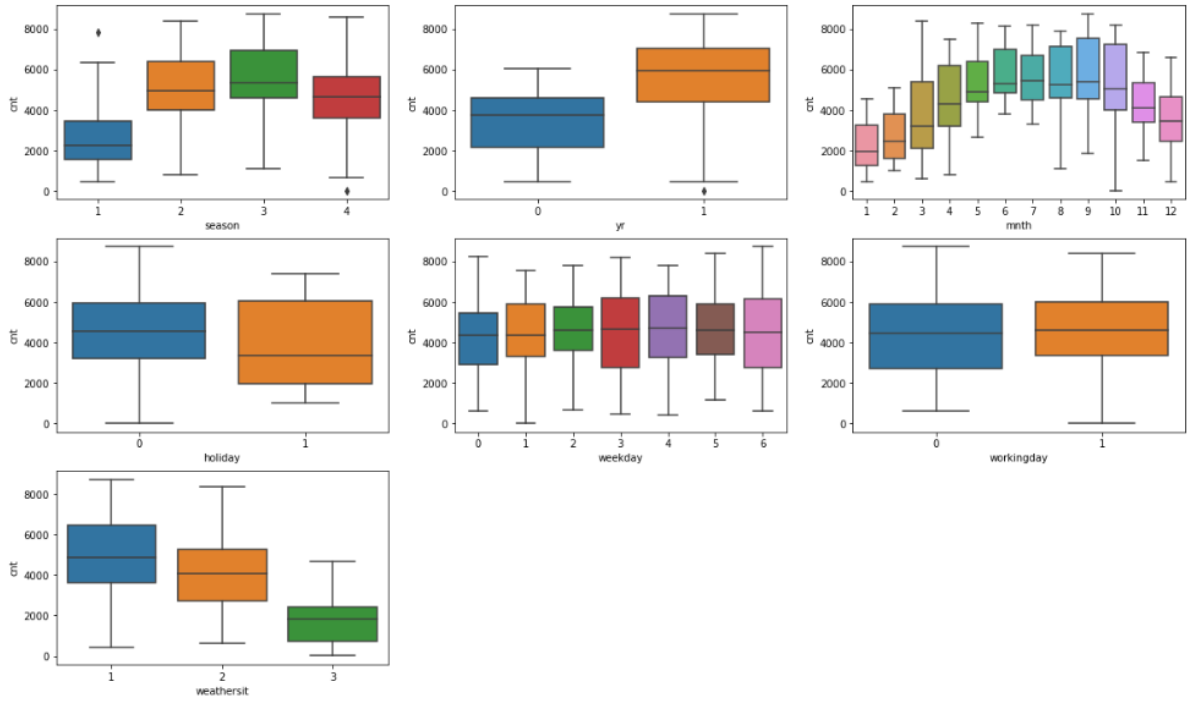
Assignment based Subjective Questions:

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

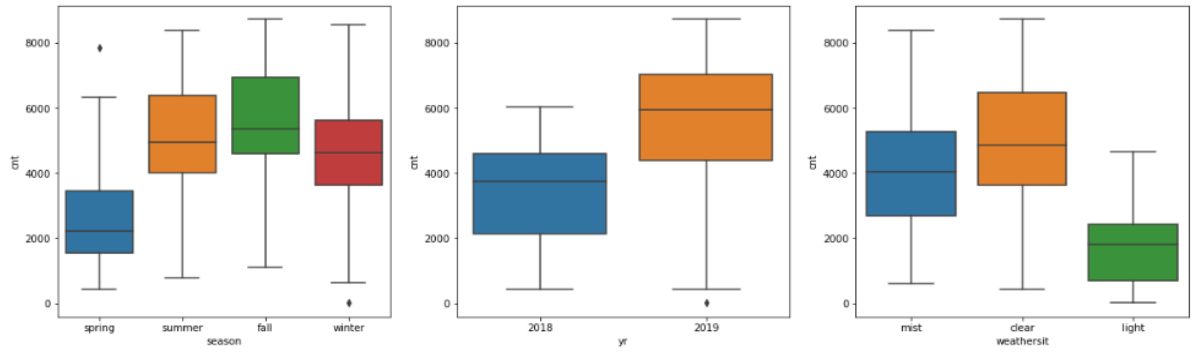
The categorical variables given in the dataset has significant correlation with the dependent variable 'cnt'.
Some observations about the effect of categorical variable on the dependent variable 'cnt' are:

1. The demand for bikes increased in the year 2019 than that was in the year 2018.
2. The demand for bikes increased when the weather was clear (weathersit = 1).
3. The demand for bikes increased during the fall season (season = 3) and fallen in spring (season =1).

