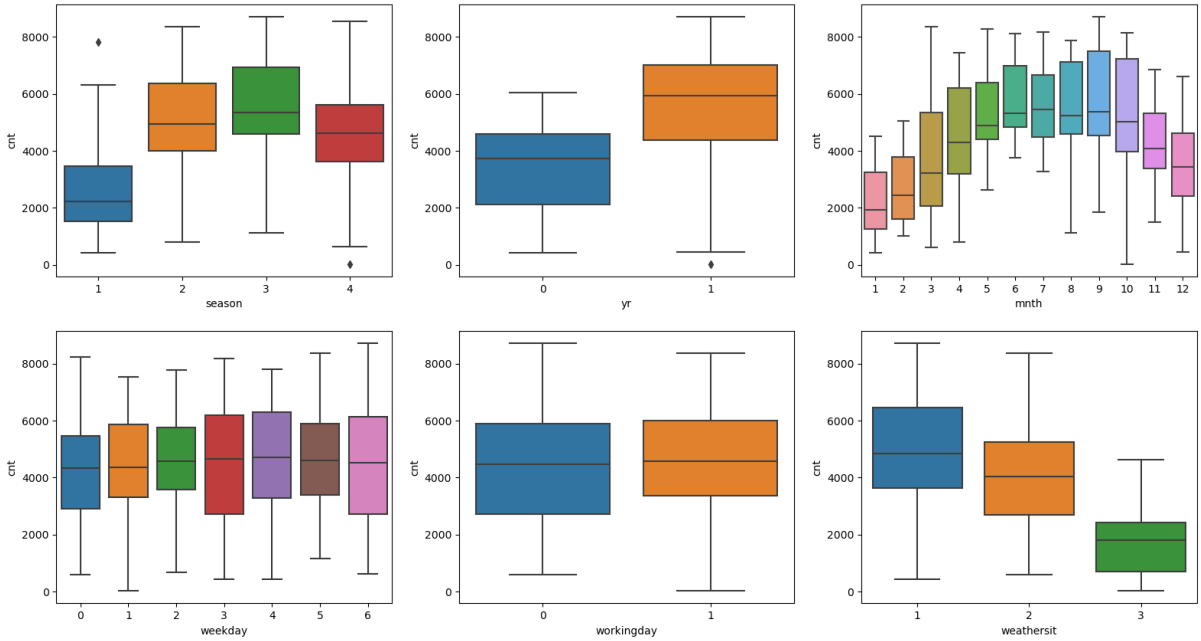
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans:

In the dataset, we have 6 categorical variables namely- ‘yr’,’workingday',’mnth', 'weekday', 'season' & 'weathersit'.

We used Box plot (refer the fig below) to study their effect on the dependent variable (‘cnt’) .



The inference that we could derive were:

season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

weathersit: Almost 67% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

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

Ans:

- If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.
- **drop_first=True** is important to use, as it helps in reducing the extra column created during dummy variable creation.
- Hence it reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:

Looking at the pair-plot, has highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

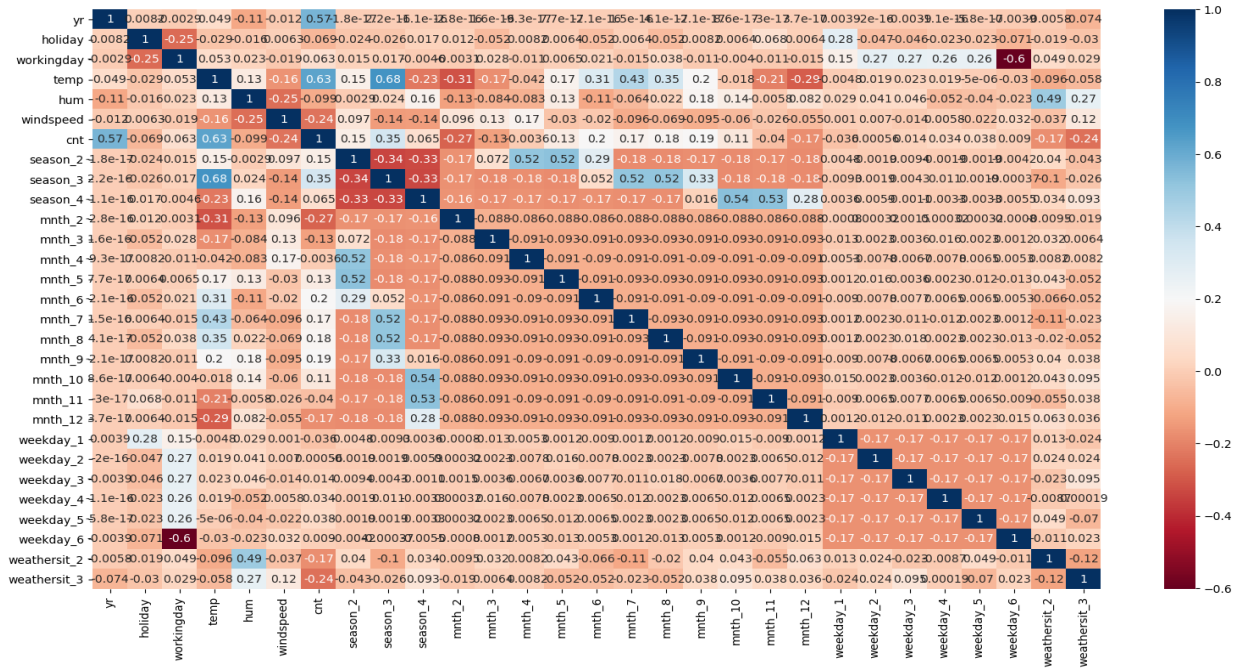
Linear regression assumptions after model building and being able to verify and act upon them is especially important. Below are the pointers:

1. Absence of Multicollinearity

   Multicollinearity refers to the fact that two or more independent variables are highly correlated (or even redundant in the extreme case).

   Verification:

   - **Pairwise correlations** could be the first step to identify potential relationships between various independent variables.
   - A more thorough method, however, would be to look at the Variance Inflation Factors **(VIF).**
   - Checking **p- value**



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- Temperature (temp) - A coefficient value of '0.5499' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5499 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of '-0.2871' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.2871 units.
- Year (yr) - A coefficient value of '0.2331' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2331 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans:

- Linear regression is a machine learning algorithm. The output variable to be predicted is a continuous variable. Eg: price of house
- Regression is the most commonly used predictive analysis model.
- There are several main reasons people use regression analysis:
    - To predict future economic conditions, trends, or values
    - To determine the relationship between two or more variables
    - To understand how one variable changes when another change.

There are two types of linear regression under this module:

- Simple linear regression
- Multiple linear regression

1. Simple Linear Regression

The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

2. Multiple Linear Regression

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Multiple regressions are based on the assumption that there is a linear relationship between both the dependent and independent variables. It also assumes no major correlation between the independent variables.

KEY TAKEAWAYS

- ✓ Regression analysis is a common statistical method used in finance and investing.
- ✓ Linear regression is one of the most common techniques of regression analysis.
- ✓ Multiple regression is a broader class of regressions that encompasses linear and nonlinear regressions with multiple explanatory variables.

3. Explain the Anscombe's quartet in detail. (3 marks)

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.
- It was constructed in 1973 by statistician 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.
- This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

3. What is Pearson’s R? (3 marks)

Ans:

- Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson’s.
- Pearson’s correlation (also called Pearson’s R) is a correlation coefficient commonly used in linear regression.
- This correlation coefficient is designed for linear relationships and it might not be a good measure for if the relationship between the variables is non-linear.
- The other correlation coefficient is Spearman's R which is used to determine the correlation if the relationship between the variables is not linear.
- So even though, Pearson's R might give a correlation coefficient for non-linear relationships, it might not be reliable. For example, the correlation coefficients as given by both the techniques for the relationship y=x^3 for 100 equally separated values between 1 and 100 were found out to be:

  Pearson’s R =~0.91
  Spearman’s R= ~1

- And as we keep on increasing the power, the Pearson's R value consistently drop whereas the Spearman's R remains robust at 1. For example, for the relationship Y=X^10 for the same data points, the coefficients were:

  Pearson’s R≈0.66
  Spearman’s R≈1

- So, the takeaway here is that if you have some sense of the relationship being non-linear, you should look at Spearman's R instead of Pearson's R. It might happen that even for a non-linear relationship, the Pearson's R value might be high, but it is simply not reliable.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units
- When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:
  1. Ease of interpretation
  2. Faster convergence for gradient descent methods
- There are two types of scaling:-
  1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

     $$x = \frac{x - mean(x)}{sd(x)}$$

  2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

- It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

4. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Ans:

- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Ans:
   - 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.
   - 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.
   - A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

     The advantages of the q-q plot are:
   - The sample sizes do not need to be equal.
   - Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.