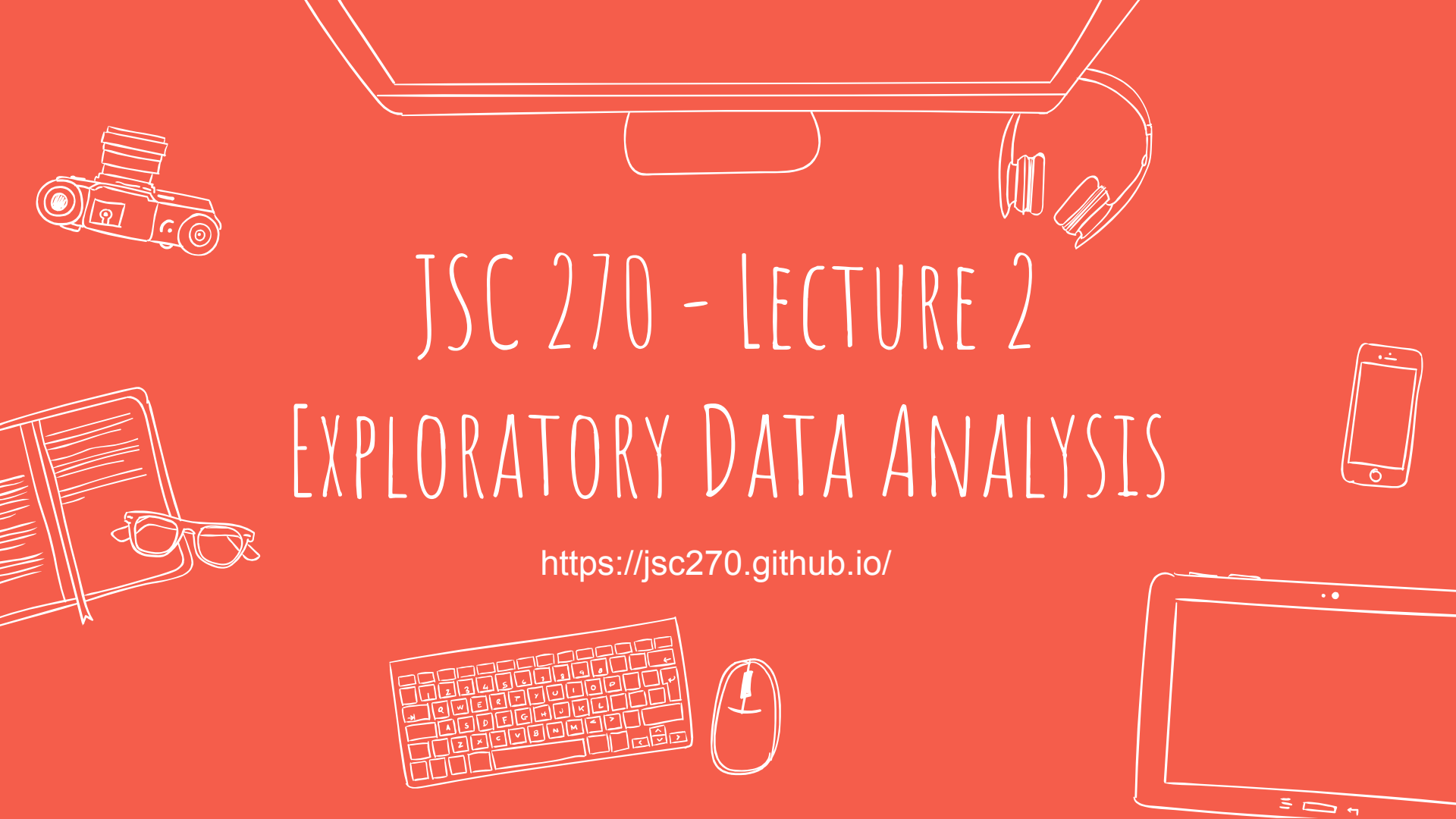


JSC 270 - Lecture 2
Exploratory Data Analysis

https://jsc270.github.io/

Announcement 1

New Office hours for TAs

Matt Edwards: Tues 9a-10a
Lauren Erdman: Thurs 4p-5pm

Announcement 2

Guest lecturer JSC370: Ben Allison. Tuesday, Jan 19 from 3:10pm to 4:00pm

Ben Allison: Principal Machine Learning Scientist at Amazon. Mr Allison is one of the architects / managers of Amazon's recommender system, which I think is one of the world's largest.

Ben will be talking about the ins and outs of his job, and building and deploying models in real-world settings. Your class is also welcome to stick around for the lecture afterwards.

The zoom link is https://utoronto.zoom.us/j/89713225921

Meeting ID: 897 1322 5921

Exploratory Data Analysis (EDA)

A definition: descriptive and graphical summaries of the data used to understand your data, relationships among variables and problems (missingness, outliers, other data entry errors) that are in the data

Benefit: allow to explore data without making any assumptions

Data

Subject

Sample

Variable/Feature:

Intelligence

Durability

Powers

Strength

Speed

....

Marvel character dataset: https://www.kaggle.com/dannielr/marvel-superheroes?select=charcters_stats.csv

Data

Subject

Sample

Variable/Feature:

Cap size

Cap color

Habitat

Poisonous

Odor

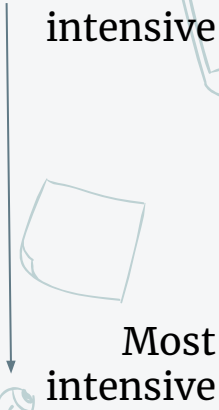
....

