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?

    - We can see the trend of higher number of bikes rented in fall season
    - Rainy and snowy weather attracted considerably lesser number of bookings as expected
    - Weekends have higher number of bookings than weekdays
    - Holidays curiously have lower number of bookings indicating the use of bikes mostly for everyday tasks rather than leisure
    - 2019 has higher bookings than 2018 indicating progress

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

    - It helps to reduce redundancy in relationships between binary variables
    - For example if there are three levels for a flat (semi furnished , furnished , unfurnished) then we can use two combinations for three levels and remove the first column
    - So drop_first=true removes the extra unnecessary level

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

    - Looking at the heatmap and pairplot 'temp' has highest correlation with the target variable

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

    - Check normal distribution of error residuals after model building
    - Check for absence of multicollinearity
    - Check for linear relationship between variables

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

    - Temp
    - Year
    - Winter (Season)

General Subjective Questions

6. Explain the linear regression algorithm in detail.

- Linear regression is the statistical model that helps us understand the relationship between a dependent variable (target) and a given set of independent variables.This is like a ratio proportion where if one variable increase the other variable's increase/decrease is directly explained by that
- But we have to be careful as this only explains correlation between variables and not causation as in we are not able to expain what causes the event only the relation
- A simple linear regression is explained by the equation

  Y = mX + c
  Y is the dependent variable we are trying to predict.

  X is the independent variable
  m is the slope of the regression line which represents the effect X has on Y
  c is a constant, known as the Y-intercept.

- Linear regression is of two types –

  I. Simple linear regression
  II. Multiple linear regression

- But linear regression is based on the following assumptions

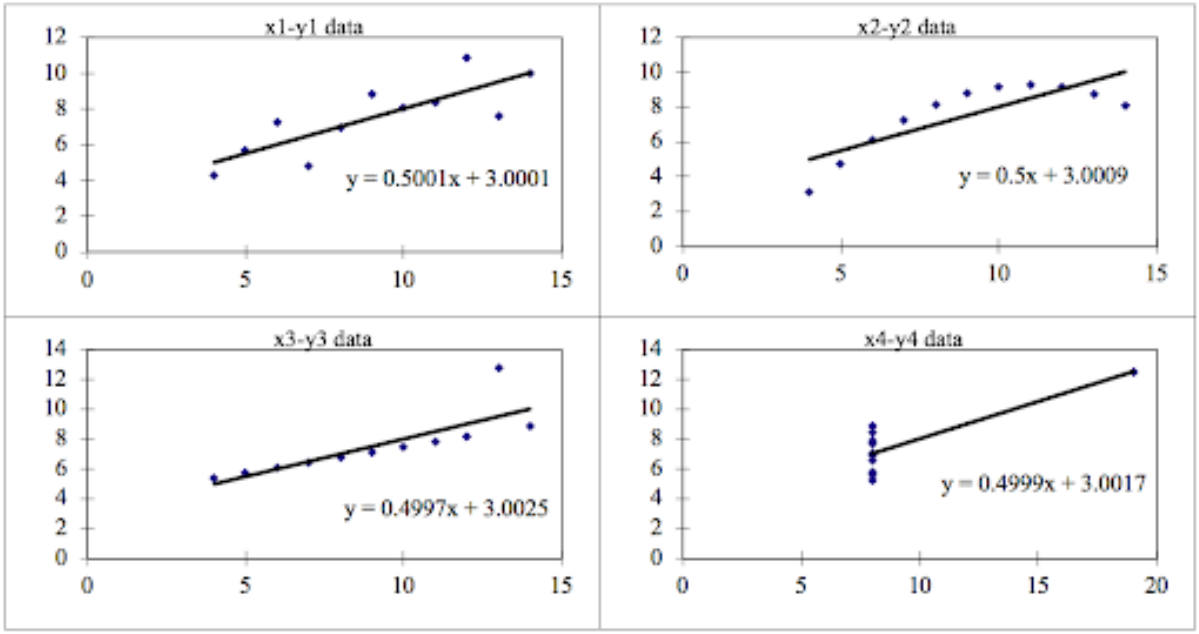
  I. Multi-colinearity – There should be no multi-colinearity between data
  II. Normal distribution – Errors terms should be normally distributed
  III. Linear relationship – Model assumes that relationship between variables is linear
  IV. Homoscedasticity - No visible pattern in residuals

2. Explain the Anscombe's quartet in detail

- "Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model". - Wiki
- "These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another". - Wiki
- Anscombe's quartet tells us the importance of visualizing data and tells us to plot the data distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.)
- Only linear regression can be considered for data with linear relationships and is not capable of handling any other data set
-

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	

- As we can observe the mean and variance are very similar among all the datsets
- But when they are plotted for the best fit line each generates a plot which will fail in any algorithm

