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: After plotting the bar chart I have come to the following conclusions:

a. There is no data recorded on extreme weather conditions. Thus no one demands for bikes in this weather. Talking about the other three categories on weather(clear,moderate and bad), we can see that as compared to moderate and bad weather, the demand for bikes on clear weather is more. And the demand has significantly increased over a year.

b. It is crearly evident that during Fall, the demand for bikes have increased and over a year the increase is clearly visible.

c. There is only a slight increase in the demand on working days than on holidays.

d. The last four days of week has more bookings than the starting days.

e. During Holidays, the demand for bikes is comparatively.

f. June, August, September seem to have more number of booking than the rest of the months in the year.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
ANS: The use of drop_first=True is that it reduces the redundant column while creating the dummy variable. Suppose we have three categories in our variable, so if the value is not from the first or second value then it is by default the third category. Thus, this reduces the correlations created by dropping one variable which is not that necessary.

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

4. How did you validate the assumptions of Linear Regression after building the model on the
training set? (3 marks)
ANS: The Assumptions of Linear regression that I have validated in my model are:

Linear relationship between variables: The variables have visible linearity among them.

Error terms distribution: The error terms are normally distributed.

Independency of Error terms: The error terms are independent of each other.i.e. there is no auto correlation

Homoscedasticity: Each term has constant variance.

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: The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

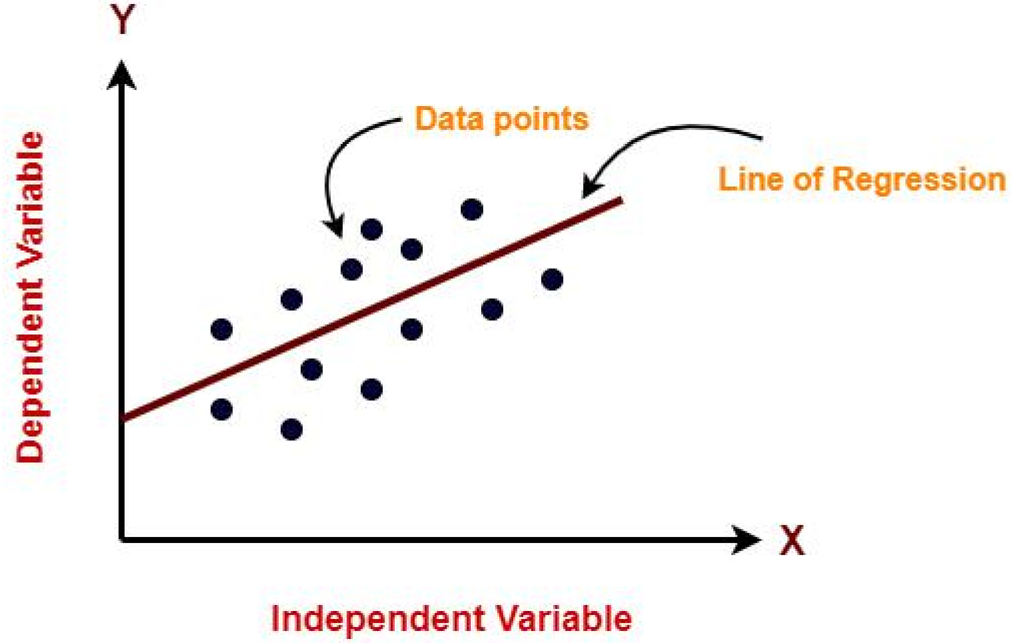
- temp
- hum
- clear

General Subjective Questions

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

ANS: Linear regression is a type of supervised learning method. Supervised learning is an algorithm which is trained on input data that has been labeled for a particular output.

1. Linear regression performs the task to predict a target variable based on the given independent variable(s). This technique finds out a linear relationship between a dependent variable and the other given independent variables.



2. The goal of linear regression is to perform predictive analytics and it is done by making the machine learn the science of generating a best fitted line that will very well generalize how the test set will be evaluated, and how the fitted line will be able to accurately estimate new datasets.
3. **There are two types of linear regression:**
   a) Simple linear regression: In this model, we have only one independent variable. Equation of a simple linear regression is $Y = c + mX$, where Y is the dependent variable , X is the independent variable , m is the slope of the line and c is the y-intercept.
   b) Multiple Linear regression: In this model, we have more than one independent variable. Formula for Multiple linear regression is : $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\epsilon$
4. **LINEAR REGRESSION ALGORITHM**:
   a) Reading and understanding the data: In this step, we will import all the important library which are required for model building like pandas, numpy, seaborn, sklearn, statsmodels and matplotlib.pyplot. Then read the dataset and clean the dataset by doing EDA.
   b) Visualizing the data: Here we will visualize the numerical and categorical datasets separately and draw some conclusions looking at those plots.
   c) Data Preparation: Here we will convert the categorical variables in to numerical variables with the help of dummy variables.

d) Train-Test split: Here we will split our original dataset into train set and test set in 70-30 ratio or 80-20 ratio. The train dataset will be used to create the model whereas the test dataset will be used for testing the train set predictions.

e) Building a model: Here we will; build out linear regression model by using any of the below three mentioned methods:

   i. Forward Selection: We start with null model and add variables one by one. These variables are selection on the basis of high correlation with target variable. First we select the one, which has highest correlation and then we move on to the second highest and so on.

   ii. Backward Selection: We add all the variables at once and then eliminate variables based on high multicollinearity (VIF>5) or insignificance (high p-values).

   iii. RFE or Recursive Feature Elimination is more like an automated version of feature selection technique where we select that we need "m" variables out of "n" variables and then machine provides a list of features with importance level given in terms of rankings. A rank 1 means that feature is important for the model, while a rank 4 implies that we are better off, if we don't consider the feature.

