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)

- Yr: Median of cnt is increasing with year on year basis
- Mnth: Median of cnt rise in first 5 months; remain mostly stable during month6 to month 10 and drops in month 11 & 12
- Holiday: Median is higher for non-holiday
- Weekday: Median is almost same across weekday
- Working day: Median is almost same for not working or working day
- Weathersit: Median is dropping from season 1 to 3; no data is available for 4

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

For N level categorical variables, it is enough to have (N-1) variables since Nth variable become default if value does not lie in N-1 level. Also, objective is to keep less number of features for Multiple Linear Regression model as model is penalized for high number of features

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

Temp and atemp has highest correlation

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

Following factors are taken care:

1. VIF score is less than 5
2. Residual is normally distributed
3. Also validated all features which has p-value of more than 0.05 are dropped

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Following features contribute to highest:

1. Month of the year (positive)
2. Year since launch of service (positive)
3. And snow weather condition (negative)

General Subjective Questions

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

- **Linear Regression** is a prediction algorithm which explains the relationship between dependent (target variable) and independent variables (viz predictors). There are two types of Linear Regression
    1. Simple Linear Regression
    2. Multiple Linear Regression
- **Simple Linear Regression:** It is a basic Linear Regression algorithm which can explain relationship between a dependent variable and ONE INDEPENDENT variable using a STRAIGHT LINE

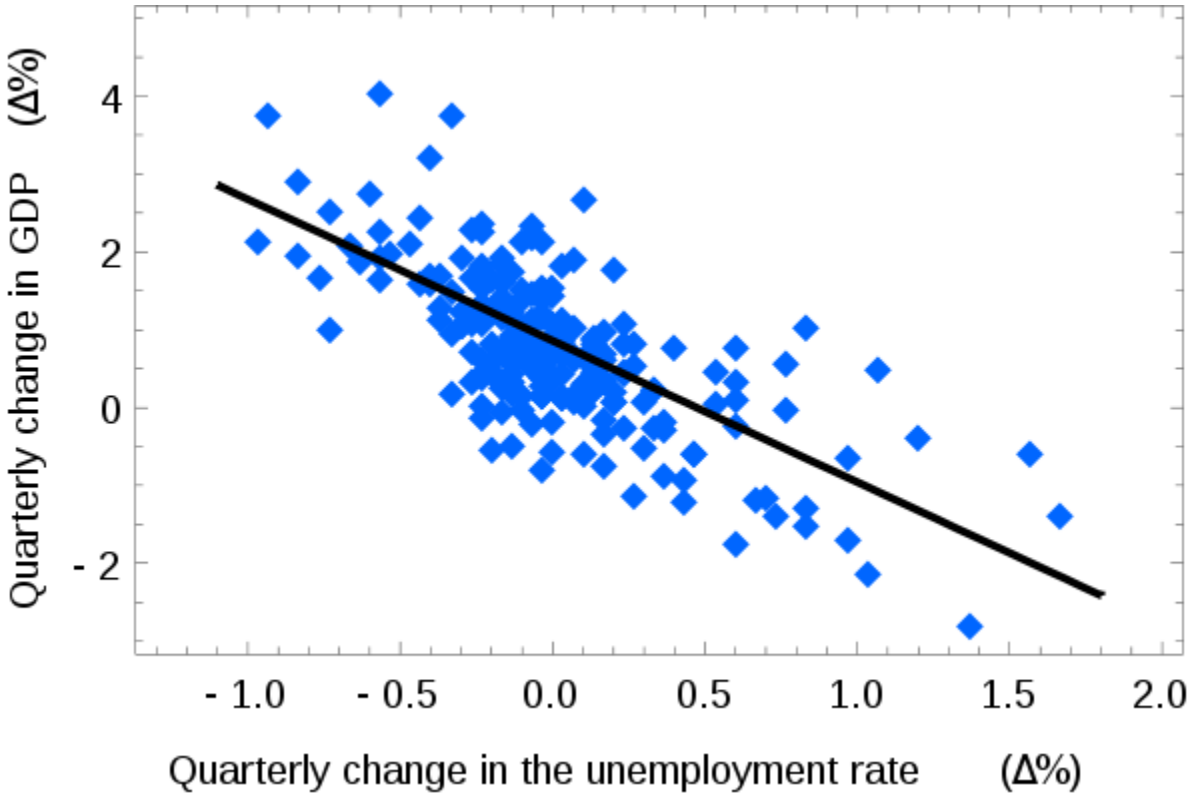


Fig Source: https://en.wikipedia.org/wiki/File:Okuns_law_quarterly_differences.svg

- Regression line is explained by following equation: y = ß0 + ß1 * X
    - Where ß0 = Intercept
    - ß1 = slope
-

- **Best Fit Regression Line** is derived by minimising the expression of **Residual Sum of Squares (RSS)** which is equal to the sum of squares of residual for each data point.

