CMPT 733 – Big Data Programming II

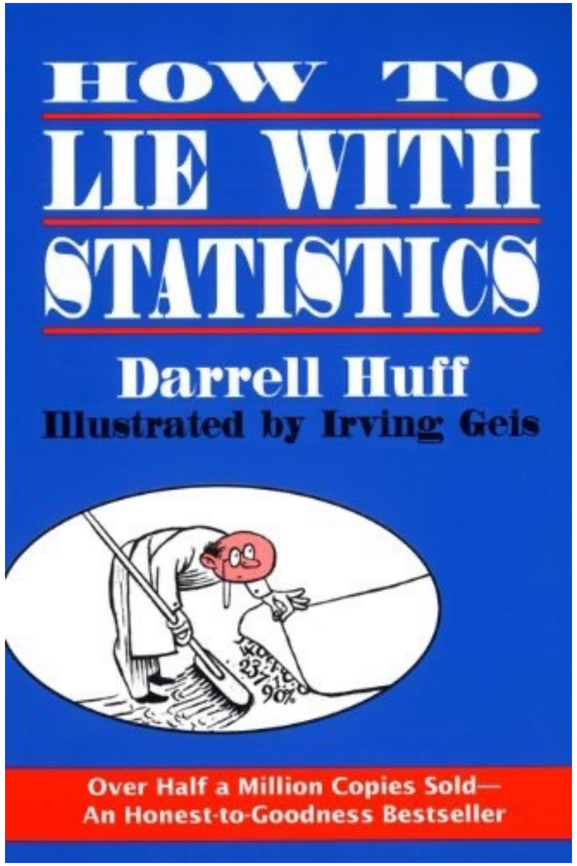
Statistics (I)

Instructor Steven Bergner

Course website https://sfu-db.github.io/bigdata-cmpt733/

Slides by Jiannan Wang & Steven Bergner

Why Should You Care?



> “ There are three kinds of lies:
> lies, damned lies, and statistics ”

Simpson's paradox

Is UC Berkeley gender biased?

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

~~YES!~~

Simpson's paradox

Is UC Berkeley gender biased?

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

NO!

Women tended to apply to competitive departments with low rates of admission

Outline

Statistical Thinking

Descriptive Statistics

Inferential Statistics

Outline

Statistical Thinking

Descriptive Statistics

Inferential Statistics

Statistical Thinking

1. Data is just a sample

2. You goal is to infer a population

3. Think about how to go “backwards” from the sample to the population

Example 1. Image Classification

Is it a dog or a cat?

Dataset: 1000 images collected from the Web

Without Statistical Thinking

Treat the 1000 images as the population

> Train a model on the data

> Evaluate a model on the same data

> Model accuracy: 95%

With Statistical Thinking

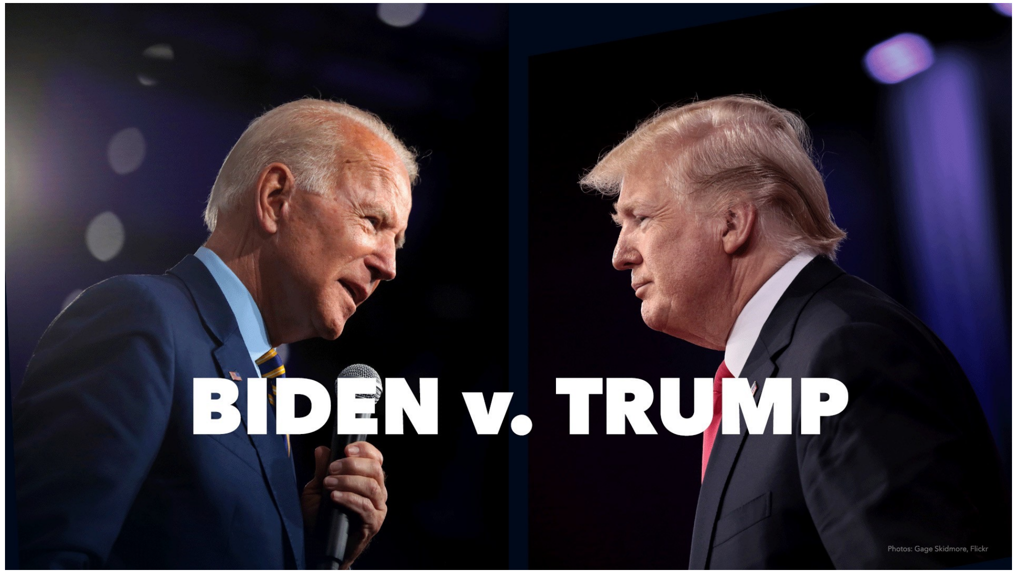
What is the population?
- All the images in the Web