Mushroom tabular dataset: https://archive.ics.uci.edu/ml/datasets/Mushroom

Getting Data

Normally in your DS projects u will be either

Given data
Required to download data
Required to scrape data of the web



Types of Data

- Categorical
  - Nominal
  - Ordinal

- Numerical
  - Discrete
  - Continuous

Categorical: Nominal

Two or more categories where the ordering doesn't matter

1. province - {Ontario, Quebec, ...}
2. hair color - {black, brunette, blonde,...}

Categorical: Ordinal

- Ordinal - order matters but not the difference between values

1. income {low, medium, high}
2. education {elementary school, high school, some college, college grad}

Binary Data

Data that has two states

Example: bit {0,1}
Surgery: {yes, no}

Is binary data nominal or ordinal?

Numerical Data: Discrete

Numerical data: countable number of values

Can’t measure but can count

Example 1:

Number of times apples were bought at a store per day

Example 2:

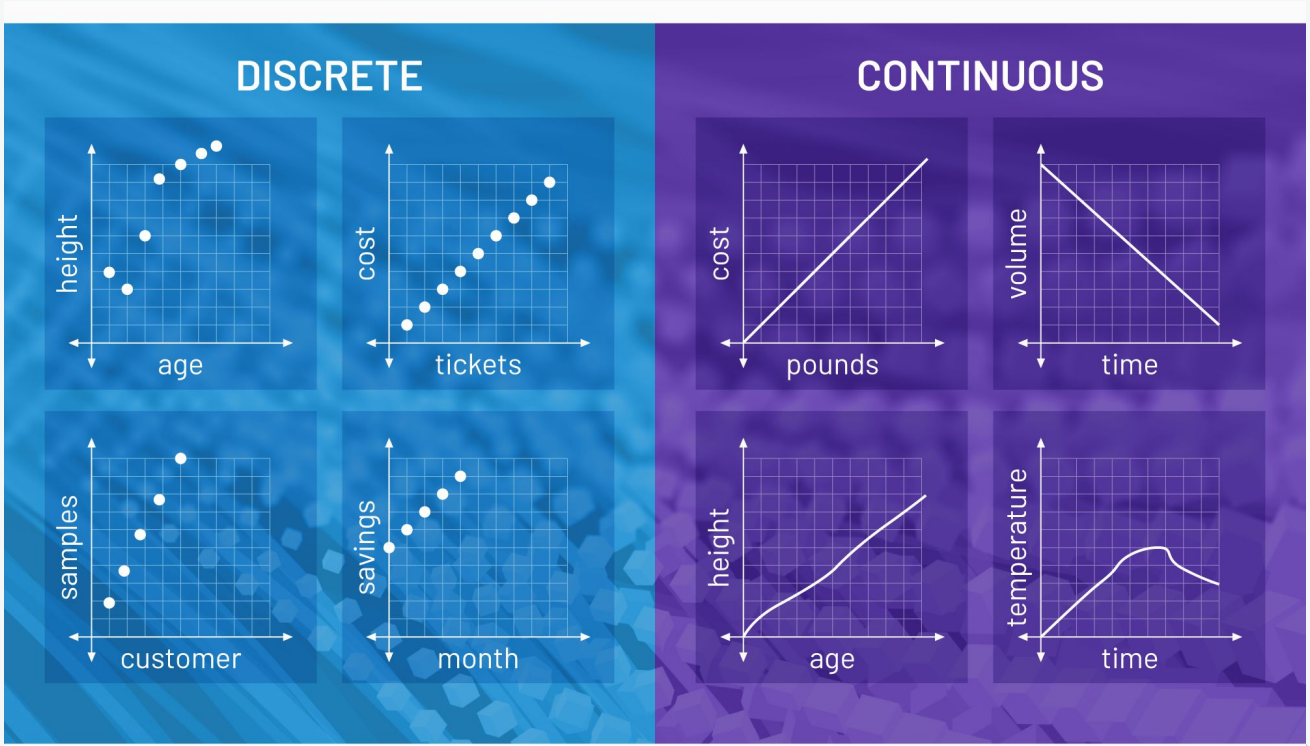
Number of people in a Data Science program

Numerical Data: Continuous

Represents measurements - can be measured but not counted

Example: age, weight, height, income

Numerical: Discrete vs Continuous



https://learn.g2.com/discrete-vs-continuous-data

What type?

Salary $55k, $77.2k, $86.1k, ...

What type?

Salary $55k, $77.2k, $86.1k, ... - continuous

Salary $25k, $26k, $40k,... (rounded)

What type?

Salary $55k, $77.2k, $86.1k, ... - continuous

Salary $25k, $26k, $40k,... (rounded) - discrete

Salary $10k-$19k, $20k-29k, $30k-$39k...

