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)

→ Year : In 2019 the booking has increased considerably from 2018

→ Season : Fall season has the maximum number of bookings followed by summer. Spring has the least number of bookings.

→ Holiday : The 25th quartile is lower for holidays compared to working days. A working day attracts more bookings.

→ weathersit: Clear weather had the highest number of bookings. Light snow and rain had lowest number of bookings

→ Month: Maximum bookings are between may and october. The bookings are lowest in the months of Nov, Dec, Jan, Feb

→ days of the week: Lowest bookings are observed on Sunday. The rest of the weekdays have similar trends of b

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

If there are 3 variables present, it is obvious that if an entry is not the first or second variable, then its the third. So we don't need an extra column for that variable.

Drop_first = True, reduces a column during dummy variable creation

Syntax example:

```
months_df=pd.get_dummies(bike.mnth,drop_first=True)
```

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

'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? (3 marks)

1. Residuals are normally distributed and their mean is zero.
2. Error terms are constant (homoscedastic).
3. Residuals are independent of each other.
4. Multicollinearity between variables is under check using VIF.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top 3 features which are positively correlated:
Temp, year, winter

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a supervised machine learning method used to find the relationship between an dependant variable and set of independent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the sum of distance between the actual data points and the predicted data points.

There are 2 types of linear regression algorithms

Simple Linear Regression

– Single independent variable is used.

- $Y = \beta 0 + \beta 1 X$ is the line equation used for SLR.

Multiple Linear Regression

- Multiple independent variables are used.
  - $Y = \beta 0 + \beta 1 X1 + \cdots + \beta pXp + \in$ is the line equation for MLR.
  - $\beta 0$ = *value of the Y when X* = 0 (*Y intercept*) o $\beta 1, \beta 2, \ldots, \beta p$ = *Slope or the gradient*.
  - Cost functions – The cost functions helps to identify the best possible values for the $\beta 0, \beta 1, \beta 2, \ldots, \beta p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable.
  - The straight-line equation is $Y = \beta 0 + \beta 1 X$. The prediction line equation would be $Ypred = \beta 0 + \beta 1 xi$ and the actual Y is as Yi.
  - While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used. o $ei = yi - ypred$ is provides the error for each of the data point.
  - OLS is used to minimize the total e2 which is called as Residual sum of squares. o RSS = $\sum (yi - ypred)\ 2\ n\ i$=1
  - Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.
  - Assumptions of linear regression:
    - Residuals are normally distributed and their mean is zero
    - They are independent of each other

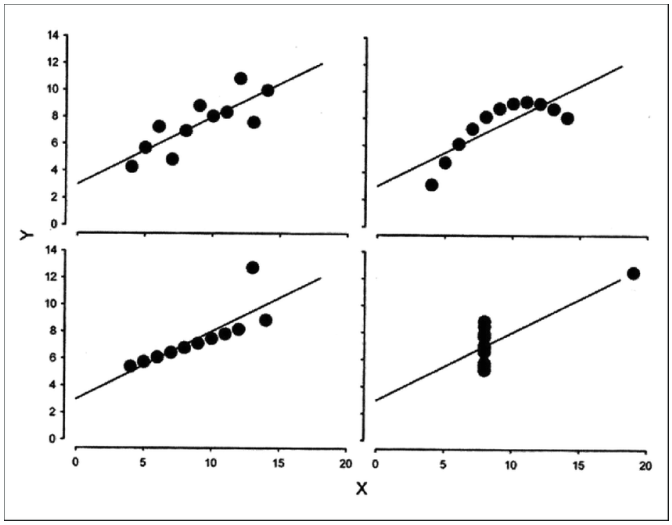
- Error terms are homoscedastic, that is they are constant

- Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

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

Anscombe's quartet has four datasets that have the same variance, mean, R squared, correlations, linear regression lines but when they are plotted on graph, the data set greatly differ from each other. The datasets were created by Francis Anscomne in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y



I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
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

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. It quantifies the strength and direction of a linear relationship between two continuous variables.

It ranges from -1 to 1, where:

1 indicates a perfect positive linear relationship,

0 indicates no linear relationship,

-1 indicates a perfect negative linear relationship.

In other words, Pearson's R helps assess how well a straight line can describe the relationship between two variables.

The formula for Pearson's correlation coefficient is:

$$r = [n(\Sigma xy) - \Sigma x \Sigma y] / \sqrt{[n(\Sigma x2) - (\Sigma x)2][n(\Sigma y2) - (\Sigma y)2]}$$

In this formula, x is the independent variable, y is the dependent variable, n is the sample size, and Σ represents a summation of all values.

