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:

- We have **categorical variables** such as yr(year), holiday, workingday, weathersit, season, weekday, and mnth(month).

i. yr(year) vs cnt:
   a. 2018 mean is around 4k, and 2019 mean 6k.
   b. Insight: This means in 2019, more people will rent bikes.

ii. cnt vs. holiday:
   a. **Not holiday** have a mean of 4.5k, and **holiday** means around 3.3k.
   b. Insight: Most people borrow the bike on a non-holiday than a holiday.

iii. cnt vs. workingday:
   a. **Nonworking days** and **working days** have a similar mean and 75th percentile, but nonworking days have **higher upper whiskers**.
   b. Insight: Which means there is a slight chance that people will borrow bikes more on a nonworking day.

iv. cnt vs weathersit:
   a. Insight: We can see that people don't prefer to rent on thunderstorms most of the time**.** And Most **people rent it in clear weather.**

v. cnt vs season:
   a. Most people tend to rent bikes in the **fall and summer.**
   b. And people prefer to **rent less in spring.**

vi. cnt vs weekdays:
   a. Mean is the same for all weekdays.
   b. But Wednesday has more chance of booking less the other weekday.
   c. And Monday has a high probability of renting a bike.

vii. cnt vs mnth:
   a. people prefer to rent a bike between May to October.
   b. And gradually decrease from December to Feb.

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

Answer:

- When we create a dummy variable for a categorical variable which have more than two values as n, what it does is that it make n new columns for that categorical variable in the form of 1s and 0s.
- For example: low, medium, and high.

Low	Medium	High
1	0	0
0	1	0
0	0	1

- It will create a three-variable low, medium, and high. But it's evident that if one is medium and the second is high, then the third will be low. So dropping the first column will make low as 00, medium 10, and high will be 01. And helps to reduce the correlations created among dummy variables.

Medium	High
0	0
1	0
0	1

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

Answer:

- **temp** and **atemp** have the highest correlation with the target variable.

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

Answer:

- For building the model, I use p-value, VIF, R-square, and Adj. R-square.
- For validation,
- I check whether **the error is normally distributed or not**. If it's not, then the p-value we have obtained during the model building will become unreliable.
- **Error are independent of each other.** And errors follow no pattern and are independent of each other.
- There is a **linear relationship between X and y.**

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:

1. Temperature shows the highest coefficient of around 0.400700, which means when the temp is increased by one unit, the rental bike increases by 0.400700
2. Light Rain + Scattered clouds show a negative coefficient of around 0.291683; when this variable increases by one unit, the rental bike will decrease by 0.291683.
3. yr shows the 2nd highest coefficient of approximately 0.235183, which means when the yr is increased by one unit, the rental bike increases by 0.235183
4. windspeed shows a second-highest negative coefficient of around 0.155351; when the windspeed variable increases by one unit, the rental bike will decrease by 0.155351.

General Subjective Questions:

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

Answer:

- Linear Regression is a machine learning model which solves a regression task. And it is based on supervised learning. It is used for predictive analysis and shows the relation between a continuous variable. It establishes the relationship between the independent variable (X) and the dependent variable(y).
- The best fit line is described as:
  - y = B0 + B1X + E
  - where y is the dependent variable
  - x is the independent variable
  - B0 is slop
  - B1 is the coefficient of X
  - E is the error
- If the independent variable (X) is just one dimension/variable, it's called **simple linear regression.** And the equation is y = B0 + B1X +E
- If the independent variable(X) is more than two dimensions/variable, then ist called **multiple linear regression.** And the equation is y = B0 + B1X1 + B2X2+…+BnXn + E
- To check the best fit line, we have to minimize the error, and to prevent the error; we use the R-square for simple LR and Adj. R-square for multi LR.
  - R-square = 1- {Residual Sum of Squares (RSS) / Total Sum of Squares(TSS)}

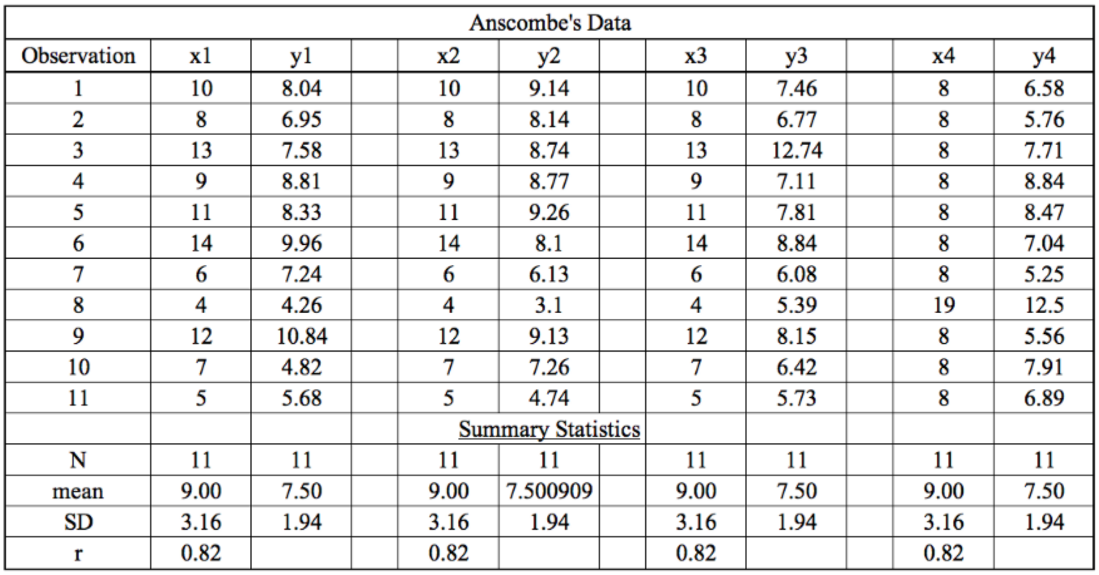
$$R_{adj}^{2} = 1 - \left[\frac{(1 - R^{2})(n - 1)}{n - k - 1} \right]$$

- n is the sample size
- k is a number of the independent variable.

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

Answer:

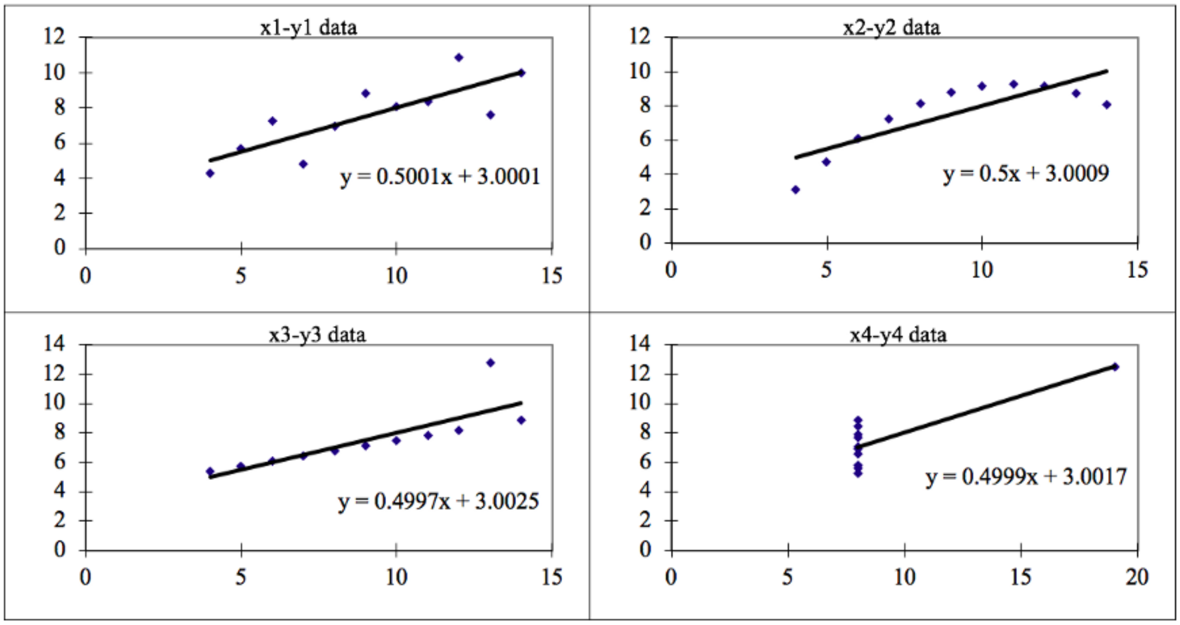
- Anscombe's quartet comprises four nearly identical datasets in simple descriptive statistics like mean, standard deviation, and count. But have a very different distribution and appears very different when plotted on the graph.
- Francis Anscombe discovered it in 1973 to illustrate the importance of plotting the graphs before analyzing and building the models. And to show the effects of observation on statistical properties and the impact of outliers on statistical properties. So four datasets with 11 data-point have identical statistical statements like mean, variance, correlation, and count.
- After that, he told the council to analyze using only descriptive statistics. And found out all have almost the same statistics.



