Assignment-based Subjective

Question 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: Inferences :

Autumn season have the highest demand for bike

Spring has the lowest demand of bike

As compared to the previous year for the next year the demand of bike is increased.

Sept have the highest demand while Jan is the lowest.

Demand for the bike is low when there is a holiday.

The Good weather hit has highest demand

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

Answer: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Example :

Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and not semi furnished, then It is obvious it is unfurnished. So, we do not need 3rd variable to identify the unfurnished.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Answer: The feature “temp” has highest correlation. It is very well linearly related with target “cnt”

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

Answer: Following methods were checked

Error terms are normally distributed with mean 0.

Error Terms do not follow any pattern.

Multicollinearity check using VIF(s).

Ensured the overfitting by looking the R2 value and Adjusted R2. 5.

Question 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: Features “holiday”, “temp” and season “hum” are highly related with target column, so these are top contributing features in model building.

General Subjective Questions:

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

Answer:

1. Linear Regression Algorithm is a machine learning algorithm based on supervised learning.

2.Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

3. Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables.

4. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Example :

Some More Examples of Linear Regression Analysis:

Prediction of Umbrella sold based on the Rain happened in Area.

Prediction of AC sold based on the Temperature in Summer.

During the exam season, sales of Stationary basically, Exam guide sales increased.

Prediction of sales when Advertising has done based on High TRP serial where an advertisement is done, Popularity of Brand Ambassador, and the Footfalls at the place of holding where an advertisement is being published.

Sales of a house based on the Locality, Area, and price.

Mathematically, we can write a simple linear regression equation as follow y ~ b0 + b1*x Where y is the predicted variable (dependent variable), b1 is slope of the line, x is independent variable, b0 is intercept(constant). It is cost function which helps to find the best possible value for m and c which in turn provide the best fit line for the data points.

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

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built.

They have very different distributions and appear differently when plotted on scatter plots. Each dataset consists of eleven (x,y) points.

The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

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

Answer:

Pearson's R is a numerical summary of the strength of the linear association between the variables.

If the variables tend to go up and down together, the correlation coefficient will be positive. Pearson's r measures the strength of the linear relationship between two variables. Pearson's r always between -1 and 1.

If data lie on a perfect straight line with negative slope, then r = -1.

Positive correlation indicates the both the variable increase and decrease together.

Negative correlation indicates the one the variable increase and the other variable decrease and vice versa.

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

Answer:

1. Scaling is a method to normalize the range of independent variables.

2. It is performed to bring all the independent variables on a same scale in regression.

3. If Scaling is not done, then regression algorithm will consider greater values as higher and smaller values as lower values.

4. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc. Example Weight of a device = 500 grams, and weight of another device is 5 kg.

In this example machine learning algorithm will consider 500 as greater value which is not the case. And it will do wrong prediction. Machine Learning algorithm works on numbers not units.

5. So, before regression on a dataset it is a necessary step to perform.

6. Scaling can be performed in two ways:

Normalization: It scale a variable in range 0 and 1.

Standardization: It transforms data to have a mean of 0 and standard deviation of 1

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

Answer: When there is a perfect relationship then VIF = Infinity whereas if all the independent variables are orthogonal then to each other then VIF = 1.0. Means if a variable is expressed exactly by a linear combination of other variable then it is said that VIF is infinite.

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

Answer:

1. 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.

2. It helps to determine if two data sets come from populations with a common distribution.

3. Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Importance of Q-Q plot in Linear Regression:

1. Two datasets/sample can be of different size.

2. Q-Q plot can detect outliers, shifts in scale, location, symmetry etc. simultaneously.

3. One of the important assumptions of Linear Regression is that the residual of the model is normally distributed. This can be assessed using Q-Q plot.

Example of Q-Q plot:
Here, first 5000 normally distributed random points are generated.

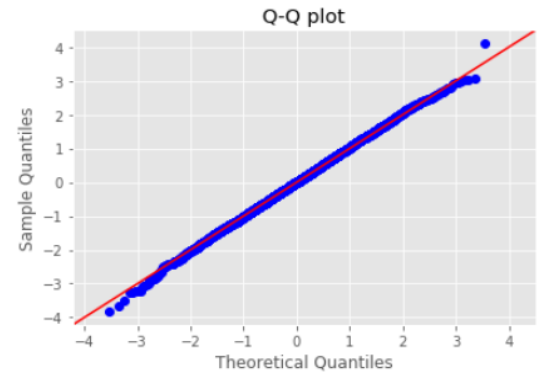
Then the random points are fed into the Q-Q plot. The blue dots representing the random points are aligning with 45 degree reference straight line in red. This re-confirms the test_data is actually normally distributed. (Left figure)

As test_data is shifted in location, the blue dots have shifted on the left side of the 45 degree reference line.

Thus Q-Q plot can show different statistical aspects. (Right figure)

```
import numpy as np
import statsmodels.api as sm
from matplotlib import pyplot as plt

test_data = [np.random.normal() for i in range(5000)]
sm.qqplot(np.array(test_data), line='45')
plt.title("Q-Q plot")
plt.show()
```



Left Figure

```
import numpy as np
import statsmodels.api as sm
from matplotlib import pyplot as plt

test_data = [np.random.normal() for i in range(5000)]
sm.qqplot(6 + np.array(test_data), line='45')
plt.title("Effect of shift in Location on Q-Q plot")
plt.show()
```

Right Figure