Pearson's R is sensitive to outliers and assumes a linear relationship, so caution should be taken when applying it, especially in cases where these assumptions may not hold.

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

In linear regression, scaling refers to the process of transforming the input features by multiplying them with a constant factor or applying some other scaling method. The purpose of scaling is to ensure that all the features contribute equally to the computation of the regression coefficients.

When the features in a linear regression model are on different scales, it can lead to issues. For example, a feature with a larger scale might dominate the learning process, and the coefficients may be heavily influenced by that particular feature. Scaling helps in preventing such dominance and ensures that the model is not skewed towards any specific feature.

Common methods of scaling include:

Min-Max Scaling: is the simplest and most consistent method in rescaling. The range of features to scale in [0, 1] or [−1, 1].Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Normalization is sensitive to outliers because it is based on the minimum and maximum values of the data.

(Z-score normalization): This method scales the features to have a mean of 0 and a standard deviation of 1. The formula for standard scaling is:

$$\text{New value} = (x - \mu) / \sigma$$

- Scaling is particularly important when using algorithms that rely on distances between data points, such as gradient descent.
- Standardization is less sensitive to outliers since it uses the mean and standard deviation, which are less influenced by extreme values.

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

- VIF is the variance inflation factor that checks for multicollinearity between the independent variables in linear regression method. If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. 〖VIF〗_1=1/(1-R_1^2)
- A general rule of thumb is that if VIF > 10 then there is multicollinearity.
- One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then more terms may need to be dropped.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

It is used to check following scenarios:

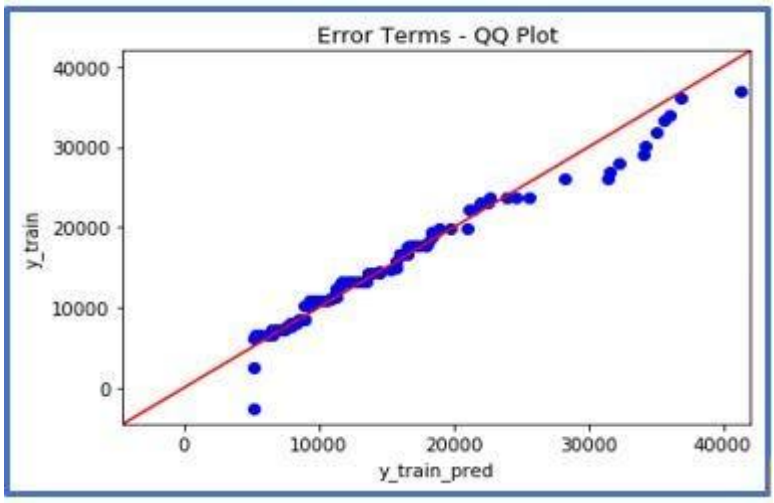
If two data sets —

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

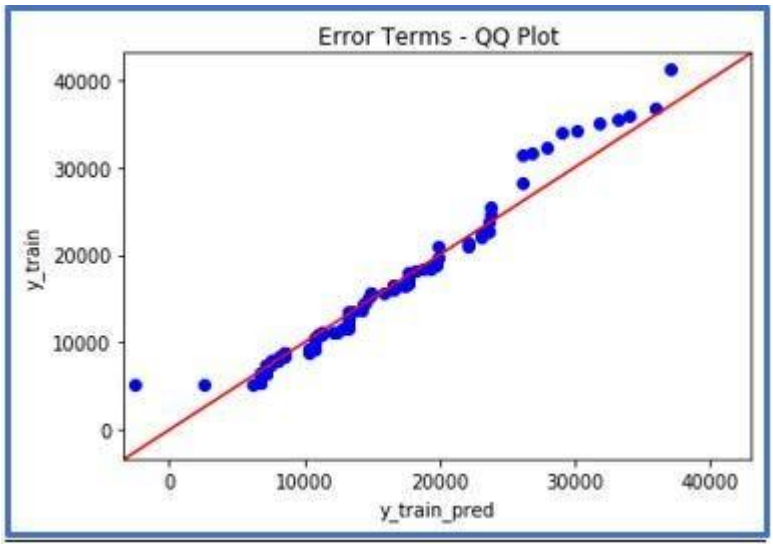
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.Below are the possible interpretations for two data sets.

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis