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?***

   As per the analysis, if we see particularly at one of the categorical variable season then we can see (from the plot) that the trend remains same for both the years (2018 & 2019). Having said that, season: fall has the maximum number of rentals for both the years.

   Also, the sales (basically cnt) are less on holidays: This might be because we have a dataset of rentals, and the people usually do not commute by bike on the holidays.

   Year 2019 has more overall rentals compared to the previous one, hence we can say the business is growing.

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

   Let's understand this with an example:
   Say, I have one column Category and it has two values: **Desk** & **Chair**

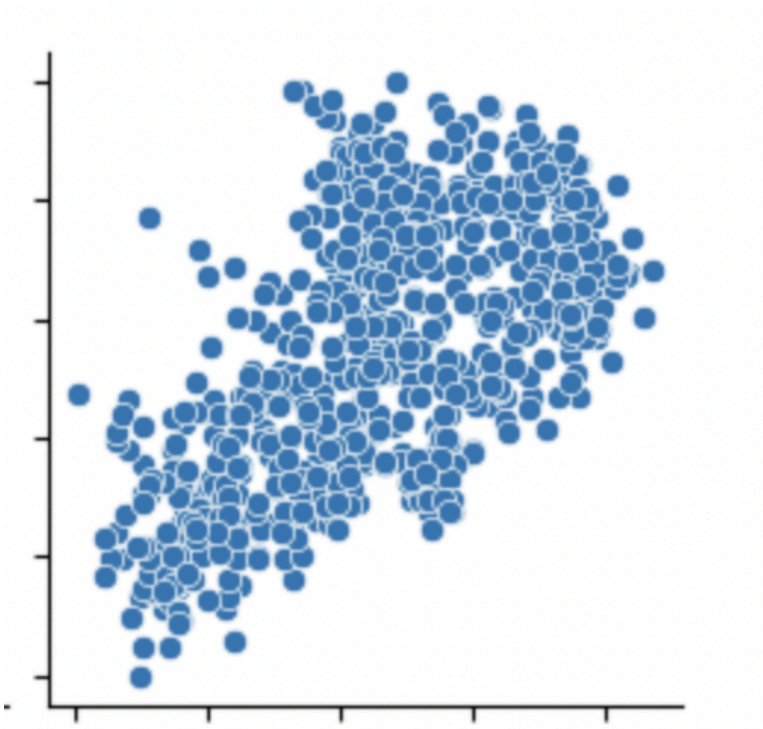
   Now, I want to create two separate columns for **Category** (with pandas.get_dummies())

   One way to create this is: One Column as **Desk** and another as **Chair** which contains values as 0 & 1 for respective rows. This method increase our dimensionality to **k** (k: number of different values in the Category)

   Another way to create is: Only one column (we can take any from Desk / Chair). This will create **k-1** dimensions. Because if we say in any row, **Desk** has value **1**, it automatically implies that **Chair** has value **0.**

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

   By just looking at the correlation graph, I'd say `temp` has the most correlation with `cnt`.



4. ***How did you validate the assumptions of Linear Regression after building the model on the training set?***

   We can check these things to validate the assumptions of Linear Regression:

   1. **Linear Relationship Validation**
      - See if the training data holds a linear relationship with the target variable.

   2. **Multivariate Normality**
      - That the residuals (observed - predicted) follows a normal distribution.

   3. **Multicollinearity**
      - That the independent variables are not highly correlated.

5. ***Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?***

   As per the final model, `temp`, `year` & `holiday` are top 3 contributing features for the predictions.

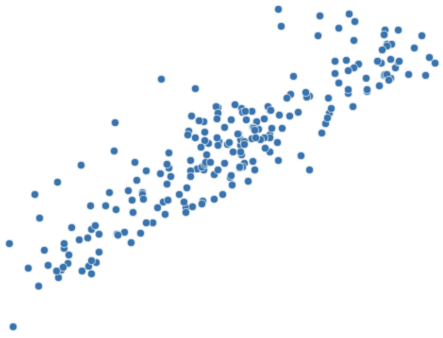
General Subjective Questions

6. ***Explain the linear regression algorithm in detail.***

   Say, you want to buy a house in your neighborhood next year. And you want to make a systematic plan of saving such that you can pay for it. For the plan, you first need to know what'd be the price that I want to save for. Now, from a super secret agency you get the data of the last 5 years of housing prices and you, being a Head of Data Science, plots the data to see the trend of the prices.

   What do you see?

   A plot like this:



   Eureka! The prices are increasing every year almost **linearly**.

   Now for the next year, you draw a **magic line** first on this data and the line will tell you the approximate price of your house for next year. Problem Solved!

How do we build this magic line?
As we know, a line equation is: **Y = mX + c**
So, we need to find out the `m` and `c`.

Here,
Y: target variable
X: training data
c: constant / intercept
m: slope of the line

Here is the algorithm for the same:

Step 1: Generate Random numbers for `m` and `c`.
Step 2: Predict Y on X with this `m` and `c`.
Step 3: Calculate **Loss** and adjust the weights accordingly.
Step 4: Repeat Step 2 & 3 until we find the optimal weights.

Loss: here refers to **(y_predicted - y_actual)** -> how much our prediction is off from actual.

7. *Explain the Anscombe's quartet in detail.*

Say you have 4 datasets with you, named as I, II, III,& IV.