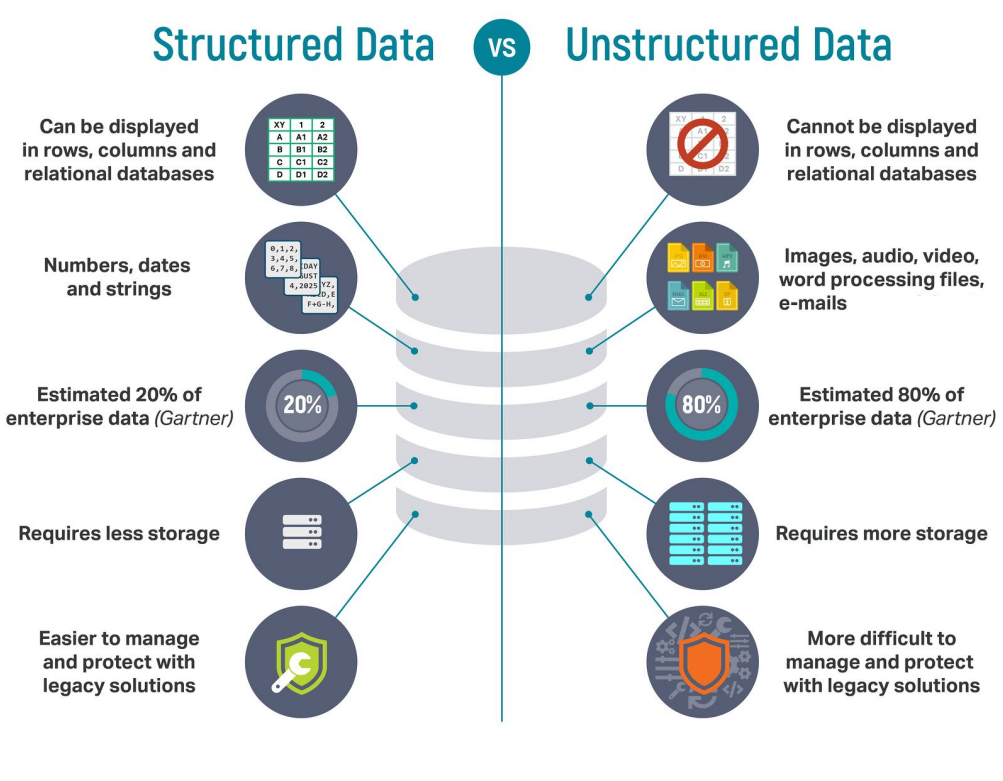
What type?

Salary $55k, $77.2k, $86.1k, ... - continuous

Salary $25k, $26k, $40k,... (rounded) - discrete

Salary $10k-$19k, $20k-29k, $30k-$39k... - ordinal

Structured vs Unstructured

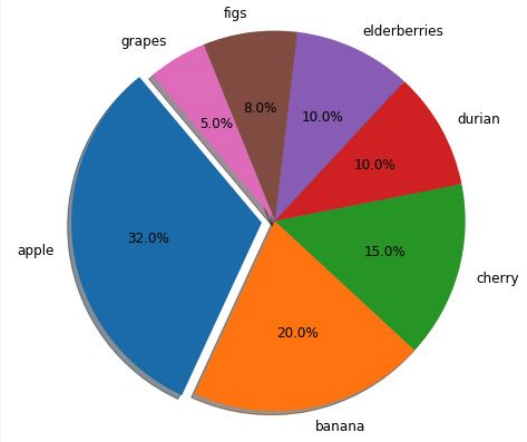


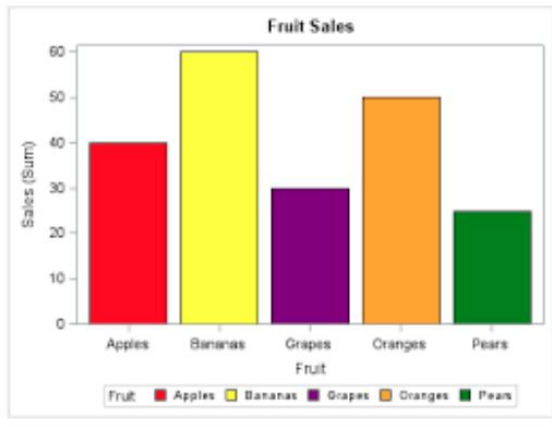
Structured data can be derived from unstructured data and often is

Why is it important?

Types of analysis you could do is different depending on the variable type

E.g. nominal or ordinal: frequency, proportion or %

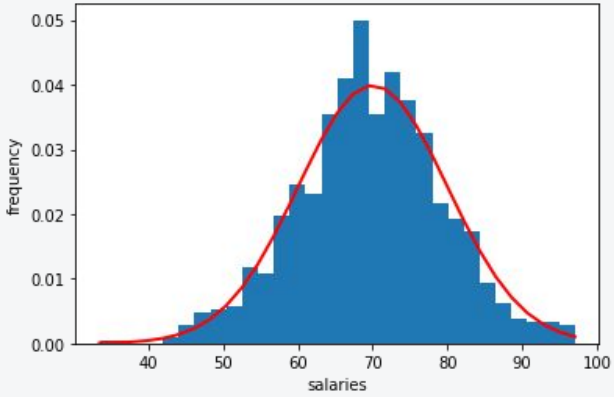




Why is it important?

