Bi-variate Data

Agenda

- Bi-variate Data
- Correlation
- Road Ahead - Regression

Till Now

- Considered a single variable at a time
  - Height, Weight, marks in exam, jet fuel, friends in FB
- Basis of understanding any statistical concepts
  - Allows us to summarize a single variable
- Often times, we collect lot of data on individual observations
  - FB data, Individual company data
- In real scenarios, changes in a variable could potentially affect another variable
  - Salary and effort
  - Winning and training
  - Smoking and deaths

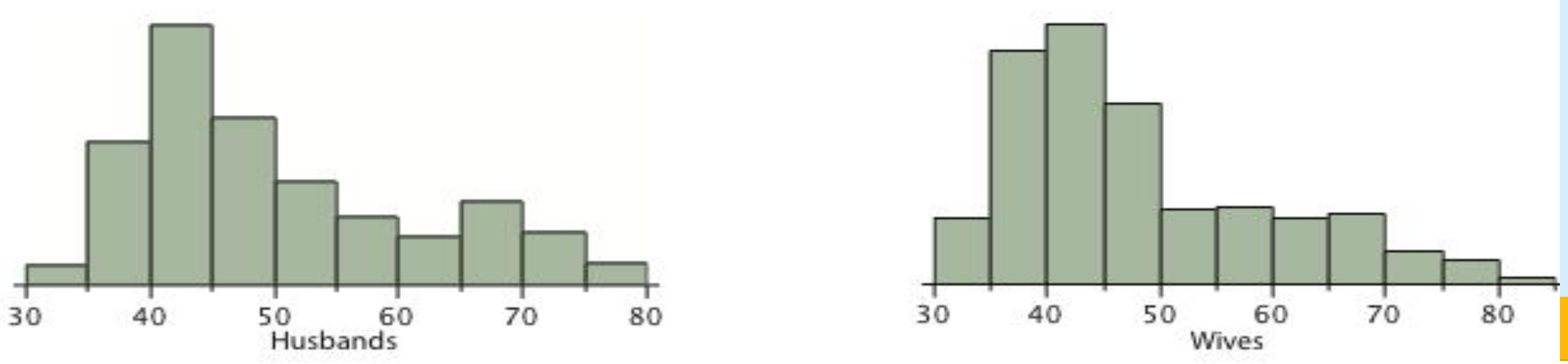
Example

- Consider following sample observations on ages

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

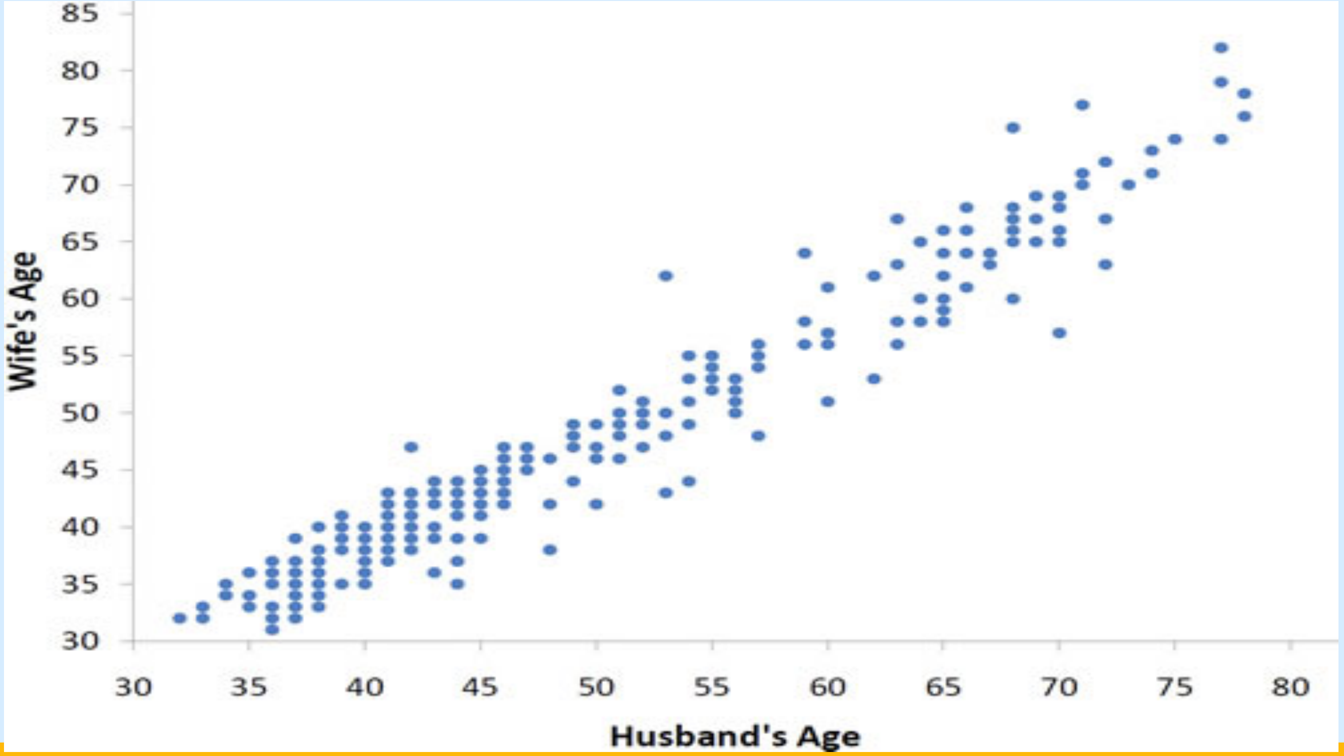
- For each variable, you can do summarization