What is your dataset?
- A sample of 1000 images drawn from the Web

What should you do?
- Split the dataset into a training dataset and a test dataset
- Train the model on the training dataset
- Evaluate the model on the test dataset

Example 2. Poll Prediction

Who will win the election?



Dataset: A survey of 100 people

Without Statistical Thinking

Treat the 100 people as the population

> Count the number of people who wants to vote for Hillary, e.g., 52

> Count the number of people who wants to vote for Trump, e.g., 48

> Hillary will win the election

With Statistical Thinking

What is the population?

- All the people who will vote in the election day

What is your dataset?

- A sample of 1000 people before the election day

Analysis result

Hillary: 52% ±3%
Trump: 48% ± 2%

Assumption: People have not changed their votes since the time of the poll

Summary

Statistical Thinking

- Sample, Population and Their Connection
- With vs. Without Statistical Thinking

Descriptive Statistics

Inferential Statistics

Outline

Statistical Thinking

Descriptive Statistics

Inferential Statistics

Descriptive vs. Inferential Statistics

Descriptive Statistics: e.g., Median

- <u>Why?</u> Aim to understand the data
- <u>How?</u> Data summarization, data visualization, etc.

Inferential Statistics: e.g., A/B Testing

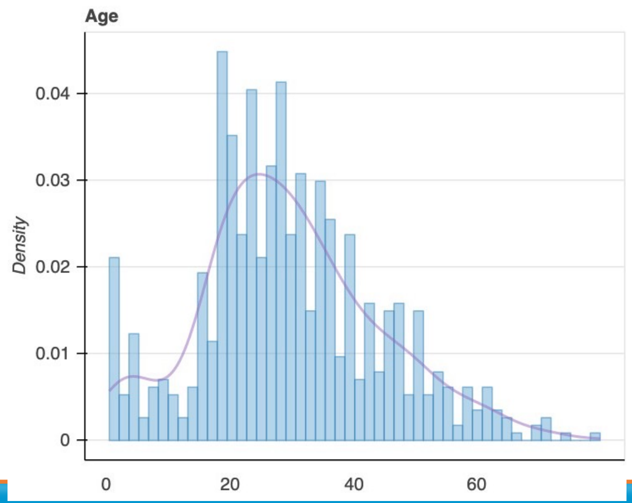
- <u>Why?</u> Aim to use the data (i.e., sample) to learn about a population
- <u>How?</u> Estimation, confidence intervals, hypotheses testing, etc.

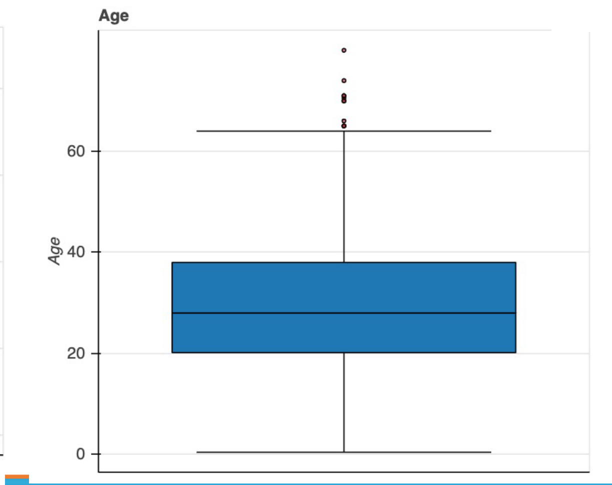
Exploratory Data Analysis (EDA)

Understand data and discover insights
via data visualization, data summarization, etc.

