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)

There are the inferences:

Season: Fall season seemed to have the highest count of bike sharings

Year: 2019 seems to have a higher amount of bike sharings

Month: October month seems to have the highest bike sharings

Holidays have a higher bike sharing

Bike sharing is significantly higher during clear weather followed by misty + cloudy days. There is absolutely no bikes borrowed when there is heavy rains with thunderstorms.

Nothing significant was observed if the bike was borrowed on a weekday or working day

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

Categorical variables cannot be used in the Linear Regression model. Hence to convert the Categorical variables into numerical values, we create dummy variables that are binary (0 and 1) representation of the value of categorical variables.

We need to use the **drop_first=True** option while creating dummy variable to eliminate redundance. If I have n different values in my categorical variable column, I will only need ‘n -1’ dummy variables to represent the n different values.

Ex: My data has the following categorical variable having 3 distinct values

Religion
Muslim
Christian
Other

It is sufficient for me to create 2 dummy variables as below

Muslim	Christian	
1	0	This row represents Muslim
0	1	This row represents Muslim
0	0	When both are 0, it means Others

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

The highest correlation with the target variable, cnt, is with the independent variable **temp**

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

   The assumptions of Linear Regression are:

   a. There should be a linear relationship between dependent and independent variables.
      Using pairplot, we checked if there were any linear relationships that existed between the independent variable and target variable **cnt**
   b. Residuals must be normally distributed.
      In the Step 4 of Residual analysis, distribution plot for error terms show normal distribution with mean at 0.
   c. Error terms must be independent.
      In Step 4 of Residual Analysis, a scatter plot of residuals show that there is no visible pattern between error terms and hence are independent.

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 the demand of shared bikes are:

   a. Temp: temperature in Celsius – higher the temperature, higher the demand
   b. Season: winter season
   c. Month : January

General Subjective Questions

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

Linear Regression is one of the Machine Learning algorithms where we predict the result/output using known parameters or variables which have a correlation with the output. It is used to predict values in a continuous range. A continuous and constant slope is made with a line which will be the best fit for all possible input values.

Linear regression falls under the category of Supervised Learning which means that you will have past data that is labeled, and prediction of a result is based on this past data.

Examples of Linear regression include, forecasting weather and temperature, predicting the score of a cohort of students in board exams, sales predictions etc.

This is the simplest Machine Learning Algorithms and can be used to solve problems in many fields easily.

We have a few assumptions for using Linear Regression:

a. There should be a linear relationship between dependent and independent variables.

b. Results obtained from an observation should not be related to previous observation

c. Residuals must be normally distributed.

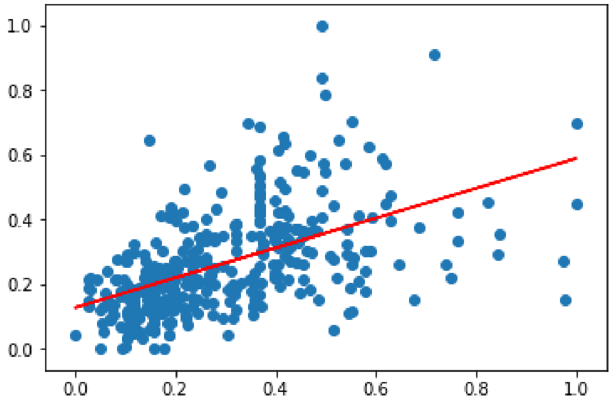
d. Error terms must be independent.

e. Data must be homoscedastic, which mean there should not be much variation between the results.

Linear regression can always be plotted and visualized as a graph to make predictions about data. This leads to creating a best fit line for all the data points. A trend line is created for Linear Regression which shows the correlation between the dependent and independent variables.

The more the data, better is the accuracy of the evaluation.

An example for the best fit line for Linear Regression is given below



If we have only one independent variable that is linearly related to the dependent variable, this is a **Simple Linear Regression**. The best fit line for Simple Linear Regression is represented modelled by the equation for a straight line as $y = mx + c$ or $y = \beta_0 + \beta_1 x_1$

The goal is to find the best value for the co-efficients, β_0 **and** β_1

Properties of Linear Regression:

a. Regression line passes through the mean of independent and dependent variables.

b. The best fit Regression line is the one that minimizes/reduces the sum of Squares of Residual (RSS – Residual Sum of Squares), which is why Linear Regression is also termed as Ordinary Least Square (OLS)

c. If the value of x is changed by 1 unit, the amount of change in y is β_1

After a Regression model is built, we will need to check the amount of variance that our model could explain. This is done using the feature R-squared.

$$R^2 = \frac{TSS - RSS}{TSS}$$

RSS = $\Sigma\left(y - y_{pred}\right)^2$ TSS = $\Sigma(y - \bar{y})^2$

Properties of R-squared

a. Range of R^2 is between 0 and 1

b. Higher the value of R^2, better is the correlation between the independent and dependent variables.

Multiple Linear Regression is an algorithm used when we have more than 1 independent variables

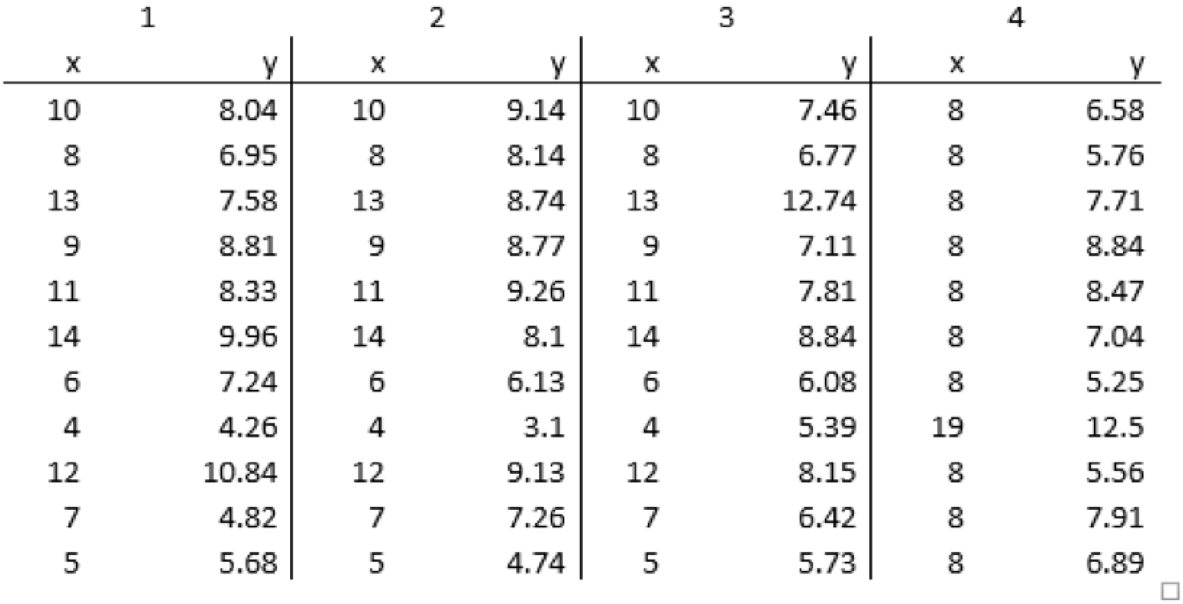
With Multiple Linear Regression comes the issue of Multicollinearity. Multicollinearity is the condition of very high correlation between independent variables. Variance Inflation Factor is the method used to identify Multicollinearity. A VIF >5 depicts high correlation, and we exclude that from our model.

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

According to the definition given in **Wikipedia**, Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of 11 x, y co-ordinates . This was constructed by a statistician named Francis Anscombe to demonstrate the importance of graphing data before we start analyzing and also to explain the effect of outliers.

Even though constructing a table with all data points might be useful, visualization gives us a better representation.

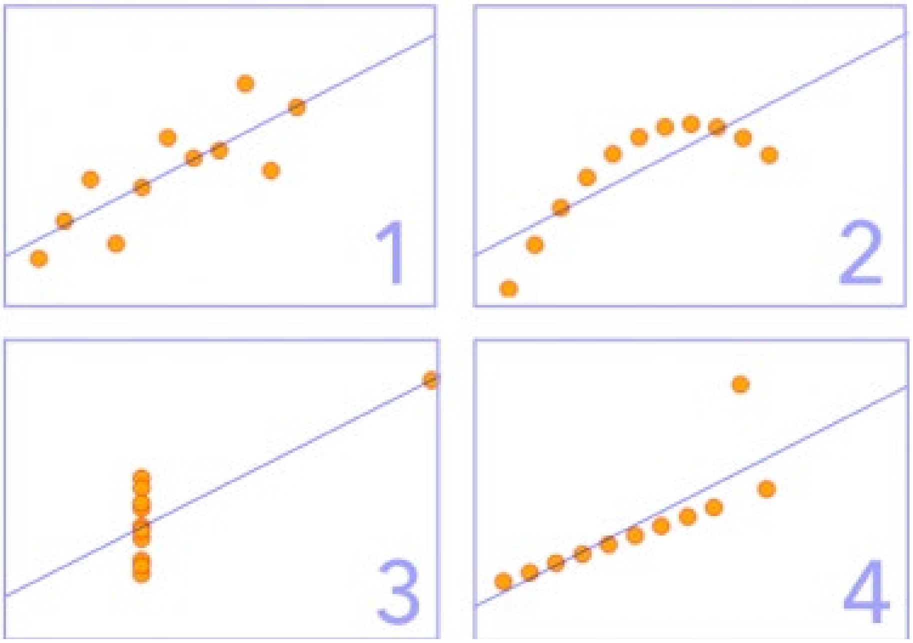
Here is the dataset



1		2		3		4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

This data set is divided into four groups where each group has eleven points with a x and a y value. From a glance at the different groups it seems like the last group is the easiest to understand, but it would be hard to say how these four groups differ from each other.

Visualization



Explanation of the graphs:

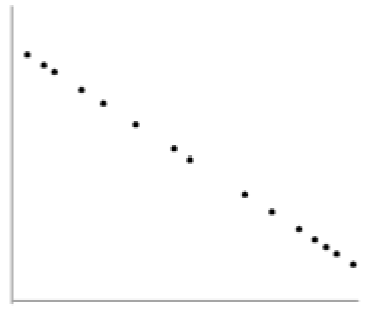
1) there seems to be a linear relationship between x and y.
2) there is a non-linear relationship between x and y.
3) there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
4) one high-leverage point is enough to produce a high correlation coefficient.

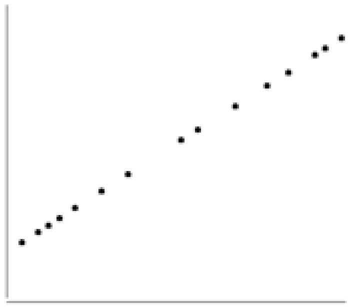
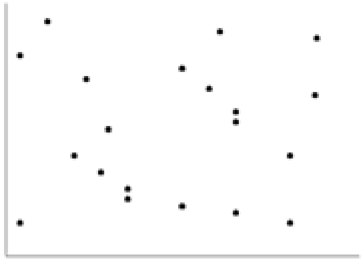
3. What is Pearson's R? (3 marks)

Pearson's R also known as Pearson's correlation co-efficient (r) is the measure of the correlation between two variables. This statistic tool is only applicable on continuous data ranges having linear relationship and hence widely used in Linear Regression.

The range for Pearson's correlation co-efficient is between -1 and +1.

A scatter plot gives a clear understanding of the co-efficients

	Here the correlation co-efficient, r = -1, negative correlation

	r = +1, positive correlation
	r = 0, no relation ship/correlation

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

Scaling is a technique used in Linear Regression to standardize or normalize the independent variables within a fixed range

We might have various kinds of numerical data in our data set and each variable might have different ranges. Ex. One of the variable may range from 0 to 1, while the other may range between 500 and 10000. In such cases, the Gradient Descent might take a longer time to converge to the local minimum to find the optimum value of the co-efficients $\boldsymbol{\beta_0}$ **and** $\boldsymbol{\beta_1}$.

To avoid this, and for the algorithm to perform faster, we use Feature Scaling. Scaling is affects only the co-efficients and does not impact the parameters such as t-statistic, F-statistic, p-values, R-squared, etc.

We either Normalize (also known as MinMax Scaling) or Standardize the data.

Standardisation brings all the data into a standard normal distribution with mean 0 and standard deviation/ variance = 1.

MinMax scaling, brings all the data in the range of 0 and 1.

The formulae in the background used for each of these methods are as given below:

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

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

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

VIF is a measure or index of multicollinearity of a variable in a given dataset. In a Regression model which has been built using several independent variables, some of these variables might be interrelated, due to which the presence of that variable in the model is redundant. To address this issue, we use the value of VIF to determine whether or not to retain that variable.

The value of VIF can range in between 1 and Infinity.

The formula for VIF is given by

$$\boldsymbol{VIF = \frac{1}{1 - R^2}}$$

VIF will have the value of Infinity when the value of R-squared = 1, this means that the variable is highly correlated with other variables in the data set.

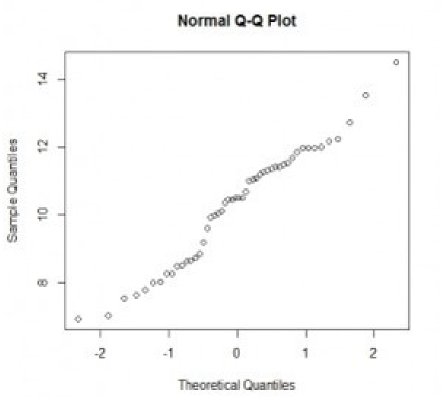
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

A Q-Q plot, or Quantile-Quantile plot is a graphical tool that helps us in assess and infer if the data came from some theoretical distribution, such as normal or exponential or uniform distribution.
In Linear Regression, if the training and test data set is obtained separately, we can use the Q-Q plot to confirm if both data sets are from the population with common distribution.

Q-Q plots take our sample data, sorts it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution.
If the points fall pretty closely along the line, the data are normal.

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.

Example of a Q-Q plot is shown below



To be able to run the code that creates a Q-Q plot, you need to install and load the package stats.