Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer.

1. Bike demand in the fall is the highest.
2. Bike demand takes a dip in spring.
3. Bike demand in year 2019 is higher as compared to 2018.
4. Bike demand is high in the months from May to October.
5. Bike demand is high if weather is clear.
6. The demand of bike is almost similar throughout the weekdays.
7. Bike demand doesn’t change whether day is working day or not.

Q2. Why is it important to use drop_first=True during dummy variablecreation?

Answer.

1. It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables
2. Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, some don’t need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it.
3. It is also used to reduce the collinearity between dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer. atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer. By checking following parameters:-

1. Residual Analysis and distribution plot of errors.
2. Homoscedasticity.
3. Multicolinearity.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer.

1. Year : With every yr demand is only increasing.

2. Temp : Positively Related with coeff of 0.5 means pleasant temp month will give boom to business.

3. Windspeed : With increase in speed demand decreases means poor weather has effect on the business . During those days bike can be put for servicing.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Answer.

1. A method of finding the best linear relationship between the independent and dependent variables.

2. It is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables.

3. Uses Sum of Squared Residuals Method.

4. Assumptions for linear regression are :-
   a) There is a linear relationship between the dependent and independent variables.

b) About Residuals.
   i) Error terms are normally distributed.
   ii) Have a mean value of zero.
   iii) Have the same (but unknown) variance, σ^2 .
   iv) Residual terms are independent of each other.

c) About the estimators:
   i) The independent variables are measured without error.
   ii) No multicollinearity in the data.

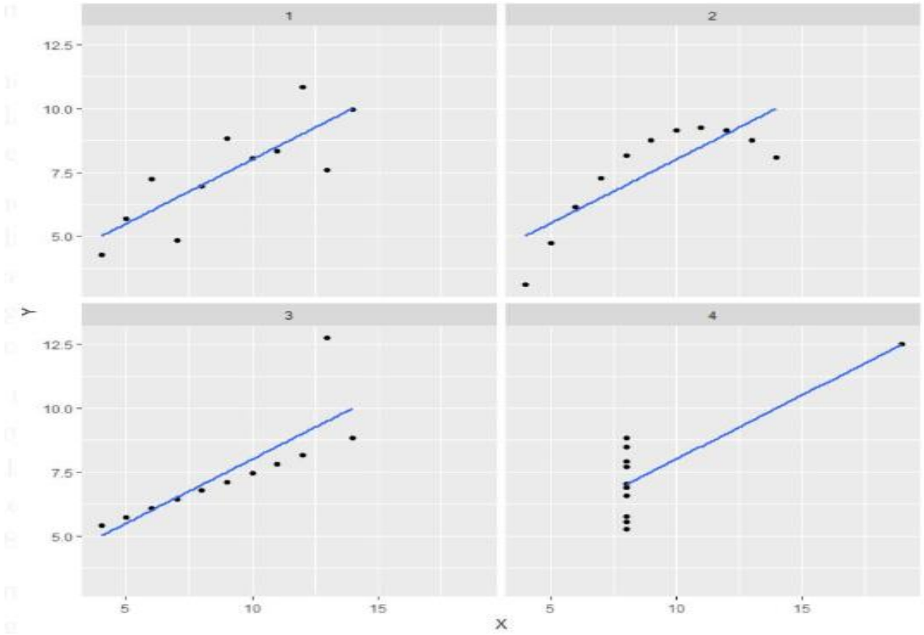
Q2. Explain the Anscombe's quartet in detail.

Answer.

1. It comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

2. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3. Graphs for the same are as under :-



4. Explanation of this output:-

a) In the first one(top left) scatter plot shows a linear relationship between x and y.
b) In the second one(top right) shows a non-linear relationship between x and y.
c) In the third one(bottom left) shows there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
d) Finally, the fourth one(bottom right) shows one high-leverage point is enough to produce a high correlation coefficient.

5. It is 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.

Q3. What is Pearson's R?

Answer.

1. **Pearson's r i.e.** Pearson's Correlation Coefficient**, also referred as bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. .
2. It is the covariance of the two variables divided by the product of their standard deviations. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

3. Requirements for Pearson's Correlation Coefficient: :-
   a) Scale of measurement should be interval or ratio.
   b) Variables should be approximately normally distributed.
   c) The association should be linear.
   d) There should be no outliers in the data.

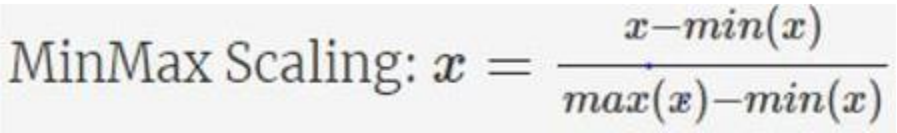
Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer.

1. Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

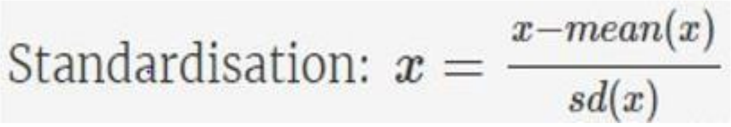
2. When collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. So, we have to do scaling to bring all the variables to the same level of magnitude. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

3. Normalization/Min-Max Scaling:
   a) It brings all of the data in the range of 0 and 1.
   b) sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
   c)



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

4. Standardization Scaling:
   a) It replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
   b) sklearn.preprocessing.scale helps to implement standardization in python.
   c)



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

5. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer.

1. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

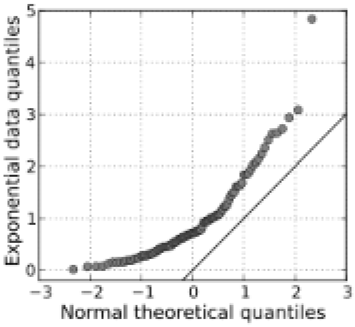
2. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer.

1. Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

2. A Q Q plot showing the 45 degree reference line:



3. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

4. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.