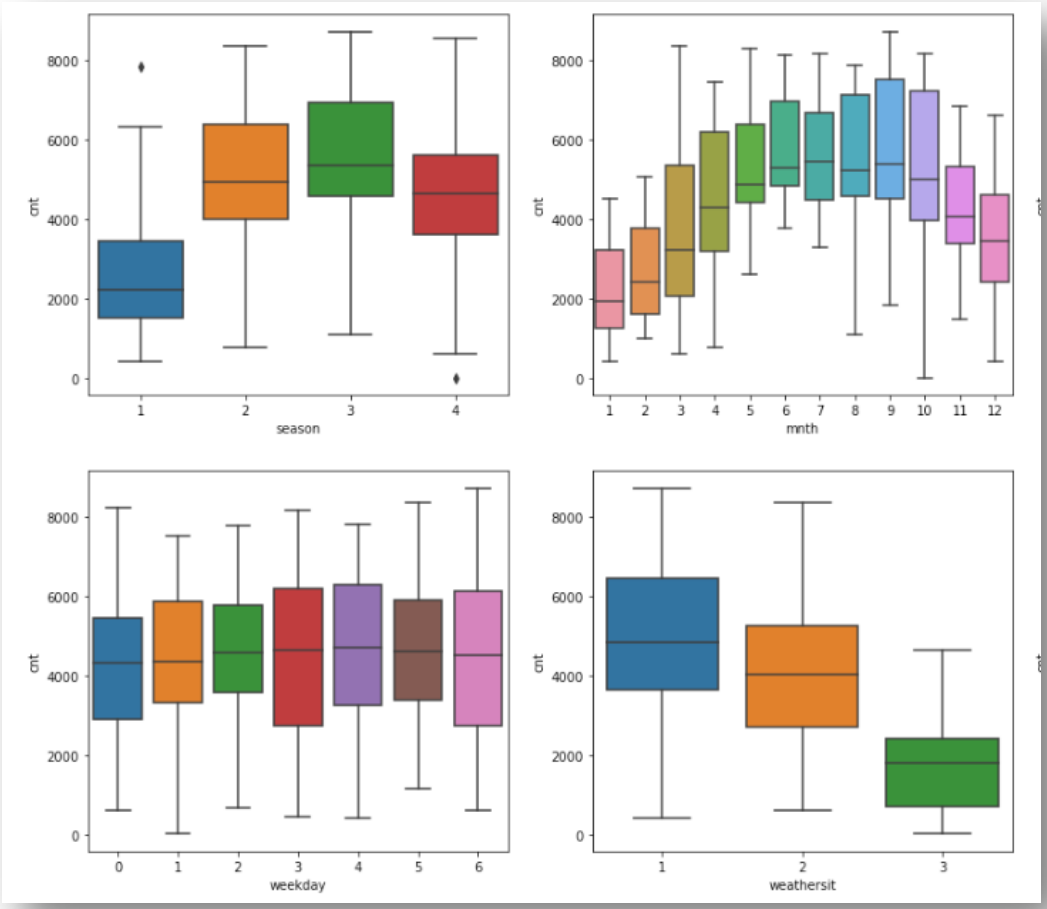
Assignment-based Subjective Questions

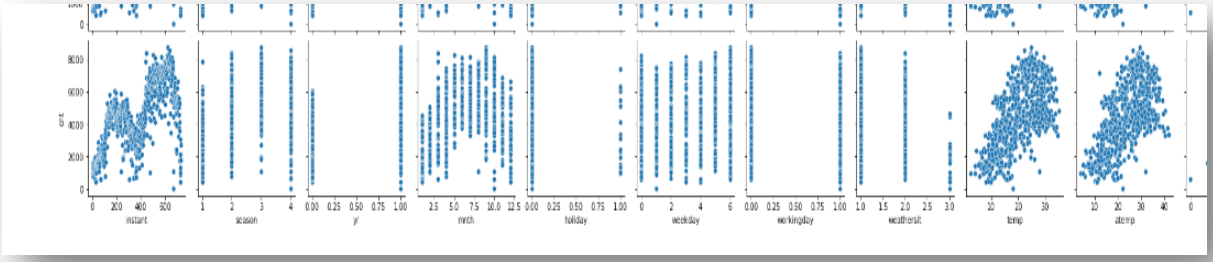
Q1 - From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
A1 – Categorical variables have good enough effect on dependent variables. Below is the box-plot of the categorical variables.
Season 1:spring, 2:summer, 3:fall, 4:winter **–** We see that during summer and fall, more number of bikes are rented out.
Month – Similarly, during the month of May to October high number of bikes are rented out.



