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: From the best fitted line linear equation, I can see that the bad weather like Cloudy Misty Weather and Light Snowy Rainy Weather negatively effects the demand for bike sharing while seasons Summer and winter have a positive impact. Th demand for bike sharing seems to be positive in the month of September.

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

Ans: During dummy variable creation, we drop the first value because the model can infer the dropped column based on the values of the other columns. So, when we are dealing with categorical variables, it is good if we drop them so that we have lesser variables to deal with.

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

Ans: From the equation, temperature seems to have the highest correlation with the total demand for bike sharing. My model predicts that to be 0.55

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

Ans: These are the assumptions of simple linear regression

a. Linear relationship between the target variable and the final dependent variables identified from the exercise.
b. We have checked if the error terms are normally distributed by plotting a graph.
c. We have checked if the dependent variables are not collinear.
d. Error terms should have constant variance which is also called as homoscedasticity.

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: From the final linear regression equation, the top 3 features which contribute to shared bikes is 1. Temperature, 2. Light snowy rainy weather (negative) and 3. year_2019.

General Subjective Questions

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

Ans: Linear regression is a form of predictive modelling technique which tells us the relationship between the target variable also called as dependent variable and its predictors or independent variables.

a. Simple Linear Regression: This is a basic linear regression which uses a straight line that passes through a dependent variable and one independent variable. We create a scatter plot of these two variables.

Then we try to find the so called a best fit line by minimising the Residual Sum of Squares.

Residual Sum of Squares is square of the difference of the predicted y value to the actual y value.

The strength of the best fit line is calculated by 2 metrics:
1. R2 or coefficient of Determination
2. Residual Standard Error (RSE)

R2 or Coefficient of Determination
To understand the R2 statistics value, there is another term which needs to be understood. TSS or Total Sum of Squares. If ymean is the mean of all the y values then the sum of all the squares of the difference of individual y values and ymean is the TSS. TSS gives us the deviation of all the points from the mean line

TSS = sum(square(yi – ymean))

R2 = 1 – (RSS/TSS)

So, for the best fit line, the R2 value will be higher compared to another line which is draw through the values of predicted x and y.

Assumptions of Simple Linear Regression:

a. There is a linear relationship between the dependent and the independent variables.
b. Error terms are normally distributed at each of the values on the linear regression best fit line.
c. Error terms are independent of each other.
d. Error terms have a constant variance which is also called as homoscedasticity.

There is NO assumption on the distribution of X and Y.

Analysing the Residuals: Apart from checking if the error terms are normal, we have to also check if the error terms have a constant variance by plotting a graph.

Hypothesis Testing of the slope or the Beta coefficient.
We need to perform hypothesis test on the Beta coefficient. The Null and Alternate hypothesis are:

Ho : Beta coefficient = 0
H_A : Beta coefficient != 0

We use the t-distribution with (n-2) degrees of freedom. The p-value is then calculated to determine if the coefficient is significant or not.

If p-value is less than 0.05 then the null hypothesis is rejected and we conclude that the Beta coefficient is significant.

The parameters to assess a model are:

1. t statistic: p-value less than 0.05
2. F statistic: higher the value, more significant the model
3. R-squared: value ranges between 0 and 1 and more close to 1, the higher the significance of the model.

Understanding Multiple Linear Regression: We try to understand the relationship between one dependent variable and multiple independent variables. We want to find all the independent significant variables that can best determine the dependent variable.

1. Model will fit a hyperplane instead of a line.
2. Coefficients will still be obtained by minimizing the sum of squares.
3. Linear regression assumption hold true.

Some new things to keep in mind.

1. Adding more variables is not helpful. We should narrow down to the least number of variables which give us a good fit. Or else, the problems can be
    a. Overfit where the models end up memorizing the data points in the training set and for a test set, the accuracy will drop.
    b. Multicollinearity: Effect of independent variables on each other. When one variable can explain another variable, better to get rid of one of them.

Multicollinearity affects the Interpretation and Inference. Interpretation that change in Y, when all others are held constant apply? And for Inference, coefficients swing widly and signs can invert. P-values are therefore not reliable.

But Multicollinearity doesn't affect the precision of the prediction and goodness of fit such as R-squared.

There are two ways to detect Multicollinearity.

1. Correlations: Drawing a heatmap of the correlations.
2. Variance Inflation Factor (Vif): Generally a VIF of less than 5 is desirable and greater than 10 is definitely high.
    Vif = 1 / (1 – square(Ri))

To deal with multicollinearity,

a. Dropping variables:
    a. Drop variables one by one which are highly correlated with others.
    b. Pick the variable which can interpret the business well.
b. Create new variables:
    a. Add interaction features
    b. Variable transformations

Feature Scaling: If the values are many and are on different scales, it becomes very difficult to compare the coefficients. So we need to scale them to the same scale.

We can scale features using

a. Standardizing: the variables are scaled so that their mean is zero and standard deviation is one.
b. MinMaxScaling: All the values are converted in such a way that they fall between 0 and 1.

Handling Categorical variables: Categorical variables need to be converted to dummy variables. If there are n categorical values, then create n-1 categorical variables.

We use AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) values for selecting categorical variables. BIC penalises for adding more variables.

Feature Selection:
There are 2 approches.

1. Manual Selection:
    a. Add Features(Forward): Add features one by one and check the strength of the linear regression and proceed ahead.
    b. Remove Features(Backward): Add all features and start removing them and then analyse the best fit.
2. Automated Approach:
    a. Use Recursive Feature Elimination.
3. There is a third hybrid approach which uses both automated and manual selections.

