Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

We have performed analysis on categorical variables and mostly all the variables have an impact on the bike rentals. Below are a few observations:

Yr: Year 2019 has more bookings compared to 2018. This trend might continue as we progress forward.

Mnth: The demand for bike rental is high starting from May and almost until October.¶

Season: During summer and fall seasons the rental of bikes is high. Spring season has the least bike bookings.

Weathersit: People might be mostly avoiding going out during rain as we observe huge decline in bike rental during rain**.**

Holiday: Rentals of bike are more during non-holiday days. Might be people prefer to stay home during holidays.

WorkingDay: This doesn't affect the booking that much.

WeekDay: Bookings are higher during the weekdays.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

There are few benefits of using drop_first=True:

a) We can get rid of a dummy variable creation as can be predicted from the other variables Example: Let's say it is a cricket match result, where we are considering win and loss, when both are zero then automatically it will be drawn. So no need of having drawn value here separately.

b) This is something known as avoiding perfect multicollinearity. Which helps us in interpreting the model's intercept, where the removed column becomes the base when we calculate coefficients for the remaining columns/variables.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Here the target variable is cnt. By observing the pair plots we can say that **temp** and **atemp** have high correlation with cnt. It is a strong positive correlation which m

---

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

a) With the VIF values of final model we can confirm that there is no multicollinearity.

b) Residual distribution should follow normal distribution and when we create a scatter plot for X vs Y graph, we can see that data points are around a straight line, which confirms that the assumptions are met as it means there is linear relationship between both the dependent and independent variables.

c) Error terms are normally distributed. There will be no autocorrelation between the residuals.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

a) Temp is highly contributing to the demand for shared bikes. While plotting the pair plot also we saw a straight line between temp and cnt.

b) Winter: During winter season we see a positive coefficient, which means increase in demand for bike rental as possibility of winter season coming around increases.

c) Yr: Year has a good positive coefficient with cnt. Which means as years go by the possibility of renting bikes increase.

d) Rain: During the rain we observe a strong negative coefficient, which is on the opposite side.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear Regression algorithm determines the relationship between target dependent variables against one or more independent variables.

a) Simple linear Regression: $y=\beta0+\beta1x+\epsilon$

X -> independent variable
Y -> dependent variable
β0 -> y-intercept of the line
β1 -> The slope of the line
ϵ -> The error term

b) Multi linear Regression: $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$

where x1,x2,…,xn are the independent variables.

Assumptions of linear Regression:

i) The relationship between y and x is linear.
ii) Error terms are independent of each other and normally distributed.
iii) Observations are independent of each other.
iv) The variance of residuals is constant across all independent variables.

The goal of linear regression is to find the best fit line that minimizes the difference between actual and predicted values of y.

Linear regression uses a cost function to find whether this best fit line is suitable for the data or not. That cost function here is known as a mean squared error.