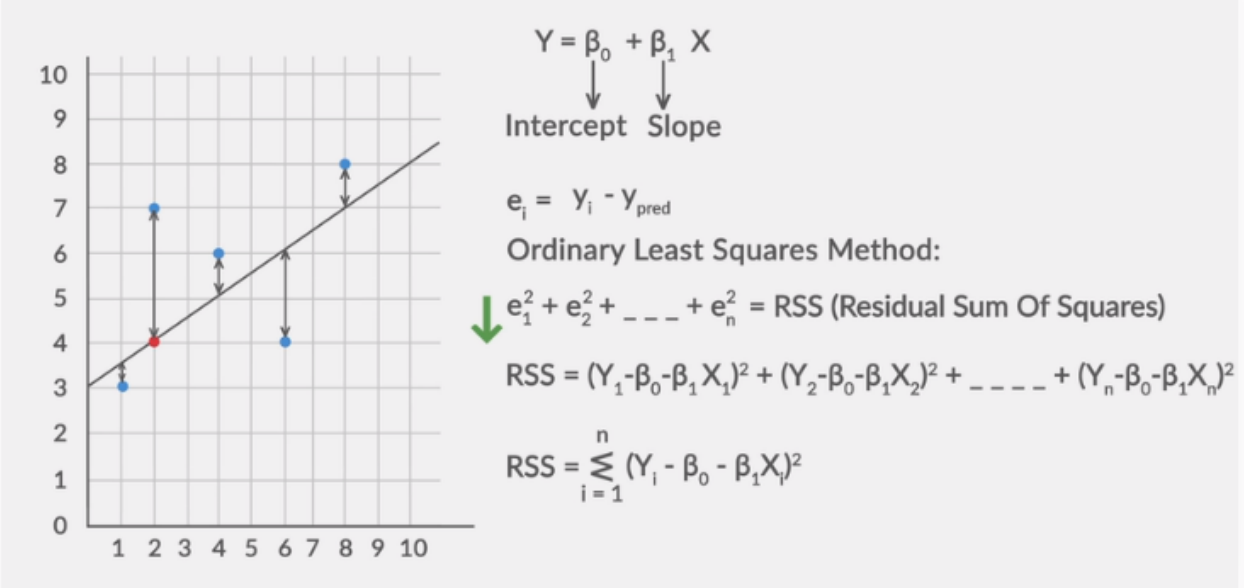


Fig source: https://miro.medium.com/max/1400/1*a0hsu_IdXOKOq88Re4xYtQ.png

- Multiple Linear Regression: It is an algorithm that explains relationship between one dependent variable and several independent variable
    - It is extension of Ordinary Least Squares (OLS) regression since it involves more than one explanatory variable

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Where Y = output variable
X = Corresponding input variables
ß1 & others = Corresponding slope co-eff
ß0 = intercept constant

- **Best fit line is derived using**
    - By selecting features (predictors) which can explain target variables in more effective way
    - Feature selections are done by taking care multicollinearity (a condition where there is strong relationship between independent features

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

Constrain of Statistical information:

- Assuming we have have 4 different sets and we perform descriptive analysis
- Outcome of Descriptive analysis, shows similar values
- Based on above, we may believe all 4 data sets has similar properties

Anscombe's Quartet:

It is the modal example to demonstrate the importance of data visualization which was developed by the statistician **Francis Anscombe** in 1973 to signify both the importance of plotting data before analyzing it with statistical properties.

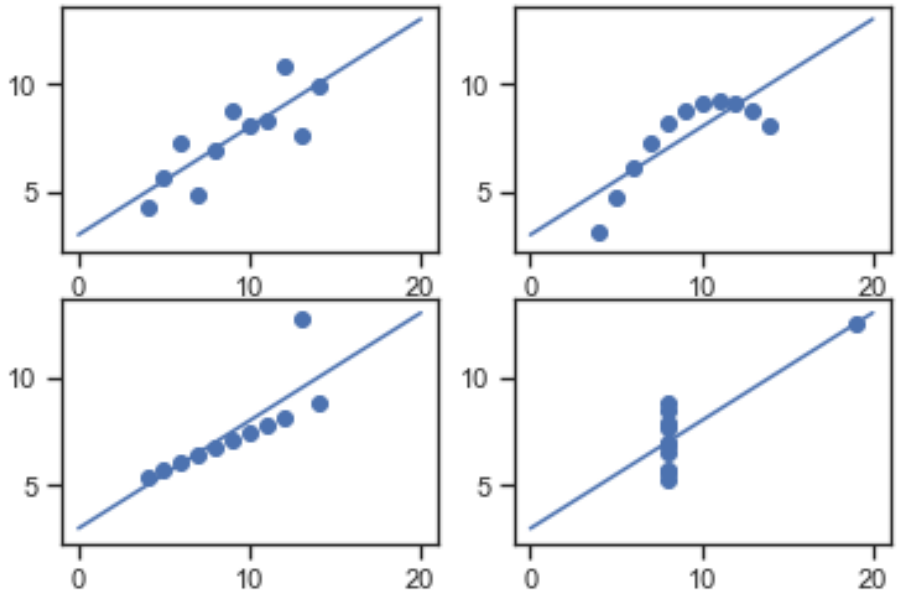
For given data:

x1	y1	x2	y2	x3	y3	x4	y4
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

Descriptive analysis outcome:

- Avg value of x = 9
- Avg value of y = 7.5
- Variance of x = 11
- Variance of y = 4.12
- Correlation coeff = 0.816
- Linear Regression: y = 3 + 0.5x

Applying visualization:

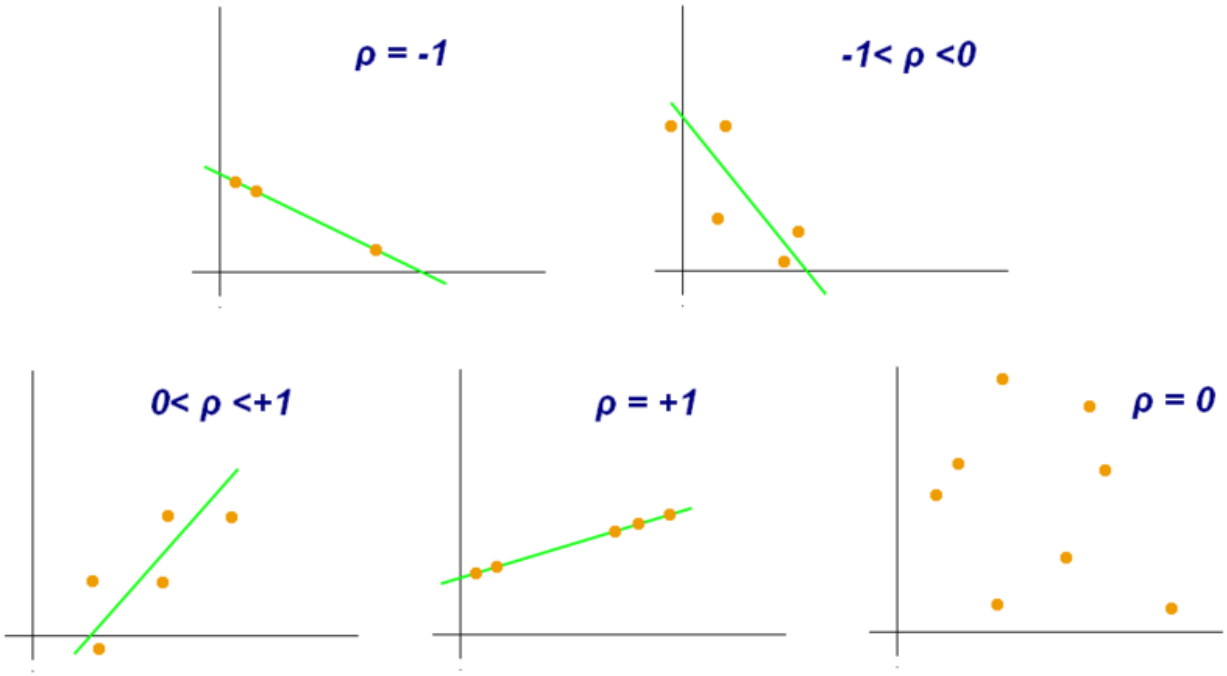


Above figure clearly explains the dataset which seems with identical statistical properties generates very different graphs.

Anscombe's quartet hence explains the significance of visualization in addition to statistical information.

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

It is a metric for measuring statistical relationship or association between two continuous variables. It is considered one of the best method for checking relationships. It provides magnitude as well as direction of relationship.



Following are the properties for Pearson's R coefficient:

- Coefficient value is in range from -1 to +1
- Independent of unit of measurement
- Values are symmetric between two continuous variables

Degree of correlation:

- Perfect: when coefficient value is either near to -1 or +1
- High: when coefficient value is in order of +/- 0.5 to +/- 1
- Moderage: when coefficient value is in between +/- 0.3 and +/- 0.49
- Low: when coefficient value is less than +/- 0.29
- No correlation: when the value is zero

Usefulness:

- High correlation independent features may require handling while building Multiple Linear Regression Model

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

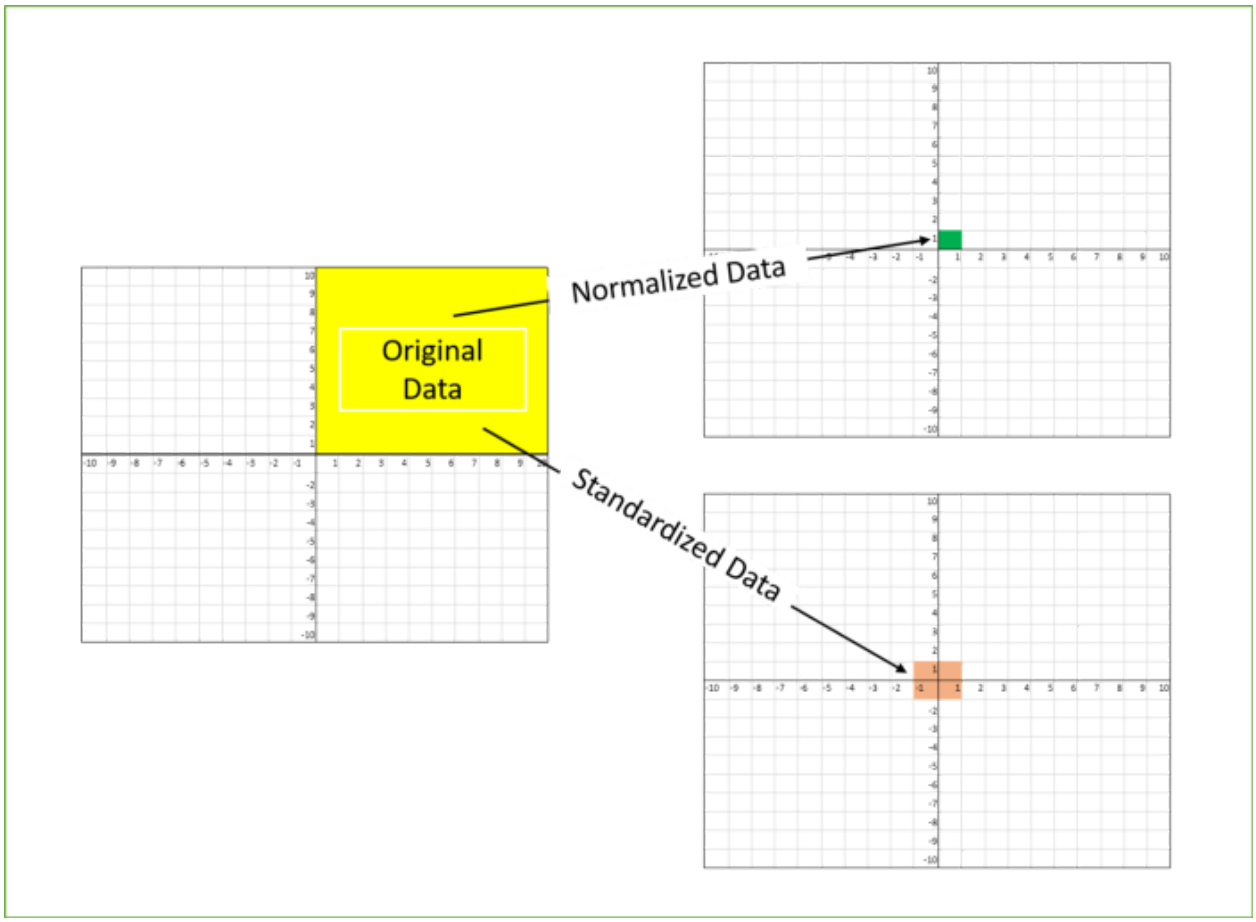
Scaling:
It is a technique to normalize the range of independent variables or features of data It is performed during the data preprocessing step in model building

Why Scaling:

1. If all features are in same range, it make easy to interpret significance of a feature irrespective of their true magnitude
2. Also improve speed for faster convergence for gradient descent method used for apt coefficient value

Normalized vs Standardized:

- Normalization is used when all values to be scaled in a given range say between 0 & 1 or -1 & 1
- Standardization transforms the data to have zero mean and a variance of 1



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

VIF score is used for checking multicollinearity among independent features being used for making Linear Regression Model. In a condition where there is PERFECT correlation (viz. when coefficient value is either near to -1 or +1), VIF score shall be infinite.

Independent features with infinite score shall require re-consideration for their selection in model building.

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

- Q-Q plot or Quantile-Quantile is a graphical tool to assess if two data sets are from populations with a common distribution.
- How to make Q-Q plot
    - Make a scatter plot using quantile of first data set against the quantiles of second data set; where quantile is a fraction or percentage of points
    - Draw a straight slope line with 45 degree angle
- Inferences from Q-Q plot
    - If points are aligned over a straight line we can conclude both data sets are from same sample population
    - High variance from straight line shall explain data is not from same population

Importance of Q-Q plot:
When performing time-series data analytics, we may need to validate if data is from same population or not before applying statistical analytics.