Understand “Age” column

Minimum	0.42
5-th Percentile	4
Q1	20.125
Median	28
Q3	38
95-th Percentile	56
Maximum	80
Range	79.58
IQR	17.875





Current EDA Solutions in Python

Solution 1: Pandas + Matplotlib

® Hard to Use

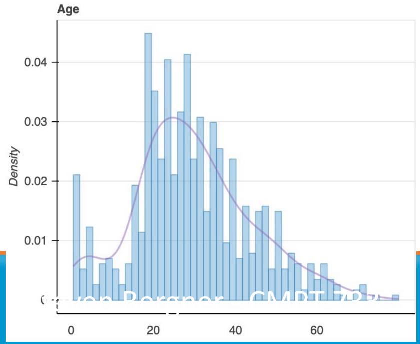
- <u>Beginner:</u> Need to know how to write plotting code
- <u>Expert:</u> Need to write lengthy and repetitive code

Understand “Age” column

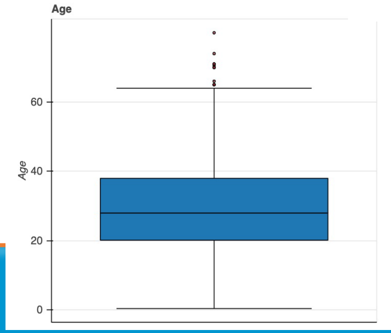
Write Code

Minimum	0.42
5-th Percentile	4
Q1	20.125
Median	28
Q3	38
95-th Percentile	56
Maximum	80
Range	79.58
IQR	17.875

Write Code



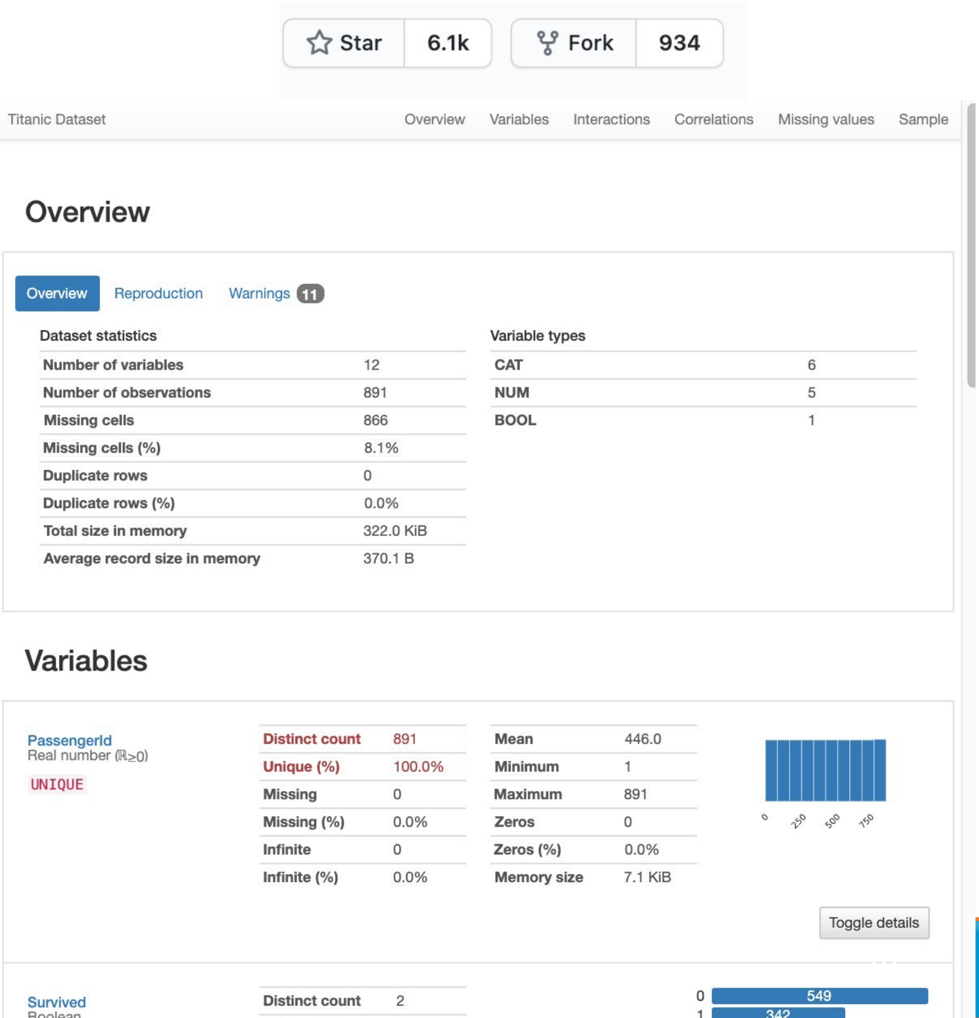
Write Code



Current EDA Solutions in Python

- Solution 2: Pandas-profiling
  - ® Slow
  - ® Hard to Customize

```
profile = ProfileReport(df, title="Pandas Profiling Report")
```