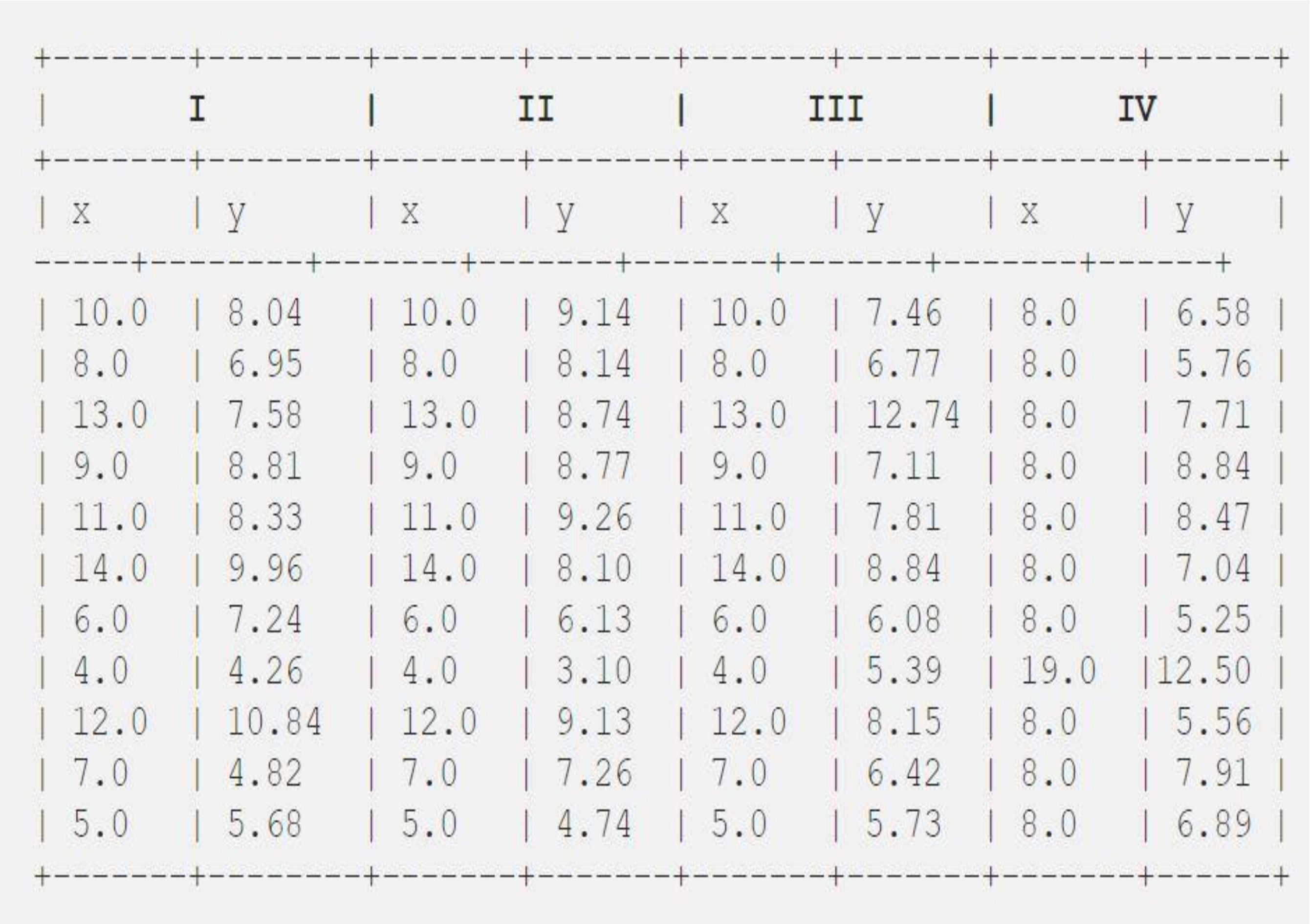
Once the coefficients are determined we can find the new data points.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet is a collection of four distinct/different datasets that have nearly identical summary statistics but have vastly different distributions and visual characteristics.

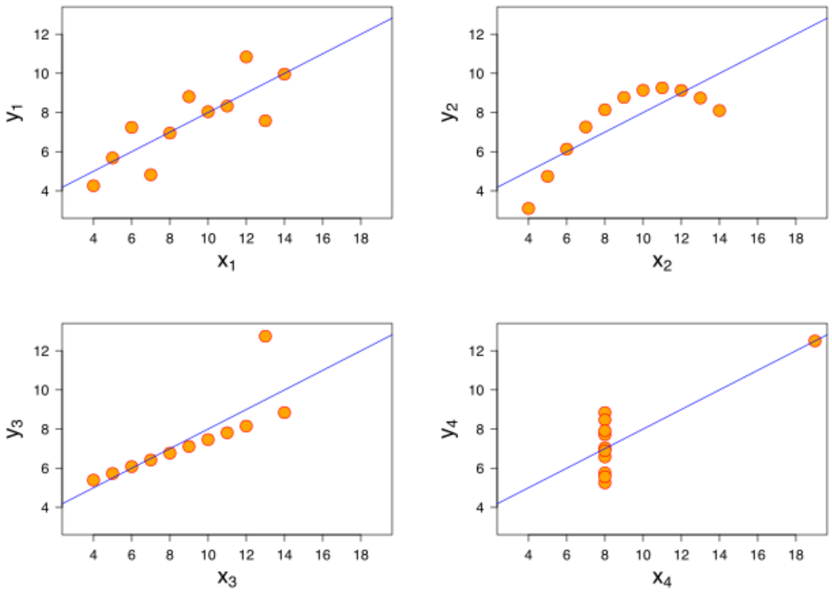
This shows the importance of visualizing data before analyzing it and limitations of depending only on summary statistics.

