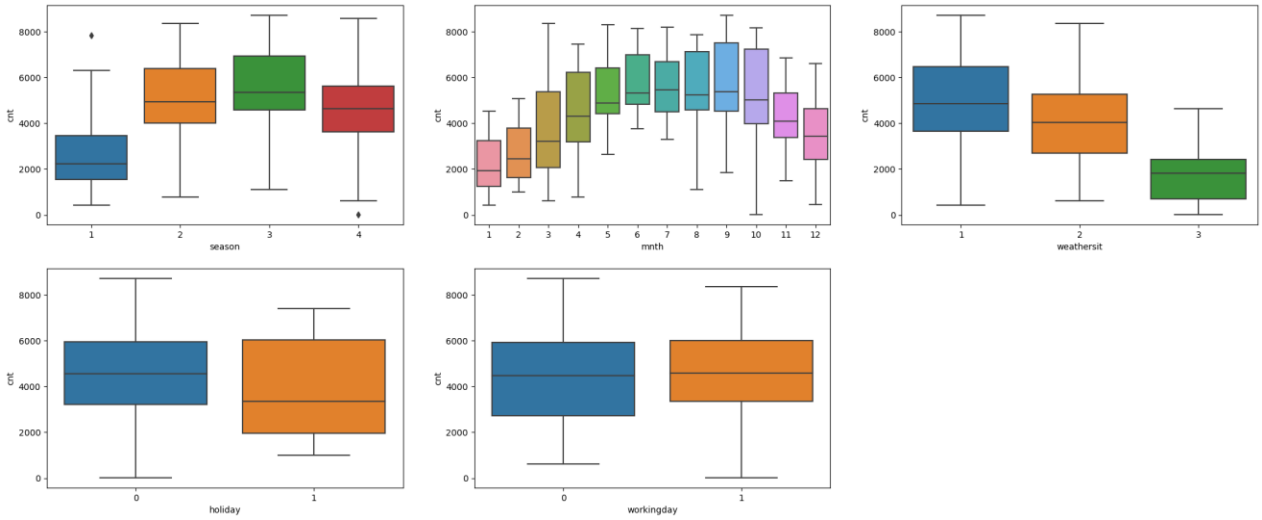
Assignment-based Subjective Questions

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



Observations

There were 6 categorical variables in the dataset.

We used Box plot (refer the fig above) to study their effect on the dependent variable ('cnt').
The inference that that could be derived is below:

1. season: Most of bike booking happens in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 of total booking. This indicates, season can be a good predictor for the dependent variable.

2. mnth: Most 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.

3. weathersit: Most of bikebooking happens 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 . This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

4. holiday: Almost all 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.

5. workingday: Almost majority 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)

In pandas.get_dummies there is a parameter i.e., drop_first allows you whether to keep or remove the reference (whether to keep k or k-1 dummies out of k categorical levels). Please note drop_first = False meaning that the reference is not dropped and k dummies created out of k categorical levels! You set drop_first = True, then it will drop the reference column after encoding .This will also avoid dummy variable trap.

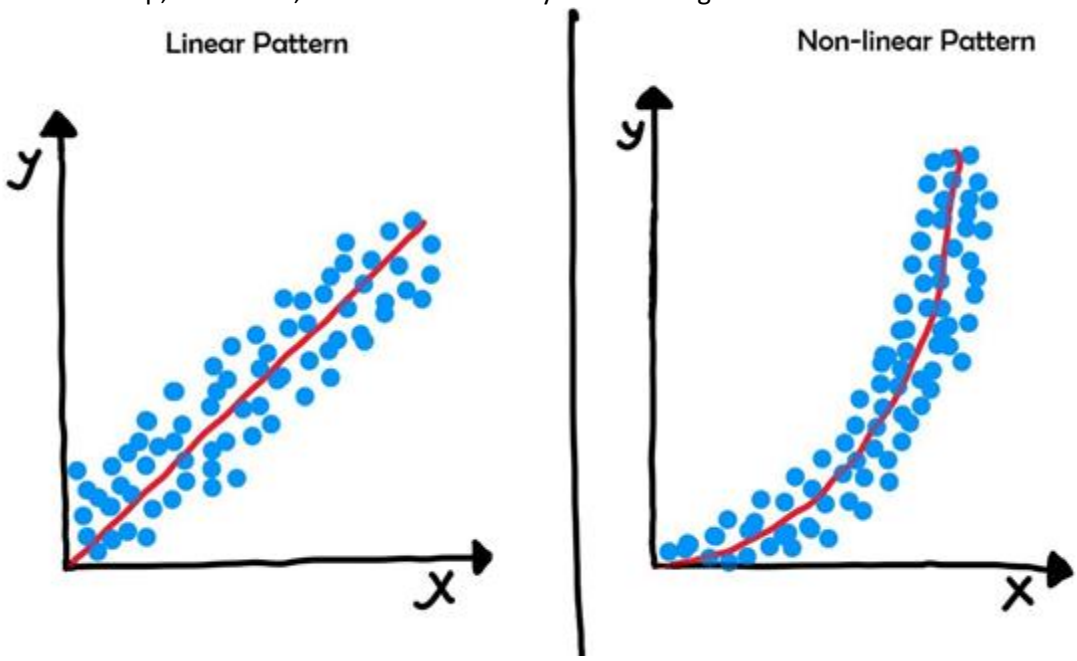
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

