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:

i. The demand of bike is less in the month of spring
ii. The demand bike increased in the year 2019 when compared with year 2018.
iii. Month Jun to Sep is the period when bike demand is high
iv. Bike demand is less in holidays than working day.
v. The bike demand is high when weather is clear and Few clouds however demand is less in case of Lightsnow and light rainfall. We do not have any dat for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so no conclusion can be drawn.

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.

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

Answer:

'temp' variable has the highest correlation with target variable 'cnt' i.e. 0.63

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

Answer:

Sr. No	Assumption	How to validate?
1.	The Dependent variable and independent variable must have a linear relationship.	A simple pairplot of the dataframe can help us see if the independent variables exhibit linear relationship with the Dependent Variable.
2.	No Autocorrelation in residuals.	Use Durbin-Watson Test. DW = 2 would be the ideal case here (no autocorrelation) 0 < DW < 2 -> positive autocorrelation 2 < DW < 4 -> negative

		autocorrelation statsmodels' linear regression summary gives us the DW value amongst other useful insights.
3.	No Heteroskedasticity	Residual vs Fitted values plot can tell if Heteroskedasticity is present or not. If the plot shows a funnel shape pattern, then we say that Heteroskedasticity is present.
4.	No Perfect Multicollinearity.	In case of very less variables, one could use heatmap, but that isn't so feasible in case of large number of columns. Another common way to check would be by calculating VIF (Variance Inflation Factor) values. If VIF=1, Very Less Multicollinearity VIF<5, Moderate Multicollinearity VIF>5 , Extreme Multicollinearity (This is what we have to avoid)
5.	Residuals must be normally distributed.	Use Distribution plot on the residuals and see if it is normally distributed.

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:

i. Spring Season
ii. Temperature
iii. Mist

General Subjective Questions

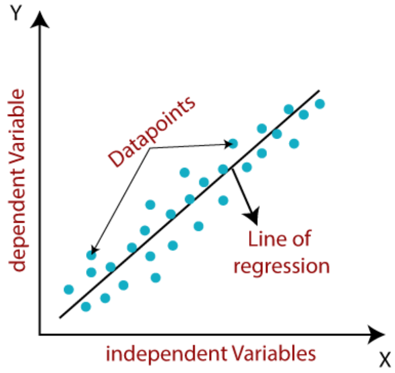
1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

i. Linear regression is one of the easiest Machine Learning algorithms.
ii. It is a statistical method that is used for predictive analysis.
iii. Linear regression makes predictions for continuous/real or numeric variables such as temperature, salary, age, etc.
iv. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.

v. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

vi. The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



vii. Mathematically, linear regression is represented as:

```
Y= B0+B1X+ε
```

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
B0= Intercept of the line (Gives an additional degree of freedom)
B1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation

viii. Types of Linear Regression

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression

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

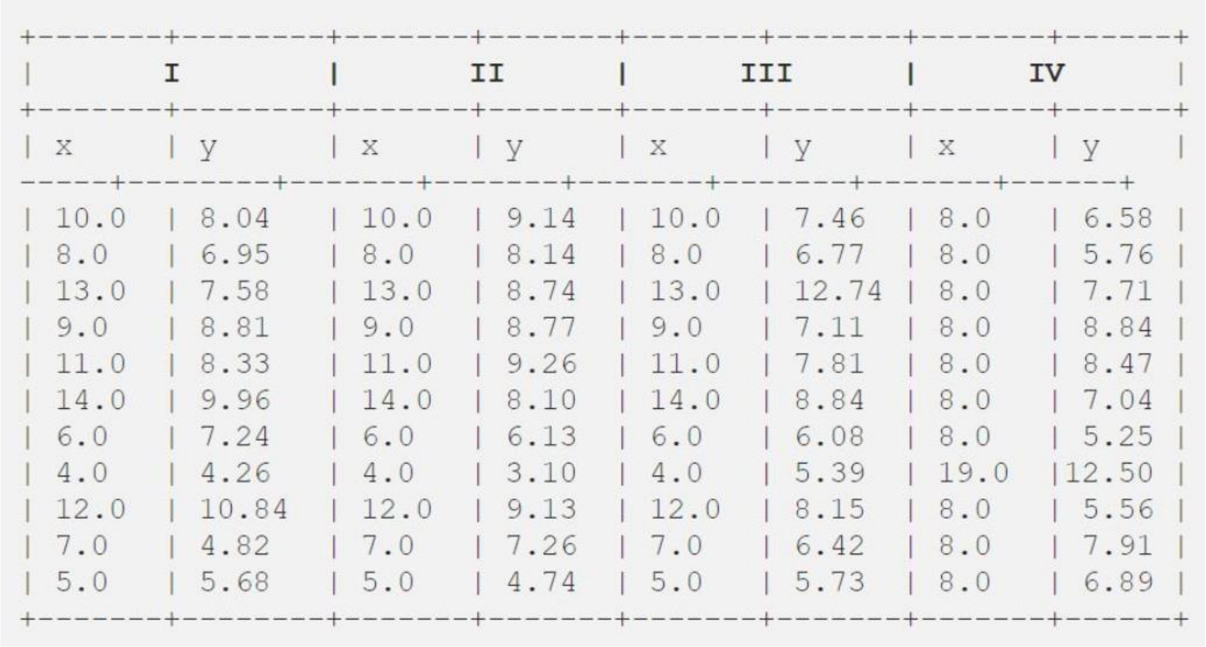
Answer:

i. Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

ii. Each dataset consists of eleven (x,y) points.

iii. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

iv. Below are the 4 datasets.



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

3. What is Pearson’s R? (3 marks)

Answer:

i. Correlation between sets of data is a measure of how well they are related.

ii. The most common measure of correlation in stats is the Pearson Correlation.

iii. The full name is the **Pearson Product Moment Correlation (PPMC)**.

iv. It shows the linear relationship between two sets of data.

v. In simple terms, it answers the question, *Can I draw a line graph to represent the data?*

vi. Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter “r” for a sample.

vii. Pearson’s coefficient is given by,

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Where,
n = the number of pairs of scores

Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

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

Answer:

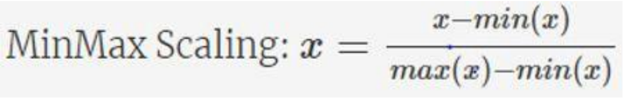
i. Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

ii. 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. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

iii. Scaling affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

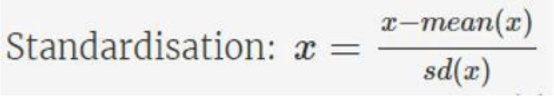
iv. There are two types of Scaling,

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



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

   b. Standardization Scaling:
      - *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** helps to implement standardization in python*



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

v. *One disadvantage of Normalization over Standardization is that it **loses** some information in the data, especially about **outliers**.*

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

Answer:

i. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.

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

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

iv. 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).

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

Answer:

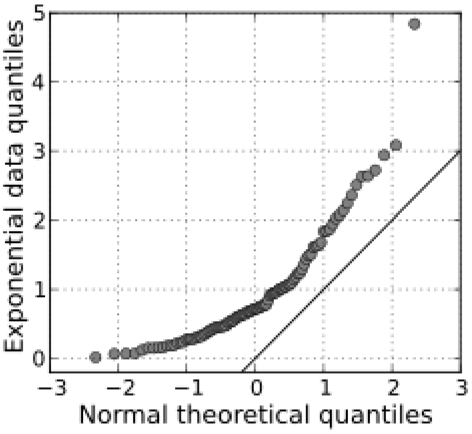
i. Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.

ii. 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.

iii. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.

iv. 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.

v. A Q Q plot showing the 45-degree reference line:



vi. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.

vii. 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.

viii. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

ix. 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.