Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about

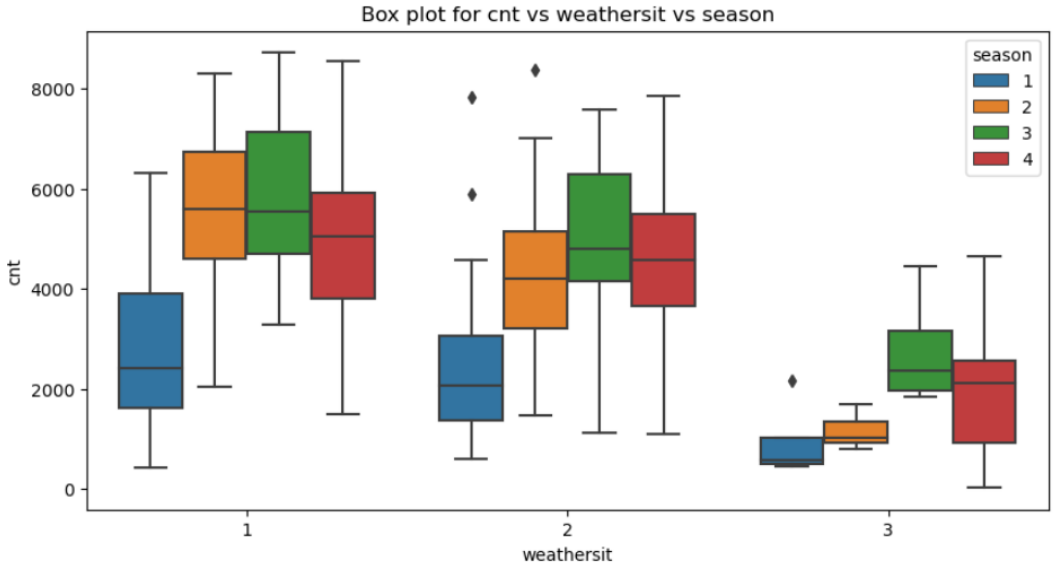
their effect on the dependent variable? (3 marks)

[ANS]:

The demand of bikes on the season 'summer' and 'autumn/fall' is more compare to 'spring' and 'winter'.

The 'weathersit' variable follows Negative impact on the demand of bikes on the Mist and Light snow when compare against clear clouds.

The overall demand of bikes is low in 'Light snow' weathersit irrespective of different seasons.



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

The dummy variable count needs to be (n-1) to avoid duplicate entries from pd.get_dummies().

Example:

For the ‘weathersit’ category variable, values are Clear, Mist and Light Snow.

Here we need only (n-1) i.e (3-1), 2 dummy variables alone needed.

Because, [weathersit_Light Snow, weathersit_Mist]

weathersit_Light Snow = 0 and weathersit_Mist=1 will correspond to Mist

Similarly, 10 will correspond to Light Snow

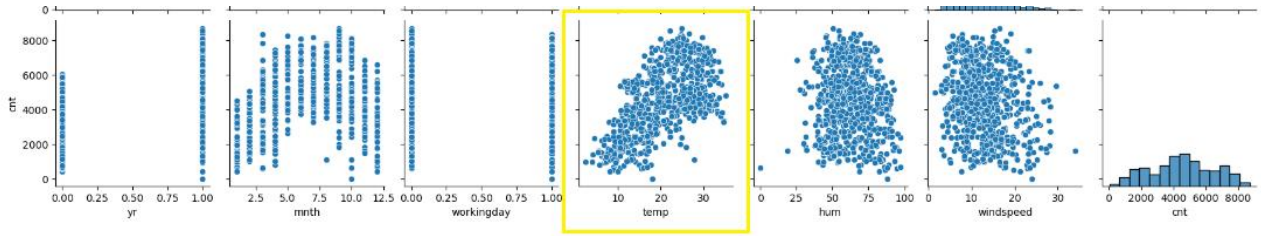
00 will correspond to Clear

If both the variables are zero – then it indirectly means it’s pointing the categorical value ‘Clear’.

Hence the pd.get_dummies() will be called with drop_first=True so, the first dummy column will be removed.

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

The ‘temp’ numerical variable has the highest correlation with the target variable ‘cnt’.