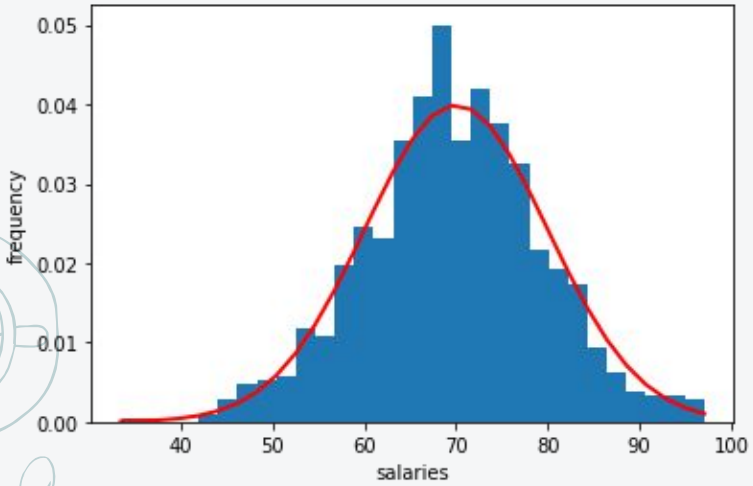
Types of analysis you could do is different depending on the variable type

E.g. distribution for continuous variables (including Gaussian fit)

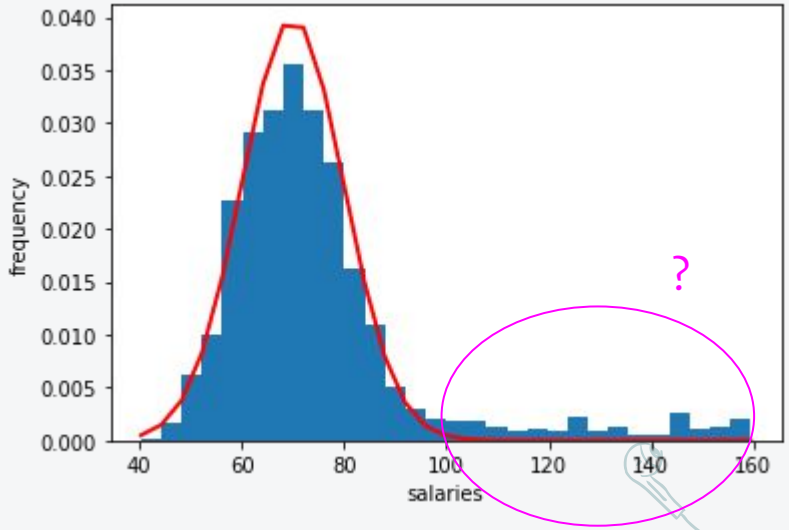


Example: salaries

Organization A



Organization B



Outlier (anomaly) Detection

Definition ?

Outlier (anomaly) Detection

Definition 1: an observation that deviates markedly from the sample it is found

Definition 2: Observations that are not conforming to the expected behavior

Outlier (anomaly) Detection

Why would we want to detect outliers?

Outlier (anomaly) Detection

Why would we want to detect outliers?
They cause issues with analysis

What do we do?

Outlier (anomaly) Detection

Why would we want to detect outliers?
They cause issues with analysis

What do we do?

1. Detect
2. Remove (?)
3. Impute (?)

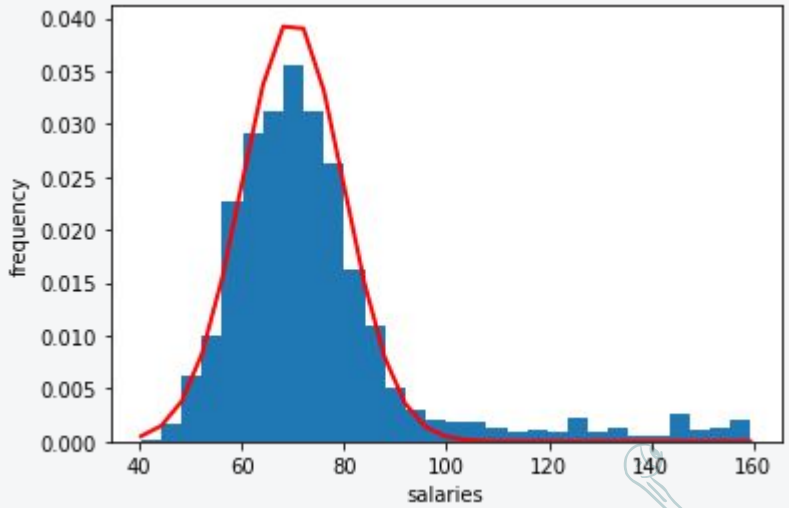
Areas of Application

- Fraud Detection
- Bioinformatics/healthcare
- Cyber Security (e.g. networking traffic)
- Trading
- Image Processing
- Many others (can you think of any?)

Detection

Eyeballing ?