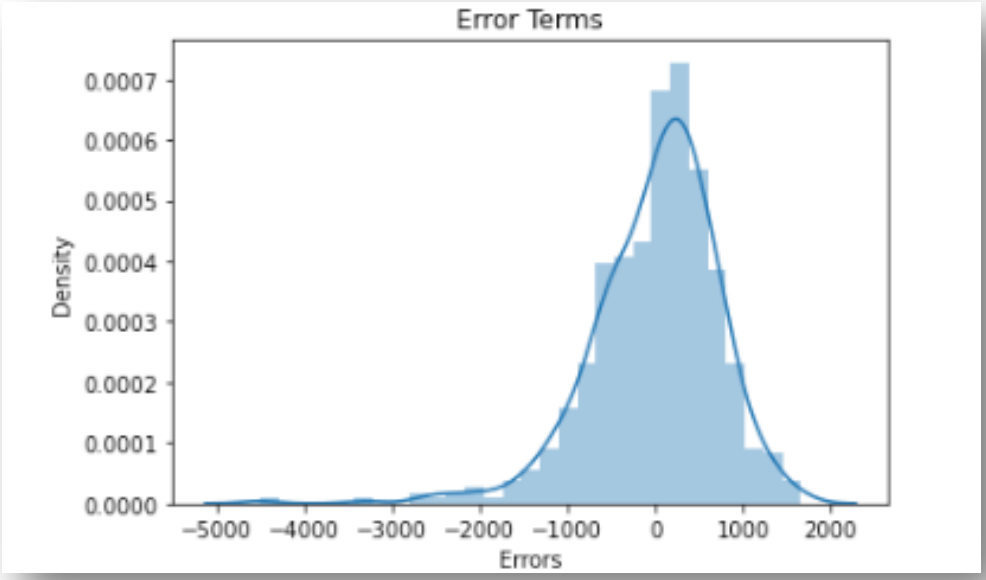
Q2 - Why is it important to use drop_first=True during dummy variable creation? (2 mark)
A2 – *drop_first=True* denotes that we drop the column created for the first value in the column for which we are creating dummy variable. This has multiple advantages:

1. Reduces the size of data-frame by one column.
2. A reduce in size, means it contributes in faster processing and optimizations.
3. Reduces the correlation created among dummy variables.

Q3 - Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
A3 – The **temp** and **atemp** variables has highest correlation with the target variable "cnt". See image for reference.



Q4 - How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
A4 – In the linear regression model we assume that the errors εiε⍰ are *independent and identically distributed (i.i.d.)* random variables. To validate the same we performed a residual analysis to check if residuals are centred around zero. See image for reference.



Q5 - Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
A5 – Top 3 features would be:

1. Year
2. Season
3. Holiday

General Subjective Questions

Q1 - Explain the linear regression algorithm in detail. (4 marks)
A1 – Linear Regression is a simplest form of statistical predictive model which is used to predict value of a variable based on another (set of) variable/s. The variable we want to predict is called the dependent variable, while the variable we are using to predict the other variable's value is called the independent or predictor variables. The effectiveness of a linear regression model is based on the ***R-square*** of a model.

While the model sounds simple at the first glance, it has a number of intricacies. The first and foremost being the assumptions of linear regression denoted by ***IID (Independent ad Identically Distributed)***. IID denotes that all errors and residuals has the same probability distribution as the others and all are mutually independent. In

linear regression or in any ML model, model is predicted on top of features we feed to the model, hence probability distribution becomes important. This is because with every new point, the probability would change, hence with all the points probability distribution should remain constant. Interestingly, this concept not only applies to the data points , but also to noise and errors.

Some of the most common cost function used in linear regression are:

1. Mean Square Error

$$\text{MAE} = \frac{1}{N}\sum_{j=1}^{N}\left(Y_j - Y_j'\right)^2$$

2. Mean Absolute Error

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{N}\left|\left(Y_j - Y_j'\right)\right|$$

3. Root mean squared error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2}$$

Where

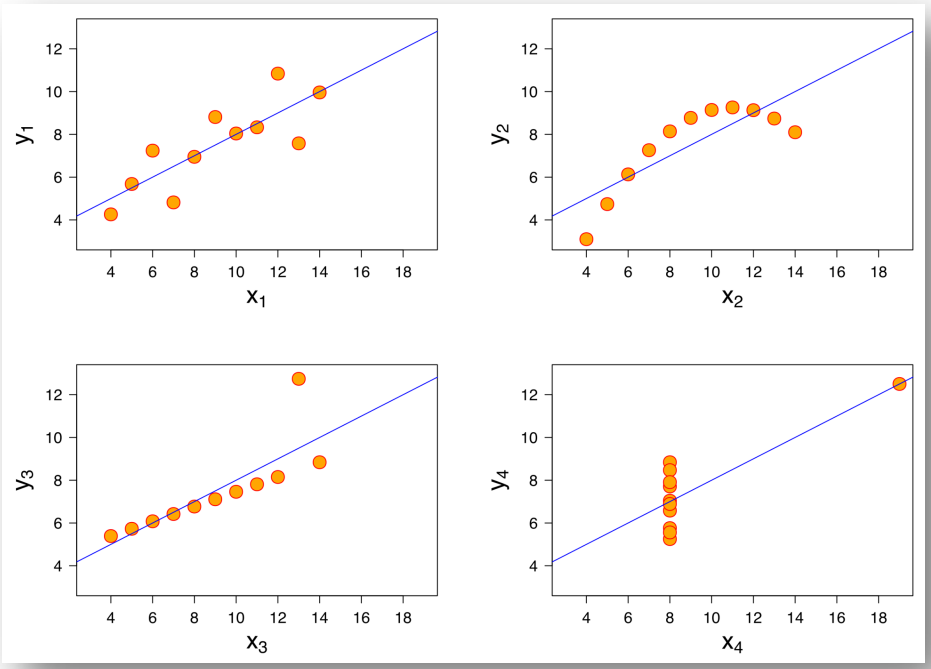
Oi=observations

Si= predicted values of a variable

n =number of observations

Q2. Explain the Anscombe's quartet in detail. (3 marks)

A2 - 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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough" There are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.



Data Set 1: fits the linear regression model pretty well.
Data Set 2: cannot fit the linear regression model because the data is non-linear.
Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Q3 - What is Pearson's R? (3 marks)
A3 – Pearson's R is a shorthand for Pearson correlation coefficient (r). The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Statistically, the Pearson correlation is appropriate when both variables being compared are of a continuous level of measurement (interval or ratio). Use the Levels of Measurement tab to learn more about determining the appropriate level of measurement for your variables.

Q4 - What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
A4 – In very simple words, scaling refers to putting the feature values into the same range. It is also known as data normalization and is generally performed during the data pre-processing step. Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling. And if we don't perform scaling, our model would lead to inaccuracies.

Standardization centres data around a mean of zero and a standard deviation of one, while normalization scales data to a set range, often [0, 1], by using the minimum and maximum values.

Normalization	Standardization
This method scales the model using minimum and maximum values.	This method scales the model using the mean and standard deviation.
When features are on various scales, it is functional.	When a variable's mean and standard deviation are both set to 0, it is beneficial.
Values on the scale fall between [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
Additionally known as scaling normalization.	This process is called Z-score normalization.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful.

Q5 - You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
A5 – A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.
If all the independent variables are orthogonal to each other, then VIF = 1.0. *If there is perfect correlation, then VIF = infinity.* A large value of VIF indicates that there is a correlation between the 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.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
A6 - In statistics, a Q–Q plot (quantile–quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions. A Q–Q plot is generally more diagnostic than comparing the samples' histograms, but is less widely known. Q–Q plots are commonly used to compare a data set to a theoretical model.[2][3] This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic. Q–Q plots are also used to compare two theoretical distributions to each other.[4] Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. 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. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.
3. The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.