Assignment-based Subjective Questions

Ankur Napa

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

Variable Name	Inferences
yr	The count increased significantly in 2019 compared to 2018.
month	The count is higher during May to October months.
season	The count is higher for Fall (Autumn) and then followed by Summer.
holiday	The count is lower during holidays.
working day	Working Day / Non-Working Day shows almost similar behaviour.
weathersit	The count is higher on Clear, Few clouds, Partly cloudy, Partly cloudy days followed by Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist days. No records found for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog weather.
weekday	For weekdays, the median is almost similar for all days.

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

It is important to use drop_first=True because it helps in reducing the extra column created during dummy variable creation.

Hence it reduces the correlations created among dummy variables.

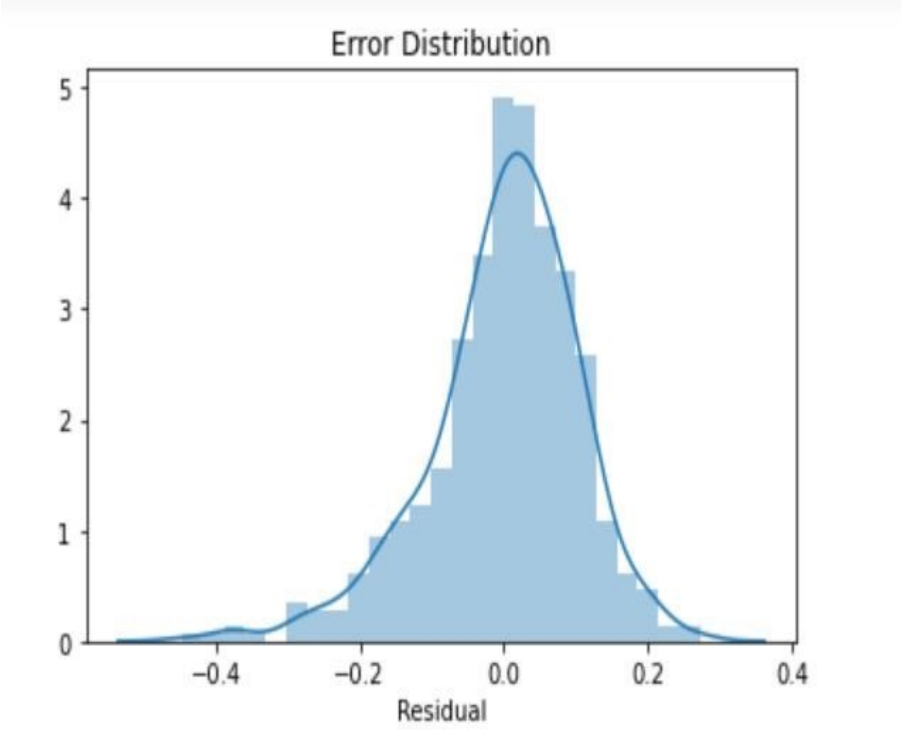
For example, we have 4 types of values in the Categorical column (season), and we want to create a dummy variable for that column. If one variable is not summer, winter and spring, then it is obviously Fall. So, we do not need 4th variable to identify the fall column.

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

Temperature(temp) and Feeling Temperature(atemp)

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

- Steps followed to validate the assumptions after building the model on the training set.
  1. Calculate the residual and see its distribution by plotting a distplot of residuals. It should give a normal distribution and should be centred around 0.
  2. Validation: Above shown image depicts a normal distribution and the residuals are distributed about mean zero.



Q5Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ranking	Feature	Correlation . Coefficient	Type of Correlation
1	Season_fall	0.312	In fall season there will be more bike hiring
2	Season_summer	0.27	In summer season there will be more bike hiring
3	Weatherit_3	0.244	It will affect the bike hire count inversely

General Subjective Questions

Ankur Napa

Q1Explain the linear regression algorithm in detail.

Linear Regression Algorithm can be explained by the below points:

- Linear Regression is a supervised (with labels) machine learning algorithm.
- In Linear Regression, the dependent variable or the target variable is continuous in nature and hence used to predict values with a continuous range e.g., Sales, weight.
- The equation for the best-fit line:

$$Y = a + bX$$

where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

Types of linear regression

Simple Linear Regression:

- Linear Regression algorithm with only one independent variable.
- Best fit line represented by the equation - y=mx+c, where x is the independent variable, y is the target variable, m is the coefficient for the variable x and c is the intercept.