DataPrep.EDA Design Goals

EDA Solutions	Easy to Use	Interactive Speed	Easy to Customize
1. Pandas + Matplotlib		©	©
2. Pandas-profiling	©		
3. DataPrep.EDA	©	©	©

Key Idea

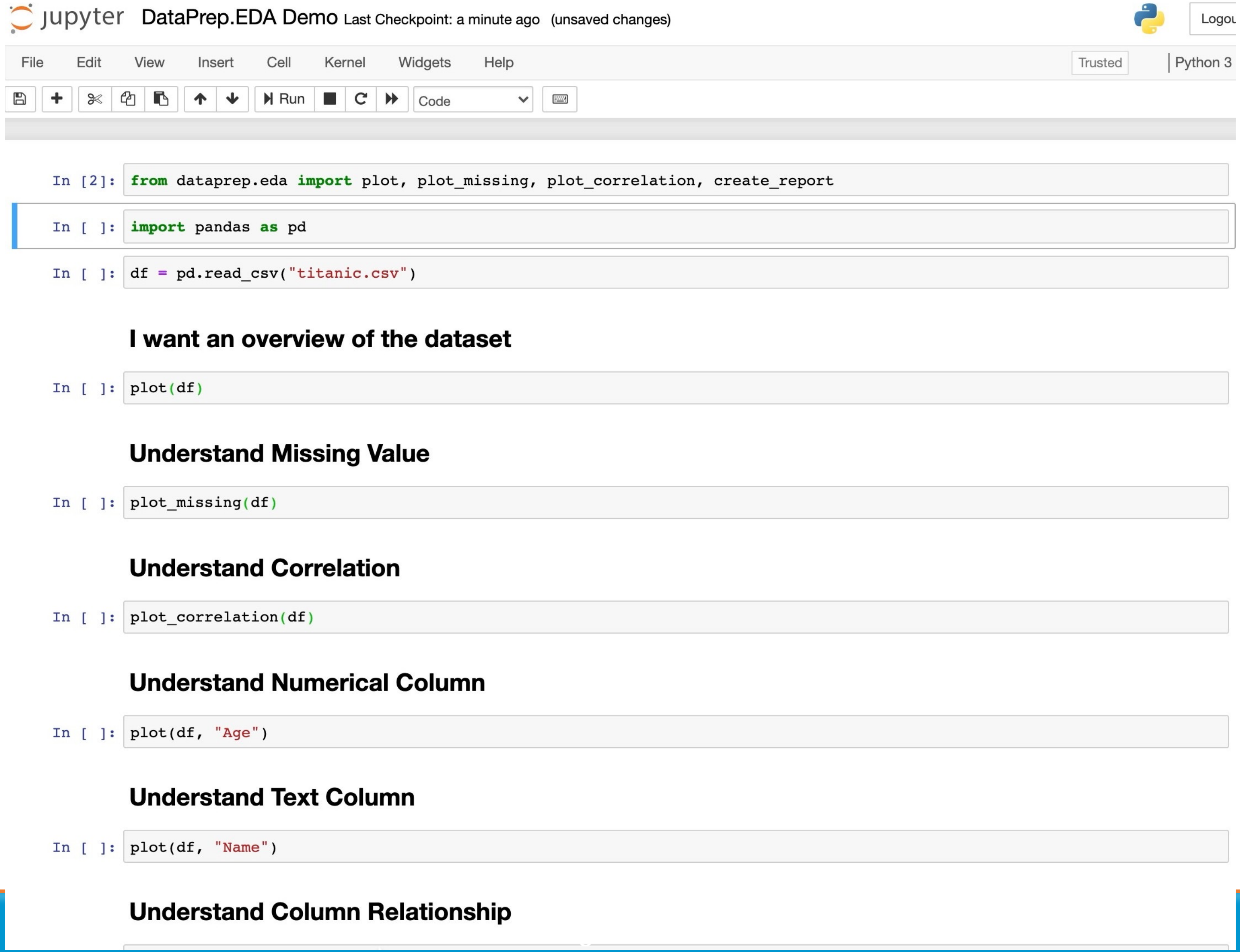
Task-Centric API Design

- Declarative
- Support both coarse-grained and fine-grained EDA tasks

Example

- plot(df): "I want to see an overview of the dataset"
- plot_missing(df): "I want to understand the missing values of the dataset"
- plot(df, x): "I want to understand the column x"
- plot(df, x, y): "I want to understand the relationship between x and y"
- …

DataPrep.EDA (Demo)



Correlation Analysis

Correlation

- It is a measure of relationship between two variables

Why is correlation analysis useful?

- For understanding data better
- For making predictions better

Case Study: How to do correlation analysis

Height and weight are correlated

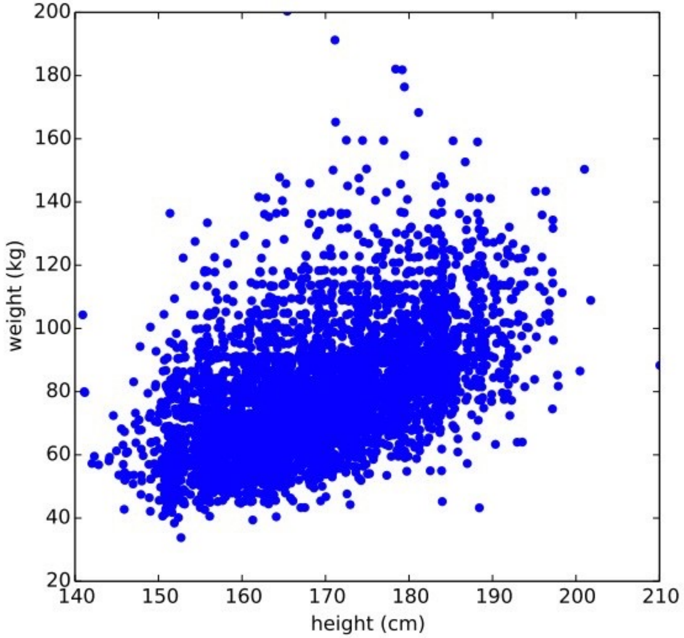
1	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0