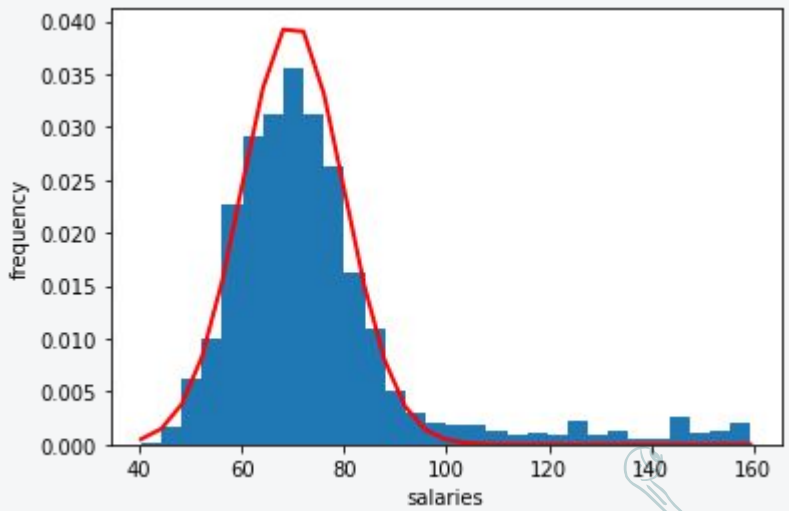
Organization B



Detection

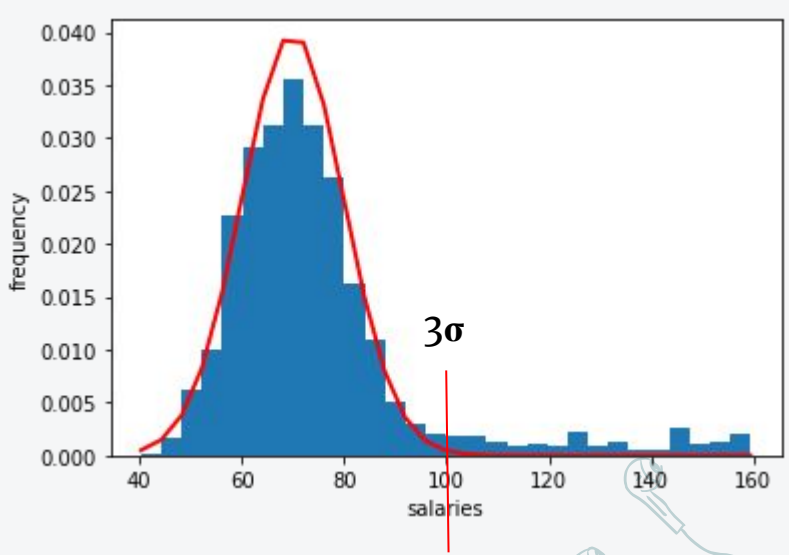
Eyeballing ?
Hard to tell...

Organization B



Detection

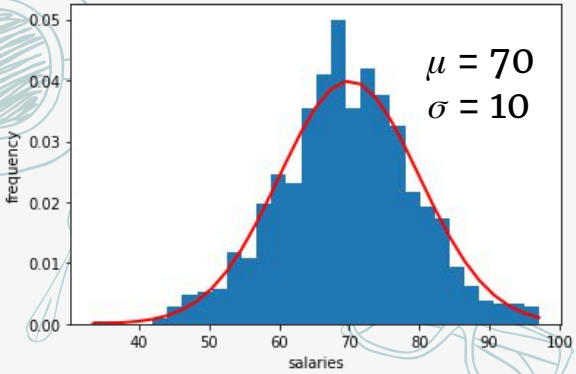
Eyeballing ?
Hard to tell...
Statistical Threshold?



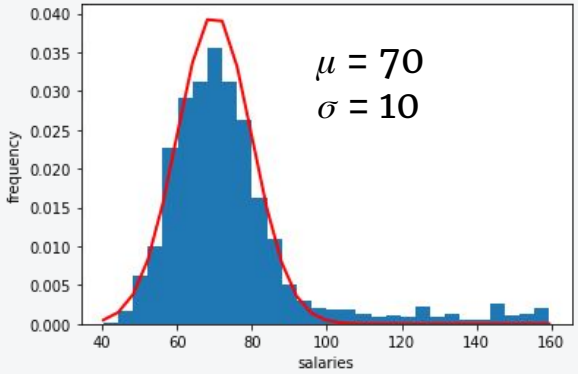
Detection

Eyeballing ?
Hard to tell...
Statistical Threshold?

Organization A



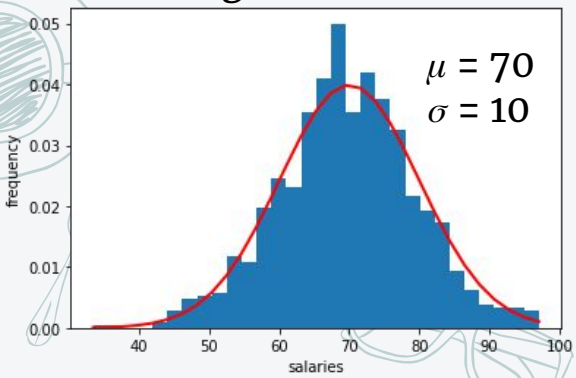
Organization B



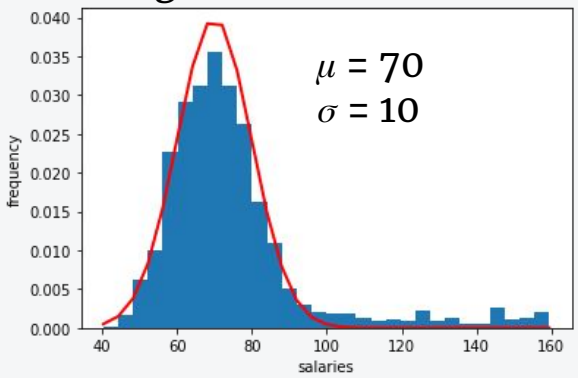
Detection

Eyeballing ?
Hard to tell...
Statistical Threshold?