Types of Linear Regression

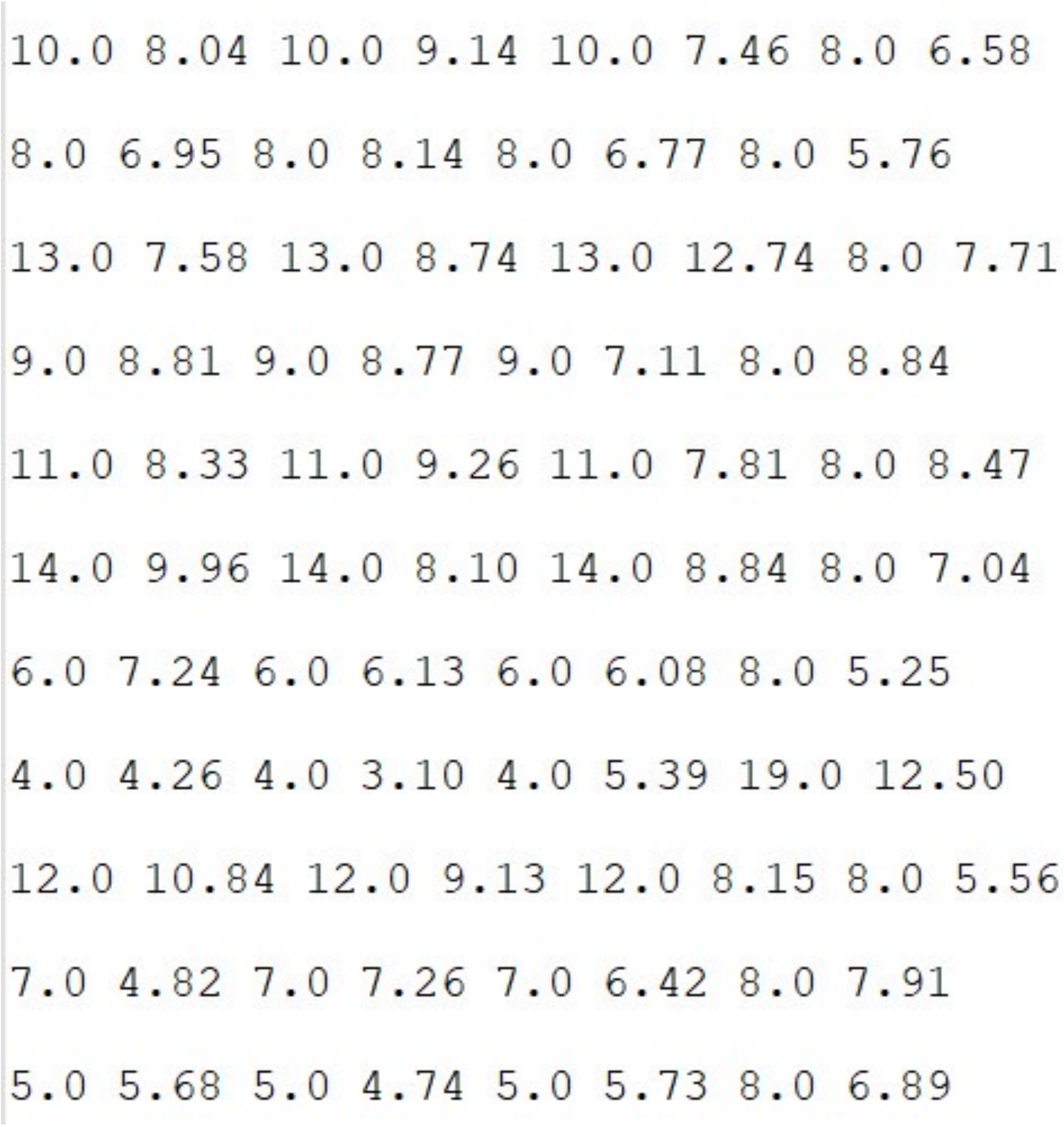
Multiple Linear Regression:

- Linear Regression algorithm with more than one independent variable.
- Best fit line represented by the equation - y= m1x1+m2x2+m3x3..mnxn, where x1, x2, x3,…,xn are the independent variables, y is the target variable, m1, m2, m3,..,mn are the coefficients for the variables x1, x2, x3 respectively and c is the intercept.