For more details, please see the box charts below.



The categorical variables given had numerical values which were explained in the dictionary. However, it was essential to impute actual values which should make the variable interpretable. Hence replaced the actual values instead of keeping the numeric values during the data analysis stage.



Q2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

During dummy variable creation in pandas, the drop_first=True is used to drop a redundant dummy variable.

For example, if a variable is represented by 3 levels (i.e., levels = k), the dummy variable creation process will create 3 dummy variables. However, to represent all the 3 levels, two dummy variables are enough (dummy variables required = k-1). Thus, dropping first column is done through parameter drop_first=True.

Example:

Original data -

	status
0	open
1	open
2	in-process
3	closed
4	closed

After Dummy variable creation (pd.get_dummies()) -

	open	**in-process**	**closed**
0	1	0	0
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	1

After dropping a redundant dummy variable (drop_first=True) -

	in-process	**closed**
0	0	0
1	0	0
2	1	0
3	0	1
4	0	1

Thus,

‘open’ will be represented by 00
‘in-process’ will be represented by 10
‘closed’ will be represented by 01

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

The ‘atemp’ variable has highest correlation with the target variable ‘cnt’ by looking at the pair plot.

Note: The variables ‘casual’ and ‘registered’ are highly correlated to the target variable ‘cnt’ but as we know the ‘cnt’ is formed by summation of ‘casual’ and ‘registered’ variables, hence they are discarded as they are not giving additional information.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The following assumptions are validated using following methods:

1. With the help of scatter plot checked the relationship between predictor variable and dependent variable, which is linear.
2. The error terms must be normally distributed. Plotted a distribution plot which shows that the mean is at 0 and error terms are normally distributed.
3. There was no correlation between the residual (error) terms by making a plot for residual vs Y_train_pred
4. The error terms must have constant variance (Heteroscedasticity). Checked using scatter plot between error terms vs Y_train_pred.
5. The independent variables should not be correlated which is checked by checking VIFs (multicollinearity).

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

After building models using two methods, i.e. starting with most significant variable and then adding further significant variables to the model, and the other method is to start with all the variables and drop off the insignificant variables from the model. Found that following three variables were contributing significantly towards explaining demand of the shared bikes:

1. Actual temperature
2. Year of operation – The more time the bike company is in the market, it is gaining demand of its bike
3. Clear Weather – When the weather is clear, the demand for the bike is more

General Subjective Questions:

Q1. Explain the linear regression algorithm in detail. (4 marks)

In linear regression, the relationship between the independent or predictor variable(s) and the dependent or target variables is found to be linear.

If the predictor variable is single then it is called as simple linear regression and when the predictor variables are two or more than two then it is called as multiple linear regression.

The linear regression is the most basic form of regression in machine learning however it is highly used in most of the real-world situations.

In the simplest terms, if the relationship predictor variable and dependent variable is found to be linear then it is possible to predict the value of dependent variable for the given predictor variable as long as it is interpolated.

There are few assumptions made for linear regression:

1. The relationship between dependent and predictor variable is linear.
2. The error terms must be normally distributed.
3. There should be no correlation between error terms.
4. The error terms must have constant variance.
5. In multiple linear regression, the independent variables should not be correlated.

