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 :
Below categorical variables have an impact on the dependent variable ‘cnt’ as :
1. yr : A coefficient value of 0.2308 indicates that a unit increase in yr increases the bike bookings by 0.2308 units
2. season_4 : i.e 'Winter' season as per data dictionary indicates bike hiring increase by 0.1487 units
3. weathersit_3 : Unit increase in weather situation3 i.e Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds decrease bike hiring by 0.1422 units
4. mnth_9 : September month sees an increase in bike hiring by 0.0990 units
5. season_2 : Summer season impacts the bike hiring with an increase by 0.0864 units

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

Answer:
We need to use drop_first = True, during dummy creation in order to avoid creation of extra column as for example : incase of traffic signal column , if we assign 1 to ‘Green’ and 0 to ‘Red’ then by default it’s understandable that if it’s not 1 or 0 then it’s ‘Yellow’, we dont need another column to specify the same.
So, for dummy creation of columns it’s recommended to use this command so that n-1 columns are created for n kind of combinations inorder to reduce correlation among the dummy variable created

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

Answer:
The pairplot clearly indicates a very high correlation between ‘temp’ & ‘atemp’ variables and ‘cnt’

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

Answer:
Residual analysis helps in validating the model’s reliability and predictability
The residual terms/Error terms for the model is distributed uniformly and are centred around 0

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:
As per the model the top 3 features are :

1. yr : A coefficient value of 0.2308 indicates that a unit increase in yr increases the bike bookings by 0.2308 units
2. temp : Unit increase in temperature increases bike hiring by 0.5977 units
3. hum : increase in humidity by a unit decreases the bike hiring by 0.2192 units

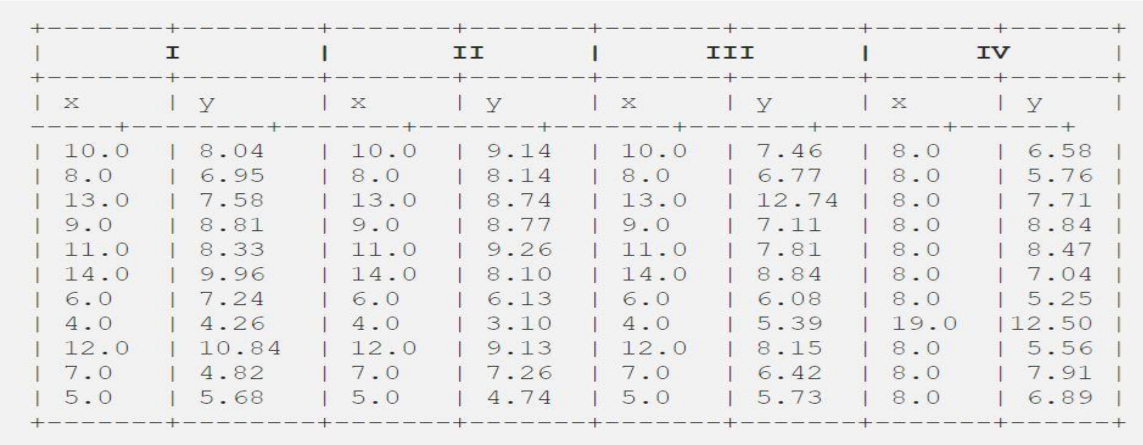
General Subjective Questions

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

Answer: Below is the alogorithm:
1. EDA – loading, understanding, missing value check, outliers treatment, imputation of missing values if necessary, visualizing the data, identifying the target variable and independent variables
2. Checking the correlation of the independent variables with target variable to understand if linar regression model is applicable
3. Creating dummies for Categorical variables
4. Scaling the variables using either MinMax scaling or normalisation
5. Dividing the data into train and test set
6. Creating the X_train and y_train
7. Selecting the feature either manually or using RFE incase number of features is greater than 15
8. Adding constant/intercept for the X_train while using statsmodels for model creation
9. Creating the model
10. Viewing the model summary
11. Checking the values of R-squared, Adjusted R-squared, F-statistics and p-value
12. Checking the VIF for the features
13. Feature elimination based on high p-value and high VIF
14. Create the model by Forward/Backward/Stepwise feature selection/elimination
15. Once model is ready, perform Residual Analysis
16. Calculate residual based on y_train – y_train_pred
17. Check the distribution of residuals: it should be normally distributed and mean around 0. There should not be any specific pattern in the residuals
18. Based on sucess of residual analysis, proceed to apply the model on the test set
19. Predict and Apply the model on the test set and check the R-squared and Adjusted R-squared for the test set
20. The R-squared and Adjusted R -squared for the test set should be close to the one for the train set for the model to be decent one