	Mean	Standard Deviation
Husbands	49	11
Wives	47	11

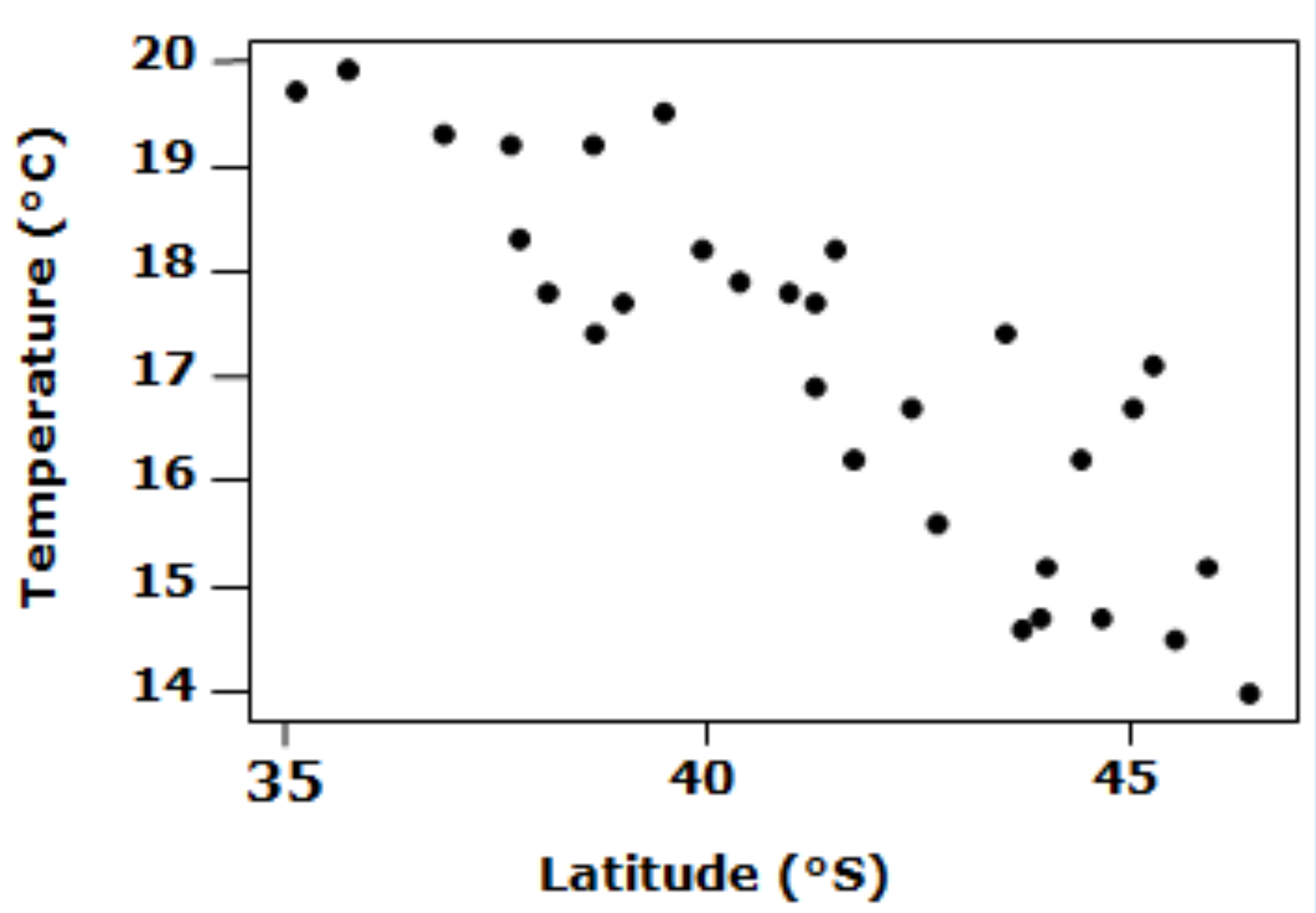


Example

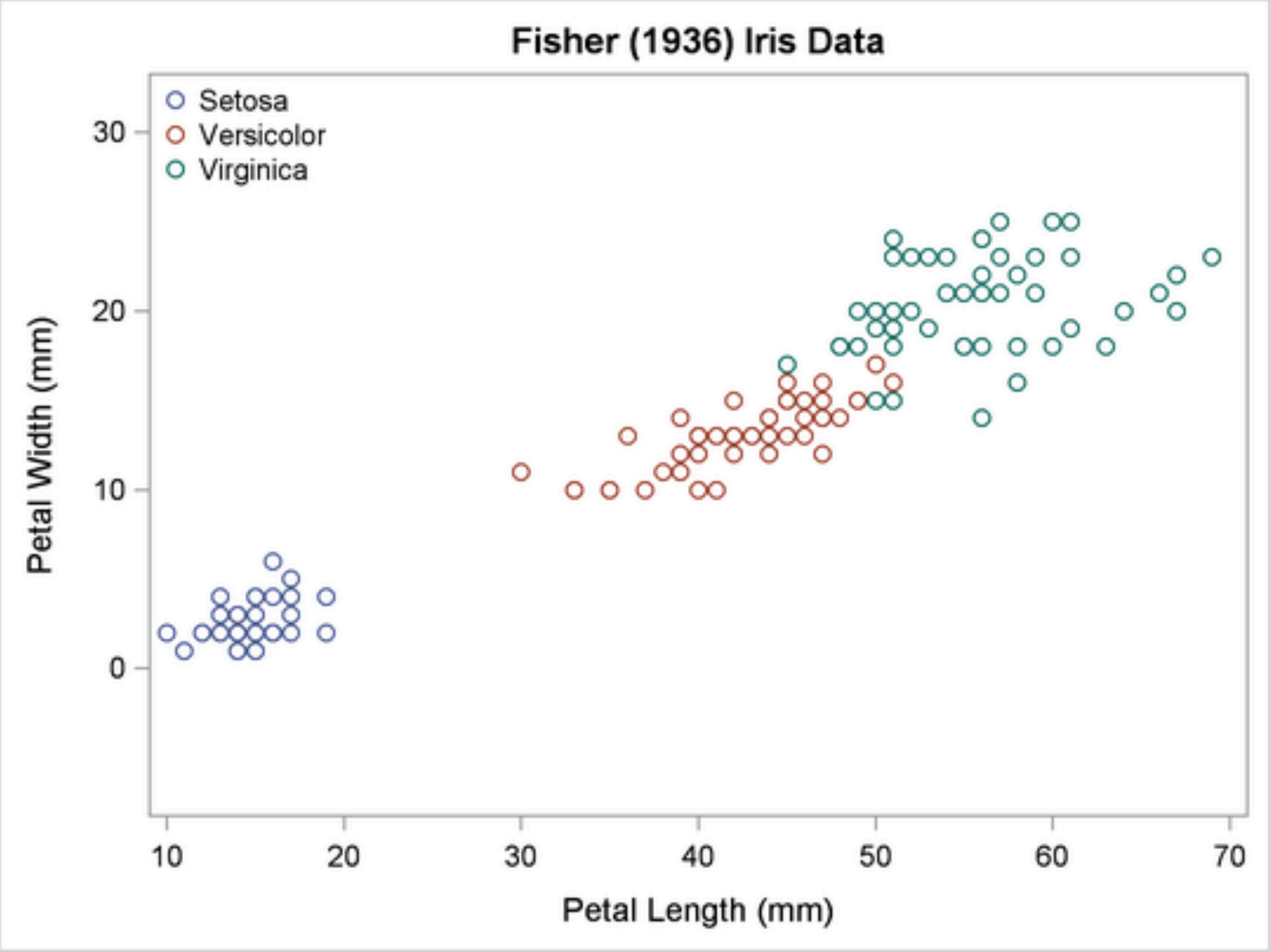
- What about following questions:
  - On average do women have younger or older husbands?
  - Overall relation between age of women and men