Each dataset must have:

- Same mean
- Same variance
- Same statistics (overall)

BUT, when we plot them they look a whole different.
These type of datasets are known as: **Anscombe's quartet**

Example of such dataset:

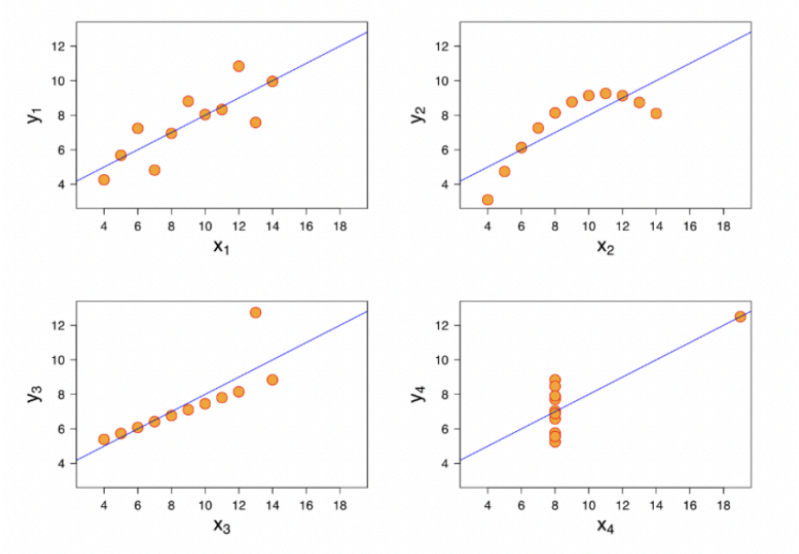
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Mean (x): *9*
Mean(y): *7.5*

Variance (x): *11*
Variance(y): *4.125*

Correlation (x & y): *0.816*

PLOT(s) of the datasets



8. ***What is Pearson's R?***

It is a measure of correlation between two variables. Also known as Pearson's correlation coefficient.

It ranges from +1 to -1.

If R > 0: it means, when one variable increases, the second also increases.
If R < 0: means, when one decreases, the second increases.
If R = 0: no linear relationship between them.

Formula for calculating:

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

Limitations:

1. Only measures the linear relationship.
2. Sensitive to outliers.

9. ***What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?***

A data pre-processing technique which adjusts the range / distribution such that they are on a comparable scale.

Why is scaling performed?

1. **Convergence Speed:** Because the values are in a comparable scale, the algorithm can converge easily compared to the large scale values.
2. **Equal Contribution:** In the algorithm, if we have too much difference in scales of values, then the model can be biased towards the higher scale.

What is the difference between normalized scaling and standardized scaling?

Normalized scaling

Transforms the data into a fixed range, typically from [-1, 1]. It is used when the model requires input bound values. Also, it can be used when data does not have any outliers and or the outliers are in the specific range.

Standardized Scaling

Transforms the data to have mean = 0, and standard deviation = 1.

It is used when the data follows normal distribution. It is also used for algorithms like Linear Regression, Logistic Regression where the assumption is that the data is following normal distribution.

10. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Yes, when we took the variable registered & casual into account, we have seen that the VIF is infinite for these. Because the cnt variable is the combination of these two.

This can happen due to the following reasons:

Dummy Variable

When dummy variables are included in the model it creates perfect collinearity with the target variable. Hence, the collinearity between the dummy variable and target variable ~1.

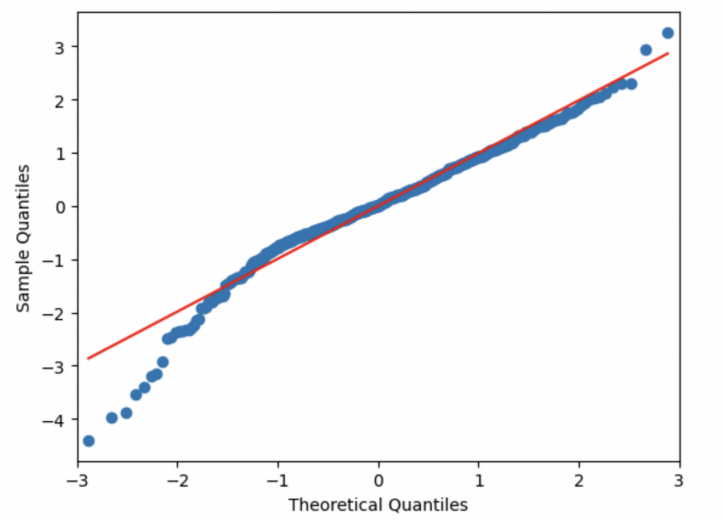
Because, VIF = 1 / 1 - R(i)^2

And the R(i) for the dummy variable is 1, hence the VIF **infinity.**

11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Q-Q (Quantile-Quantile) plot represents whether different datasets come from populations with a common distribution.

Here an example how a Q-Q plot looks:



Quantiles are points in your data below which a certain percentage of the data falls. For example, the median is the 50th percentile (or 0.5 quantile), meaning 50% of the data is below this value.

Importance of a Q-Q Plot

Residual Analysis: If the residuals follow a normal distribution, then the Q-Q plot of the same will show the points in a straight line.

Skewness: If the points deviate from the line at one end, it suggests skewness in the data.

Improving Model: If the Q-Q plot shows deviations from normality, it may suggest the need for transformations of the dependent variable or the use of different models.