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 concluding a regression model for Boom Bikes, it has been found that the dependent variable ie. the sum total of rentals is directly proportional to the independent variables- **temperature and the weather conditions.**

   The immediate proof is that during the month of November to February there is a **negative correlation**, i.e. during this month, the rentals were low, whereas September has a **positive correlation** with the rental counts. Probably, September of every year would be a peak time to issue schemes by the Boom Bikes for more registered users.

   Again when the weather is foggy or cloudy or there are chances of light snow, we again observe a **negative correlation**. Another significant correlation that we can't ignore is the year, there is a definite rise in the popularity of Boom Bikes over a year which led to the increase in rentals in the following years.

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

   Ans: The command is commonly used along with the get_dummies function of pandas to avoid variable redundancy. The difference is, even more, permanent for variables with comparatively smaller cardinality. For this data set, we won't need a dummy for the years 2018 or 2019 separately as a dummy for 2019 is enough to explain the incidences on behalf of 2018. Hence, we use drop_first = True for our convenience.

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

   Ans: According to the pair plot **temperature** has the highest correlation with the variables.

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

   Ans: The benefit of using **statsmodels** library is the in-built features like **params** and **summery** to get details of the linear regression model from where we can check the **R-squared** value and **F-static** for model validity whereas, for each variable, **coefficient**, and the **p-value** is also mentioned in the table which gives us an idea about the significance of the features that we included in the model.

   And we dropped off the features under conditions where-
   the p-value is >0.05 and the VIF is greater than 10 or utmost greater than 5.

   * The process of dropping the features is remodeling is done until we have got a perfect set of p-values and VIFs for each feature.

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

   Ans: The 3 contributing factors are-
   the month of the year, climate and temperature

General Subjective Questions

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

   Ans: Linear regression algorithm in machine learning is a category of supervised learning which performs regression task to predict the most probable value of the dependent variable based on its relationship with the independent variable.

   If we dive deeper to get a statistical view, linear regression
   - At each x (independent variable), it tries t derive the best estimate of y (independent variable
   - model predicts a single value each time, from which we can observe a uniform distribution of each term.

   Further, to ensure that the algorithm captures the entire distribution of error and make sure it's visible when we collecting the value of y. Hence, there exist a set of assumptions in linear regression-

   - A linear relationship between x and y
   - Error terms are normally distributed
   - Error terms are independent of each other
   - Error terms have a constant variance

   Given that

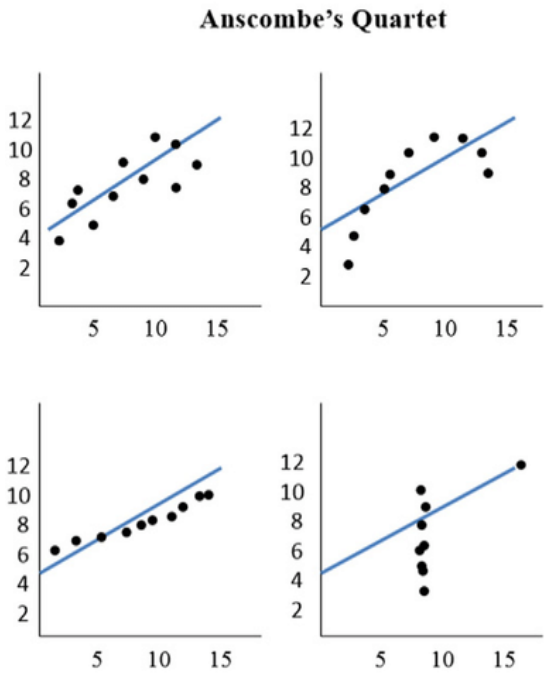
   - there are no assumptions in the distribution of x and y
   - the assumptions are truly for explaining linearity between x and y
   - and the error terms have to have a normal distribution

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

   Ans: Anscombe's Quartet a data set of four that has identical statical parameters (mean, variance, sample variance, and correlation) but possesses some peculiarities that fools the regression model when building on the data sets. when these data points are separately plotted for each of the four data sets, they show different distributions and appear differently when plotted on scatter plots.

   This snap from the original paper published by Anscombe explains it all-

Anscombe's Quartet of Different XY Plots of Four Data Sets Having Identical Averages, Variances, and Correlations



Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

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	5.76
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	8.81	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	7.26	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

Source: Adapted from Anscombe (1973, pp. 19-20)

1

These inferences were first concluded by statistician Francis Anscombe to assert the importance of data visualization before starting with analysis or model building and to break the dependency only on statistical properties.