By looking at the pair plot temp variable has the highest (0.63 approx.rounded to two decimal places) correlation with target variable 'cnt'

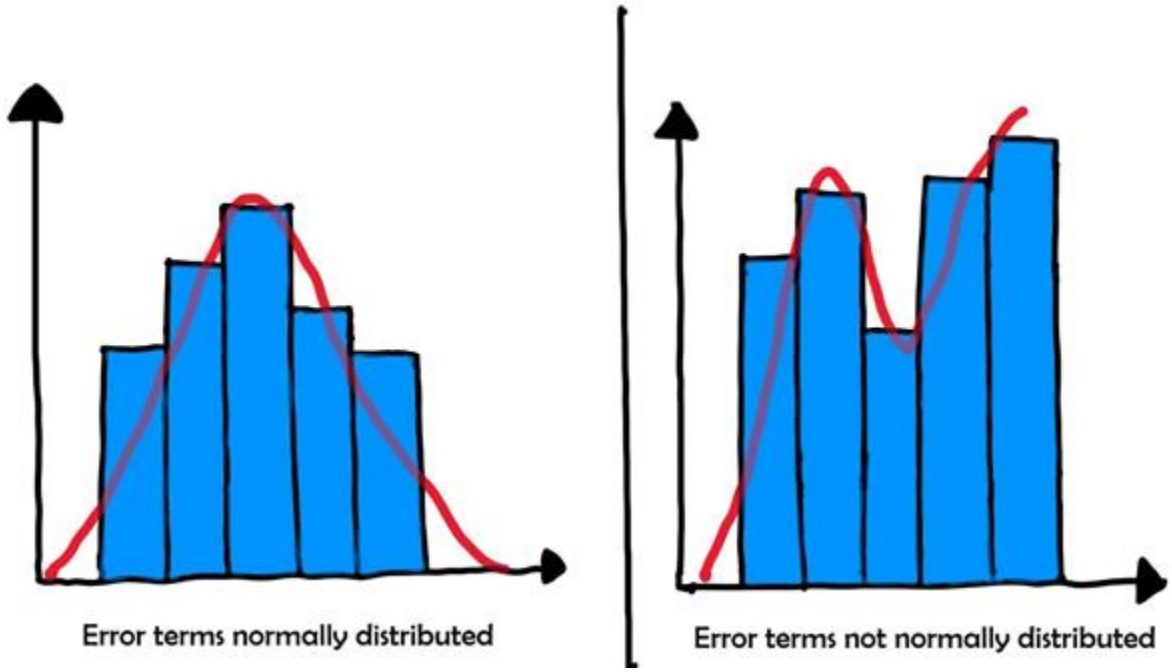
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

At the time of building a linear model, we assume that the target variable and predictor variables are linearly dependent. But, apart from these, below are few assumptions in linear regression model:

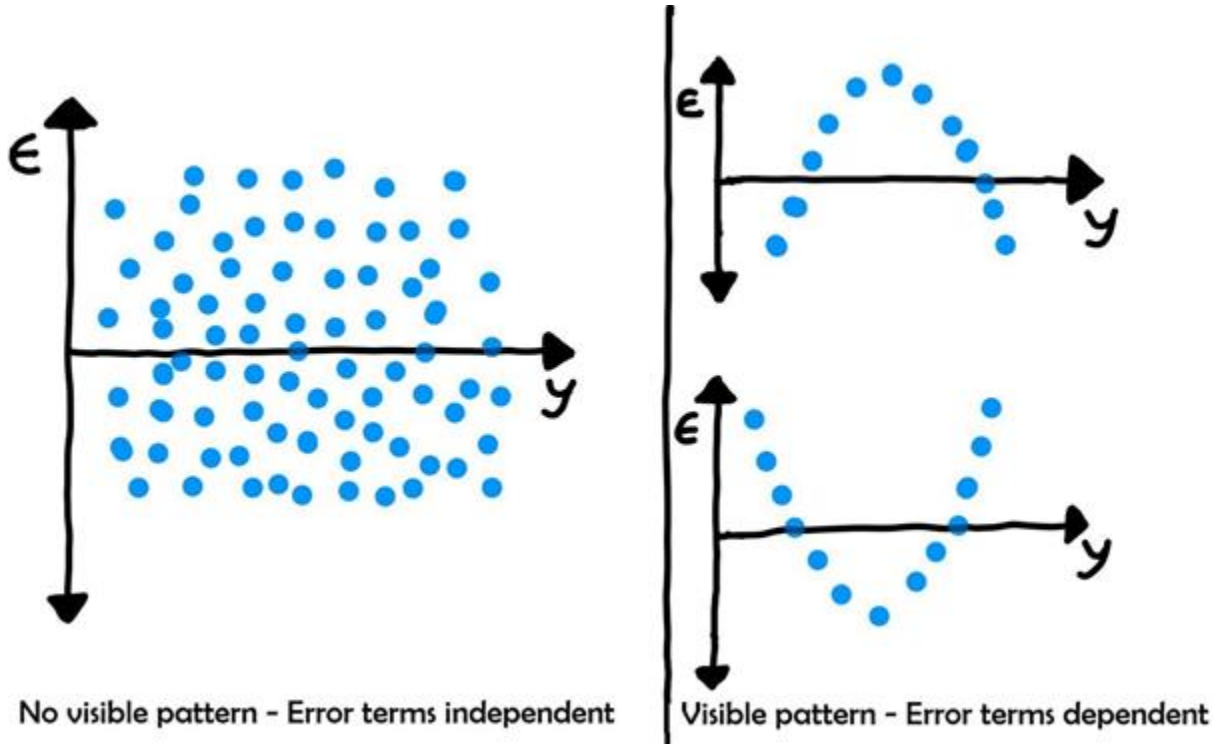
1. Linear relationship between X and y: X and Y should always display some sort of a linear relationship; otherwise, there will not be any use of fitting a linear model between them.



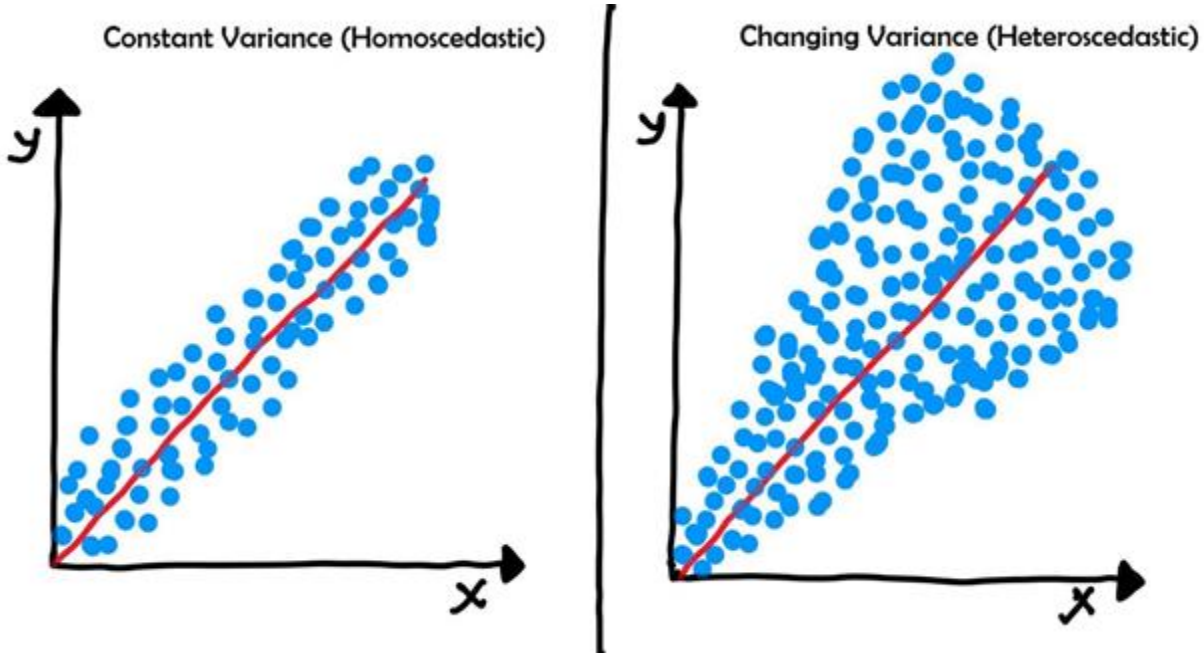
2. Normal distribution of error terms: It represents the assumption of normality. Which exhibits that error terms generally follow a normal distribution with mean equal to zero in most cases.



