Name- Rishav Kumar
Email ID-: rishavworld01@gmail.com
Cohort: - DSC27
ASSIGNEMENT: BIKE SHARING ASSIGNMENT

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?**

   After analysis (EDA) of the categorical variables from the dataset. We can infer that:

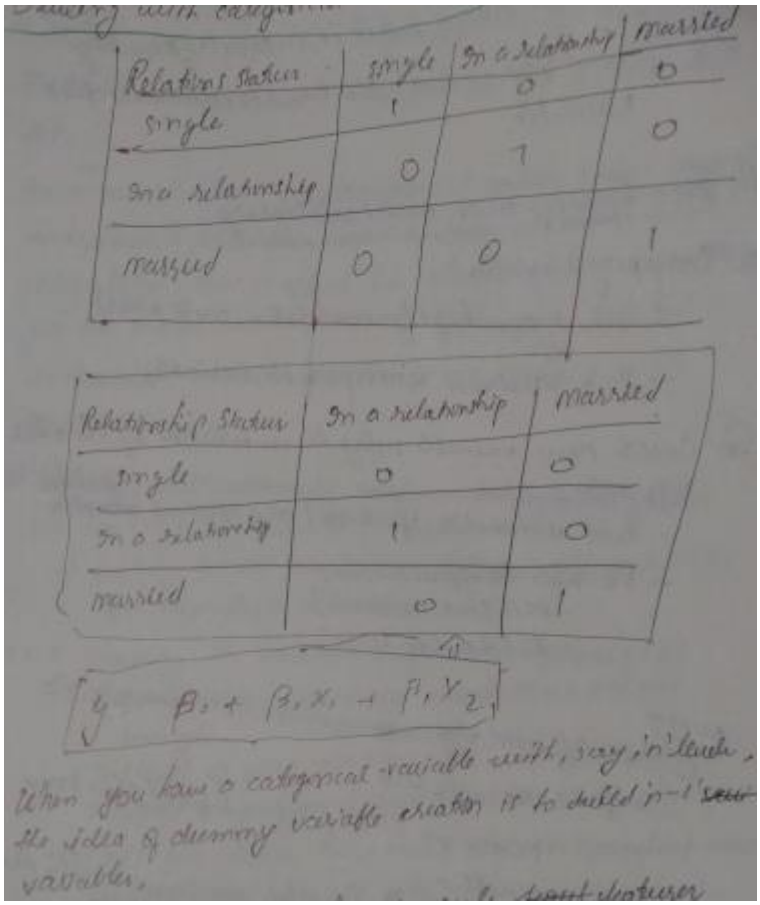
- The numerical variable 'registered' has the highest correlation with the target variable 'cnt'
- Temp and atemp are highly correlated and can affect the model.

- Demand Increases in the month of May, June, Aug & sept. In July, there is slight decrease in demand can be seen.
- There is very less demand of shared bikes in spring season
- Demand surges high in the year 2019 in comparison to its previous year
- Clear, few clouds, partly cloudy, partly cloudy seems to be favorable atmosphere for the uses of shared bikes
- In holidays, company can expect to get increase in demand of shared bikes

   This implies that the categorical features were helpful in explaining a greater proportion of variance in the dataset. The above knowledge we can imply while building the model.

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

   It is advisable to use drop_first=True, otherwise we will get a redundant feature i.e., dummy variables might be correlated because the first column becomes a reference group during dummy encoding. This in turn can affect the building of the linear regression model.



3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    I. If we consider all the features, the numerical variable 'registered' has the highest correlation with the target variable 'cnt'. And its correlation value is 0.95
    II. Temp and atemp are highly correlated to each other with a correlation value of 0.99

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

    After building the model on the training set, I carried out the following analysis: - 1. A test for normal distribution of error terms(residuals) by visualizing a distribution plot of the error terms. 2. Eliminations and inclusion of independent variables into each model based on VIF and p-values to avoid multicollinearity.

    After building the model on the training set, I carried out

    1. Residual analysis on the trained data set to check whether the error terms are normally distributed or not by plotting histogram on error terms.
    2. Plotted scattered plot in between y_test and y_pred to check whether the model is able to predict well or not
    3. Checked R- Squared and the Adjusted- R- Squared of both the train and test dataset for the final validation of the model.

