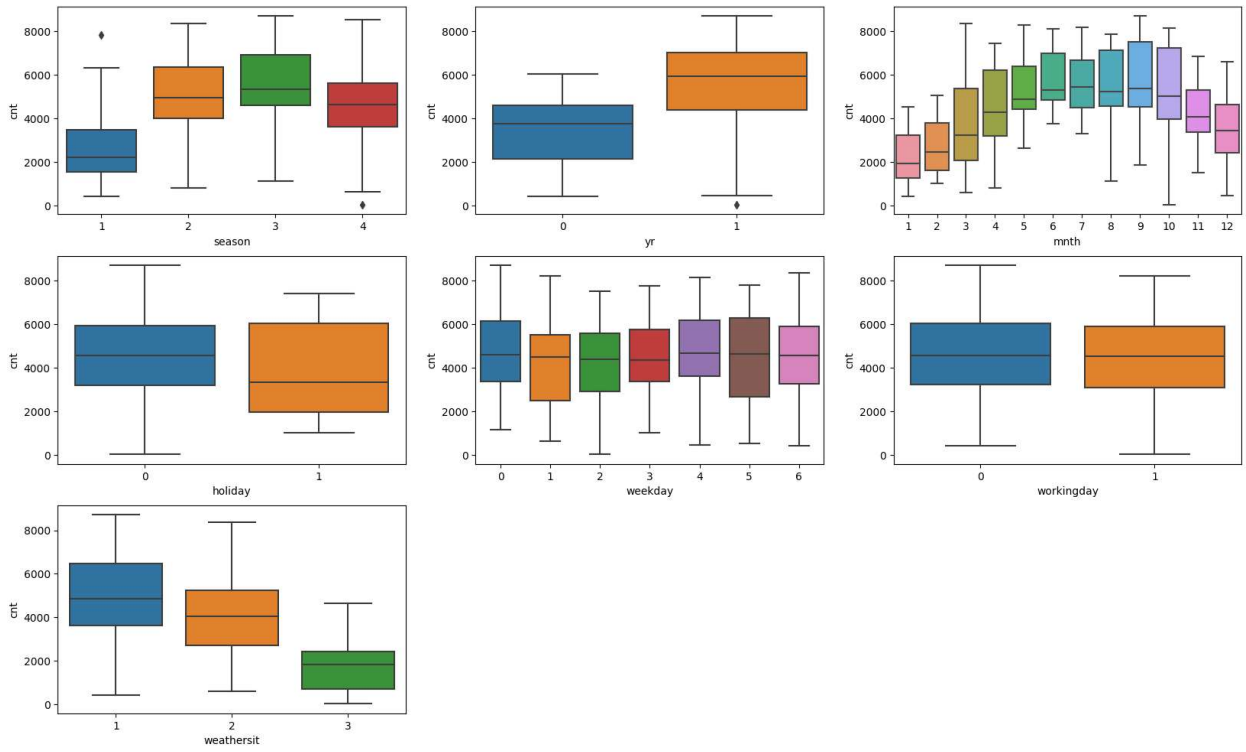
Assignment-based Subjective Questions

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

Ans: There are 7 categorical variables in dataset, we used box plot to study the effect on dependent variable cnt as shown below.



Season : Most of the booking had happened in season 3 followed by season 2 and season 4. This indicate that season is good predicator for dependent variable cnt.

Yr: We have seen increased demand in year 1 (2019), this shows significant growth year over year in the demand.

Mnth: We have seen a trend in the booking for month 5, 6, 7, 8 and 9 so it can be good predicator for the dependent.

Holiday: Max number of booking we can see when there is no holiday which means the data is biased so this can not be good predicator for the dependent.

Weekday: Weekday shows very close trend, this can or can not be good predicator, let the model decide this.

Workingday: Almost same number of booking we can see on working day and non working day, we are not yet sure if this can be a good predicator or not, will decide based on model.

Weathershit: High number of booking had happened during the weathershit1 followed by 2, this indicates weathershit does show some trend towards bike booking and can be good predicator for the dependent variable.

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

Ans: drop_first = True is important to use as this will eliminate the extra column while creating the dummy variables.

For example, in our bike sharing case study we have column season, when we create the dummy variable, it creates 4 season:
Season_spring
Season_summer
Season_fall
Season_winter

So when we look at the data when there is no spring, fall and winter then it will be summer automatically so there is no point having extra variable. After dropping the summer it will reduce the correlations between these dummy variable.

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

Ans: The numerical variable atemp has highest correlation 0.65 with target variable followed by temp. Both of the parameters cannot be used in the model due to multicolinearity. We will decide which parameters to keep based on VIF and p-value w.r.t other variables

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

Ans: We will validate the assumptions of Linear Regression by plotting the distplot of the residuals and validate if this is normal distribution or not. VERY LOW Multicollinearity between the predictors and the p-values for all the predictors seems to be significant. The Coefficient values from the model of all the variables are not equal to zero which means we are able to reject Null Hypothesis F-Statistics is used for testing the overall significance of the Model: Higher the F-Statistics, more significant the Model is. The Residuals were normally distributed after plotting the histogram . Hence our assumption for Linear Regression is valid. VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 5.

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

Ans: As per our final Model, the top 3 predictor variables that influences the bike booking are:
Actual Temperature (temp) - A coefficient value of '0.5069 indicated that a unit increase in temp variable increases the bike hire numbers by 0.5069 units.
Weather Situation 3 (weathersit_3) - A coefficient value of '-0.2791' indicated that, w.r.t Weathersit1, a unit increase in Light Precipitation variable decreases the bike hire numbers by 0.2791 units.
Year (yr) - A coefficient value of '0.2343' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2343 units. So, it's suggested to consider these variables utmost importance while planning, to achieve maximum Booking

General Subjective Questions

1. Explain the linear regression algorithm in detail.
Ans: Linear Regression is a machine learning algorithm based on supervised learning. Linear regression basically predict a dependent variable value (y) based on a given independent variable/variables (x/Xi). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model. y = a1 +a2.x here, a1 is intercept a2 is the coefficient of x x: input training data y: labels to data Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but 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. Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

The four datasets can be described as

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

3. What is Pearson's R?

Ans: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

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

Ans: Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. In order to solve this issue, we have to do scaling to bring all the variables to the same level.
1- Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
2- Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ). sklearn.preprocessing.scale helps to implement standardization in python.

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

Ans: The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.
VIF = 1/1-R^2
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 . The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

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

Ans: Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
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.

The slope tells us whether the steps in our data are too big or too small . for example,
if we have N observations, then each step traverses 1/(N-1) of the data. So we are seeing how the step sizes (a.k.a. quantiles) compare between our data and the normal distribution.
A steeply sloping section of the QQ plot means that in this part of our data, the observations are more spread out than we would expect them to be if they were normally distributed. One example cause of this would be an unusually large number of outliers (like in the QQ plot we drew with our code previously).