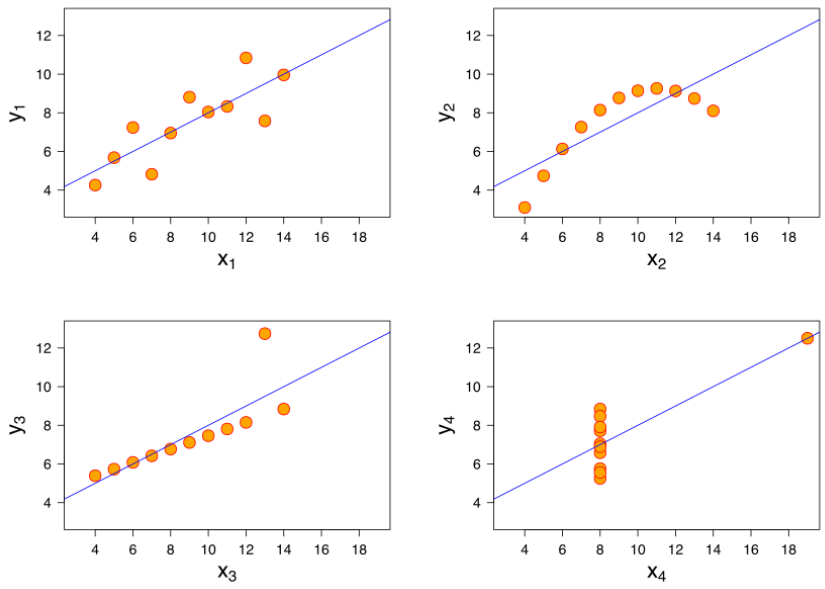
The steps to follow to build linear regression model:

1. Reading and understanding the data
2. Data preparation
3. Splitting the data into Training and Testing sets
4. Build a linear model
5. Residual analysis of the train data
6. Making prediction using the final model and test data
7. Evaluate the model

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

In the year 1973, the statistician Francis Anscombe demonstrated both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, but have different distributions and appear very different when graphed. Please see below:



He presented his findings in a paper using 4 datasets consisting of eleven (x, y) points.

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

He presented this paper to counter the impression among statistician that “numerical calculations are exact, but graphs are rough”.

Q3. What is Pearson’s R? (3 marks)

The Pearson’s correlation coefficient – also known as Pearson’s R, the bivariate correlation or simply called as the correlation coefficient.

Pearson’s R is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and 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.

1 mean two variables are positively correlated. -1 mean that they are negatively correlated.

For a population:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

The formula for ρ can be expressed in terms of mean and expectation. Since[10]

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

where:

- σ_Y and σ_X are defined as above
- μ_X is the mean of X
- μ_Y is the mean of Y
- $\mathbb{E}$ is the expectation.

The formula for ρ can be expressed in terms of uncentered moments. Since

$$\begin{aligned}
\mu_X &= \mathbb{E}[\,X\,] \\
\mu_Y &= \mathbb{E}[\,Y\,] \\
\sigma_X^2 &= \mathbb{E}[\,(X - \mathbb{E}[X])^2\,] = \mathbb{E}[\,X^2\,] - (\mathbb{E}[\,X\,])^2 \\
\sigma_Y^2 &= \mathbb{E}[\,(Y - \mathbb{E}[Y])^2\,] = \mathbb{E}[\,Y^2\,] - (\,\mathbb{E}[\,Y\,])^2 \\
\mathbb{E}[\,(X-\mu_X)\,(Y-\mu_Y)\,] &= \mathbb{E}[\,(X-\mathbb{E}[\,X\,])\,(Y-\mathbb{E}[\,Y\,])\,] = \mathbb{E}[\,XY\,] - \mathbb{E}[\,X\,]\,\mathbb{E}[\,Y\,],
\end{aligned}$$

the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{\mathbb{E}[\,XY\,] - \mathbb{E}[\,X\,]\,\mathbb{E}[\,Y\,]}{\sqrt{\mathbb{E}[\,X^2\,] - (\mathbb{E}[\,X\,])^2}\,\sqrt{\mathbb{E}[\,Y^2\,] - (\mathbb{E}[\,Y\,])^2}}.$$

For a sample:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Rearranging gives us this formula for r_{xy}:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

where n, x_i, y_i are defined as above.

This formula suggests a convenient single-pass algorithm for calculating sample correlations, though depending on

Rearranging again gives us this[10] formula for r_{xy}:

$$r_{xy} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2}\sqrt{\sum_i y_i^2 - n\bar{y}^2}}.$$

where $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above.