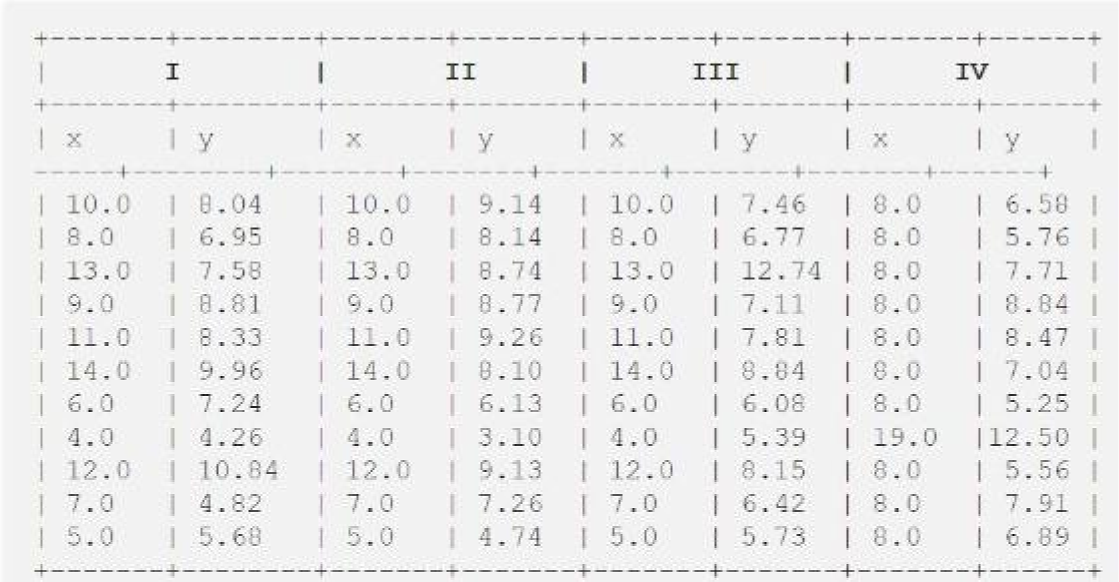
f) Residual analysis of the train data: It tells us how much the errors (y_actual — y_pred) are distributed across the model. A good residual analysis will signify that the mean is centred around 0.

g) Making predictions using the final model and evaluation: We will predict the test dataset by transforming it onto the trained dataset

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

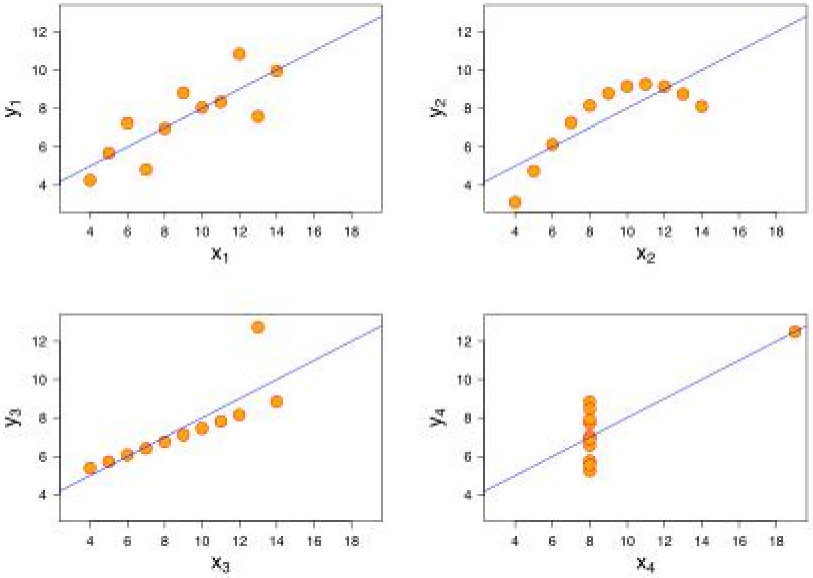
ANS: The Anscombe's quartet comprises of 4 datasets that look very similar to each other but when they are drawn on the graph, they look quite different. This quartet is basically made to demonstrate the importance of graphing the dataset and also the effect of outliers on the appearance of dataset.

It was created in 1973 by the statistician Francis Anscombe. The dataset is as follows:



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

The graphical representation of the above datapoints is as follows:



As we can clearly see that data points are identical when examined using simple summary statistics, but vary considerably when graphed

3. What is Pearson’s R? (3 marks)

Ans: 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. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

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

ANS:

What is Scaling:

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.

Why is Scaling performed:

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.

Difference between Normalized scaling and standardized scaling:

Normalisation	**Standardisation**
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)

ANS: Whenever there is perfect correlation between two independent variables then we get VIF value as infinite.An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variable.The value of VIF is calculated as:

$$\mathbf{VIF_i} = 1/(1 - R_i^2)$$

Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

ANS: The Q-Q plot is also known as a quartile-quartile plot. It is a graphical technique for determining if two datasets come from populations with a common distribution by creating a scatter plot that plots two sets of quartiles against one another If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

The advantages of the q-q plot are:

a. The sample sizes do not need to be equal.

b. Many distributional aspects can be simultaneously tested.