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?
    - The demand for bikes is highest during the fall and lowest during the spring.
    - The demand for bikes has increased to a great extent from 2018 to 2019.
    - The demand for bikes in highest in September and lowest in January.
    - The demand for bikes is high on Monday, Friday and Saturday.
    - The demand for bikes is high when the weather is clear and it is low when the raining or snow.

2. Why is it important to use drop_first=True during dummy variable creation?
    - It is important to use drop first= True during dummy variable creation, because it automatically removes the extra variable created during the dummy variable creation, and it reduces the correlation between dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    - ‘temp’ variable has 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?
   I have validated the assumptions of linear regression after building the model on the training set in the following manner:
    - Normality of error terms: The error terms should be normally distributed. The dist plot of our model is normally distributed.
    - Multicollinearity Check: There should be no significant collinearity among the variables. The variables having VIF > 5 and p value > 5% have been excluded from the model.

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

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   - Linear regression may be defined as the statistical model that analyses the linear relationship between 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.

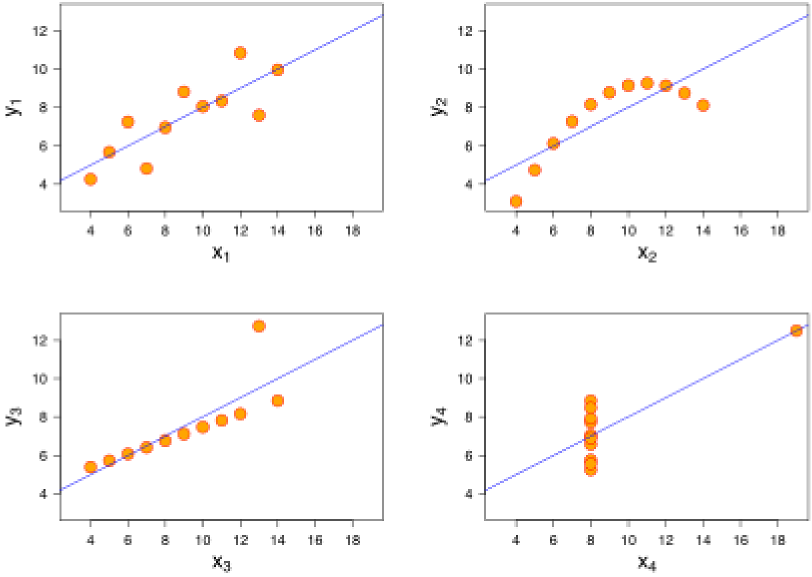
     m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.
   - Assumptions –

     The following are some assumptions about dataset that is made by Linear Regression model :
     - ➢ Multi-collinearity - Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
     - ➢ Auto-correlation - Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
     - ➢ Relationship between variables - Linear regression model assumes that the relationship between response and feature variables must be linear.
     - ➢ Normality of error terms - Error terms should be normally distributed.
     - ➢ Homoscedasticity - There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail.

   - 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.

- For all four datasets:



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where *y* could be modelled as gaussian with mean linearly dependent on *x*.
- For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- 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.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson’s R?

The Pearson correlation coefficient (*r*) is the most widely used correlation c oefficient and is known by many names:

- Pearson’s *r*
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   - Scaling variables involves adjusting their range or magnitude without changing their distribution.
   - Feature scaling is performed in linear regression to ensure that all input features are on a similar scale, preventing dominance by features with larger magnitudes. This allows the optimization algorithm to converge

faster and more accurately, leading to better performance and stability in the model.

- **Standardizing:** The variables are scaled in such a way that their mean is zero and standard deviation is one. Standardization is very useful if data has varying scales and the algorithm assumption about data having a gaussian distribution.
- **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. It is useful when data has varying scales and the algorithm does not make assumptions about the distribution. It is a good technique when we did not know about the distribution of data or when we know the distribution is not gaussian.

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

    - If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation 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.
    The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from 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 and 70% 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 this reference 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 data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful 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