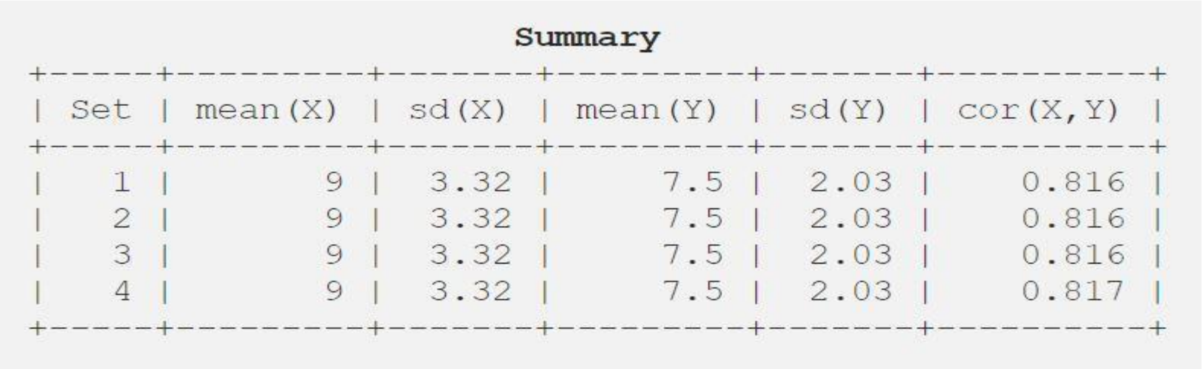
2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer:
Anscombe’s quartet comprises four datasets of eleven points which seems to look statistically similar but each plot for these datasets appears different graphically.



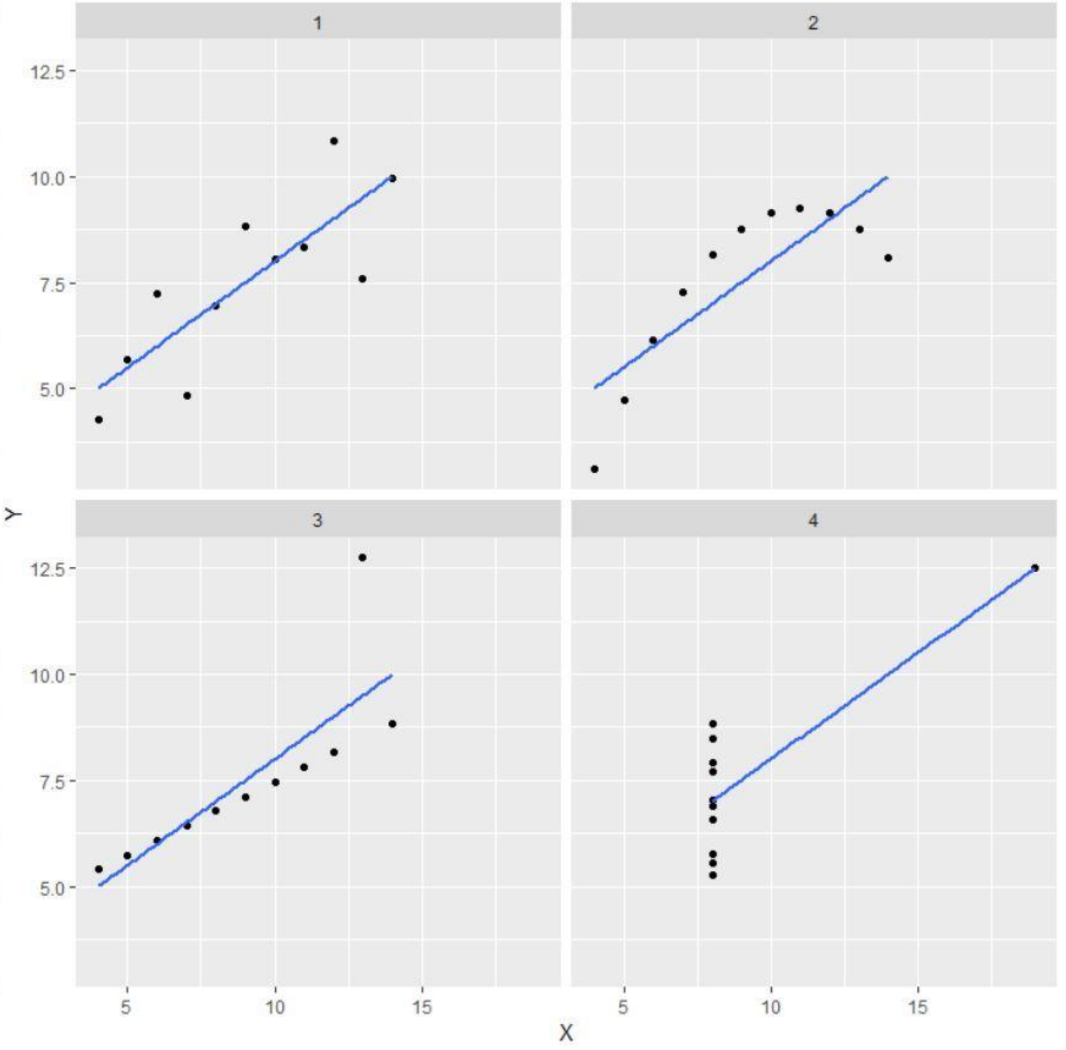
```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+-------+-------+-------+-------+-------+-------+------+
| 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 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

The basic statistical analysis with mean, standard deviation and correlation of these values of X & Y is as below:



