Assignment-based Subjective Questions and answers

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: The effect of few of the categorical variables was quite significant on the dependent variable with their median (and even the entire distribution) varying significantly across categories. A few notable variables that showcased this trait were season variable, year, weather and month.

On the other hand, the rest of the categorical variables didn't have much distribution difference across its categories like holiday, working day and weekday variables.

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

Answer: It's important to use drop_first attribute because when dummies are created for all the categories present 1 condition out of the N can be represented by the False of all other flags, i.e. to represent N categories we need N-1 flags only hence, first condition is redundant.

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 highest correlation is observed between temperature and demand (cnt) but this excludes the other demand variables (casual and registered). If those are included the highest correlation is between the registered and cnt variable.

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

Answer: I validated the assumptions of Linear Regression by performing few basic checks along the data:

1. For linearity between dependent and independent variables I plot a scatterplot between the variables to observe any linearity in relations.
2. For homoscedasticity checked the plot for residuals vs predictions for patterns
3. For normality across error, I plot a histogram of residuals to check for the distribution.
4. For multicollinearity I calculated VIF for each predictor and removed the highly correlated one.

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: The 3 major variables would be temperature, year and snowy weather

General Subjective Questions

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

Answer: The linear regression algorithm is the mathematical technique which is used to determine a relationship between dependent and independent by assuming that dependent variable varies linearly with independent variables with a degree/factor of β.

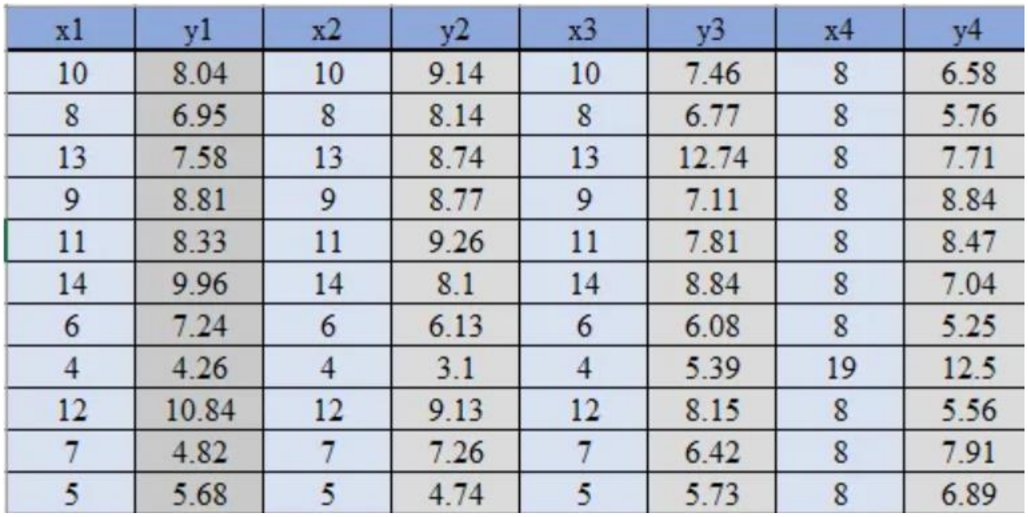
The major goal of the linear regression exercise is to fit the best line from a hyperplane of variables minimizing the error between the actual and the predicted.

There are some assumptions that linear regression algorithm operates with:

- The relationship between dependent variables and independent variables should be linear
- Distribution of error terms should be normal
- Error terms should be independent
- Error terms should have constant variance

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

Answer: Anscombe's quartet is a demonstration of 11 data points across 4 data sets with the same descriptive statistics but different graphical representations. This emphasizes the importance of visualizing the data on top of the descriptive statistics associated with the data sample.



x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

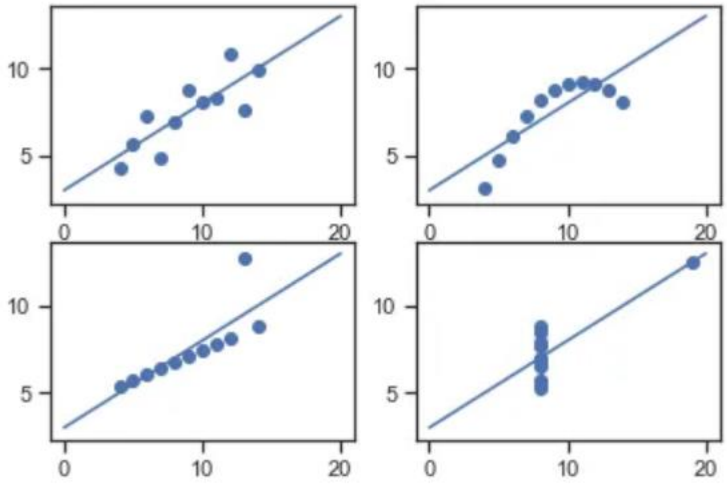
Mean x= 9

Mean y= 7.5

Variance of x= 11

variance of y= 4.12

but when graphically plotted the values look like following:



- Dataset 1: consists of a set of (x,y) points that represent a linear relationship with some variance
- Dataset 2: shows a curve shape but doesn't show a linear relationship
- Dataset 3: looks like a tight linear relationship between x and y, except for one large outlier
- Dataset 4: looks like the value of x remains constant, except for one outlier as well.

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

Answer: Pearsons's correlation coefficient is the measure of the linear relation ship between 2 continuous variables. The magnitude of the pearson's coefficient varies between -1 to 1 with the following interpretation:

- R=1 perfectly linear relation i.e. increase of one variable will increase the other variable in a defined proportion.
- R= -1 perfectly inverse relation i.e. increase of one variable will decrease the other variable in defined proportion.
- R=0 no linear relationship.

This is widely used in statistics and machine learning to determine the linearity of 2 variables/features with each other.

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

Answer: Scaling is the processing technique used in machine learning to adjust the range of a given numerical variable within defined limits. This ensures that all the data points maintain the predefined distribution and enables improved performance/better algorithm convergence for machine learning purposes.

The major difference between normalized scaling and standardized scaling is:

- Normalized scaling scales data in a set range while standardization transforms the data with mean =0 and SD=1 without any range boundary
- Normalized scaling is sensitive to outliers while standardization is not.

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

Answer: The case where VIF becomes infinite is usually the one where perfect multicollinearity exists i.e. R^2 tends to 1. This is only possible in places where one variable is an exact representation of one or more predictor variables making the value of R^2 as 1 or nearly 1 which will make the VIF to be 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: Q-Q plots are quantile to quantile plots which is a graphical representation of the theoretical quantile distribution vs the quantile distribution of the actual data. These are predominantly used to check if two sample of data came from same population or check for the normal distribution of the data. Following are the advantages of Q-Q plot:

- They enable comparison datasets of different sizes without requiring equal sample sizes
- As they are dimensionless, we can create these across samples of different units or scales
- It easily detects departures from assumed distributions which helps in assessing distributional assumptions, identifying outliers, and understanding data patterns