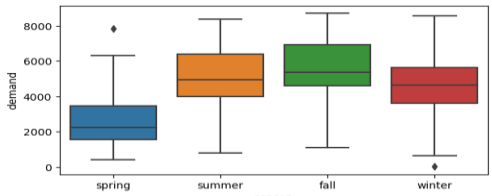
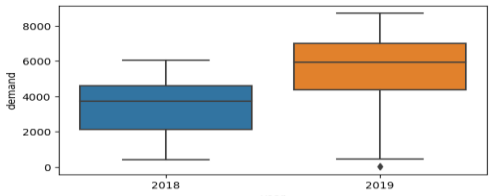
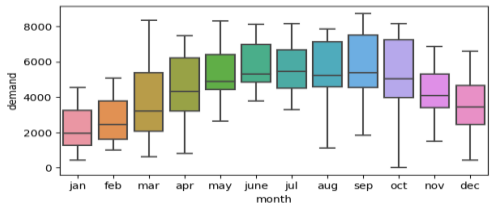
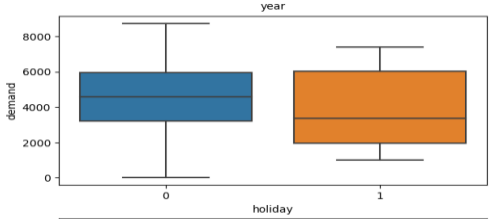
Assignment-based Subjective Questions

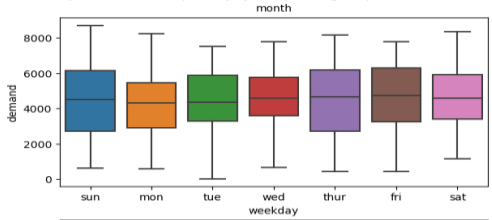
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    a. My target variable 'cnt' has been renamed to 'Demand'. So, a few observations made on the effect of categorical variables on the dependent variable 'Demand' are:
        i. The Fall season had the most Demand generated.
        ii. Demand had significantly gone up in 2019 as compared to 2018.
        iii. June to September months had the most Demand for bikes.
        iv. If it is a Holiday bike Demand is more.
        v. Over the Weekdays there is mostly similar Demand for bikes
        vi. Though the median for Demand is similar even if it is a working day or not, the larger set of Demand is when it is not a Working day.
        vii. If the Weather Situation is Clear or there are only a few clouds or it is Partly cloudy then the Demand for bikes is significantly high, whereas there is no Demand when there is Heavy Rain with falling Ice Pallets and Thunderstorms along with Mist or Snow with Fog
        viii. As it approaches fall from Summer, Demand increases. It becomes steady demand during fall and drops as it advances in winter.
        ix. Sunday sees a constant high demand irrespective of any season except spring. And demands usually flow in from Friday to Sunday.
        x. The Fall season sees demand irrespective of whether it's a working day or a holiday.

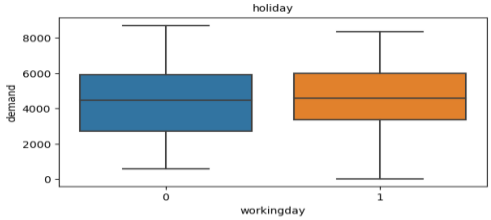

demand
0
2000
4000
6000
8000
spring
summer
fall
winter
season

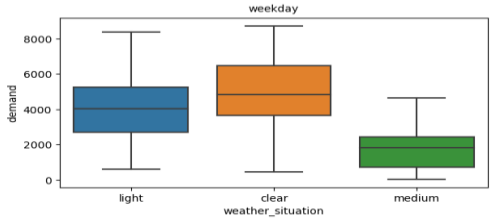

demand
0
2000
4000
6000
8000
2018
2019
year

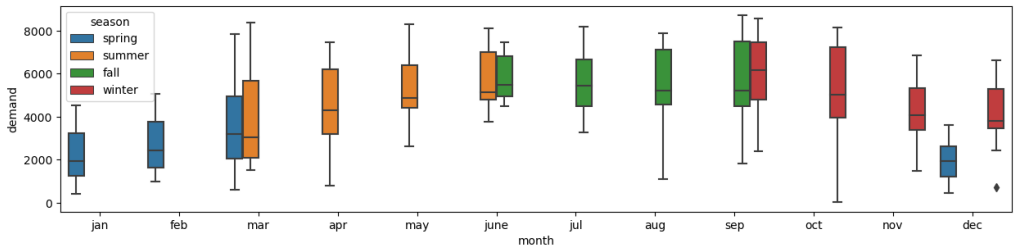

demand
0
2000
4000
6000
8000
jan
feb
mar
apr
may
june
jul
aug
sep
oct
nov
dec
month

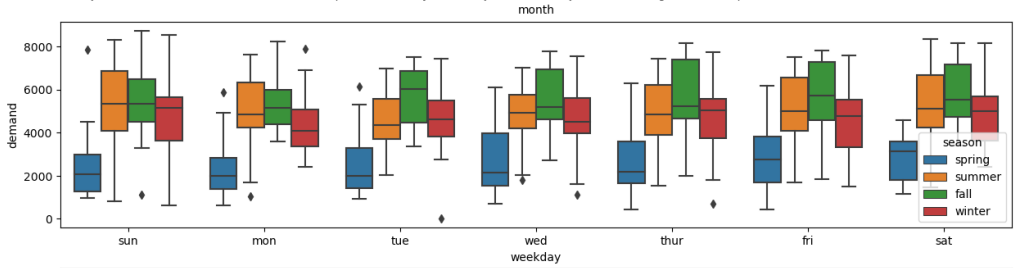

demand
0
2000
4000
6000
8000
0
1
holiday

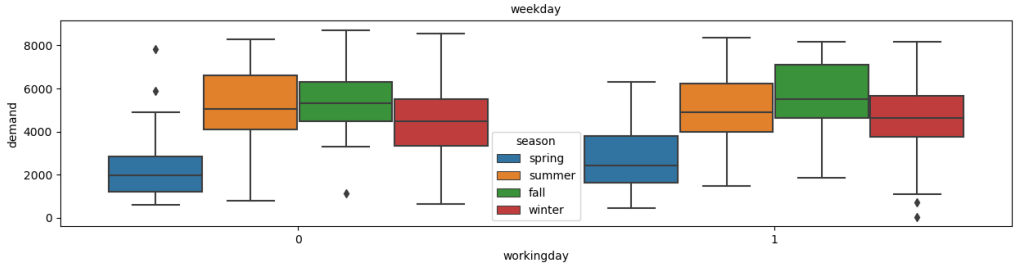

demand
0
2000
4000
6000
8000
sun
mon
tue
wed
thur
fri
sat
weekday

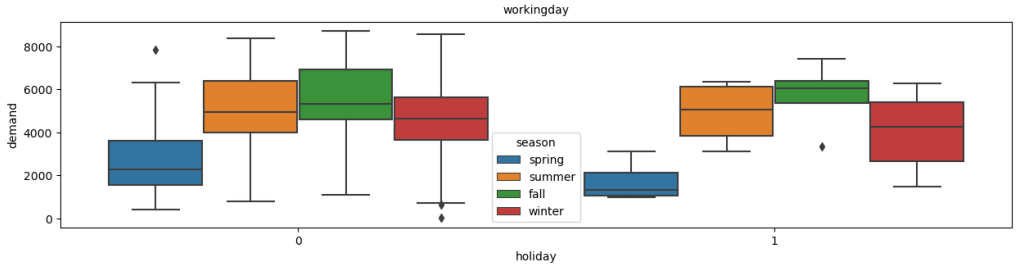

demand
0
2000
4000
6000
8000
0
1
workingday


demand
0
2000
4000
6000
8000
light
clear
medium
weather_situation

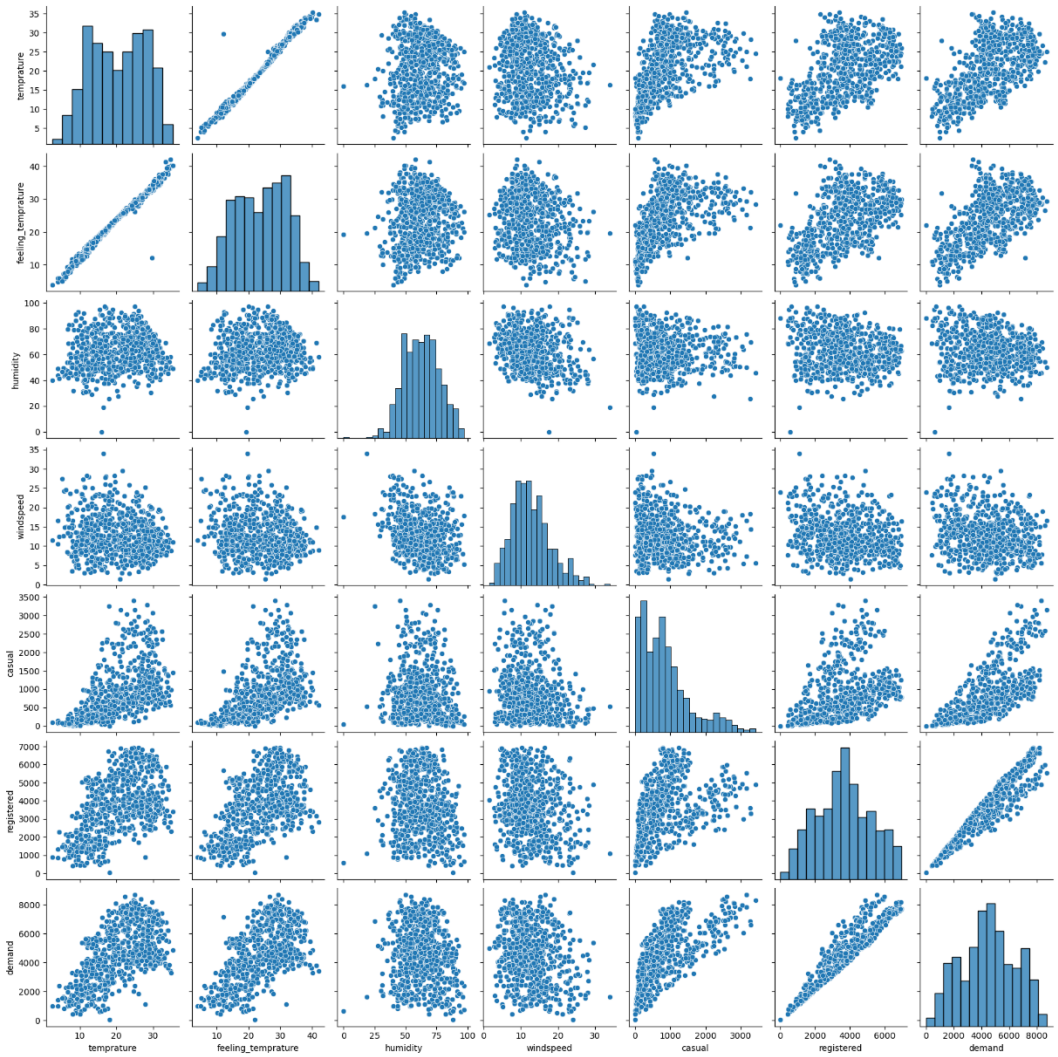

season
spring
summer
fall
winter
demand
0
2000
4000
6000
8000
jan
feb
mar
apr
may
june
jul
aug
sep
oct
nov
dec
month


demand
0
2000
4000
6000
8000
sun
mon
tue
wed
thur
fri
sat
weekday
season
spring
summer
fall
winter

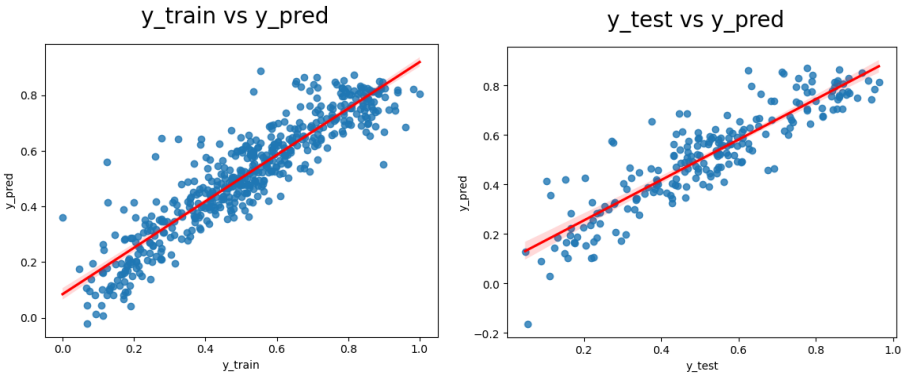

demand
0
2000
4000
6000
8000
0
1
workingday
season
spring
summer
fall
winter


demand
0
2000
4000
6000
8000
0
1
holiday
season
spring
summer
fall
winter

2. Why is it important to use drop_first=True during dummy variable creation?
    a. It is important to use drop_first=True during dummy variable creation because it creates 1 less dummy variable of the feature selected to represent the same data, i.e., if the feature has n-type of values in it then after drop_first=True there will be n-1 number of dummy variables created for the feature explaining the same data with one less column created. This enables us to reduce some number of features to have a better model as well as reduces the chance of multi-collinearity.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    a. Looking at the pair-plot among the numerical variables, 'registered' has the highest correlation with the target variable 'Demand'. But if they are dropped then 'temperature' has the most correlation with 'demand'.



4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. The following observations made while doing the analysis validate the assumptions of Linear Regression after building the model on the training set:
        i. Prob(F-Statistics) is 1.62e-186 (less than 0.05), this shows the overall model is significant and there exists a linear relationship between demand and the set of predictor variables.
        ii. Plotting histogram of errors obtained for actual demand and calculated demand, a normal distribution is obtained. This validates that errors should be normally distributed
        iii. Scatter plot for actual demand and calculated demand shows there is homoscedasticity, i.e., error terms have constant variance

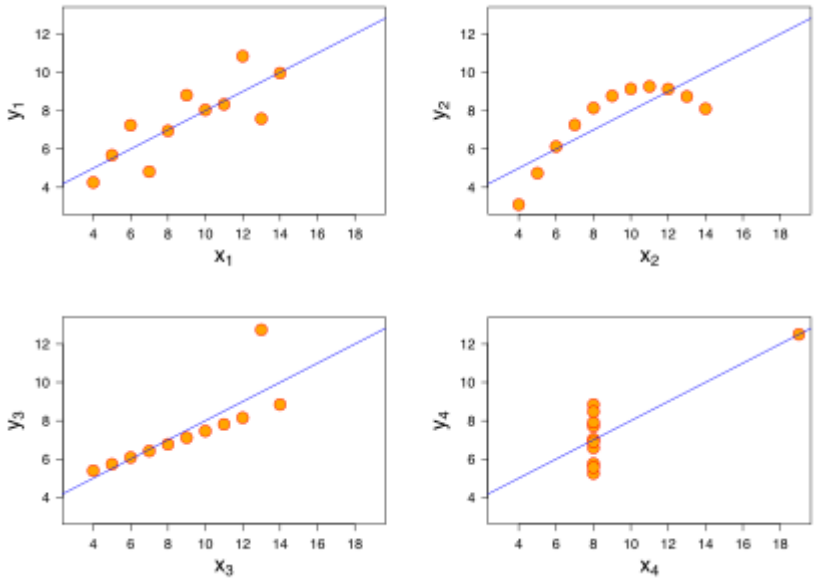


        iv. Previously by eliminating highly correlated values using heatmap, VIF, and p-values it is ensured that multi-collinearity doesn't exist. Hence, it is validated that the features are independent.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. Top 3 features contributing significantly towards explaining the demand of shared bikes are:
        i. Temperature (coefficient: 0.3903)
        ii. year_2019 (coefficient: 0.236430)
        iii. month_sep (coefficient: 0.066495)

General Subjective Questions

1. Explain the linear regression algorithm in detail.
    a. Linear regression algorithm is a supervised machine learning approach that explains the relation between dependent and independent variables using a straight line. As this is a regression learning, the target variable is continuous in nature.
    b. Use-cases for this approach are predicting demand for a bike-sharing company based on some features available, predicting house prices based on related features available, predicting the height of a person based on other body features available, etc.
    c. The model aims to find the best-fitted straight line that can generalize all the features taken into consideration for predicting the target variable, so the prediction is closest to the actual value.
    d. The objective of finding the best-fit line is done by reducing the error within the predicted and actual target values (this is defined as the cost function). The line giving the least error is the best-fit line.
    e. The minimization of error terms is achieved by deploying optimization techniques like:
        i. Closed form method: the most straightforward method of optimization, where we equate the derivative of the cost function to 0.
        ii. Gradient Descent method: an iterative approach to reduce the cost function. We move in the opposite direction of the derivative.
    f. It can be further classified as:
        i. Simple Linear regression: here there is one independent variable used for predicting the target or dependent variable. The equation is given by, **y = beta_0 + beta_1*x.**
        ii. Multiple Linear Regression: here there are more than one independent variable used for predicting the target variable. The equation is given by, **y = beta_0 + beta_1*x_1 + ... beta_n*x_n** (n is the number of features).

2. Explain the Anscombe's quartet in detail.
    a. Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics such as the same mean, standard deviation, and regression line yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.
    b. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.
    c. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".
    d. The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.



Source: Wikipedia

3. What is Pearson's R?
    a. Pearson's R or Pearson's Correlation Coefficient is a measure of linear correlation between two sets of data.

b. 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.
c. It gives the direction of the relationship between the dependent and independent variable in Linear Regression.
    i. If the value of R is -1, then they are negatively correlated.
    ii. If the value of R is 0, then they are not correlated.
    iii. If the value of R is 1, then they are positively correlated.
d. By squaring them we get the Coefficient of Determination (R^2). It gives the strength of the relationship between the dependent and independent variable in Linear Regression.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    a. Scaling is a part of Data Pre-Processing in Data Analytics and Machine Learning, which is applied to all numeric independent variables to bring the data within a particular range.
    b. It helps in speeding up the calculations in an algorithm as the range of data to which the data is scaled becomes small.
    c. In a real-world scenario, the collected data set contains features highly varying in magnitudes, units, and range. If scaling is not done then the algorithm only takes magnitude into 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.
    d. It can be broadly 2 types:
        i. Normalized Scaling: It brings all of the data in the range of 0 and 1. The most popular method is Min-Max scaling, where the data is scaled using the minimum and maximum values of the features taken into account.
        ii. Standardize Scaling: Standardization replaces the values by the feature's values to Z scores. This brings all of the data into a

standard normal distribution which has mean (μ) zero and standard deviation one (σ).

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    a. A large value of VIF indicates that there is a correlation between the variables.
    b. If there is a perfect correlation, then VIF = infinity.
    c. This happens because if we notice the formula for VIF, given by $1/(1-R^2)$, where R^2 is found on the feature VIF is performed. This means when data is too much correlated the VIF value will increase. When R^2 reaches 1 (perfect correlation), the denominator becomes 0, and 1/0 is infinity.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    a. Q-Q plots are also known as Quantile-Quantile plots.
    b. They plot the quantiles of a sample distribution against the quantiles of a theoretical distribution.
    c. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, or exponential.
    d. The power of Q-Q plots lies in their ability to summarize any distribution visually.
    e. QQ plots are very useful in determining:
        i. If two populations are of the same distribution
        ii. 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.
        iii. Skewness of distribution
    f. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.