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)

⇨ Season, Year , Holiday , Weekdays , Weathersit , Month were the categorial variables in the dataset. A Bar- Graph was used to visualize this variables

1. **Season :** Here we can see that in fall there is more booking then other season and Spring has low booking in both the years
2. **Month :** The bar graph illustrate that the most number of bookins has been done between the months May to Oct. and the least number of bookings has been done in Jan and Feb in both the years
3. **Weathersit :** The bar chart decipetates that the most of the booking has been done on clear weather conditions Whereas least booking has benn done on Light snow rainy weather
4. **Weekday :** Bar graph shows that Thursday to Sunday have more bookings as compared to other days
5. **Holiday :** Bar graph illustrate that the less bookings are when there is no holiday.
6. **Year :** Bar chart shows that the Number of bookings has increased from the year 2018 to 2019.

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

⇨ Your dummy variables will be correlated if you don't remove the first column (redundant). This may have a negative impact on some models, and the effect is amplified when the cardinality is low. Iterative models, for example, may have difficulty convergent, and lists of variable importance's may be distorted. Another argument is that having all dummy variables results in multicollinearity between them. We lose one column to keep everything under control

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

⇨ The temp and atemp are highly correlated with each other. However there is no such variable which have high co relation with each other.

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

⇨
1. The distribution of residuals should be normal and centered around 0.
2. We test this residuals assumption by producing a distplot of residuals to see if they follow a norm distribution or not.
3. The residuals are scattered around mean = 0 as seen in the diagram above.

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 three predictor variables that influence bike booking, according to our final Model, are:

   1. Temperature(temp): With a coefficient of 0.5173, a unit increase in the temp variable increases the number of bike rentals by 0.5173 units.
   2. Weather Situation 3 (weathersit_3): With a coefficient of '-0.2828, a unit increase in the Weathersit3 variable reduces the number of bike hires by 0.2828 units as compared to Weathersit 1.
   3. Where weathersit_3 = Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain +Scattered clouds
   4. Year(yr): With a coefficient of 0.2324, a unit increase in the yr variable increases the number of bike rentals by 0.2324

General Subjective Questions

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

⇨ Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model. Linear regression is based on the popular equation "y = mx + c".
It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term. Regression is broadly divided into simple linear regression and multiple linear regression

1. Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.
2. Multiple Linear Regression :MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

Yi = Bo+B1x1+B2x2+...+BpXp
Where,
B1 = coefficient for XI variable

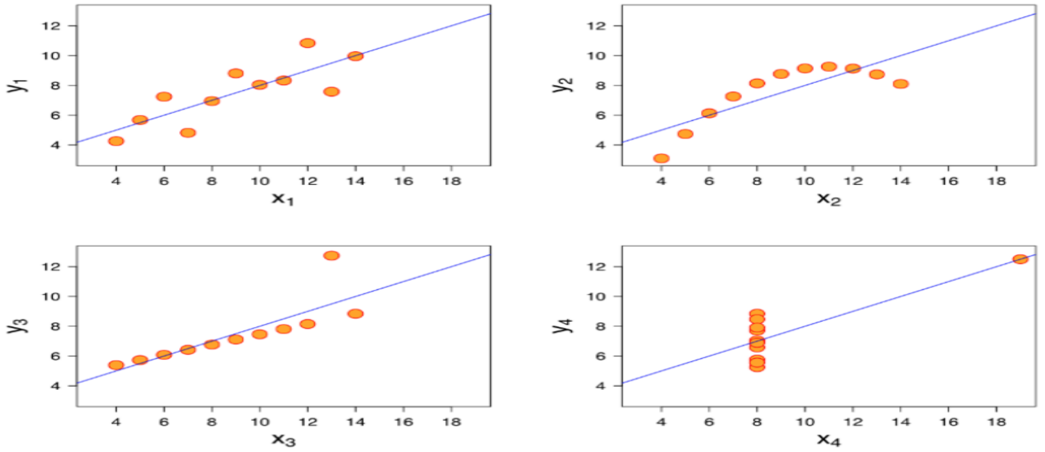
B2= coefficient for X2 variable

B3 = coefficient for X3 variable and so on...

BO is the intercept (constant term).

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

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph .It was developed to emphasize both the importance of graphing data before analysing it and the effect of outliers and other influential observations on



Statistical Properties:

1. The first scatter plot (top left) appears to be a simple linear relationship.
2. The second graph (top right) is not distributed normally; while there is a relation between them it's not linear.
3. In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816
4. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

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 correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

r = 1 means the data is perfectly linear with a positive slope

r = - 1 means the data is perfectly linear with a negative slope

r = 0 means there is no linear association

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

⇨ Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

1. **Normalization** is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

2. **Standardization**, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization

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

⇨ **VIF - the variance inflation factor : -** The VIF indicates how much collinearity has increased the variance of the coefficient estimate. (VIF) is equal to 1/(1-Ri^2).
VIF = infinity if there is perfect correlation. Where R-1 denotes the R-square value of the independent variable for which we want to see how well it is explained by other independent variables. - If an independent variable can be completely described by other independent variables, it has perfect correlation and has an R-squared value of 1. As a result, VIF = 1/(1-1) provides VIF = 1/0, which is "infinity."

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

⇨ The quantiles of the first data set are plotted against the quantiles of the second data set in a q-q graphic. It's a tool for comparing the shapes of different distributions. A scatterplot generated by plotting two sets of quantiles against each other is known as a Q-Q plot. Because both sets of quantiles came from the same distribution, the points should form a line. That’s a fairly straight line.
The q-q plot is used to answer the following questions:

1. Do two data sets come from populations with a common distribution?
2. Do two data sets have common location and scale?
3. Do two data sets have similar distributional shapes?
4. Do two data sets have similar tail behavior.