These are the 4 datasets that he took for example.



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

When we plot graphs for them:



The statastics for all 4 data sets are the same:

a) Same mean and variance for both x and y

b) Same linear equation y=3+0.5x

c) Same correlation coefficient of 0.816

d) Same R-squared values of around 0.66.

Question 8. What is Pearson's R? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Pearson's-R (Pearson correlation coefficient) measures strength and direction of a linear relationship between 2 continuous variables.

The formula is:

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

Where xi and yi are individual data points and x- and y- are mean values.

Pearson r can have values between -1 and 1:

i) r =1, positive linear relationship between x and y.

ii) r =0, No linear relationship between x and y

iii) r =-1, negative linear relationship between x and y.

Pearson's r only calculates linear relationships And for smaller samples x and y should be normally distributed.

Variance of y should be constant across all values of x.
The one limitation is Pearson’s r is sensitive to outliers.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling ensures consistency across features as part of preprocessing by adjusting the distribution of data.

Scaling is performed to:

i) Reduce Bias towards certain features.
ii) Normally scaled data is easier to interpret.
iii) Scaling can help models converge faster by balancing its scales.
iv) It makes sure that each feature contributes proportionally.

There are 2 types of scaling:

a) Normalization/Min-Max Scaling:
Fits data in specific range like -1 to 1 or 0 to 1.

$$X' = \frac{X - Xmin}{Xmax - Xmin}$$

X’ is normalized x

Preserves the original distribution of data within the range which helps algorithms requiring bounded data.

But Normalized scaling is sensitive to outliers.

b) Standardization/ Z-score scaling:
It transforms data to make standard deviation of 1 and mean of 0

$$z = \frac{x - \mu}{\sigma}$$

$$\mu = \text{Mean}$$
$$\sigma = \text{Standard Deviation}$$

Z is standard value.
X is the score value.

This is useful for models that consider data to be normally distributed.

Less sensitive to outliers compared to normalization scaling.

But it is not a good scaling approach for algorithms which require bounded data

.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

VIF calculates the measure of collinearity in the set of multiple regression variables.

VIF is infinite means there is a predictor variable which is perfectly linearly dependent on another variable or a combination of few variables.

Here the predictor can be easily predicted through these other variables.

VIF can be infinite when R-squared values is 1(**perfect correlation**) as the formula of VIF is: (1/1-R2).

To solve this, we need to drop the variable which is causing the multi collinearity.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Q-Q plot is a tool that helps us to tell if a set of data came from some theoretical distribution. Also helps us to determine if 2 data sets came from population of common distribution.

In the case of linear regression, we will have training and testing data on which once we apply Q-Q plot we can derive that they both came from populations with common distribution.

It is a plot of quantiles of first data set against the second data set.

Importance of Q-Q plot:

a) Here the sample sizes do not need to be equal.
b) Can confirm if residuals follow normal distribution.
c) It can confirm whether skewness is similar or different in 2 distributions.

The Q-Q plot can visually summarize any distribution.

Many distributional aspects like shifts in scale, location, symmetry change and outliers' presence can be detected from this plot.