Source: *Think Stats -- Exploratory Data Analysis in Python*

Idea 1. Visualization

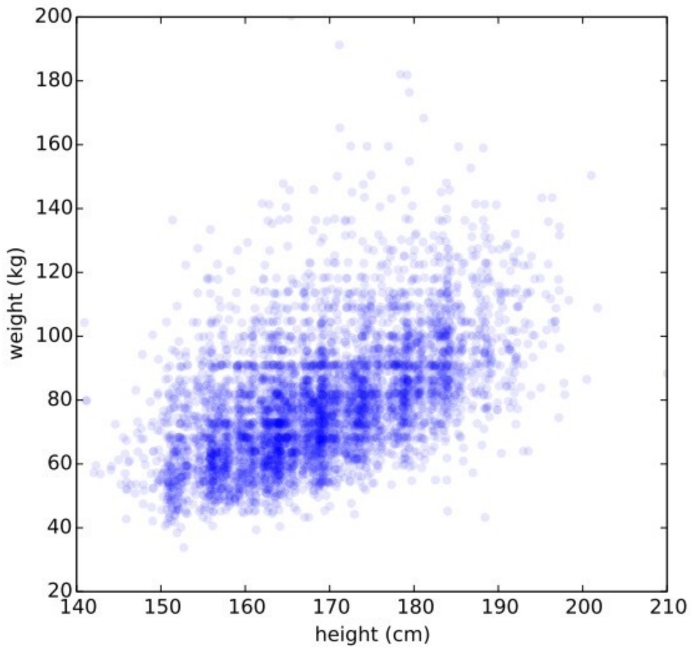
Scatter Plot

1	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



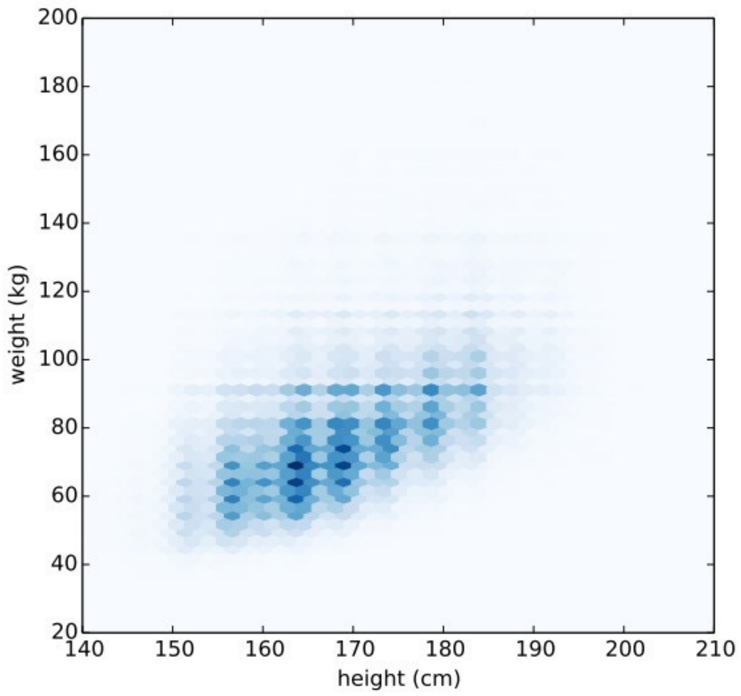
Scatter Plot (with transparency)

	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



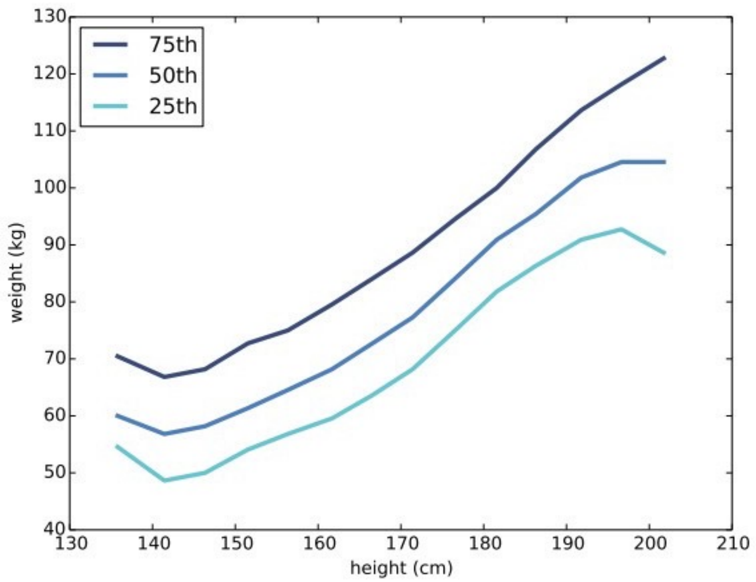
Hexbin Plot

1	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



Characterizing relationships

	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



Idea 2. Correlation Coefficient

Covariance

Covariance is a measure of the tendency of two variables to vary together.

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

$$\mathrm{cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$$

Hard to interpret

113 kilogram-centimeters

Pearson's correlation

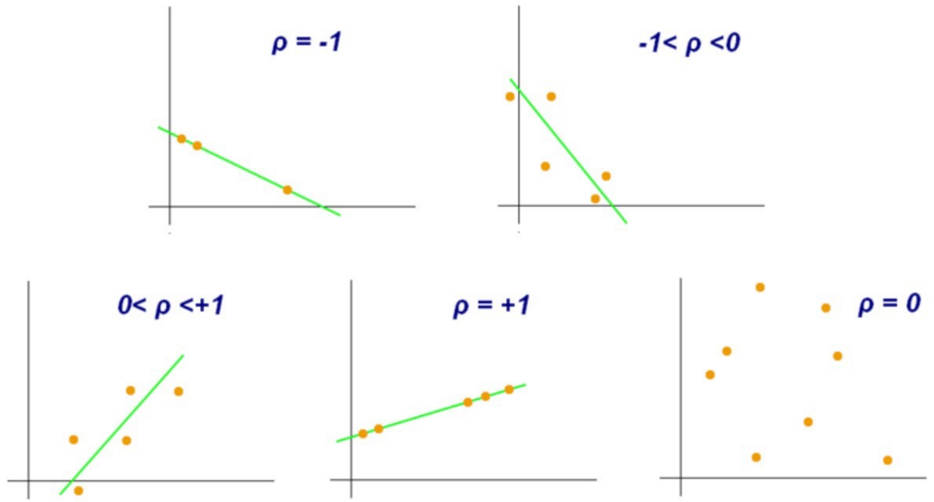
What about non-linear relationship?