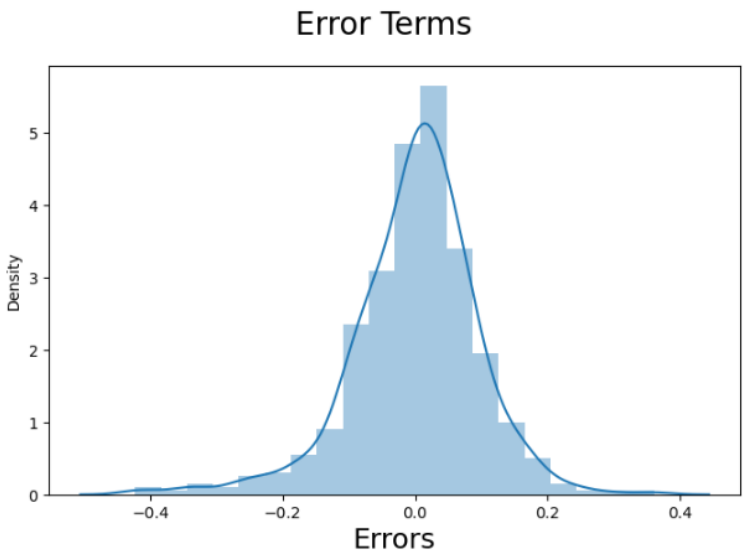
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Residual Analysis of the train data helps to validate the assumption of linear regression by predicting the target variable.

Compare the predicted value with actual value in training dataset.

Plot a scatter plot between predict against actual - if the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.

Plot the histogram of the error terms – it should follow normal distribution.



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

The top 3 features contributing significantly towards explaining the demand of the shared bikes are ‘temp’, ‘weathersit_Light Snow’, ‘yr’, ‘windspeed’.

Here the ‘temp’, ‘yr’ are positively co-related with target variable (cnt) whereas ‘weathersit_Light Snow’. ‘windspeed’ are negatively corelated.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   [ANS]:
   The linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning.
   It is both a statistical algorithm and a machine learning algorithm.

   Linear regression is a basic and commonly used type of predictive analysis. These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

   The simplest form of the regression equation with one dependent and one independent variable is defined by the formula
   y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

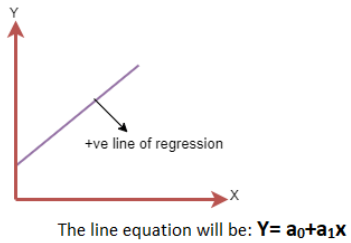
   In machine learning - it will be denoted as, Y= β0*X + β1 + ε
   β0 and β1 are two unknown constants representing the regression slope, whereas ε is the error term.

   For more than one independent variable,
   Y= β0*X0 + β1X1 + β2X2+…… βnXn+ ε

   Note: As the number of predictor variables increases, the β constants also increase correspondingly.

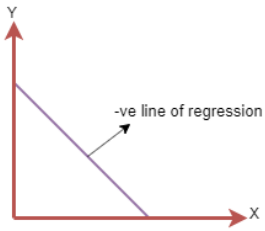
   Linear Regression Line:
   A linear line showing the relationship between the dependent and independent variables is called a regression line.

   Positive Linear Relationship:
   If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.
   Linear Regression in Machine Learning



The line equation will be: $\mathbf{Y= a_0+a_1x}$

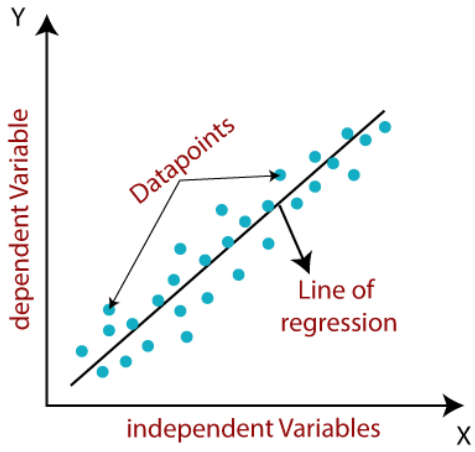
   Negative Linear Relationship:
   If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.
   Linear Regression in Machine Learning.



The line of equation will be: $Y= -a_0+a_1x$

Finding the best fit line:
The main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

Residuals: The distance between the actual value and predicted values is called residual.
If the observed points are far from the regression line, then the residual will be high, and so cost function will high.
If the scatter points are close to the regression line, then the residual will be small and hence the cost function.
The ideal graph:



Cost function:
It optimizes the regression coefficients or weights. Cost function quantifies the error between predicted and expected values and present that error in the form of a single real number.

