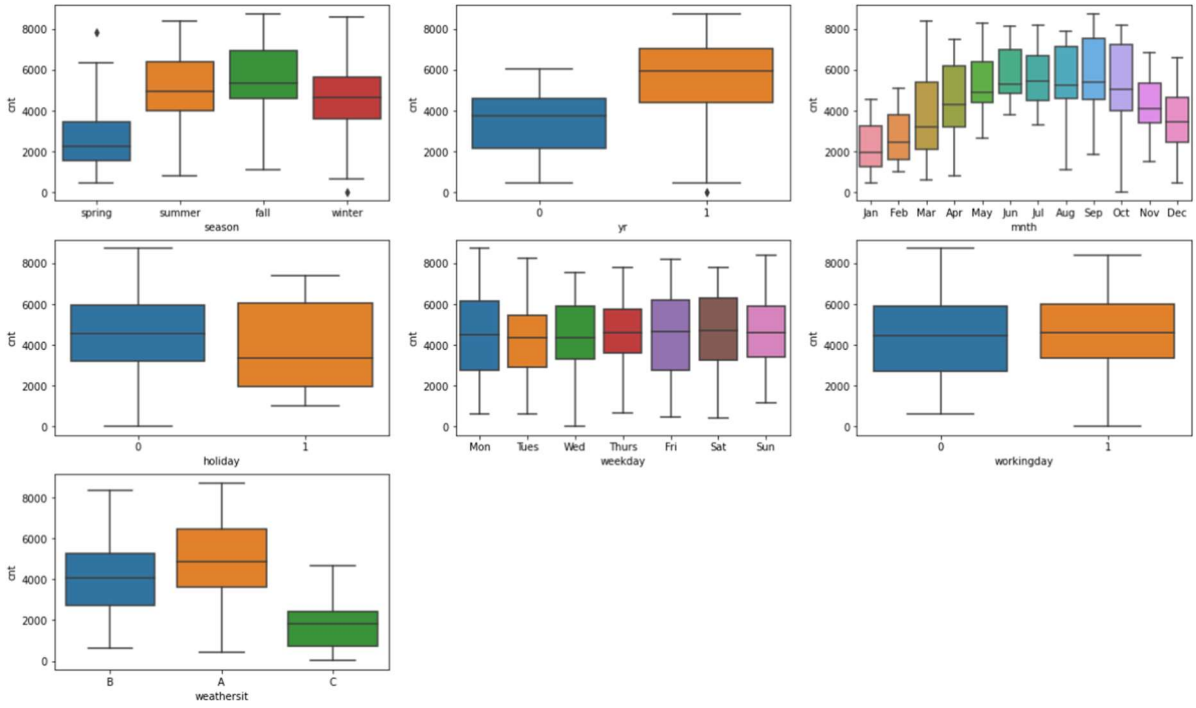
Assignment-based Subjective Questions

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 variables which have been identified as categorical variables are: season, year, month, holiday, weekday, workingday, weathersit. Potted boxplots to analyse categorical variables.



From the above graphs following are the inferences:

1. The demand for bike is high in fall and summer. It is comparatively low in spring.
2. The demand for bike was higher in 2019 as compared to 2018
3. The demand for bike is highest in Sep and lowest in Jan.
4. The demand for bike on holidays is low than compared with non holidays.
5. The demand for bike is higher on clear days than compared with cloudy and rainy days.
6. Working day and weekday does not have much impact on the count.
7. There is no record on Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog.