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

    The top 3 features that significantly explain the demand of the shared bikes are: -
    1. 'temp'- Temperature in Celsius
    2. 'yr'- year (0: 2018, 1: 2019)
    3. 'winter'- A subcategory of 'season'.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm which falls under the category of supervised learning.
It models a target(y) prediction value based on the independent variables(x). This regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

Linear Regression Formula:

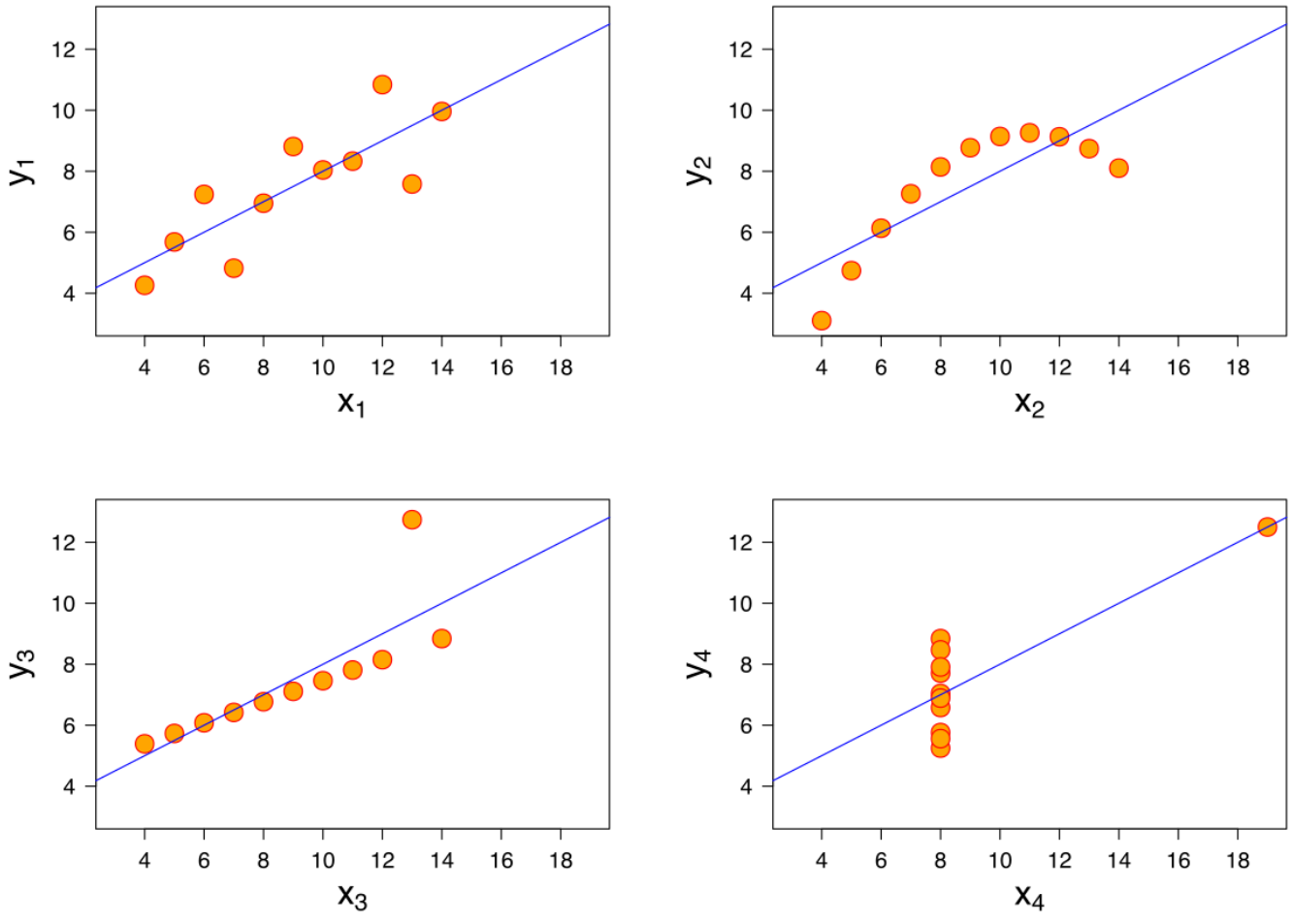
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

Linear Regression can be used to draw a trend line which can then be used to confirm or deny the relationship between the attributes.
"More the previous data, more will be the accurate prediction."

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises of four data sets that have nearly identical descriptive statistical data and I graph appears differently
For reference see the below graph:



3. What is Pearson's R?

Pearson's R is the strength of the linear association between the variables. Its value varies between -1 and +1.

- 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

Formula of Pearson's R

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the process of standardising the dataset in between the fixed range.
It is very important to rescale to handle highly varying magnitudes or values or units of the independent features.

Scaling is of two types:

i. Standardization
ii. Normalization/Min-Max Scaling

Normalization means scaling the data in between 0 and 1
Standardization means scaling the data to have the mean of 0 and standard deviation of 1.

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

VIF is the measure of the amount of multicollinearity between the dataset variables or multiple regression variables. Its full form is Variance Inflation Factor (VIF).
VIF formula: 1/(1-R^2)

If VIF is infinite then R^2 has to be 1.
R^2=1 signifies that 100% of the variance of the data is explained by the regression line.
A **high VIF** indicates that the associated independent variable is highly collinear with the other variables in the model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

Q-Q plot is a graphical tool to measure both sets of quantiles truly come from some theoretical distribution such as a Normal or exponential.
A Q-Q plot is the scatter plot created by plotting two set of data. If both the sets of the data came from the same distribution then we will see a line the plot which is roughly straight.