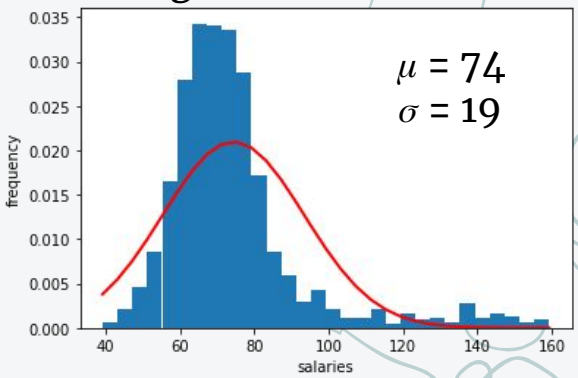
Organization A



Organization B



Organization B



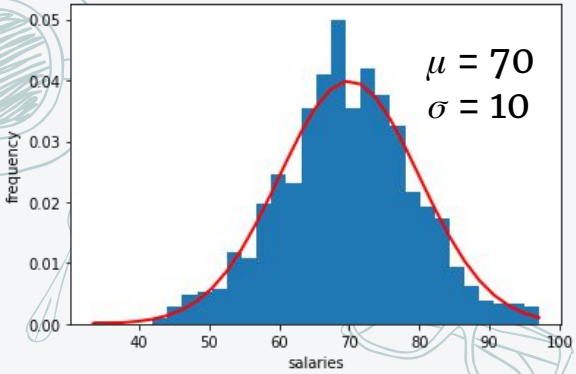
Detection

Eyeballing ?

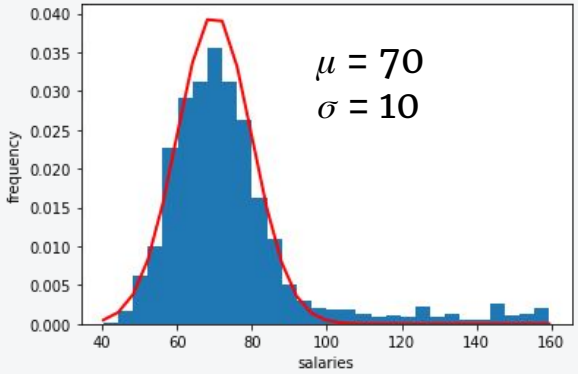
Hard to tell...

Statistical Threshold? Outliers affect the model!!!

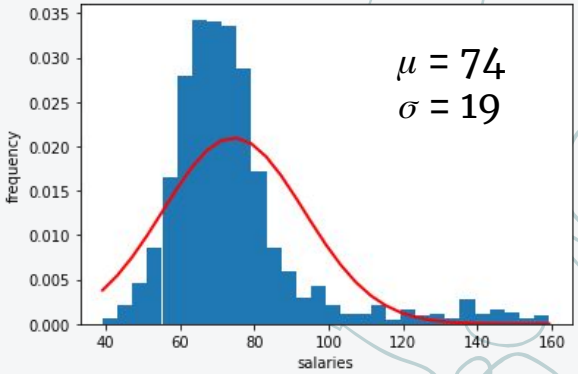
Organization A



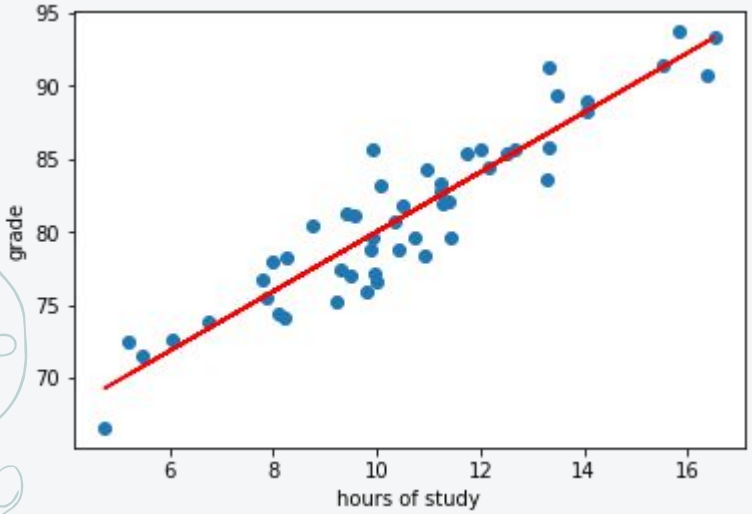
Organization B

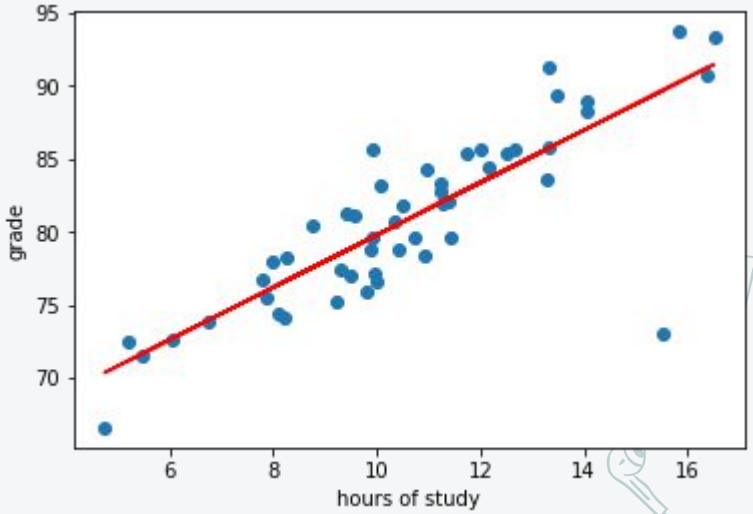


Organization B



Another example of outliers

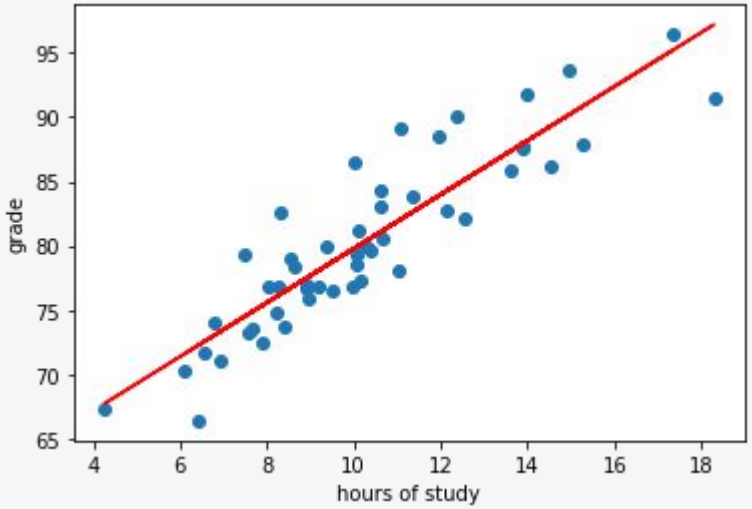




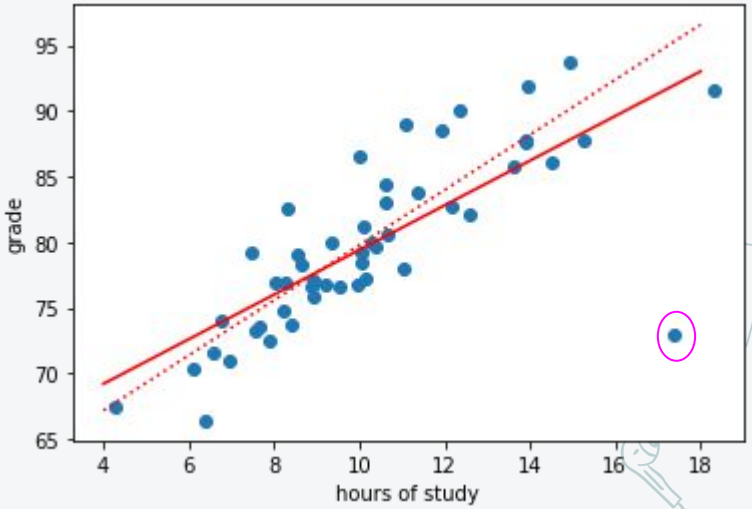
Another example of outliers

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

Pearson correlation $r = 0.9$

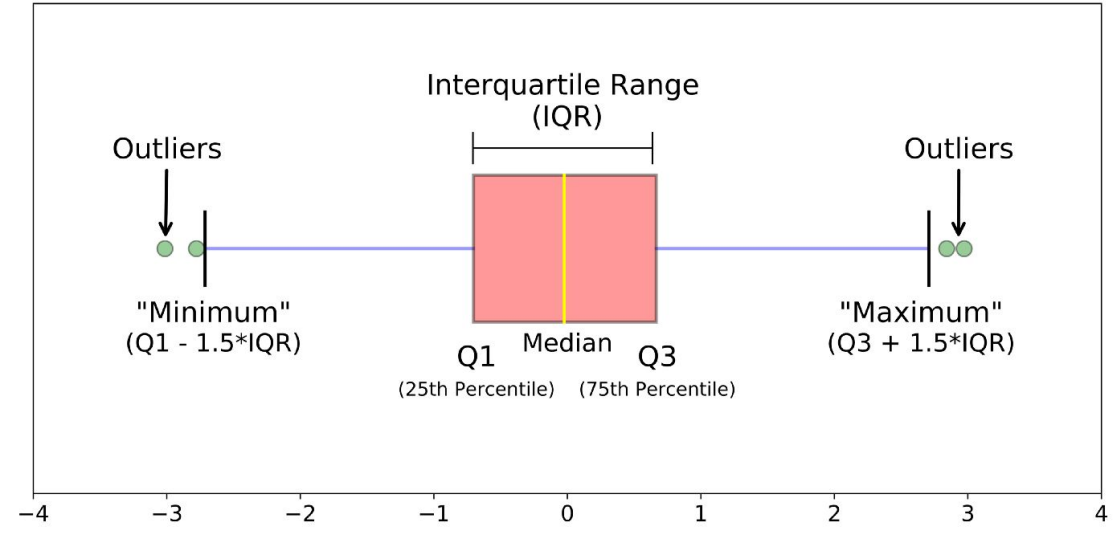


$r = 0.76$



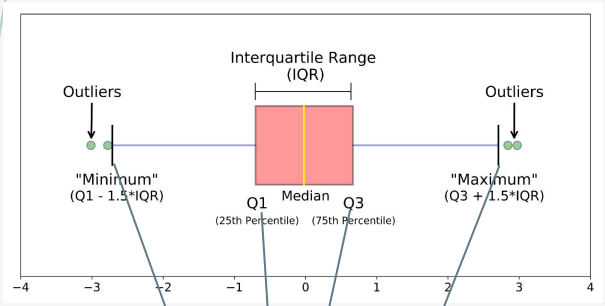
How to make sure your analysis is less affected by outliers?