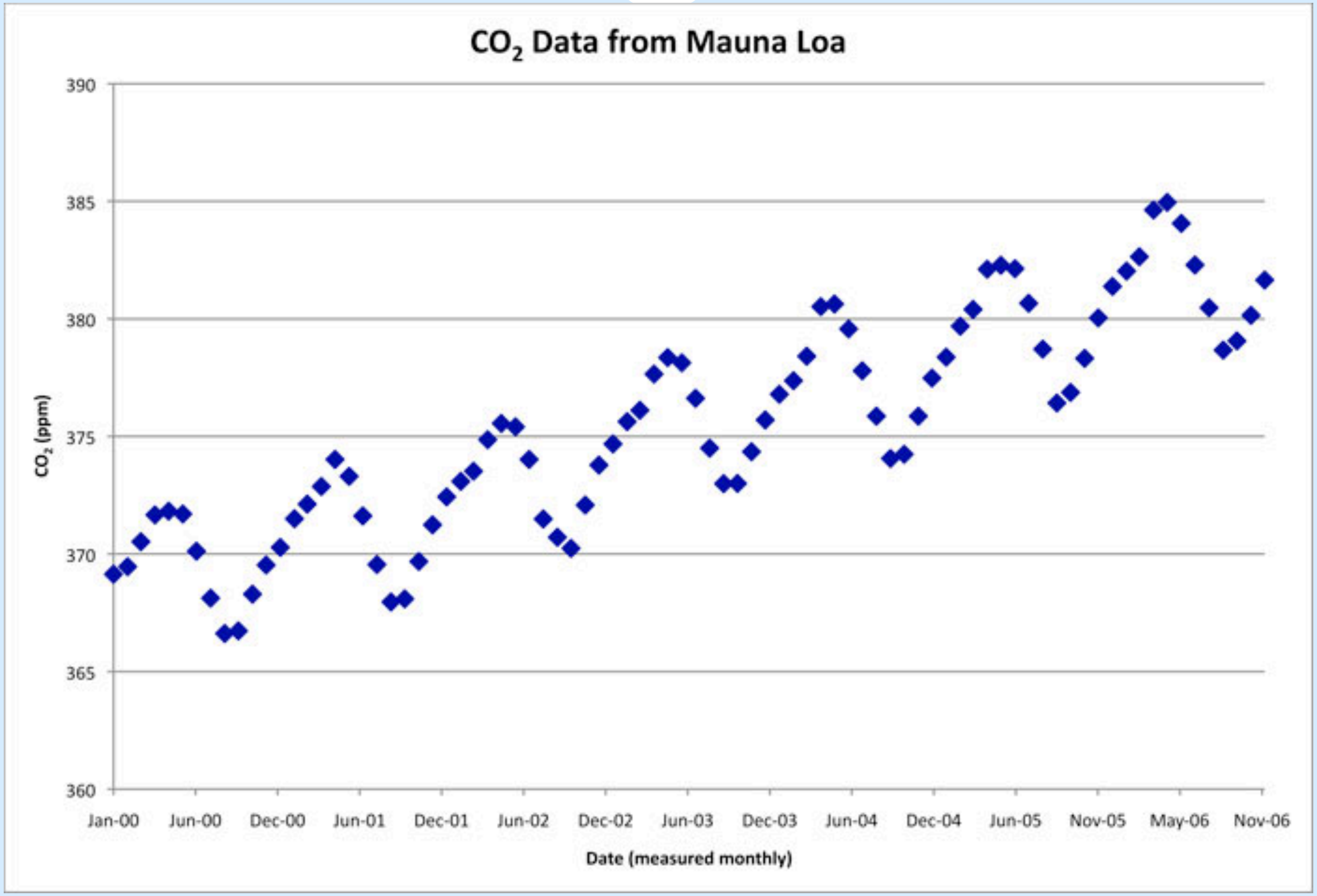
Few more examples

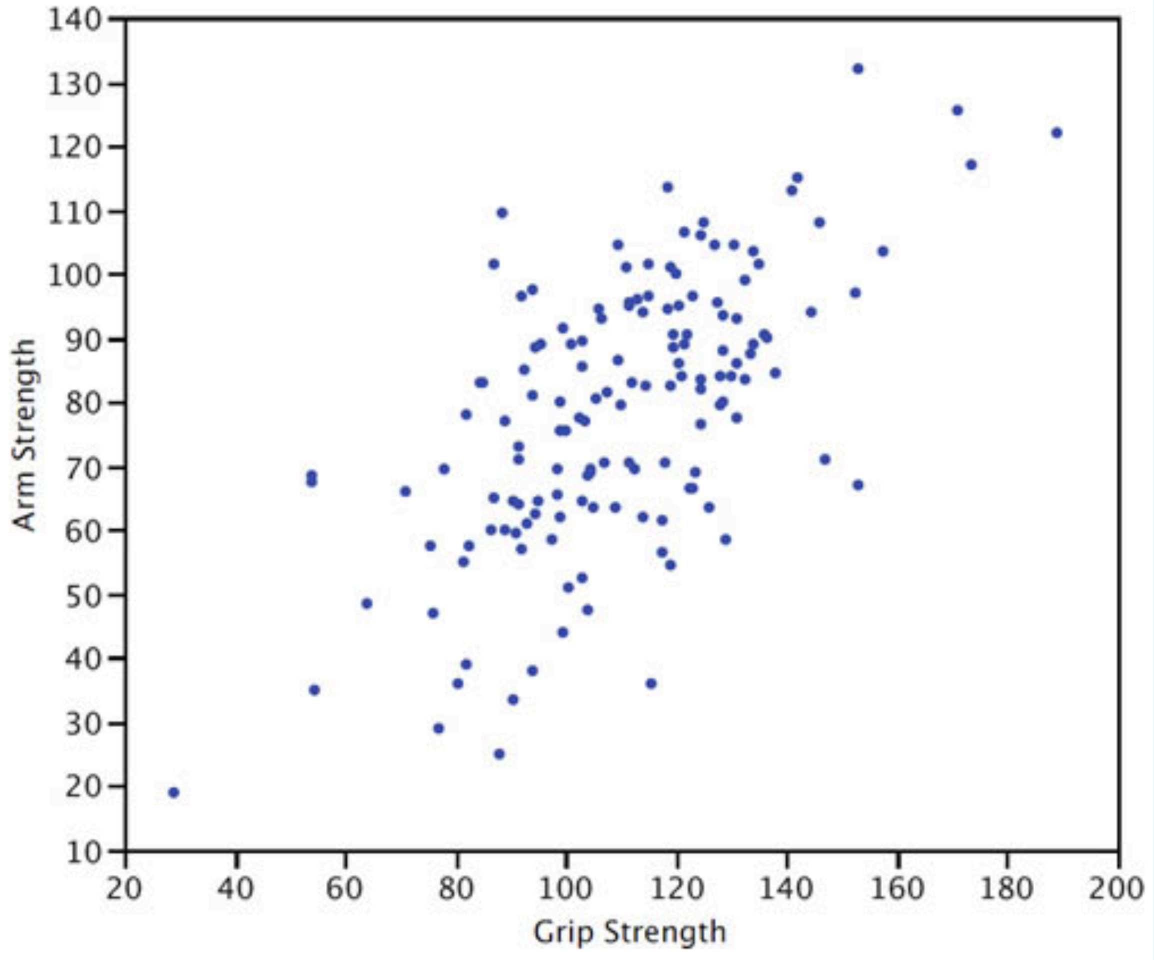


Few more examples



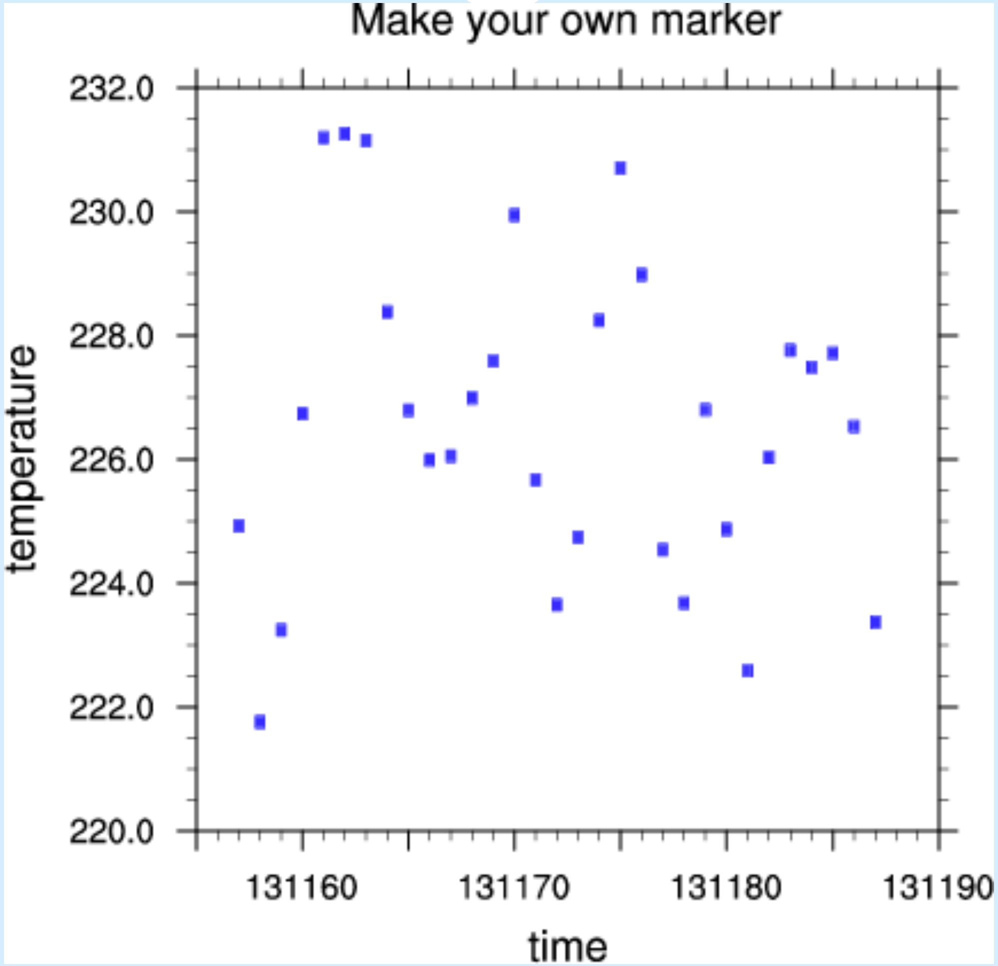
What about this case?





Arm Strength
Grip Strength
140
130
120
110
100
90
80
70
60
50
40
30
20
10
20
40
60
80
100
120
140
160
180
200

How about this?



x	y
1.0	2.5
2.0	3.9
