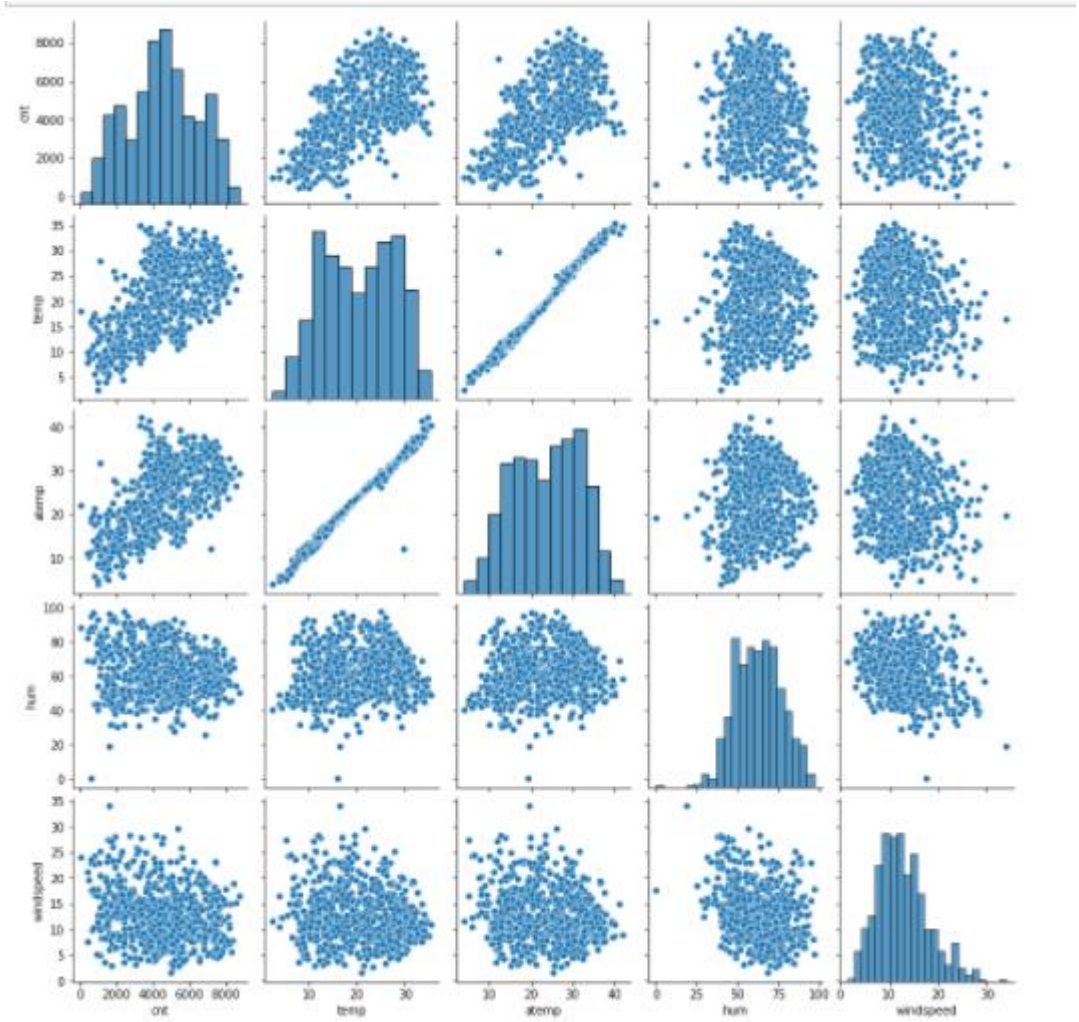
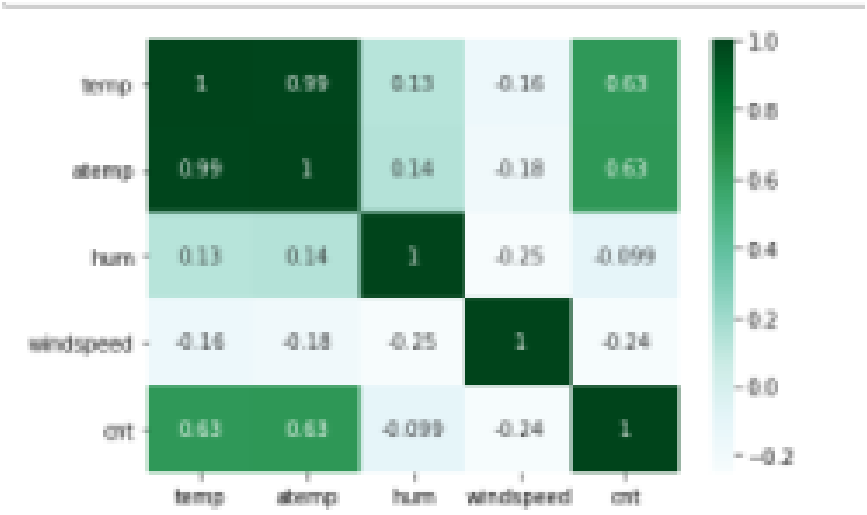
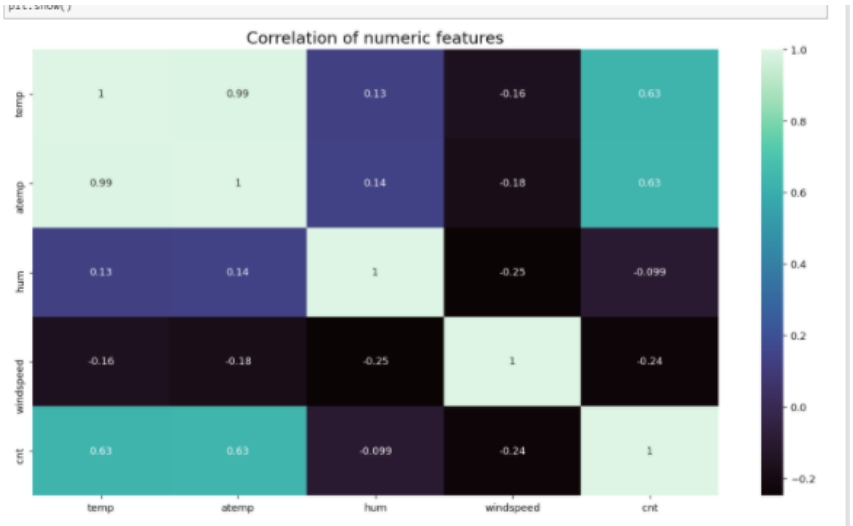
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    a. The categorical variables were season, weather, holiday, mnth, yr and weekday and were visualised. With the following influences
        i. Season – fall had the highest count vs spring, which had the lowest. Winter and summer were moderate
        ii. Highest count in clear day when compared to winter and rain
        iii. Holiday reduced the demand and count
        iv. The rentals increased from 2018 to 2019
2. Why is it important to use drop_first=True during dummy variable creation?
    a. If drops are not done for the first column then the variables will be correlated. This could have adverse effect on inferences and models as the effect is better when the data is pruned to smaller dataset.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    a.