Pearson's correlation is a measure of the linear relationship between two variables

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Easy to Interpret

- [-1, 0) ➔ Negative Correlated
- (0,+1] ➔ Positive Correlated
- -1 or +1 ➔ Perfectly Correlated



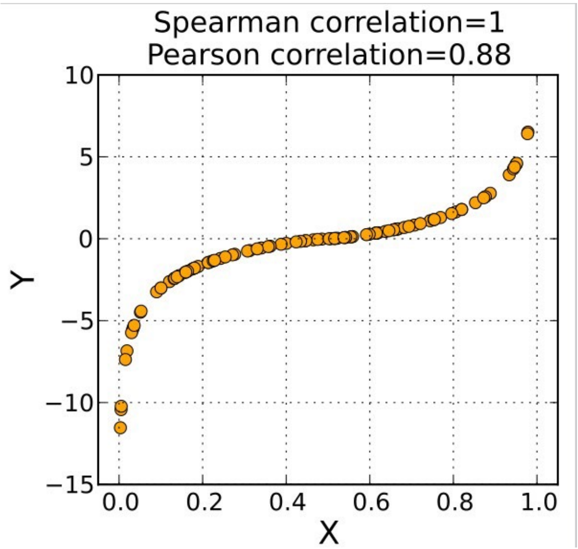
Spearman's rank correlation

Spearman's rank correlation is a measure of monotonic relationship between two variables

$$r_s = \rho_{\mathrm{r}_X,\mathrm{r}_Y} = \frac{\mathrm{cov}(\mathrm{r}_X, \mathrm{r}_Y)}{\sigma_{\mathrm{r}_X}\sigma_{\mathrm{r}_Y}}$$

Advantages

- Mitigate the effect of outliers
- Mitigate the effect of skewed distributions



Summary

Statistical Thinking

Descriptive Statistics

- Descriptive vs. Inferential Statistics
- Exploratory Data Analysis with DataPrep
- Correlation Analysis

Inferential Statistics

Outline

Statistical Thinking

Descriptive Statistics

Inferential Statistics

- Estimation

Estimation

Problem statement

- Estimate a numerical value associated with a population

Examples

- Estimate the percentage of the people in the US who will vote for Biden
- Estimate the median annual income of all households in the US

Example: Median Annual Income

How to estimate the median annual income of all households in the US?

- Randomly select 10,000 households from the US
- Report their median annual income: 50,000USD

- BUT, we need to report something like

50,000 ±500 USD

A Naïve Solution

- Randomly select 10,000 households from the US
- Report their median annual income

Repeat this process for 100 times

50,000 49,600 50,200 ⋯ 49,200

You have to survey 1,000,000 million households in total!

A Smart Solution: Bootstrapping

Key Idea: Resampling

- Sample with replacement from the original data sample

Population: 1, 1, 8, 2, ... 3, 3

Sample: 3, 8, 1, 8, 3

Resample: 8, 3, 3, 3, 1

A Smart Solution: Bootstrapping

- Randomly select 10,000 households from the US
- Draw a resample from the 10,000 households
- Report the median annual income of the resample

Repeat this process for 100 times

You do NOT need to survey any new household. ©

Notes on Bootstrapping

- Start with a large random sample (at least 30)

- Replicate the resampling procedure as many times as possible (more than 1000 times)

- Does not work for min/max

Conclusion

Statistical Thinking

- Sample, Population and Their Connection
- With vs. Without Statistical Thinking

Descriptive Statistics

- Descriptive vs. Inferential Statistics
- EDA with DataPrep.eda
- Correlation Analysis

Inferential Statistics

- Estimation and Bootstrapping