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)**
   Answer:
   - **season** - Most of the bookings happened in fall season and followed by summer, winter and spring. This shows that most bookings happen at fall season.
   - **mnth** - High bookings were found in months May, June, July, August and September with median >4000 bookings per month.
   - **weekday** - Is having almost similar trend throughout. This may or may not be a good predictor variable.
   - **weathersit** - Most of the bookings happened in weathersit 1(clear, few clouds, partly cloudy). This shows most people prefer to use bikes on clear day and on a clear day we can expect an increase in hiring bikes.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Answer:
   - It’s important to use **drop_first = True** because it helps in reducing the extra column created during dummy variable creation. In this way, it reduces the correlations among dummy variables.
   - If we have 3 (k) types of values in a Categorical column, if we are creating dummy variable, 3-1 (k-1) dummy variables are needed. It is enough to represent the whole data.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   Answer:
   - ‘temp’ and ‘atemp’ variables have the highest correlation with the target variable.

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

   1. Normality of error terms

      - Error terms should be normally distributed

2. Multicollinearity check
    - There should be insignificant multicollinearity among variables.
3. Linear relationship validation
    - Linearity should be visible among variables
4. Homoscedasticity
    - There should be no visible pattern in residual values.
5. Independence of residuals
    - No auto-correlation

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   Answer:
    - Temp - A unit increase in the variable temp increases the bike hire by 0.546436 units.
    - weathersit_3 - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (negative correlation)
    - yr (Year)

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Answer:
    - Linear regression may be defined as the statistical model that analyses the linear relationshipbetween a dependent variable with given set of independent variables.
    - Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
    - Mathematically the relationship can be represented with the help of following equation –
        - $y = mx + c$
            - Here, Y is the dependent variable we are trying to predict.
            - X is the independent variable we are using to make predictions.
            - m is the slope of the regression line which represents the effect X has on Yc is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

- Positive Linear Relationship:
  - A linear relationship will be called positive if both independent and dependent variable increases.
- Negative Linear relationship:
  - A linear relationship will be called positive if independent increases and dependent variable decreases.

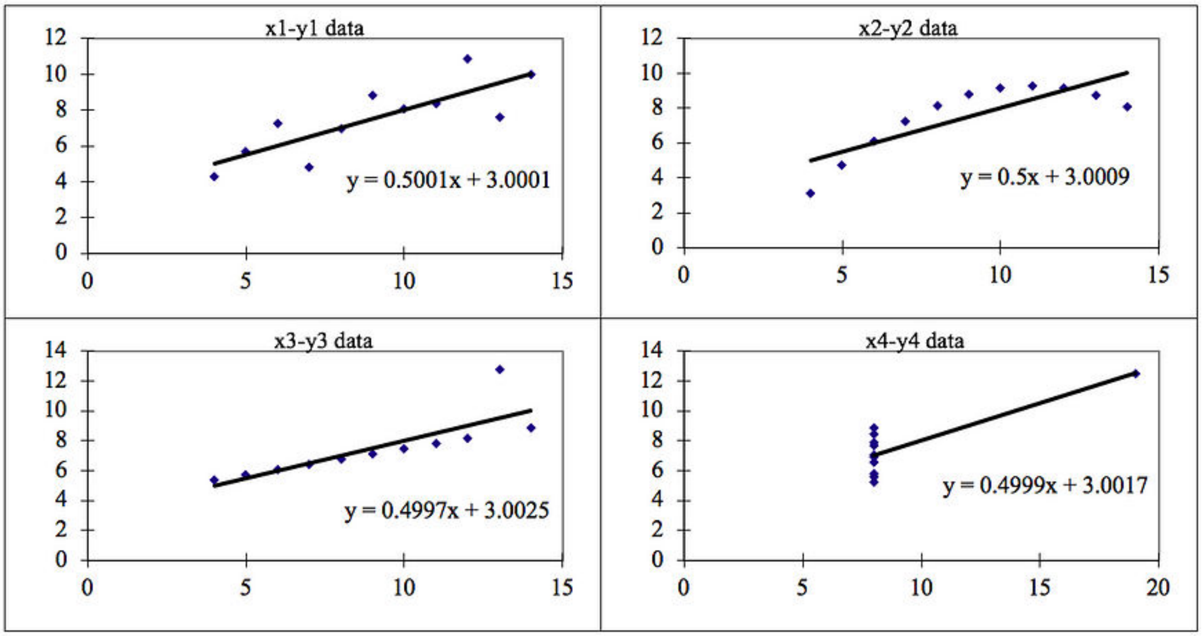
- Linear regression is of the following two types –
  - Simple Linear Regression
  - Multiple Linear Regression

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

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



The four datasets can be described as:

1. **Dataset 1:** this **fits** the linear regression model pretty well.

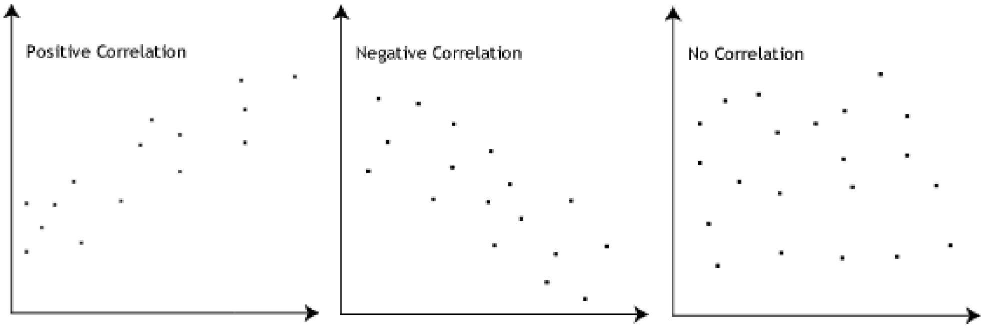
2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.

3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

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

- 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.



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

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.
- It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
- 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, regardless of the unit of the values.

Normalized scaling

- Minimum and maximum value offeatures are used for scaling.
- It is used when features are of differentscales.
- Scales values between [0, 1] or [-1, 1].
- It is really affected by outliers.
- Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

Standardized scaling

- Mean and standard deviation is used forscaling.
- It is used when we want to ensure zeromean and unit standard deviation.
- It is not bounded to a certain range.
- It is much less affected by outliers.
- Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Answer:

- If there is perfect correlation, then VIF = infinity.
- A large value of VIF indicates that there is acorrelation between the variables.
- If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
- When the value of VIF is infinite it shows a perfect correlation between two independent variables.
- In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity.
- To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer:

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets comefrom populations with a common distribution.

Use of Q-Q plot:

- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset.
- By a quantile, we mean the fraction (or percent) of points below the given value.
- That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and70% fall above that value.
- A 45-degree reference line is also plotted.
- If the two sets come from a population with the same distribution, the points should fall approximately along thisreference line.
- The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

- When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified.
- If so, then location and scale estimators can pool both datasets to obtain estimates of the common location and scale.
- If two samples do differ, it is alsouseful to gain some understanding of the differences.
- The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.