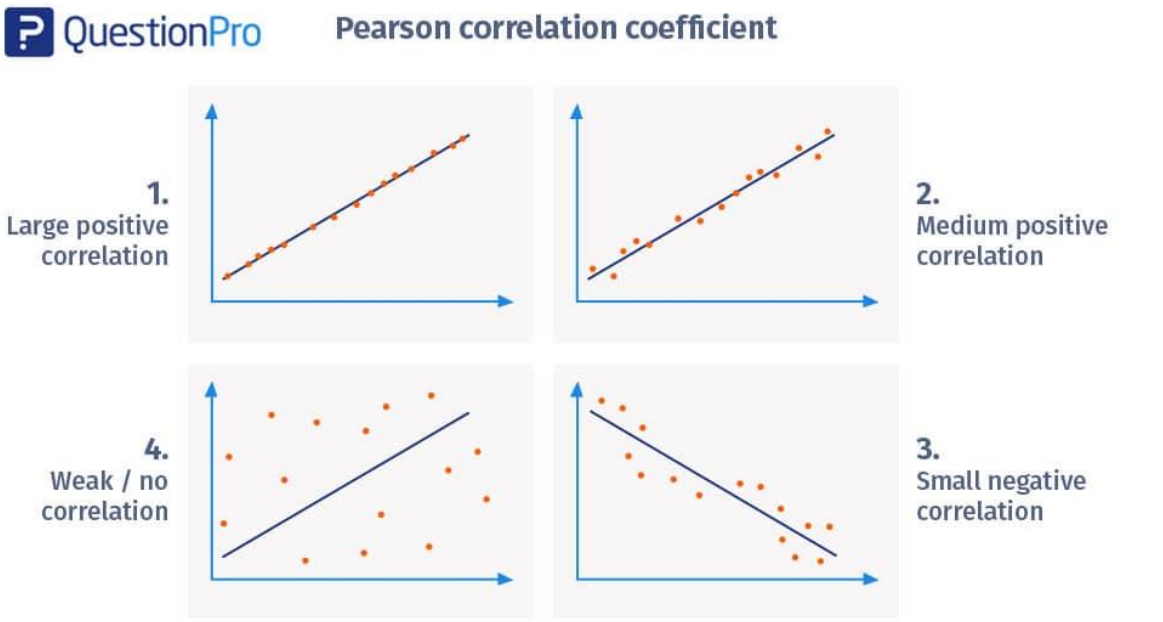


- “Data Set 1: fits the linear regression model pretty well”.
- “Data Set 2: cannot fit the linear regression model because the data is non-linear”.
- “Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model”.
- “Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model” - wiki.

- Hence this quartet is for us to conclude that it always easy to fool a regression algorithm and before attempting to build any kind of model it is always prudent to perform visualization and data cleaning

3. What is Pearson's R?

- Pearson's R or the correlation coefficient is used to measure the relation between any two variables
- This is usually used to measure linear correlation
- The value is a number between -1 and 1 . -1 indicates a strong negative relationship and 1 indicates a strong positive relationship
- For example is there a relationship between the temperature outside and the amount of bikes booked in the Bike Sharing case study



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

- Scaling is a technique that is used to standardize all the values used in a training data set so that there won't be any disparity in units causing the model to behave abnormally
- It is performed during pre-processing step to process all features which are highly variant in their magnitudes
- For example in one of my errors for the case study the temperature values were really higher than count and humidity and they were all having different unit of measures which threw a VIF of 400 for temp and 2 for hum when I first ran the model
- The algorithm will give wrong predictions in this case and hence once feature scaling is performed all the units are brought to the same magnitude

Differences :

- Normalization uses max and min values to perform scaling. Standardization uses mean and standard deviation
- Values fall between [0,1] and [-1,1] for normalized and values can be anything for a standardized scale
- We uses MinMax scaler from Scikitlearn for normalized scaling and StandardScaler for standardized scaling
- 'Normalization is helpful when feature distribution is not that clear' – wiki. 'Standardization is helpful when feature distribution is clear'

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

    - A very large value for VIF indicates that the independent variables are highly correlated and VIF infinity essentially implies there is a perfect correlation
    - Because the R2 value for a perfect correlation is 1 and according to the formula for vif

    $$VIF = 1/ (1 – R2)$$

    - Since R2 is 1 the VIF will reach infinity
    - To avoid this we have to remove the variables which are causing this internal multicollinearity

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

    - This is a technique used to determine if any two data sets come from populations with the same distributions or not

    Uses of Q-Q plot :

    - A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions.
    - A q-q plot is used to plot the quantiles of the first set against the second
    - If we say 25% of the data falls below that particular value at that quantile point the. 75% values will be above that point
    - A reference line will be plotted and then if two sets come from same population then data clustering will align with this line
    - The more spread out the data the more the probability that they are from different populations

    Importance of Q-Q plot :

- Basically this is useful in a scenario where we create the training and testing data sets from a set of data and then we can use this to find out whether the data for both comes from same population or not
- Residuals following a normal distribution is one of the assumptions of linear regression and we can verify this assumption using Q-Q plots.
- Also other factors like outlier positions and skewing of any particular values in a dataset