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

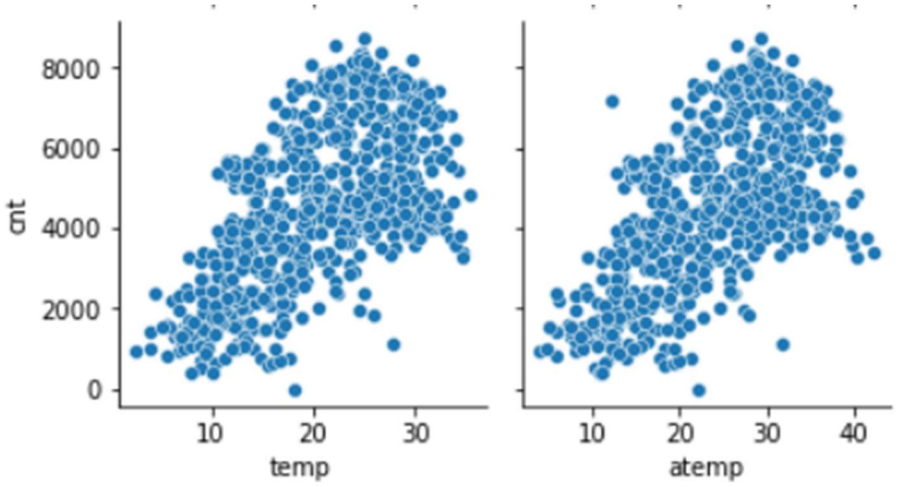
Ans: This is known as one hot encoding. It is a technique which converts categorical data into a form which is understandable by ml model. If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables. It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

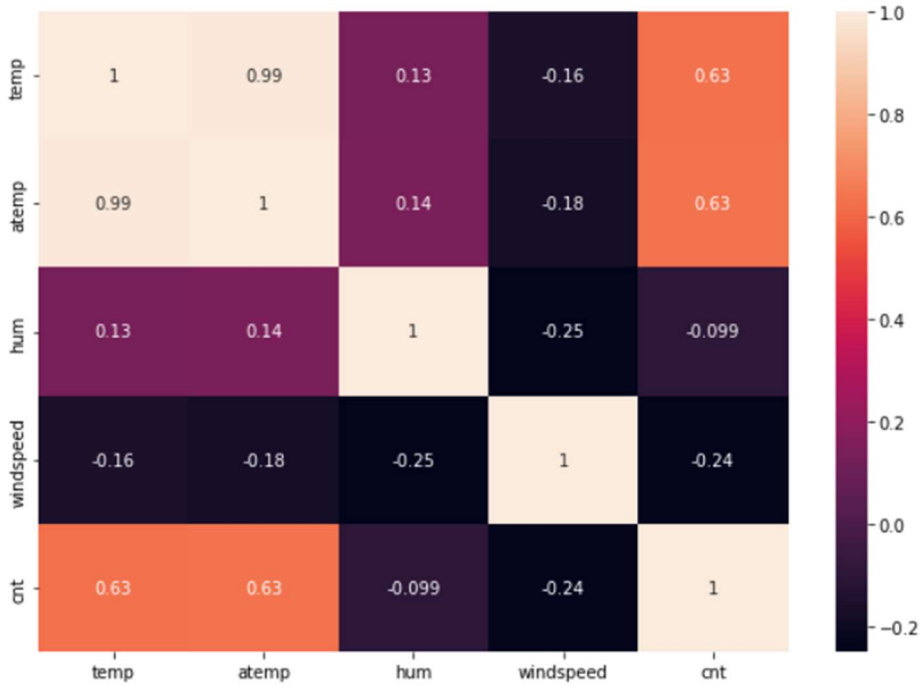
For example, the given dataset has season categorical variable with values spring, summer, fall, winter. If we create dummies by dropping fall column then values will look like below. Here, 000 represents Fall season.

Season	Season_spring	Season_summer	Season_winter
Spring	1	0	0
Summer	0	1	0
Winter	0	0	1
Fall	0	0	0

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

Ans: By looking at the pair plot among the numerical variables, 'temp' and 'atemp' are the two numerical variables which have highest correlation with the target variable 'cnt'. From the correlation plot also we see that 'temp' and 'atemp' have 0.63 value which shows highest correlation.