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

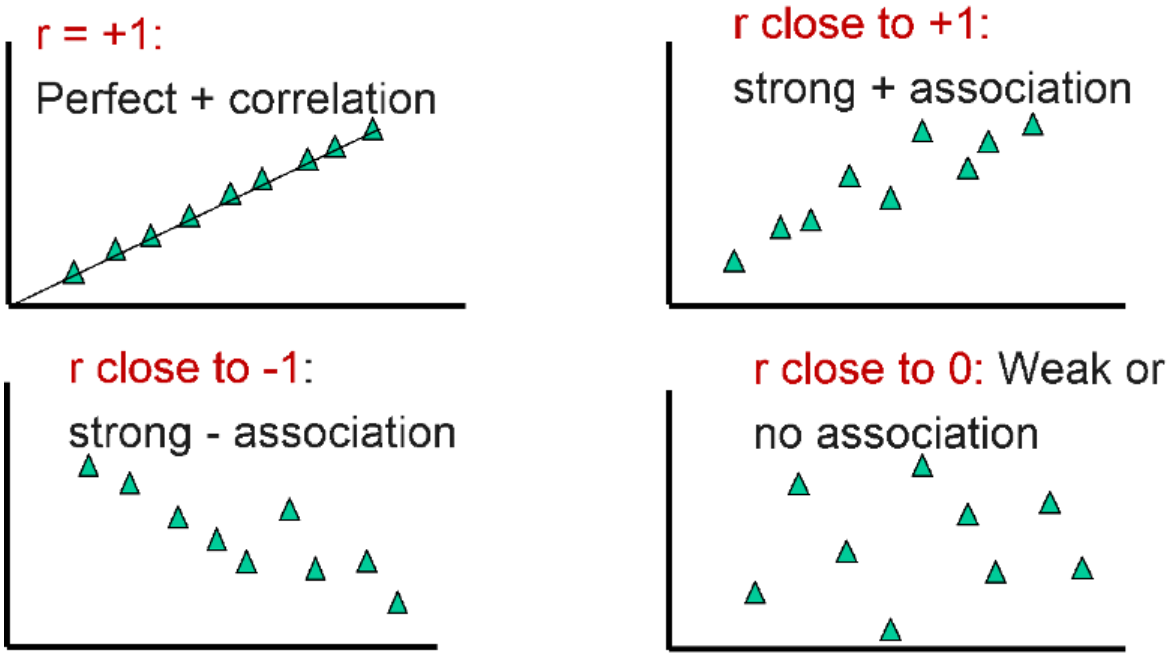
Ans: Pearson's R is the most popular type of correlation co-efficient of the several others which is commonly used in linear regression. Pearson Correlation Coefficient (PCC) is primarily used in the feature selection process and defined as the covariance of x (independent variable) and (dependent variable) upon .standard deviation of x and y

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]}}$$

This correlation coefficient asserts

- how strongly x and y are correlated in quantity
- and whats the direction of the relationship, ie positive or negative.

The PCC value always ranges between +1 and -1 =, which is perfectly represented in the following diagram, showing how the data distribution looks when r (PCC)=1, or r = close to +1, or r = close to - 1 or r = nearing to 0.



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

Ans: Feature scaling is another important step for data preparation for variables with numerical data points. When we have large set of variables data points measured at completely different metric scales or units, then scaling is applied to even out the values within the range of 1-0 or a data set whose mean is 0 with a standard deviation of 1.

In a crux, scaling is performed -

- for ease of interpretation
- for Faster convergence for gradient descent methods

There are two scaling methods

- **Standarded scaling:** The variables are scaled in such a way that their mean is zero and standard deviation is one.

- **Normalized Scaling:** The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

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

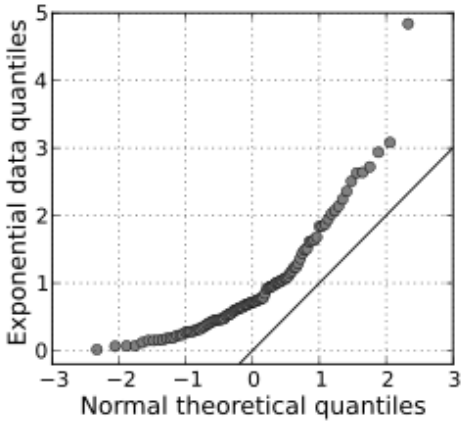
Ans: Yes, it did occur at one instance where the variable 'yr' (from the **Boom Bike** case study) showed an absolute collinearity with the feature '2019' and VIF value got fixed

soon after ‘yr’ was dropped from making predictions for the model.

This a pure case of an instance when, R-squared value becomes ‘0’ as per formulae of VIF ie. (1/1-R^2) when the denominator gets the value 0, automatically the VIF value becomes infinite. It asserts that when the VIF value is infinite then there is a perfect correlation between the variables.

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

Q-Q Plots or Quantile-Quantile plots are observed to check weather the data is normally distributed. The plot has two axes, one which is generated from the data set or **Exponential Data Quantile** Vs the quantiles generated from the theoretically derived distribution of the data or **Normal Theoretical Quantiles**.



NB: *Quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities .*

It takes the representation of a scatter plot and the best fit line drawn for us to conclude - whether the data points hugs the line or are found aloof. This decides which model to choose - a normal distribution or a uniform distribution, thus guides us to chose what is the best fit for the data. The Q-Q plot also comes in handy when we are comparing the derived data -set, to compare model outcomes.