Gradient Descent:
A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

For model evaluation:
R-squared is a statistical method that determines the goodness of fit.
The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.

Assumptions of Linear Regression:

- Linear regression assumes the linear relationship between the dependent and independent variables.
- The model assumes either little or no multicollinearity between the features or independent variables.
  Because due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables.
- Homoscedasticity is a situation when the error term is the same for all the values of independent variables.
- Linear regression assumes that the error term should follow the normal distribution pattern.
- No autocorrelations - If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model.

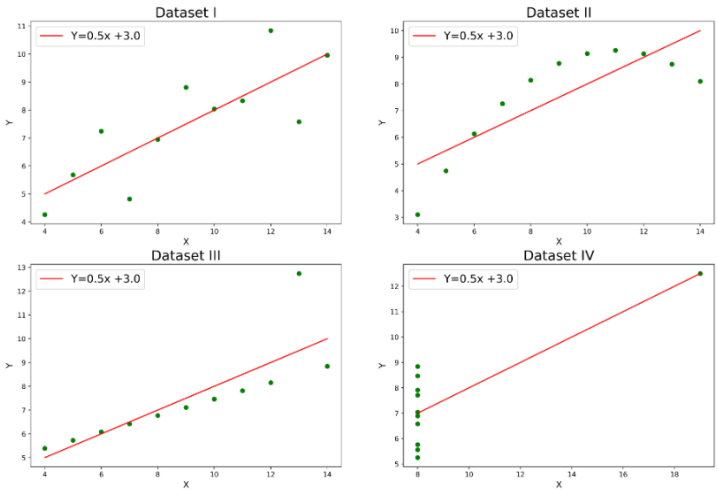
2. Explain the Anscombe's quartet in detail. (3 marks)

[ANS]:

A set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

It tells us about the importance of visualizing data before applying various algorithms to build models.

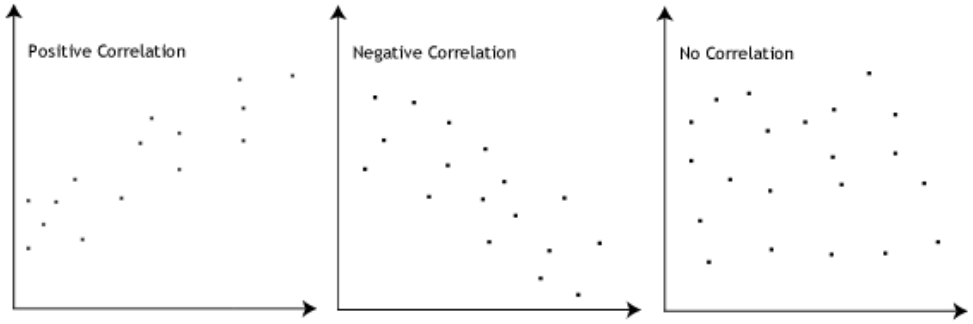
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



3. What is Pearson's R? (3 marks)
   [ANS]:
   Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables.
   The English mathematician and statistician Karl Pearson is credited for developing many statistical techniques, including the Pearson coefficient, the chi-squared test, p-value, and linear regression.

   We can categorise the type of correlation by considering as one variable increase what happens to the other variable:
   Positive correlation – the other variable has a tendency to also increase.
   Negative correlation – the other variable has a tendency to decrease.
   No correlation – the other variable does not tend to either increase or decrease.

   The starting point of any such analysis should thus be the construction and subsequent examination of a scatterplot. Examples of negative, no and positive correlation are as follows.



   - The Pearson coefficient is a mathematical correlation coefficient representing the relationship between two variables, denoted as X and Y.
   - Pearson coefficients range from +1 to -1, with +1 representing a positive correlation, -1 representing a negative correlation, and 0 representing no relationship.
   - The Pearson coefficient shows correlation, not causation.

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

[ANS]:

It's a data Pre-Processing step which is applied to independent variables to normalize the data within a particular range.

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. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It also helps in speeding up the calculations in an algorithm.