3. Independence of error terms: It explains that the error terms should not be dependent on one another. It means, there should not be any meaningful distribution between independent variable and error term.



4. Constant variance of error terms: This assumption says that the variance should not increase or decrease as the error values change. Also, the variance should not follow any pattern as the error terms change.



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

As per the analysis and final model, the top 3 features contributing significantly towards of share bikes are the below:

• temp(Positive correlation)

• Light Snow (negative correlation).

• year (negative correlation).

General Subjective Questions

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

Regress means "the act of going back" and Regression means “returning to a former state”. In Regression, we plot a graph between the variables which best fit the given data points. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).

To calculate best-fit line linear regression uses a traditional slope-intercept form. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.

It is mostly used for finding out the relationship between variables and forecasting.

• Regression estimates the relationship among variables for prediction.

• Regression analysis helps to understand how the dependent variable changes when some of the independent variables are varied, while the other independent variables are held fixed.

• It determines the relationship between one dependent variable and a few other independent variables. • Regression analysis also estimates the Optimum value, Conditional Expectation, Quantile, Probability Distribution etc. of the dependent variable given the independent variables.

• Dependent variables are also called as regress and, endogenous variable, response variable, measured variable or criterion variable.

• Similarly, independent variables are also called as regressors, exogenous variables, explanatory variables, covariates, input variables or predictor variables.

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

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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

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

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

Anscombe's quartet

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

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

3. What is Pearson's R? (3 marks)

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data.

It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. 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

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)

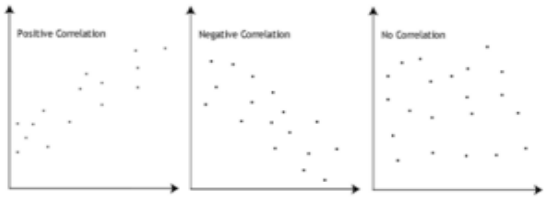
• r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

• 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



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r =correlation coefficient
- x_i =values of the x-variable in a sample
- $\bar{x}$ =mean of the values of the x-variable
- y_i =values of the y-variable in a sample
- $\bar{y}$ =mean of the values of the y-variable

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

It 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.

	Normalization	Standardization
•	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
•	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
•	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
•	It is really affected by outliers.	It is much less affected by outliers.
•	Scikit-Learn provides a transformer called **MinMaxScaler** for Normalization.	Scikit-Learn provides a transformer called **StandardScaler** for standardization.
•	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
•	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
•	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

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)

The value of VIF is calculated by the below formula:

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

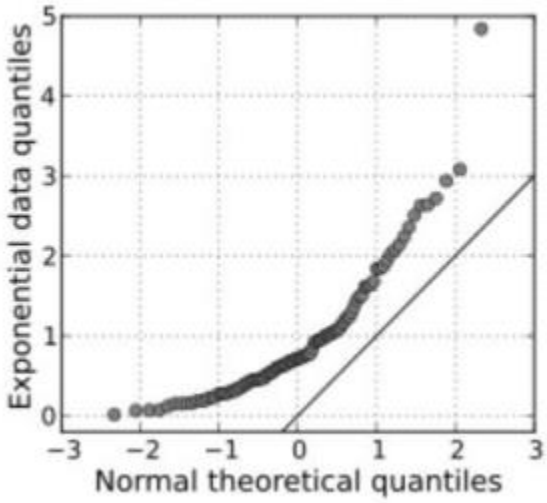
Where, 'i' refers to the ith variable.

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

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. It is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

A Q-Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

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