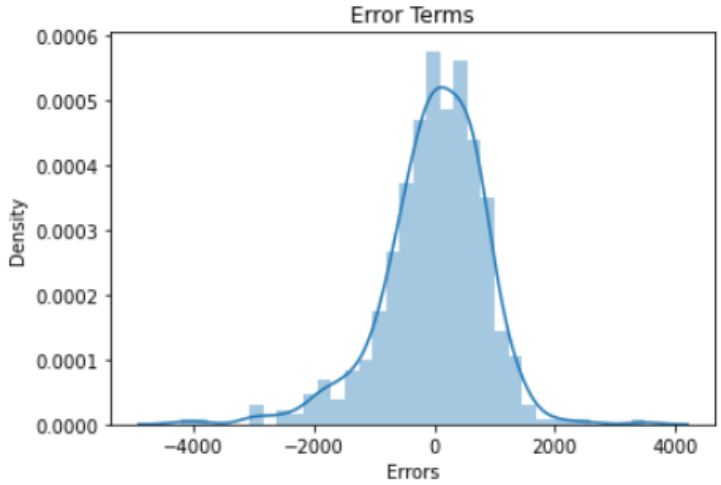






According to heatmap and pairplots temp and atemp are highly correlated

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



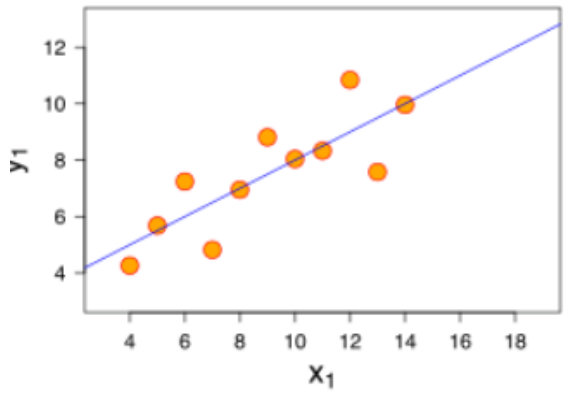
The graph looks normally distributed

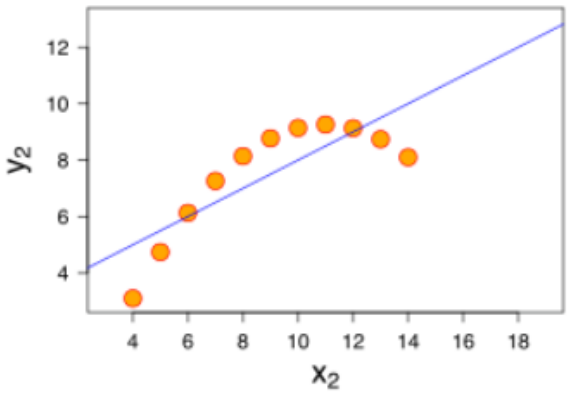
The above diagram shows that residuals are distributed normally therefore mean = 0. This helps in validation and assumption

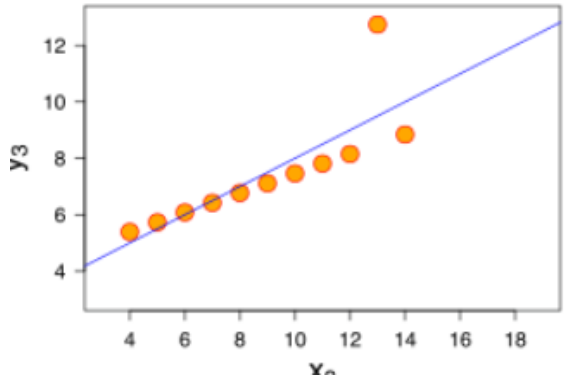
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. The temp, year and weather

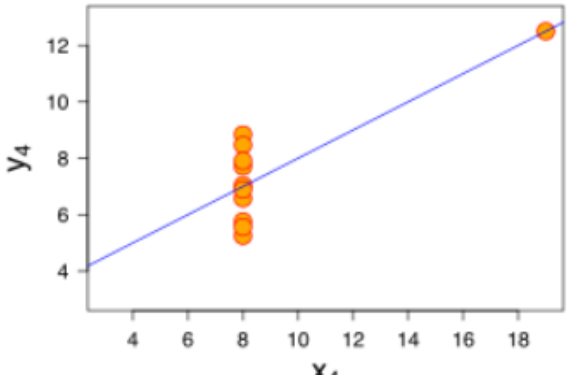
General Subjective Questions

1. Explain the linear regression algorithm in detail.
    a. Linear Regression is a type of supervised Machine Learning model. It si the basic regression model with equation y= mx+c. Assumes that there is linear relationship b/w dependent variables. Regression is divided in 2 SLR and MLR
2. Explain the Anscombe’s quartet in detail.
    a. 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.









3. What is Pearson's R?
    a. 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
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    a. Feature scaling is to normalize and standardize
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    a. Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    a.

    The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
    A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.