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

- So this tells that before applying any algorithm or building model, it's essential to visualize it first. And tell us that the data feature must be plotted to see the distribution, which can help us see the outliers, linear separation of the data, diversity of the data, and more. Also, to use Linear Regression, we have to know that the data has a linear relationship.

- When the dataset is plotted by scatter plot, they realize that some of the datasets cannot be plotted using linear regression, which had fooled them.



- **Dateset1:** There is a linear relationship between x and y.
- **Dateset2:** Data is not linear.
- **Dateset3:** It has an excellent linear relationship, but because of outliers, It got distorted. That shows that linear regression can't Handel outliers.
- **Dateset4:** The high-leverage point is enough to produce a high correlation coefficient.

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

Answer:

Pearson's R Correlation checks the relationship between two variables (bi-variant) to know the strength of the relationship. The value varies from -1 to +1. And the direction of the relationship is indicated by the sign of the coefficient. -ve shows a negative relationship, and +vs indicates a positive relationship.

Assumptions:

a. Normally distributed: both the variable should be normally distributed
b. Linear relationship: both should be linearly related to each other.
c. No outlets: there shouldn't be any outlets because it's sensitive to outlets.
d. Homoscedascity: The error term is the same across all the values of independent variables.
e. Both the variable should have the same number of rows.

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

Answer:

- What is:
  - Scaling is a part of data preprocessing applied on the dataset to normalized in a particular range, and it helps to make the calculation faster for the algorithm.
- Why is:
  - Dataset is a collection of numbers in general, and it comes with different units, magnitude, and range. So higher/larger the number, the more superiority it has in some sort in the model building. Because the machine learning algorithm works on the number and doesn’t know what represents then numbers if the scale is not done, it takes magnitude, not the unit, so it is incorrect. To solve this issue, we used scaling to bring variables to the same level. However, scaling affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
- Normalized scaling/ Min-Max scaling:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

  - It brings all the data in the range of 1 and 0. And it is sensitive to outliers.
  - **sklearn.preprocessing.MinMaxScaler** is used to apply Min-Max scaling.

- Standardized scaling:

$$x_{new} = \frac{x - \mu}{\sigma}$$

  - It brings all the data into its standard normal distribution center around its mean(0) and one standard deviation.
  - If the data is not normally distributed, then it's not a good scale to use.
  - **sklearn.preprocessing.scale** is used to implement standardization.

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

Answer:

$$\text{VIF} = \frac{1}{1 - R^2}$$

As we can see from the formula, R2 is the only variable that affects VIF, and if R2=1, it means it's perfectly correlated, and it causes the VIF to be infinity. So Basically, it means the two variables have a perfect correlation. If the VIF is high, then it indicates that there is a correlation. If the VIF=5, the variance of the model coefficient is inflated by a factor of five due to the presence of multicollinearity. If the VIF is greater than ten, then there is multicollinearity. And between 5-10 is moderate, less than five will be less chance, and one will be no multicollinearity.

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

Answer:

Quantile-Quantile plot is a graphical tool to help us evaluate if the one row is plausibly
Came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. It is used to compare the shape of the distribution by checking the graph plots properties such as scale, distribution, location, and skewness the same or different from the two distributions.
If all point of quantiles lies on or close to a straight line at an angle of 45 degrees from x -axis, then it means the distribution is similar. If all point of quantiles lies away from the straight line at an angle of 45 degrees from the x-axis, the distribution is different.
This can be used to check whether:

- The training and test dataset is from the same distribution or not.
- The dataset follows theoretical distribution such as a Normal, exponential, or Uniform distribution