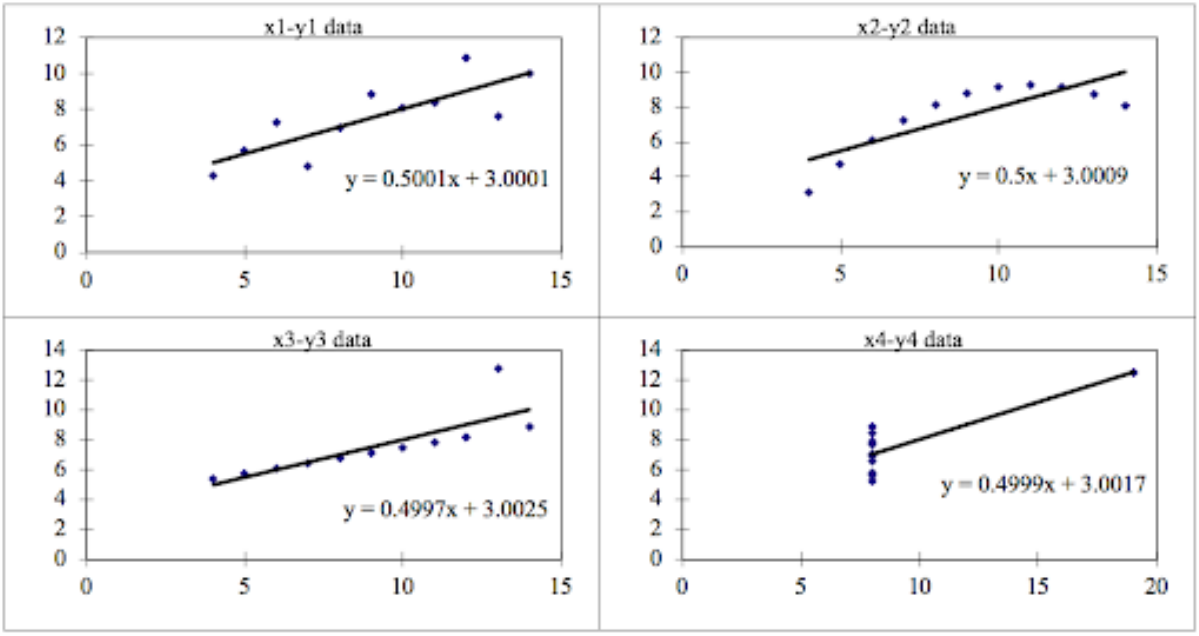
2. <u>Explain the Anscombe's quartet in detail. (3 marks)</u>
   Ans: Anscombe's quartet was explained by statistician Francis Anscombe to elaborate the importance of plotting data before one analyses it and builds a model. These four data sets have nearly the same statistical observations which provide the same information about the variance, mean for each of the x and y co-ordinates. The difference is evident only when they are plotted.

   He explains that data can be generated which has the same mean, variance, correlation co-efficient and best fit.

   I have taken a sample dataset from builtin.com/data-science/anscombes-quartet to explain the same in the below table.

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The values in the table share the same summary statistics. The difference can only be seen when you use matplotlib to plot them.



The first diagram shows that the data fits the linear regression well.
The second diagram is not a linear regression
The third diagram has outliers
The fourth diagram is nowhere close to being a linear regression.

So, this illustrates how easy it is to fool a linear regression.

3. <u>What is Pearson's R? (3 marks)</u>
   Ans: Pearson's R is also called as Pearson correlation coefficient is a measure of linear correlation between two data sets. It is the ratio between the covariance of two variables and the product of their standard deviations.

Since it is a normalized value, the value always lies between -1 and 1. If the value is positive, it means the variables are positively correlated and if it negative, the variables are negatively correlated.

When there is no correlation, the value is 0.

We can also think of Pearson correlation coefficient as a measure of how close the observation are to a line of best fit.

Pearson's correlation coefficient is a good choice when
1. Both variables are quantitative:
2. Both variables are normally distributed.
3. The data has no outliers
4. The relationship is linear.

We use it test the significance of the relationship between the 2 variables.
We formulate a null hypothesis Ho p = 0,
Alternate hypothesis Ha: p!=0

If the p-value is less than 0.05 then the variable is significant.

4. <u>What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)</u>
Ans: When there are a lot of independent variables and if they are all on different scales, this will result in different and varying co-efficient which can be misleading. So, scaling is important for two reasons.
    a. Ease of interpretation
    b. Faster convergence of gradient descent methods

There are two popular ways to scale independent variables:
1. Standardizing: The variables are scaled so that their mean is zero and their standard deviation is one.
   X = (x – mean(x))/sd(x)

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and minimum values in the data.
   X = (x-min(x))/(max(x)-min(x))

The scaling affects only the co-efficients and no other parameters like p-value, F-statistic, R-square etc.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Ans: The value VIF shows the multi collinearity between independent variables.

   The higher the value, the higher the collinearity between the variables. The infinite value would mean that the independent variables are highly predictable and are highly multicollinear. Or, it would mean that the independent variables are in a perfect linear relationship with each other.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks
   Ans: A Q-Q plot, also called as 'Quartile-Quartile plot' is a visual tool to assess whether the given dataset came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

   The data plausibly came from some normal or exponential distribution. If both the datasets of quantile came from the same distribution, we should be able to visualize the points to form a seeming straight line.

   Since they are supposed to form a straight line, we use linear regression concepts to understand if they really form a straight line. So, all the assumptions of linear regression applies to the q-q plot quantile as well.