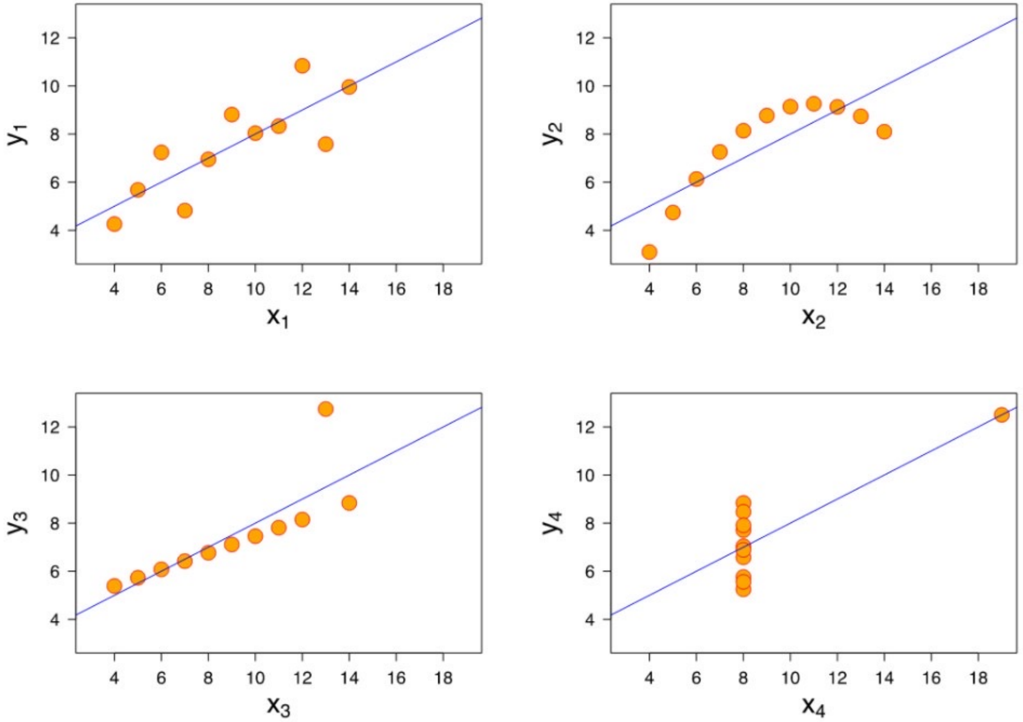
- The performance of the regression model built using these algorithms can be evaluated by using various metrics like MAE, MAPE, RMSE, R-squared etc.

Q 2. Explain the Anscombe’s quartet in detail.

- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed.
- For example,
  - Below are (x, y) pairs of 4 datasets whose summary statistics is given as:
  - The average *x* value is 9 for each dataset
  - The average *y* value is 7.50 for each dataset
  - The variance for *x* is 11 and the variance for *y* is 4.12
  - The correlation between *x* and *y* is 0.816 for each dataset
  - A linear regression (line of best fit) for each dataset follows the equation *y = 0.5x + 3*



```
10.0 8.04 10.0 9.14 10.0 7.46 8.0 6.58

8.0 6.95 8.0 8.14 8.0 6.77 8.0 5.76

13.0 7.58 13.0 8.74 13.0 12.74 8.0 7.71

9.0 8.81 9.0 8.77 9.0 7.11 8.0 8.84

11.0 8.33 11.0 9.26 11.0 7.81 8.0 8.47

14.0 9.96 14.0 8.10 14.0 8.84 8.0 7.04

6.0 7.24 6.0 6.13 6.0 6.08 8.0 5.25

4.0 4.26 4.0 3.10 4.0 5.39 19.0 12.50

12.0 10.84 12.0 9.13 12.0 8.15 8.0 5.56

7.0 4.82 7.0 7.26 7.0 6.42 8.0 7.91

5.0 5.68 5.0 4.74 5.0 5.73 8.0 6.89
```



These four datasets appear to be almost similar. But when we plot these four datasets on an x/y coordinate plane, we get the following results:

Q3 What is Pearson’s R?

Pearson’s correlation (also called Pearson’s *R*) is a correlation coefficient commonly used in linear regression.

- Its full name is the **Pearson Product Moment Correlation (PPMC).**
- It shows the linear relationship between two sets of data.
- Two letters are used to represent the Pearson correlation: Greek letter rho **(ρ)** for a population and the letter **“r”** for a sample.
- For example, scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson’s correlation between the two groups was analysed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a fairly strong relationship.

