between the two variables, the closer your answer will incline towards 1 or -1. Attaining values of 1 or -1 signify that all the data points are plotted on the straight line of 'best fit.' The strength and direction mentioned earlier signify two things:

- Strength signifies the relationship correlation between two variables. It means how consistently one variable will change due to the change in the other. Values that are close to +1 or -1 indicate a strong relationship.
- The direction of the line indicates a positive linear or negative linear relationship between variables. If the line has an upward slope, the variables have a positive relationship. A negative correlation depicts a downward slope.

In Python, the default correlation method is Pearson's R. If we want to use other correlation methods like 'Kendall' or 'Spearman', then we have to specify e.g. df.corr(method='Kendall')

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

All the datasets have many features having a wide range of values and there is huge difference in the range of these features. It becomes difficult to do the comparison between the features because of this difference in the range, this difference will create a different step size for each feature. Machine Learning algorithms use Euclidean distance as a metric to calculate the similarities which fail to give reasonable recognition. This is the

reason why scaling is important so that the data is changed in such a way that we can process the model without any problems.

Neural networks expect the data to be between 0 & 1. Hence, we apply the scaling to only those columns which do not have 'yes/no' & dummy variables. These already have values that are either 0 or 1.

There are two primary scaling techniques used. The first is standard scaling (StandardScaler()) and is calculated by subtracting the mean and dividing by the standard deviation. The second is min-max scaling (MinMaxScaler()) and is calculated by subtracting by the minimum value and dividing by the difference between the maximum and minimum values.

The difference between normalized scaling and standardized scaling is that normalization of data get highly affected by the presence of outliers, whereas there is no such effect in standardization.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

Infinite values for VIF indicate that one independent variable is explained completely by all other independent variable, it also means that variable is completely multicollinear with other variables (R-squared=1)
(1/(1-R-squared)) → (1/0) → ∞

This is the reason why we have dropped casual and registered columns from our model, since they are perfectly linearly correlated to cnt variable.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

The plot obtained when the quantiles of two variables are plotted against each other is known as a quantile-quantile plot or a Q-Q plot. This plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations. Since normal distribution is of so much importance, we need to check if the collected data is normal or not, the Q-Q plot checks the normality of skewness of data. This plot represents the z-scores of standard normal distribution along x-axis and corresponding z-scores of the obtained data on the y-axis.