3.0	3.8
4.0	4.8
5.0	4.1
6.0	7.2
7.0	5.5
8.0	7.7
9.0	7.1
10.0	7.9

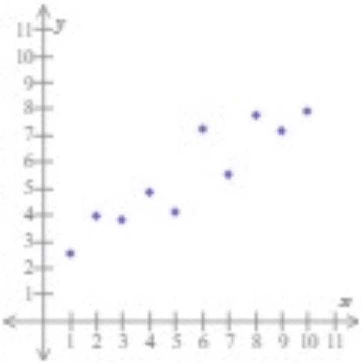


Figure 1

u	v
1.0	6.8
2.0	9.3
3.0	3.9
4.0	9.4
5.0	4.5
6.0	1.9
7.0	5.8
8.0	10.3
9.0	4.8
10.0	8.5

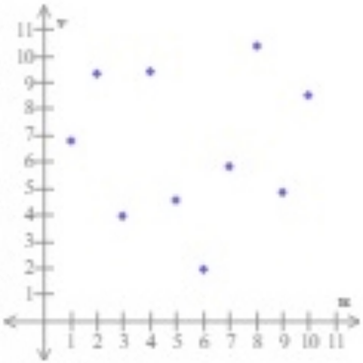


Figure 2

w	t
1.0	7.5
2.0	9.2
3.0	7.1
4.0	5.6
5.0	8.1
6.0	5.1
7.0	4.8
8.0	6.8
9.0	6.2
10.0	3.8

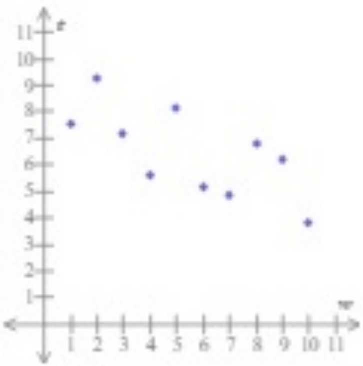


Figure 3

m	n
1.0	1.0
2.0	2.0
3.0	3.0
4.0	4.0
5.0	5.0
6.0	6.0
7.0	7.0
8.0	8.0
9.0	9.0
10.0	10.0

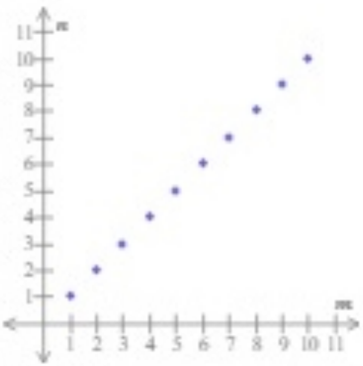


Figure 4

Bi-variate data

- Need to understand relationship between the two variables
- If one variable increases in value, what happens to other variable
  - Direction (Increase, Decrease, or no effect?)
  - Magnitude (by how much it changes)
- Basis for all predictive analytics
  - If you observe effort, can you predict salary?
  - If you observe Indian cricket team training effort, can you predict their wins?
  - If you observe humanitarian aids by UN, can you predict wars?

Covariance

- Single variable
  - Summarize by central tendency and dispersion (Variance, SD)
  - Variance – deviation from mean
- Bi-variate data
  - Observe combined deviation
  - Var(X) = $E((X - E(X))^2)$
  - Cov(X,Y) = E[(X-E(X))(Y-E(Y))]
  - Cov(X,Y) = E[XY-XE(Y)-YE(X)+E(X)E(Y)]
  - Cov(X,Y) = E(XY) – E(X)E(Y)   (Since E(E(X))=E(X))
  - Cov(X,Y) = $\frac{\sum xy}{n} - (\frac{\sum x}{n}\frac{\sum y}{n})$

Covariance

- Expected value of product of deviation of X from its mean, and deviation of Y from its mean
- Simply put, measure of how much two variables change together (not individually)
- Measure of joint variation of two variables
- Tells us how much and in which direction variables move

Example

- Consider the following dataset

X	Y	X*Y			
36	35	1260			
72	67	4824			
37	33	1221			
36	35	1260			
51	50	2550			
50	46	2300			
47	42	1974			

- E(X) = 47, E(Y) = 44 , E(XY) = 2198.42
  - Cov = 2198.42 – (47*44) = 130.42
- What can we infer?

Covariance

$$\mathrm{Cov}(X,Y) = E\Big([X - E(X)][Y - E(Y)]\Big)$$

Covariance

$$\mathrm{Cov}(X,Y) = E\left([X - E(X)][Y - E(Y)]\right)$$

+	+	+
-	+	-
-	-	+
+	-	-

Covariance Properties

- Covariance unit is product of unit of X and Y
- Cov(X,Y) <= Var(X)Var(Y)
- Covariance changes with scale.
- For any A = (X-j)/k and Y = (Y-l)/m
  - Cov(A,B) = k*m*Cov(X,Y)
  - Thus, it might be an issue to interpret covariance relatively
  - We use correlation (similar to SD in one variable case)

Example

- Consider salary and effort of Male and Female
- For female: Cov(Salary, Effort) = 562
- For male: Cov(Salary, Effort) = 434
- Can we infer Cov for female is higher than that for male?
  - No
  - We need to do standardization

Correlation

- Standardization removes the problem of comparison
- Measure of movement becomes independent of unit
- Eg: for the same dataset
  - For female, standardized: Cov(Salary, Effort) = 0.58
  - For male, standardized: Cov(Salary, Effort) = 0.72
- Now we can infer the two numbers
- Standardized measure is called correlation coefficient or rho (ρ)

Correlation

- Interpreted as degree of linear relationship between X and Y
- $\rho_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$
- Lies between [-1, +1]
- Allows us to infer both direction and strength of relationship
  - -1 => variables perfectly move against each other
  - 0 => variables do not move with each other
  - +1 => variables perfectly move with each other

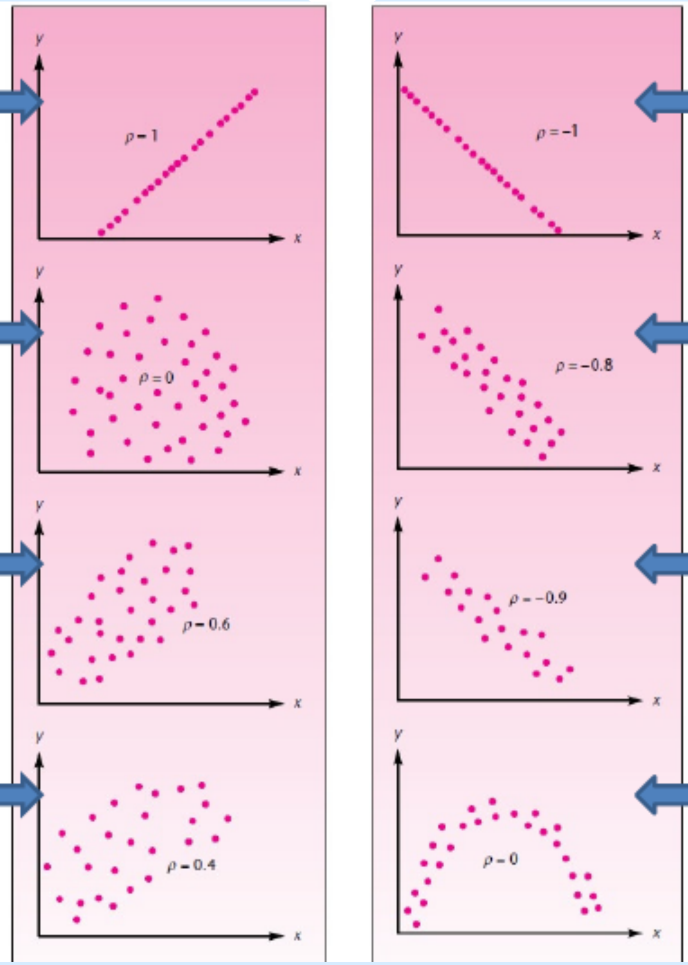
Different shapes of correlation

Perfect positive correlation. If one of X or Y increases, the other one must increase as per an exact linear relation. Similarly if one decreases, the other decreases by the same rule.

No linear relationship.

Moderate positive correlation. If one of X or Y increases, the other must increase as per a moderately strong linear relation. Similarly if one decreases, the other decreases by the same rule.

Weak positive correlation. If one of X or Y increases, the other must increase as per a weak linear relation. Similarly if one decreases, the other decreases by the same rule.



Perfect negative correlation. If one of X or Y increases, the other must decrease as per an exact liner relation. Similarly if one decreases, the other increases by the same rule.

Strong negative correlation. If one of X or Y increases, the other decreases as per a moderately strong linear relation. Similarly if one decreases, the other increases by the same rule.

Strong negative correlation. If one of X or Y increases, the other decreases as per a very strong linear relation. Similarly if one decreases, the other increases by the same rule.

No linear relationship.

Important Properties of Covariance and Correlation

- $Corr(X,X) = 1$
- $Cov(X,X) = Var(X)$
- $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
- $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y)$

Notes about correlation

- Simple and powerful concept for mathematical relationship
  - One of the earliest diagnostic tools
- However, it can be misleading
- Consider

X	-3	-2	-1	0	1	2	3
Y	9	4	1	0	1	4	9

- Corr = 0 but Y = square(X)

Notes about correlation

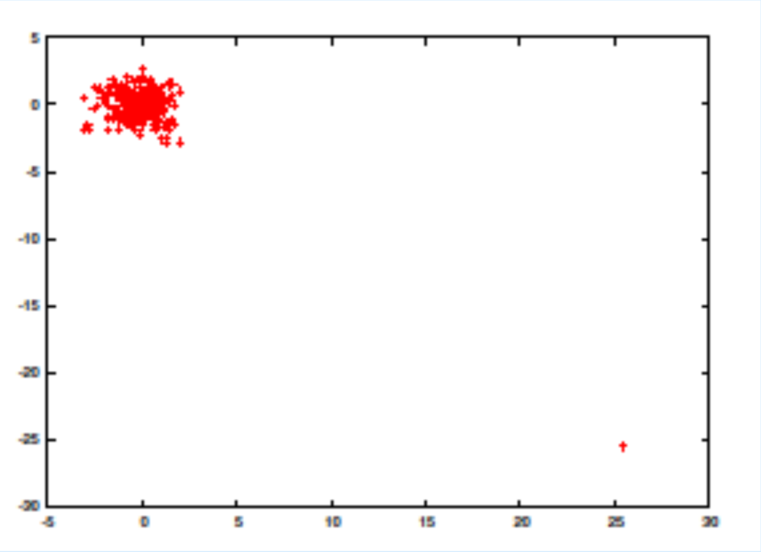
- Consider

X	.6	.2	.2	.2	.1	.1	.1	.05	.05	0
Y	2.01	2	2	2	2	2	2	2	2	2

- Corr = 0.91!!!
- Structure of data is very important (Scatter plots help here)

Impact of outlier on correlation

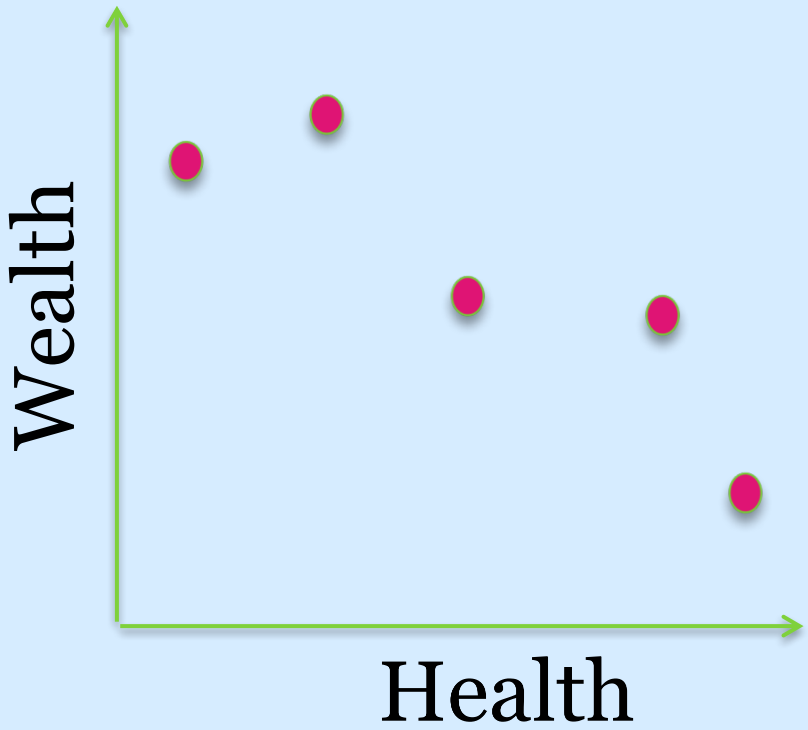
- In essence, very huge impact



- In presence of outlier, correlation is -0.75
- Otherwise, it would be 0.01

Correlation is not Causation!

- Health has a negative correlation with wealth.

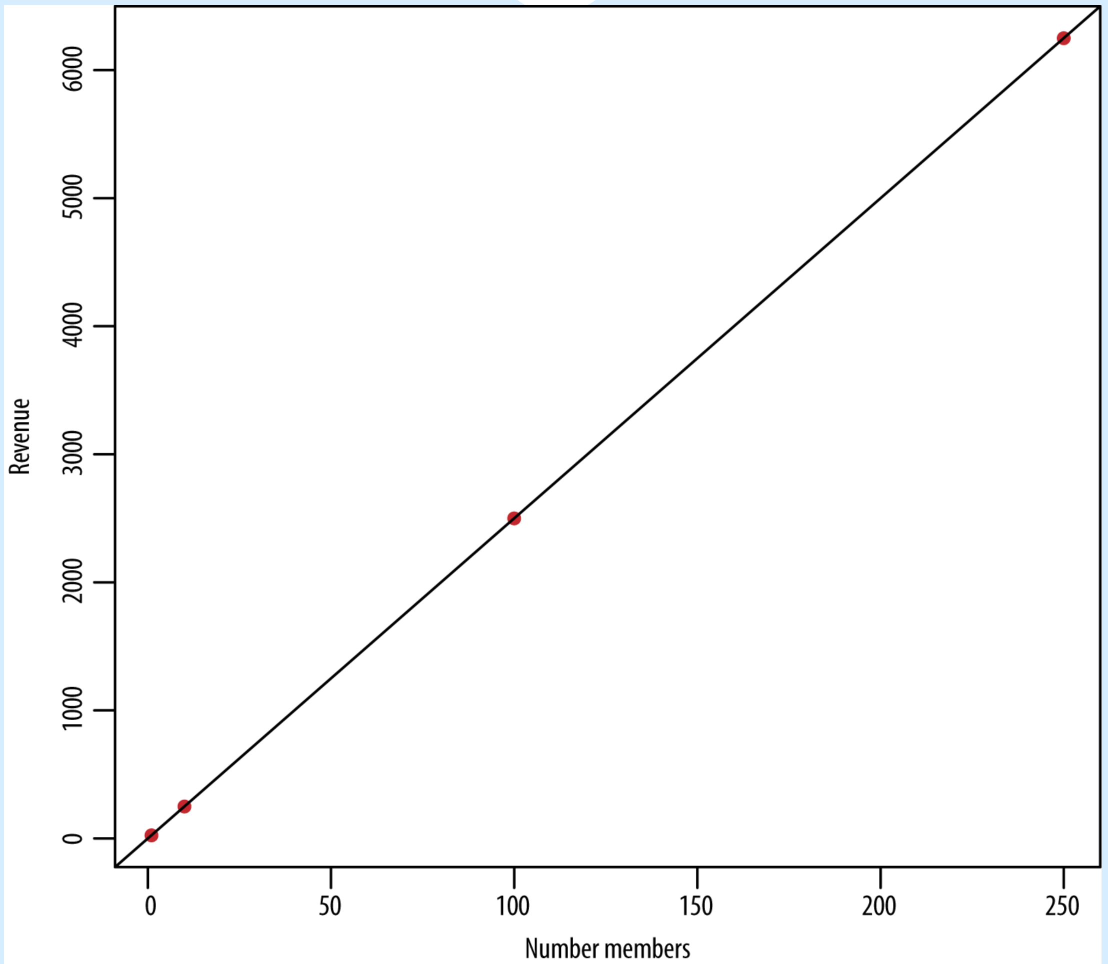


Correlation is not Causation!

- Health has a negative correlation with wealth.
- This does not mean poor health causes you to be rich. It does not mean being rich causes you to be unhealthy.
- You can predict health from wealth.
- You can predict wealth from health.

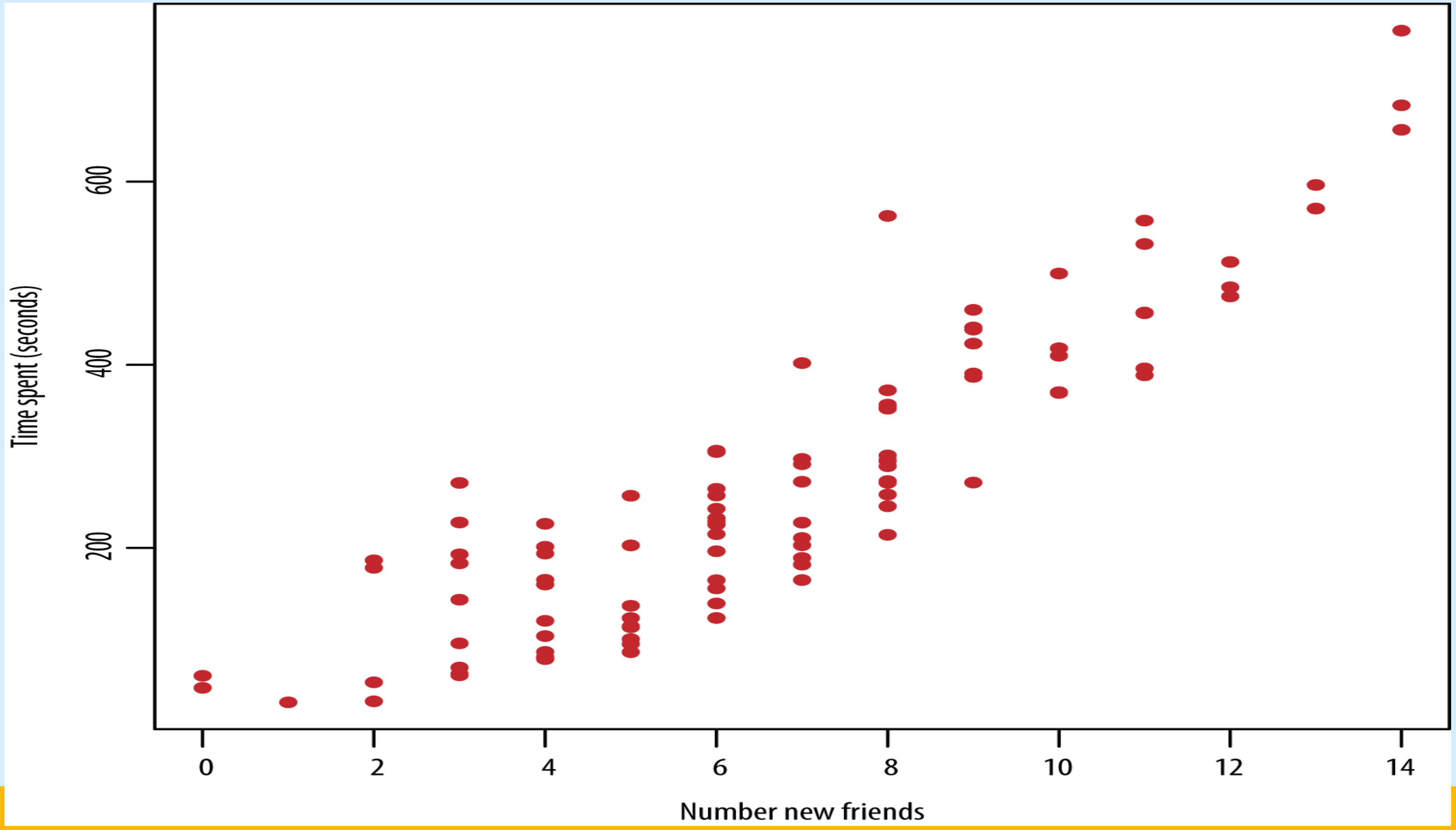
Glimpse of road ahead….Linear Regression

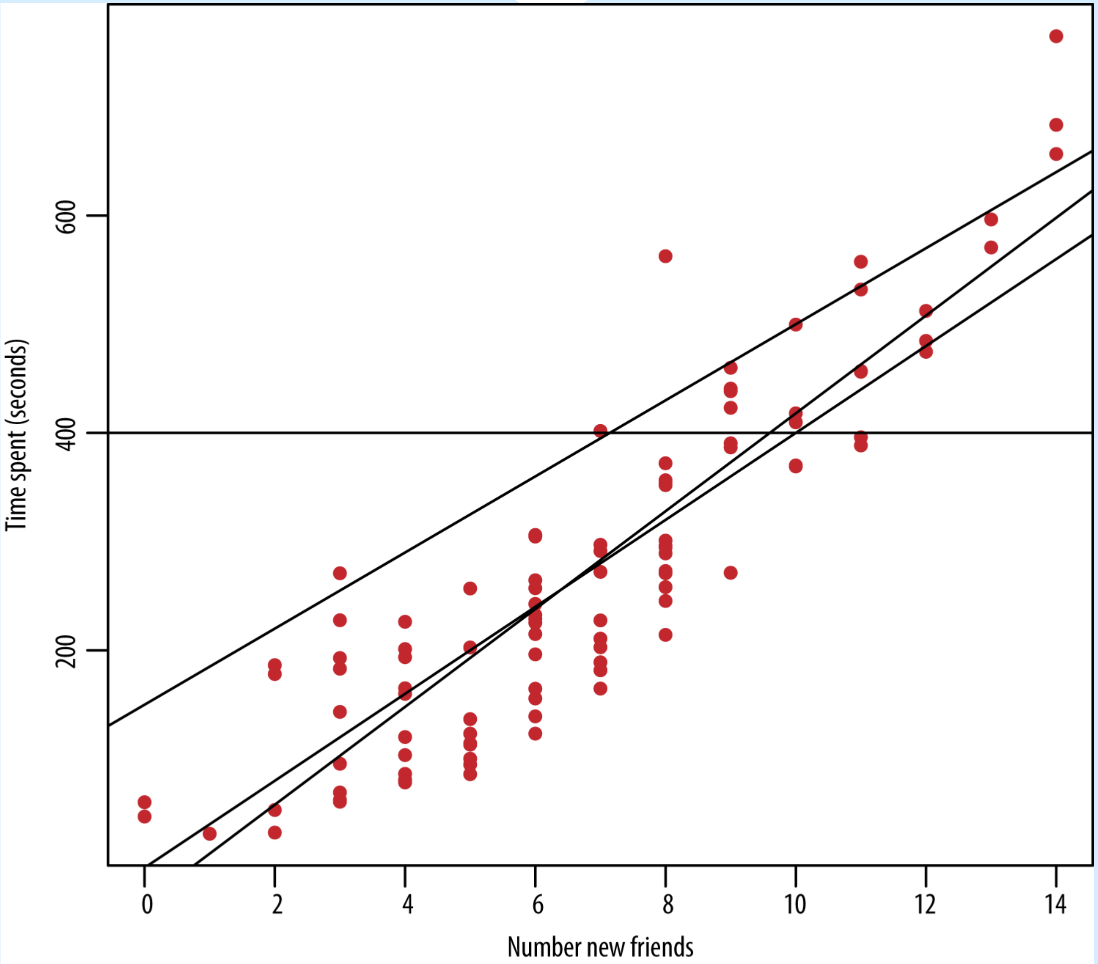
- Suppose you run a subscription service (Netflix).
- Consider following data points for months:
  - (1,25),(10,250),(20,500),(40,1000)
- Can you express a mathematical relation between x and y?
  - $y = 25x$



Revenue
0
1000
2000
3000
4000
5000
6000
0
50
100
150
200
250
Number members

- Consider the following data points





Time spent (seconds)
600
400
200
0
2
4
6
8
10
12
14
Number new friends

- You will learn about all this in coming Terms.

To Wrap It Up

- Probability & Statistics with R mini module
  - Idea was to see how can we use concepts of Probability & Stats to understand and manipulate data and its variations
  - Prob & Stats would give us the theoretical underpinnings
  - R would give us the toolset for data manipulation
- Started with Probability
  - Capture the notion of uncertain outcomes
  - Many probable outcomes but everytime we run an experiment we do not know what exactly the outcome would be
  - So we assign probability with each outcome depending on how likely the outcome is

To Wrap It Up

- Conditional Probability
  - Many events are conditional on occurrence of other events
  - Equivalent to updating our probabilities upon getting some new information
  - Sample space changes due to conditional events
  - Led us to think about Naïve Bayes theorem
    - We have prior probabilities and as we get new information our new probabilities change
    - Built a simple Naïve Bayes Machine Learning algorithm based on this concept

To Wrap It Up

- Random Variables
  - Most natural phenomena can be represented by a random variable
  - Builds on probability concepts
  - Outcomes and probability for each outcome
  - You can think of each column in your data as a random variable
  - Summarize RV – Expectation (Mean), and Spread (Variance, SD)

To Wrap It Up

- Distributions of RV
  - Nothing but a formula through which you assign probability for each possible outcome
  - Many possible distributions – we focused on Binomial, Poisson, Normal, and Uniform
  - Once we know distribution, we can assume each column of data to follow a distribution – and hence apply properties of distribution while doing analysis (Makes our lives easier)
- Statistics
  - Descriptive – describes the data
  - Inferential – Infer or predict from sample about the population

To Wrap It Up

- Descriptive Statistics
  - Summarization – Mean, SD, Skewness, Kurtosis
  - Visualization – Histogram, Line, Boxplot
- Univariate – Deals with only one variable at a time
  - Helps in understanding distribution
  - Make data transformation
  - Look for outliers
- Bi-variate
  - Statistical relationship between two variables
    - Scatter plot
    - Correlation

To Wrap It Up

- Next mini module
- Python programming – understand the de-facto programming language for Business Analytics/Data Science
- Do a lot of data analysis via case studies