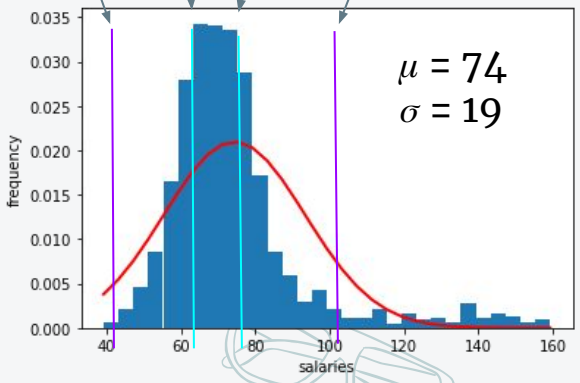
Can estimate parameters using % of data

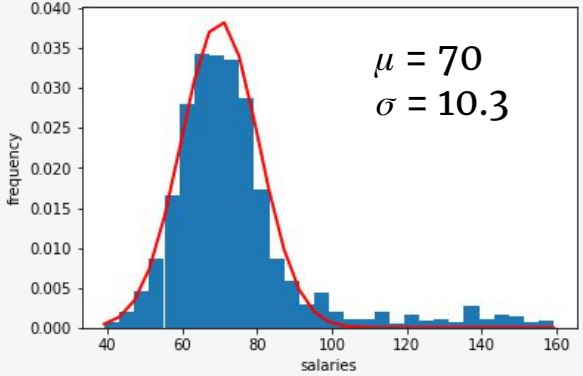


What can you do?

Can estimate parameters with % of data







Data

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20		NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni		$40k
Estelle	100	3.2	Festivus Uni	1.7	$0k
Leo	15	2.4	Festivus Uni	0	
Rachel	50	4.0	Columbia		$75k

1. Outliers

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20		NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni		$40k
Estelle	100	3.2	Festivus Uni	1.7	$0k
Leo	15	2.4	Festivus Uni	0	
Rachel	50	4.0	Columbia		$75k

1. Outliers

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20		NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni		$40k
Estelle	100	3.2	Festivus Uni	1.7	$0k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia		$75k

1. Outliers

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20		NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni		$40k
Estelle	100	3.2	Festivus Uni	1.7	$0k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia		$75k

remove

1. Outliers

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20		NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni		$40k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia		$75k

2. Missing Data

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20		NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni		$40k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia		$75k

2. Missing Data

25% missing data

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20		NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni		$40k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia		$75k

2. Missing Data

25% missing

Remove?

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20		NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni		$40k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia		$75k

2. Missing Data

25% missing

R = 65%

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20		NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni		$40k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia		$75k

2. M

R = 84%

25% missing (

R = 65%

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20		NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni		$40k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia		$75k

Imputation: with Mean

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20	3.1	NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni	2.4	$40k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia	2.4	$75k

Imputation: with Mean

$\sigma_{original}$=1.9
$\sigma_{imputed}$=1.6

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20	3.1	NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni	2.4	$40k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia	2.4	$75k

Imputation: with Mean

$\sigma_{original}$=1.9
$\sigma_{imputed}$=1.6

$R_{original}$ = 65%
$R_{imputed}$ = 56%

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20	3.1	NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni	2.4	$40k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia	2.4	$75k

Imputation: From Similar Individual

$\sigma_{original}$=1.9
$\sigma_{imputed}$=1.6
$\sigma_{similar}$ = 1.9

$R_{original}$ = 65%
$R_{imputed}$ = 56%
$R_{similar}$ = 71%

Student Name	Avg Hours Studied per week	GPA	University	Sense of Humour (0-5)	Salary
George	20	3.1 -> 2.8	NYU	3	$40k
Jerry	35	3.5	Columbia	5	$80k
Elaine	55	4.0	Columbia	4.2	$60k
Cosmo	5	2.0	City College	2	$25k
Newman	25	2.8	City College	0	$50k
Frank	35	3	Festivus Uni	2.4 -> 0	$40k
Leo	15	2.4	Festivus Uni	0	$35k
Rachel	50	4.0	Columbia	2.4 -> 4.2	$75k

Other Imputation Methods

MICE - draw from distribution of the variable, create multiple rows per individual

Regression

Random Forest

https://scikit-learn.org/stable/modules/impute.html

Reading Material

Data Types:
https://towardsdatascience.com/data-types-in-statistics-347e152e8bee

Visualization:
https://jingwen-z.github.io/data-viz-with-matplotlib-series3-pie-chart/

Reminders:

Data representation, random variables:
https://www.youtube.com/watch?v=vUpuj9KIe4k&list=PLXBDYmaCbeL8efhOZS4g9W6Z3m9_hFSnT&index=5&ab_channel=JeffLeek