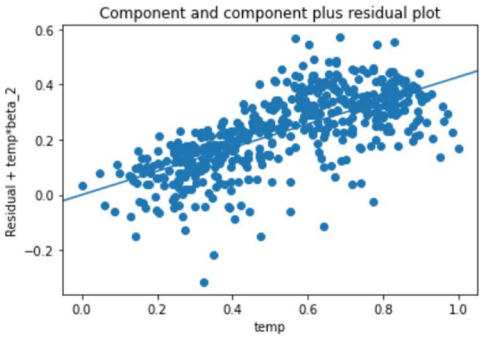


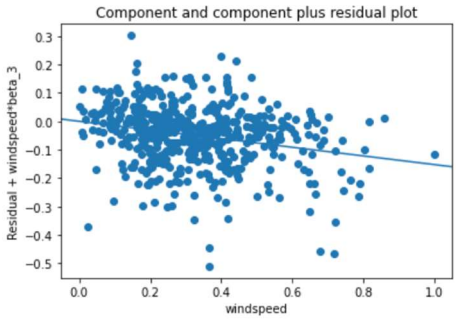


4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

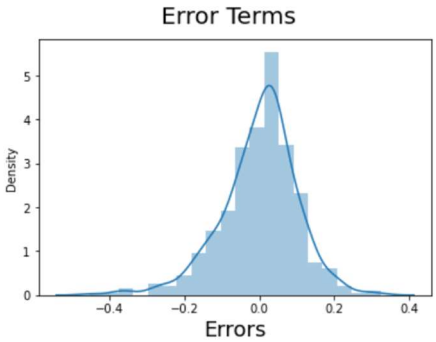
Ans: After building the model, the assumptions of Linear Regression have been validated based on below:

1. Linear Relationship
   Plotted ccpr plot to validate linear relationship assumption on two variables temp and windspeed. And found that linear relationship exists.

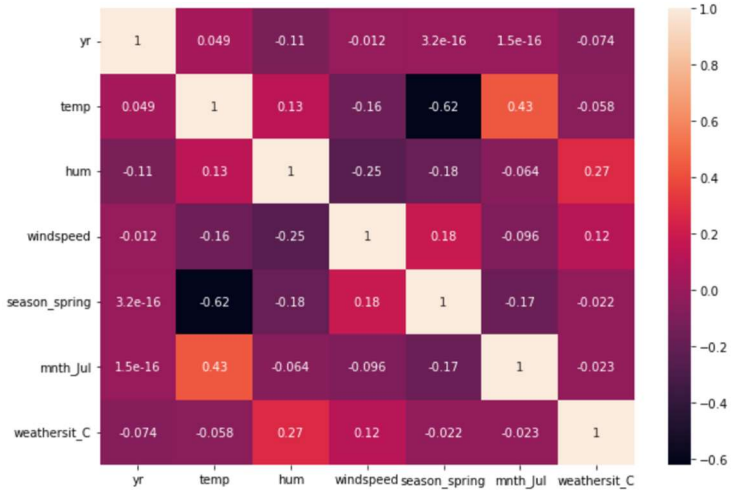




2. Normal distribution of error terms
   We validated this by performing residual analysis on the train data. And found that Error terms are normally distributed.

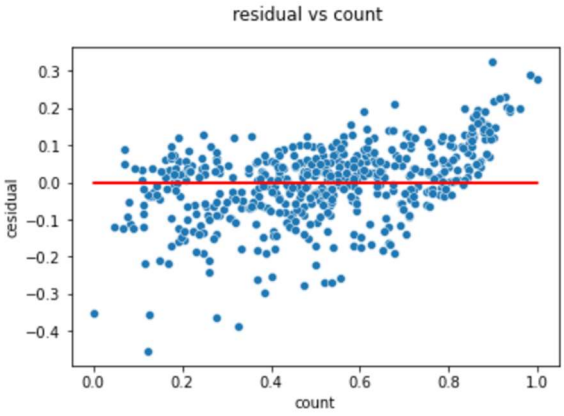


3. Little or No multicollinearity
   Plotted heatmap to understand the correlations between final predictor variables. Final predictor variables are not found correlated.



4. Homoscedasticity

   Homoscedasticity in a model means that the error is constant along the values of the dependent variable. Plotted residual vs count plot. And found that the points have constant deviation from zero-line which means homoscedasticity is present.



5. No Auto-correlation or independence

   Durbin-Watson statistic is a test statistic which is used to detect autocorrelation in residuals from the regression analysis. This statistic will always assume a value between 0 and 4. Below are the interpretations:
   DW = 2: no autocorrelation
   DW< 2: positive correlation
   DW > 2: negative correlation
   In our OLS Regression Results of the final model, the DW value is 1.895 which is approximately equal to 2. Hence, we say that there is little or no auto-correlation.

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: By looking at coefficients, top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. temp (coef = 0.4279)
2. year (coef = 0.2360)
3. weathersit_Light Snow, Light Rain (coef = -0.2413)

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Ans: Linear regression is one of the basic forms of machine learning model that analyses the linear relationship between a dependent variable with the given set of independent variables. Linear relationship means if there is any increase or decrease in the independent variable then there is increase/decrease in the dependent variable as well.

Mathematically, equation for the linear regression can be represented as,

Y = mX +C

Where,

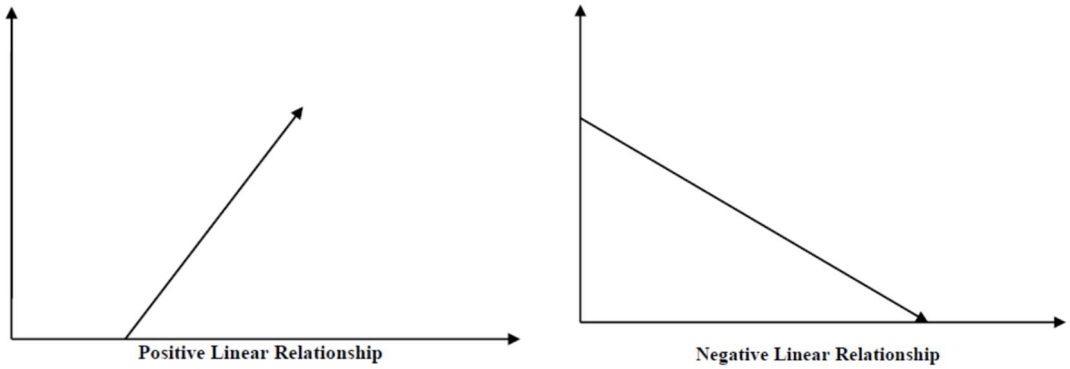
Y = dependent variable or target variable which we want to predict

m = slope of the regression line

X = independent variable

C= constant or intercept of the line If X=0 then Y = C

The linear regression relationship can be positive or negative.



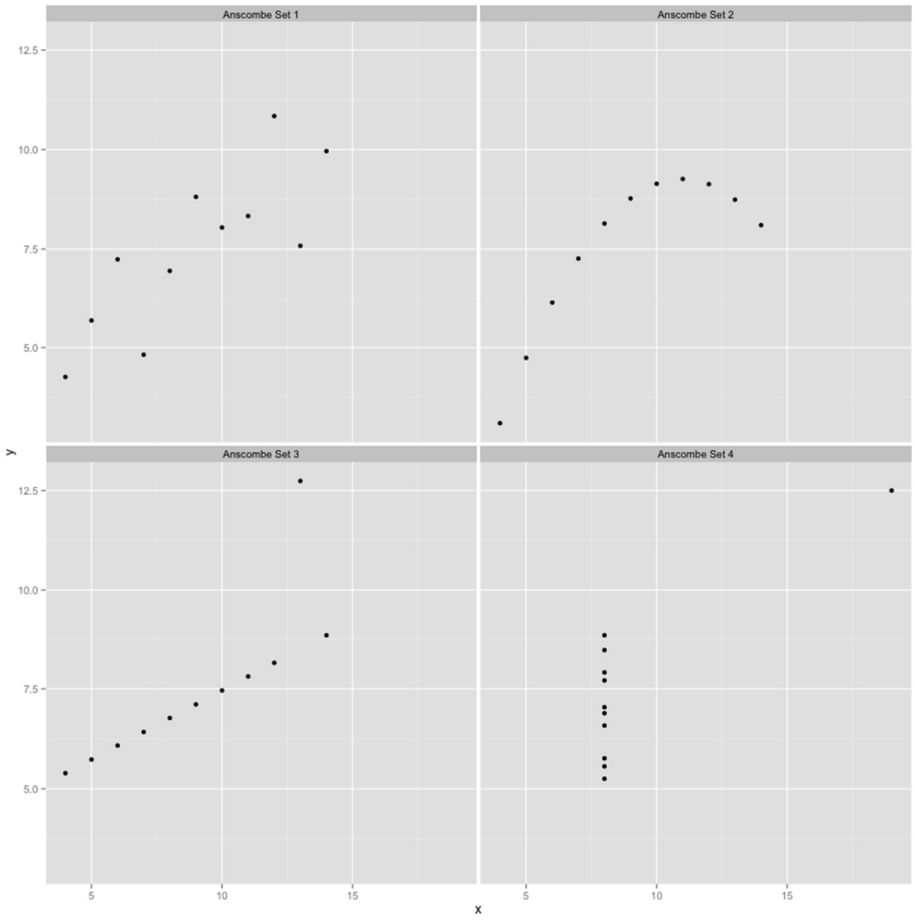
2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: 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.

These 4 sets of 11 data points are given below:

	x1	x2	x3	x4	y1	y2	y3	y4
1	10.00	10.00	10.00	8.00	8.04	9.14	7.46	6.58
2	8.00	8.00	8.00	8.00	6.95	8.14	6.77	5.76
3	13.00	13.00	13.00	8.00	7.58	8.74	12.74	7.71
4	9.00	9.00	9.00	8.00	8.81	8.77	7.11	8.84
5	11.00	11.00	11.00	8.00	8.33	9.26	7.81	8.47
6	14.00	14.00	14.00	8.00	9.96	8.10	8.84	7.04
7	6.00	6.00	6.00	8.00	7.24	6.13	6.08	5.25
8	4.00	4.00	4.00	19.00	4.26	3.10	5.39	12.50
9	12.00	12.00	12.00	8.00	10.84	9.13	8.15	5.56
10	7.00	7.00	7.00	8.00	4.82	7.26	6.42	7.91
11	5.00	5.00	5.00	8.00	5.68	4.74	5.73	6.89

After plotting these points, we can see different graphs for each set.



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

Ans: Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. 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)

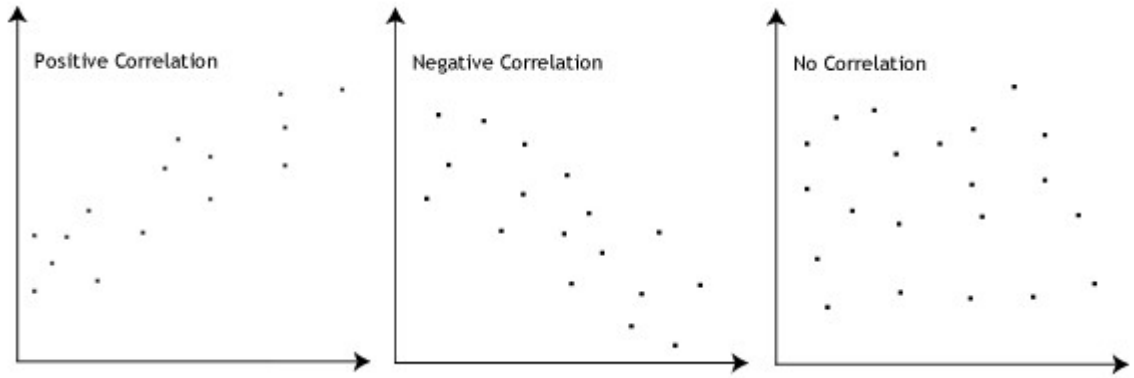
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

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



Person r formula

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

Where,
r = correlation coefficient
xi = values of the x variable in a sample
x bar = mean of x values
yi = values of y variable in a sample
y bar = mean of y values

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

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. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
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-squared, etc.

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.

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

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).

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

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

Ans: VIF is variance inflation factor that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity and is given by following formula

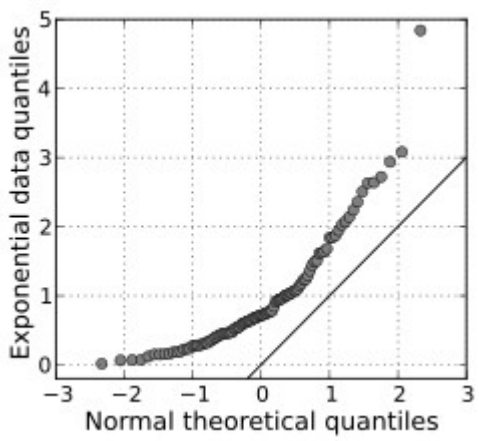
VIF = 1/1-R^2.

So, if R^2 is 1, VIF becomes infinity. This happens because of perfect correlation between two independent variables. To solve this, we have to drop one of the variables from the dataset causing the perfect multicollinearity.

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

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.

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.