Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what would you infer about their effect on the dependent variable? (3 marks)

    - The demand for bikes is higher in 2019 compared to 2018
    - There is no significant difference in the demand for bikes between a working day and holiday
    - The months of August, September and October had the highest demand for the bikes
    - The demand for bikes is high when the weather is clear and the demand is low when there is snow
    - The demand for the bikes is high during fall and summer
    - There is no significant difference in the demand for the bikes between the different days of the week and Saturday had a small increase in the demand.

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

    - Dummy variable is used to encode a categorical variable into a numeric value. For example in the bike sharing case study weathersit is categorical variable with 3 values - Clear, Mist, Snow. To make use of this in the machine learning model, this variable has to be encoded into a numerical variable.

    - The dummy variable creation function converts this into columns Clear, Mist and Snow and the column values for each value will be 1 0 0, 0 1 0, 0 0 1 . As you can see, one of the variables can be clearly predicted if we know the values of the other two columns. If Mist, Snow columns are 0, then Clear must be 1. This increases the ***multicollinearity*** between the variables.

    - The solution is to encode one variable less than the number of categorical values available. This is automatically done when you set 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)

    - The temp variable has the highest correlation with the target variable

4. How did you validate the assumptions of Linear Regressions after building the model on the training set? (3 marks)
    - No Multicollinearity - The independent variables should not be correlated with each other. This was tested using Variance Inflation Factor (VIF). All the variables used in the model have VIF <5

- Normal Distribution of Error terms - The distribution should be centered around zero and approximately normal. This was validated by doing a distribution plot of the residuals.
- R-squared measures the strength of the relationship between the model and the dependent variables. The R-squared was 0.83 indicating that 83% of the data fit the model
- P-value - A low P-value < .05 indicates the variable is a meaningful addition to the model. The P-values for all the variables are < 0.05

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
    - The top 3 features contributing significantly towards the demand of the shared bikes are
        a. temp
        b. Light snow
        c. year

General Subjective Questions

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

    - Linear regression is a supervised machine learning algorithm where a mathematical model is constructed by learning the past data and is used to predict a target variable based on other independent variables.It assumes that there is a linear relationship between the target and the independent variables.

    - Linear regression can be classified in to
        - Simple Linear regression where number of independent variable is 1
        - Multiple linear regression where number of independent variables is > 1

    - Here are the steps involved in constructing a linear regression model
        a. Understanding and Visualizing the data

            In this step we do data clean up and do exploratory data analysis and use pair plot for continuous numerical variables and box plot for categorical variables. We use heat maps to understand the relationship between the variables.
        b. Preparing the data for modeling

            - We need to replace the categorical variables with dummy data to make them numeric and useful in machine learning
            - We need to scale the numeric features to be of the same scale using Normalized or Standardized scaling methods

- The past data set is always divided into two parts Training data and Test data

c. Training the model
   - We start with Recursive Feature Elimination to automatically select a base set of features good for the model. Use stats model to get the detailed statistics.
   - Compute Variance Inflation Factor (VIF). Use combination of R2, p value and VIF to improve the model
   - From the base model, we progressively eliminate features and arrive at the right model

d. Residual Analysis
   - Do a residual analysis to check if the residuals are uniformly distributed

e. Prediction and evaluation on the test set
   - Do the predictions on the test data set.
   - Do the residual analysis on the test set
   - Compute the r2 score for the test set

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

According to the definition given in Wikipedia, Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

I obtained the data from https://query.data.world/s/6p2ntncvkzj5mnvbpkaswfilryvnrk and did statistical analysis on the data set. The four set of 11 data points when analyzed show similar results

Anscombe

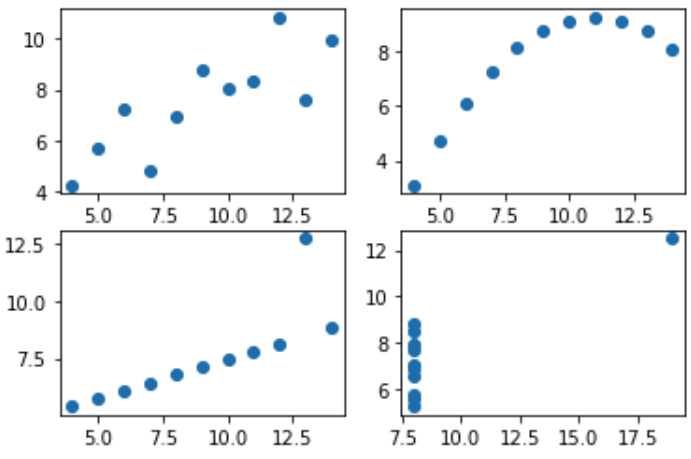
File Edit View Insert Format Data Tools Extensions Help Last edit was seconds ago

B26

	A	B	C	D	E	F	G	H	I	J	K
1		x1	x2	x3	x4	y1	y2	y3	y4		
2		10	10	10	8	8.04	9.14	7.46	6.58		
3		8	8	8	8	6.95	8.14	6.77	5.76		
4		13	13	13	8	7.58	8.74	12.74	7.71		
5		9	9	9	8	8.81	8.77	7.11	8.84		
6		11	11	11	8	8.33	9.26	7.81	8.47		
7		14	14	14	8	9.96	8.1	8.84	7.04		
8		6	6	6	8	7.24	6.13	6.08	5.25		
9		4	4	4	19	4.26	3.1	5.39	12.5		
10		12	12	12	8	10.84	9.13	8.15	5.56		
11		7	7	7	8	4.82	7.26	6.42	7.91		
12		5	5	5	8	5.68	4.74	5.73	6.89		
13											
14											
15	Mean	9	9	9	9	7.500909091	7.500909091	7.5	7.500909091		
16	Stdev	3.31662479	3.31662479	3.31662479	3.31662479	2.031568136	2.031656736	2.030423601	2.030578511		
17	Correlation	0.8164205163	0.816236506	0.8162867395	0.8165214369						
18											
19											
20											

Now when the same data set is graphically represented, it looks like

[5]:
```
from matplotlib import pyplot as plt
import pandas as pd


df = pd.read_csv("anscombe.csv")


list1 = df['x1']
list2 = df['y1']

list3 = df['x2']
list4 = df['y2']

list5 = df['x3']
list6 = df['y3']

list7 = df['x4']
list8 = df['y4']



plt.subplot(2,2,1).scatter(list1, list2)
plt.subplot(2,2,2).scatter(list3, list4)
plt.subplot(2,2,3).scatter(list5, list6)
plt.subplot(2,2,4).scatter(list7, list8)

# Function to show the plot
plt.show()
```



- The first one shows a linear relationship
- The third one shows a linear relationship with outlier
- The second one shows a non linear relation
- The fourth one shows how one point is enough to show a high correlation coefficient.

This shows the importance of visual representation in data analysis.

3. What is Pearson's R? (3 marks)
    - Pearson's correlation coefficient measures the statistical relationship, or association, between two continuous variables. It provides information about the magnitude and direction of association. It is calculated using the formula

      $$r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2}}$$

      where

      x_i is the values of x
      $\bar{x}$ is the mean of x
      y_i is the value of y
      $\bar{y}$ is the mean of y

    - The coefficient values can range from +1 to -1,
      +1 indicates a perfect positive relationship,
      -1 indicates a perfect negative relationship,
      0 indicates no relationship exists.
    - It doesn't depend on the unit of measurement and the relationship is symmetric

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

    - Scaling is the process of normalizing the range of values of the features in a dataset.
    - Datasets contain features that vary in units of measure and range. It is important to perform scaling for machine learning models to interpret these feature values on the same scale.
    - In Normalized scaling the values are scaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. Standardized scaling brings all of the data into a standard normal distribution which has mean zero and standard deviation one.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

    - When there is perfect correlation, then VIF can be infinite. In a perfect correlation, R2 =1,and VIF = 1/(1-R2) = infinity. This indicates that the variable for which the

VIF = infinity has a perfect linear relationship with the other variables. This results in multicollinearity. This/other variable need to be dropped to remove the multicollinearity

6. What is a Q-Q plot?Explain the use and importance of a Q-Q plot in linear regression (3 marks)
    - A Q-Q plot or Quantile Quantile plot is a scatterplot created by plotting two sets of quantiles against each other.
    - The Q-Q plot is used to help assess if a sample comes from a known distribution such as a normal distribution.
    - Q-Q plots can be used to check if the data in a sample is normally distributed.
    - Q-Q plots can also be used to validate that the residuals are normally distributed