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)**

   In relation to the analysis of the categorical variables and supported by the graphs, we can conclude the following:

   - Holiday: it can be observed that there is a high utilization of the service when the day falls on holidays.
   - WorkingDay: For working days, the use of the service rises, so there is an interesting demand to determine the use on those days.
   - Weathersit: Clear, Few clouds, Partly cloudy, Partly cloudy is day with more use the services and tne next use is for follows days: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark) 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   - It is important when creating dummy variables, to include this parameter to eliminate one of the variables so that they will be redundant in the analysis.

   - Regarding the analysis of the graph, we can see that the variable with the highest correlation is the variable "atemp", with 0.63

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

   - The validation of the assumptions mainly saw the multicollinearity both in the plot where "atemp" has a high collinearity and by eliminating those with p-value > 0.05 and analyzing the VIF parameter.

   - On the other hand, we analyzed whether the errors had a normal distribution function, and its homoscedastic.

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

- Temp (B = 0.480)
- Workingday (B= 0.1851)
- Season_4 (B=0.8767)

General Subjective Questions

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

- Linear regression is a way to identify a relationship between the independent variable(s) and the dependent variable

- We can use these relationships to predict values for one variable for given value(s) of other variable(s)

- It assumes the relationship between variables can be modeled through linear equation or an equation of line.

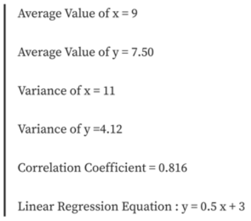
- The variable, which is used in prediction is termed as independent/explanatory/regressor where the predicted variable is termed as dependent/target/response variable.

- In case of linear regression with a single explanatory variable, the linear combination can be expressed as : response = intercept + constant * explanatory variable.

- In case of multiple linear regression with multiples explanatories variables, the linear combination can be expressed as : response = intercept + constant_1 * explanatory variable_1 + constant_2 * explanatory variable_2 + ….. constant_n * explanatory variable_n

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

- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed



Average Value of x = 9

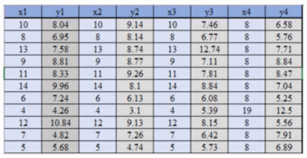
Average Value of y = 7.50

Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

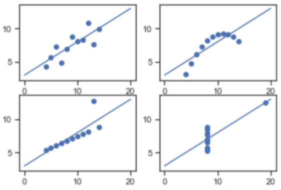
Linear Regression Equation : y = 0.5 x + 3



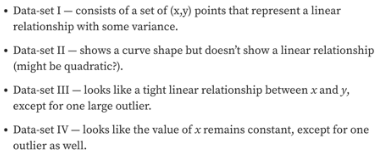
x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7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

Four Data-sets

- Examples Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data



Graphical Representation of Anscombe's Quartet



- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

- Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data

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

- Pearson's correlation coefficient is a test that measures the statistical relationship between two continuous variables. If the association between the items is not linear, then the coefficient is not adequately represented.

- Pearson's correlation coefficient is intended to indicate how closely two variables are associated with each other, therefore:

- Correlation less than zero: If the correlation is less than zero, it means that it is negative, i.e., the variables are inversely related.

- When the value of any variable is high, the value of the other variable is low. The closer it is to -1, the clearer the extreme covariation. If the coefficient is equal to -1, we refer to a perfect negative correlation.

- Correlation greater than zero: If the correlation is equal to +1 it means that it is perfect positive. In this case it means that the correlation is positive, i.e. the variables are directly correlated.

- When the value of one variable is high, the value of the other variable is also high, the same happens when they are low. If it is close to +1, the coefficient will be the covariation.

- Correlation equal to zero: When the correlation is equal to zero it means that it is not possible to determine some sense of covariation. However, it does not mean that there is no nonlinear relationship between the variables.

- When the variables are independent it means that they are correlated, but this does not mean that the result is true.

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

- Feature scaling: When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

    1. Ease of interpretation
    2. Faster convergence for gradient descent methods

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

- MinMax 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.

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

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables
- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables

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

   - A QQ plot, short for quantile-quantile plot, is often used to assess whether the residuals in a regression analysis are normally distributed or not.

   - The idea behind a QQ plot is simple: if the residuals fall along a straight line at approximately a 45-degree angle, then the residuals are approximately normally distributed.

   - If it turns out that your residuals deviate severely from the 45-degree line on the QQ plot, you may consider performing a transformation on the response variable in your regression, such as using the square root or logarithm of the response variable.