An equivalent expression gives the formula for r_{xy} as the mean of the products of the standard scores as follows:

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where:

- $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above, and s_x, s_y are defined below
- $\left(\frac{x_i - \bar{x}}{s_x}\right)$ is the standard score (and analogously for the standard score of y)

Alternative formulae for r_{xy} are also available. For example. one can use the following formula for r_{xy}:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

where:

- $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above and:
- $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (the sample standard deviation); and analogously for s_y

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

While building a linear regression model, when we optimize the cost function, if the variable's value range is different e.g., one variable is having its values between 1 and 5 and other variable's values are between 10000 and 50000 then at the backend the optimization will take longer time to process. To remove this constraint, the scaling is performed. The efficiency of the system will increase if we scale the values.

There are two types of scaling:

1. Min-Max scaling (also called as normalization)
2. Standardization (mean = 0 and std dev = 1)

In Min-Max or Normalized scaling all the variable values are fit between 0 and 1. Whereas in the standardization method the data is fit into a curve having normally distributed data with mean = 0 and srd. Dev = 1.

Normalization formula:

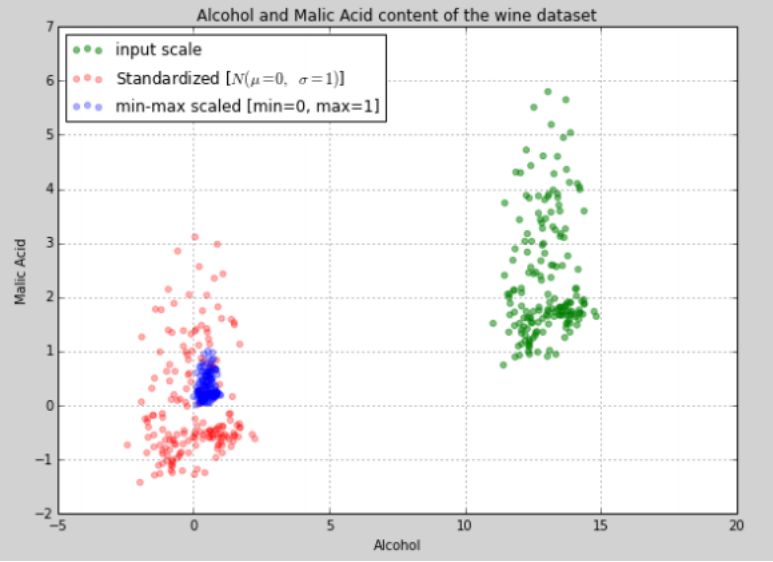
X_new = (X – X_min) / (X_max – X_min)

Standardization formula:

X_new = (X – mean) / Std Dev

SN	Normalization	Standardization
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling
2	It is used when features are of different scales	It is used when we want to ensure zero mean and unit standard deviation
3	Scales values between [0,1] or [-1,1]	It is not bound to a certain range
4	It is affected by outliers	It is less affected by outliers
5	Scikit-learn provides transformer method called MinMaxScaler	Scikit-learn provides a transformer method called StandardScalar for standardization
6	This is useful when we do not know about the distribution	This is useful when the feature distribution is Normal or Gaussian
7	It is often called as Scaling Normalization	It is often called as Z-Score Normalization

The following picture shows the input variables (green dots) are scaled using Normalization method (blue dots) and using Standardization method (red dots)



Q5. 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 = if (infinity). A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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) i.e. 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 VIF < 5 indicates that the feature is not multicollinear with other independent variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

In statistics Q-Q plot (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

The main step in constructing a Q–Q plot is calculating or estimating the quantiles to be plotted. If one or both of the axes in a Q–Q plot is based on a theoretical distribution with a continuous cumulative distribution function (CDF), all quantiles are uniquely defined and can be obtained by inverting the CDF. If a theoretical probability distribution with a discontinuous CDF is one of the two distributions being compared, some of the quantiles may not be defined, so an interpolated quantile may be plotted. If the Q–Q plot is based on data, there are multiple quantile estimators in use. Rules for forming Q–Q plots when quantiles must be estimated or interpolated are called plotting positions.

A simple case is where one has two data sets of the same size. In that case, to make the Q–Q plot, one orders each set in increasing order, then pairs off and plots the corresponding values. A more complicated construction is the case where two data sets of different sizes are being compared. To construct the Q–Q plot in this case, it is necessary to use an interpolated quantile estimate so that quantiles corresponding to the same underlying probability can be constructed.