Note:

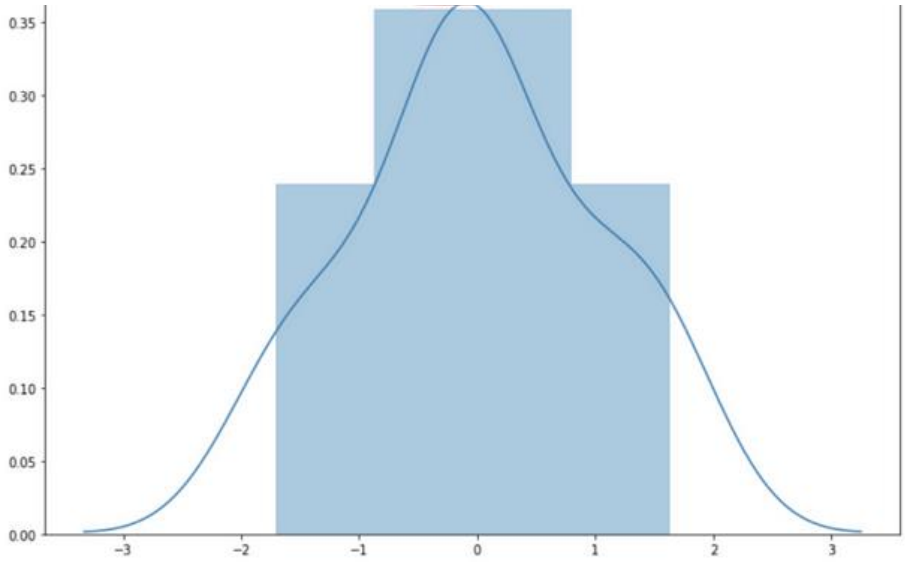
The scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Standardization Scaling:

It 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).

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

sklearn.preprocessing.scale helps to implement standardization in python.

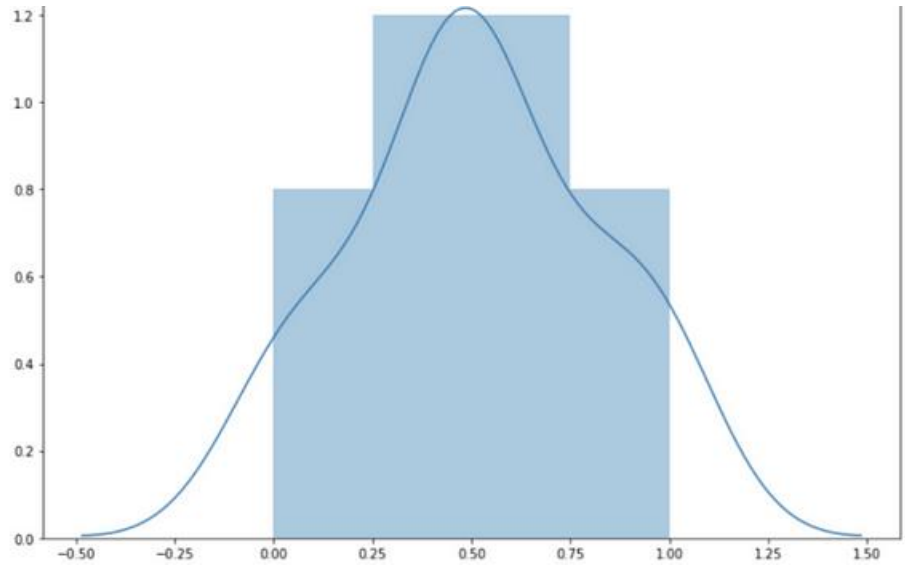


Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

One disadvantage of normalization is that it loses some information in the data, especially about outliers.



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

(3 marks)

[ANS]:

VIF is infinite 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.

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

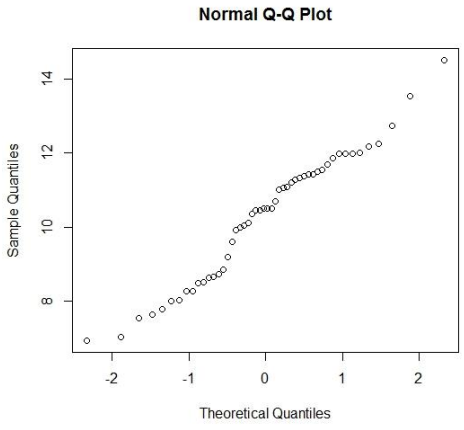
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
[ANS]:

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption.

It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a normal QQ plot when both sets of quantiles truly come from normal distributions.



The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail.
- Whether two samples have the same distribution shape.

Note:

- For a QQ plot, 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.
- The QQ plot can provide more insight into the nature of the difference than analytical methods.