```
                        Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```

The graphical representation of these 4 plots gives an idea about the understanding of below:



1. The 1st plot represents a linear relationship between x & y
2. The 2nd plot represents a non-linear relationship between x and y.
3. The 3rd plot again represents a clear linear relationship between X & Y, although the presence of 1 point in the graph indicates the case of an outlier presence in the dataset
4. The 4th plot indicates that the presence of atleast one high leverage point can result in a high correlation between X & Y as it again forms a linear representation

Thus, Anscombe's quartet signifies the importance of graphical visualization of the data prior to analysis for understanding the variance and characteristics of the data which might not be adequately explained by basic statistical components.

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

Pearson's R refers to the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It measures the linear correlation between two variables with a value that lies between -1.0 and +1.0.

It is limited to linear relationships and cannot capture non-linear relationships between two variables or differentiate between dependent and independent variables.

Pearson's R is the covariance of the two variables divided by the product of their standard deviations.

Calculating Pearson’s R:

The data should meet the below criteria:

❿ Scale of measurement should be interval or ratio

❿ Variables should be approximately normally distributed

❿ The association should be linear

❿ There should be no outliers in the data

Below is the general formula for Pearson’s R for any give sample:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\,\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}.$$

where:

n is sample size

x_{i},y_{i} are the individual sample points indexed with i

Interpretation of Pearson’s R:

1. Value between 0 and 1 represents Positive correlation i.e. When one variable changes, the other variable changes in the same direction
2. Value= 0 represents No correlation i.e. There is no relationship between the variables.
3. Value between 0 and –1 represents Negative correlation i.e. When one variable changes, the other variable changes in the opposite direction.

It graphically indicates whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative and when the slope is positive, *r* is positive.

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

Scaling refers to modifying the range of features or variables data points to a relatively smaller scale without making any changes to the shape of distribution. It helps in having all the variables in the data set distributed within the same range for better plotting , analysizing and visualization.

After the given data set is divided into Train and Test set, Scaling is first performed on the Train set using fit_transform() and then after model building and Residual analysis, the test data is scaled using transform()

1. Standardisation – the data is scaled to standard normal distribution with mean zero and standard deviation one

Standardisation: X = (x-mean(x))/sd(x)

2. MinMax scaling – the data is scaled within 0 to 1 range, min of all variables is 0 and max of all variables is 1

MinMax Scaling X = (x-min(x))/max(x)-min(x)

3. Scaling for categorical variables - Are converted to numeric format by creating dummy variables

Difference between Normalized Scaling and Standardized Scaling:

Standardized Scaling – the data is scaled to standard normal distribution with mean 0 and standard deviation 1:

Standardisation: X = (x-mean(x))/sd(x)

Normalized Scaling or MinMax scaling – the data is scaled within 0 to 1 range, min of all variables is 0 and max of all variables is 1:

MinMax Scaling X = (x-min(x))/max(x)-min(x)

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

Answer:

VIF calculates how well one independent variable is correlated to all the other independent variables combined. So basically higher the VIF , higher is the correlation among the variables.

Thus, a VIF of infinite would mean that there is a perfect correlation and the model would result in multicollienarity.

VIF is given by the below formula:

VIFi=1/(1−Ri2)

where,

VIFi : is the VIF for the independent variable

Ri2: is the R-squared of the variable

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

Answer:

Q-Q plot or Quantile-Quantile plot is a kind of probability plot that graphically displays the quantile portion of a data with respect to the same quantile portion of another data.
If the two distributions that are being compared are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

Below are the steps to plot a Q-Q plot:

1. Sorting the data in ascending or descending order.

2. Drawing a normal distribution curve.

3. Find the z-value for each segment.

4. Plot the dataset values against the normalizing cut-off points.

Q-Q plot helps to determine:

1. Whether two samples are from the same population.

2. Whether two samples have the same tail

3. Whether two samples have the same distribution shape.