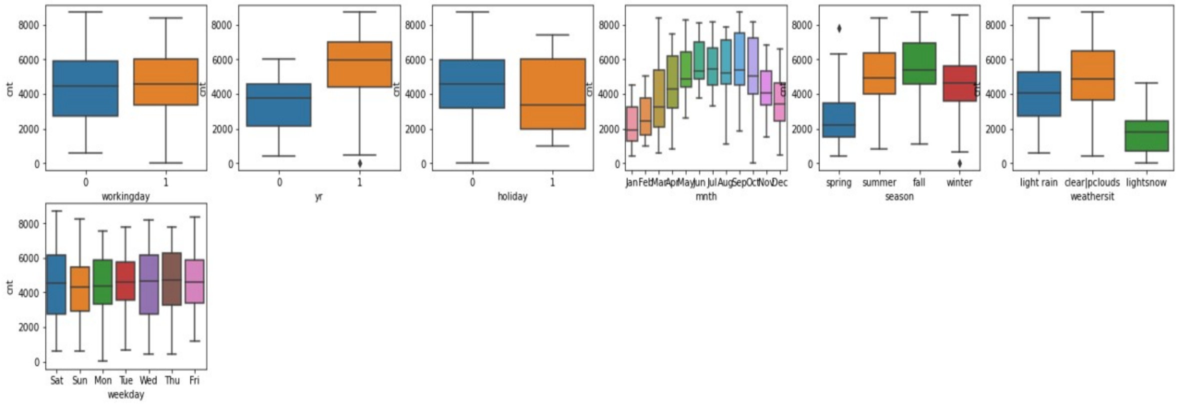
Assignment-based Subjective Questions

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



The above boxplot would help us understand the categorical variables effect on dependent variables, we can infer

1. A demand increase in year 2019 for share bike service
2. More demand in Clear and partial cloud weather conditions
3. More demand in 'fall' and 'summer' seasons
4. Saturday being demand day in a week
5. Demand spike in Sep and Oct months

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

It deletes the extra column creating during dummy variables creation which eventually reduces the correlation between dummy variables. If categorical variables have n levels, then the representation of dummy variables would be n-1.

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

‘Temp ‘ (Temperature) is the numerical variable which is having highest correlation with ‘cnt’ (target variable)

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

1. A validation must be done to check existence of linear relationship between dependent and predictors. Pair scatter plot would help to know the relationship between dependent and predictors.
2. Homoscedasticity is another assumption which needs to be validated, residuals must have constant variance irrespective of the level of the dependent variable. Residual plot and error term normal distribution would help us to validate this assumption.

3. A validation must be done to check the existence of multicollinearity between independent variables as it impacts overall interpretation of results. Correlation plot and Variation inflation factor (VIF). Acceptable VIF would be =<5 (recommended)

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

Top three features

1. Temp(Temperature)
2. Weathersit (winter, sprint etc)
3. Weekday(Sat being most)

General Subjective Questions

Explain the linear regression algorithm in detail

Linear regression is an algorithm is used to visualize the relationship between two different features and predict. It helps place the data points within the curve that helps modelling and analysing the data.

Regression model involves the values of the coefficient that are used in the representation of the data. It uses statistics to estimate coefficients. The simple liner equation wold be y= B1x+B0 or y=mx+c

y is an independent variable

m is slope

x is dependent variable

c is the constant (intercept of given line)

Simple linear regression would be focusing on one variable and multiple linear regression model would be focusing on multi variables.

Steps to perform the linear regression

1. Understand the data
2. Visualizing the data
3. Perform simple or multi linear regression
4. Model building
5. Train the model
6. Perform residual analysis
7. Prediction on test set
8. Checking R-Square and Adjusted R-Square
9. Visualizing fit on test set

Explain the Anscombe's quartet in detail.

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 demonstrates both the importance of graphing data before analysing it and the effect of outliners on statistical properties.

What is Pearson's R?

It is correlation coefficient formula which helps to find how strong the relationship between data. The formula returns a value between 1 and -1 where

1 – Indicates strong correlation (Every positive increase in one variable there is positive increase of a fixed proportion in the other variable)

0 – Indicates no correlation (Every increase there won't be positive or negative increase

-1 – Indicates negative correlation (Every positive increase in one variable there is negative decrease of a fixed proportion in the other variable)

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

1. Scaling is method to normalize the range of independent variables of data. It is generally performed during data processing. Min-max scaling is method widely used in the industry.
2. The range of all variables or features should be normalized as most of the times the variables would have different data units and range. To solve this issue, scaling must be done to bring all other variables to the same level of magnitude. It **only** affects coefficients.
3. Normalized scaling brings all the data in the range between 0 and 1. Standardization replaces the values by their Z-score. It brings all of the data into a standard normal distribution which has mean zero and standard deviation one
   MinMax Scaling

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

VIF infinity happens when there is perfect correlation between two independent variables. In this case we would get R2 = 1 which lead to infinity.

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

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.