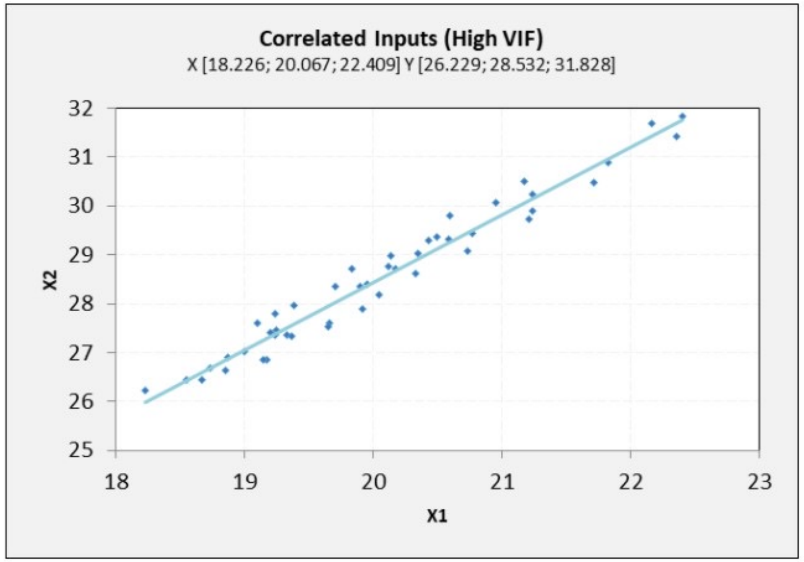
Q4What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Scaling a step of data Pre-Processing which is applied to numeric variables to normalize the data within a particular range.
- Below are the reasons why it’s been performed:
  - Helps in speeding up the calculations in an algorithm.
  - Useful in interpretation of values as unscaled values are difficult to interpret.
  - Most of the times, 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

Q5You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF = infinity shows a perfect correlation between two independent variables.

In the case of perfect correlation, we get **R2 =1**, which lead to 1/(1-R2) to equal to infinity.



To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Q6What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Quantile-Quantile (Q-Q) plot, is a graphical technique to help assess if a set of data came from some theoretical distribution such as Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Interpretation 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 data set.

- **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis.
- **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.
- **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.
- **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis.

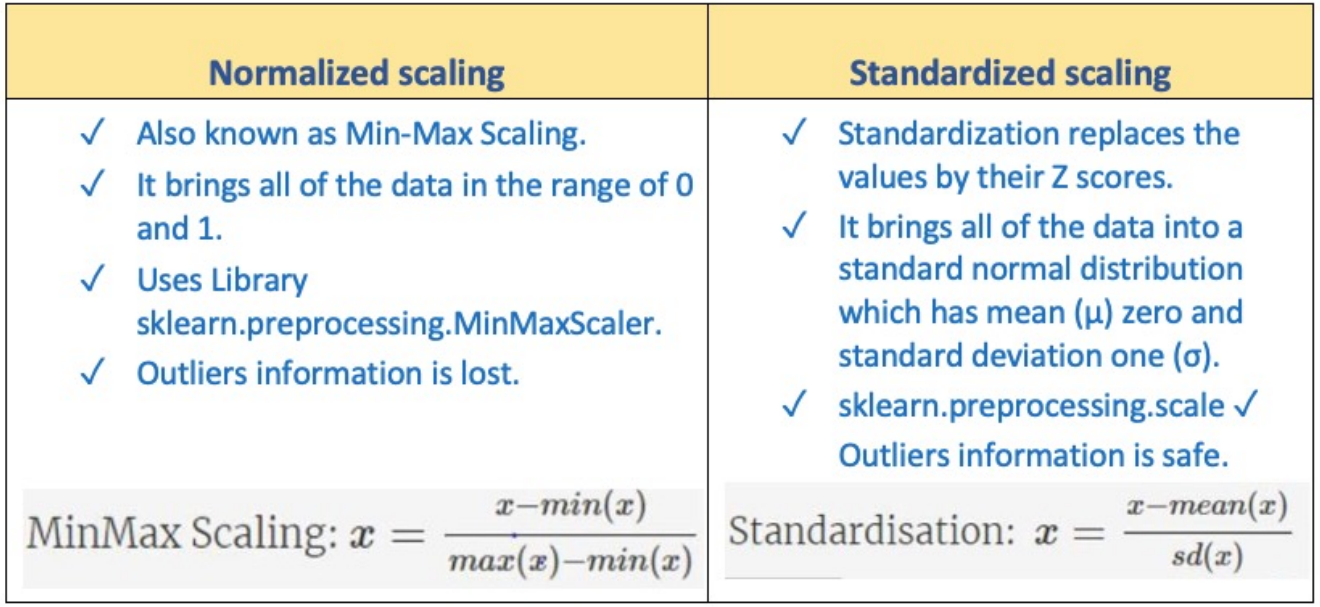
Uses of Q-Q plot:

- can be used with sample sizes also.
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

Q-Q plot in linear regression is important because:

when training and test data sets are received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Difference between Normalized and Standardized scaling:



Normalized scaling	Standardized scaling
✓ Also known as Min-Max Scaling. ✓ It brings all of the data in the range of 0 and 1. ✓ Uses Library sklearn.preprocessing.MinMaxScaler. ✓ Outliers information is lost.	✓ Standardization 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). ✓ sklearn.preprocessing.scale ✓ Outliers information is safe.
MinMax Scaling: $x = \frac{x-min(x)}{max(x)-min(x)}$	Standardisation: $x = \frac{x-mean(x)}{sd(x)}$