Descriptive Statistics & **Visualizations**

Vinayaka Gude, Ph.D.
Elon University

Agenda

01 Data Cleaning

02 Coding demo: Data cleaning using Pandas

03 Data Transformations

Missing Values with Pandas

Example DataFrame:

```
data = {"state": ["Ohio", "Ohio", "Ohio", "Nevada", "Nevada", "Nevada"],"year": [np.nan, 2001, np.nan, 2001, 2002, 2003],"pop": [1.5, 1.7, np.nan, 2.4, np.nan, 3.2]}
df = pd.DataFrame(data)
```

Delete columns: `df=df.dropna(axis=1, how='all')` or `df=df.dropna(axis=1, how='any')`

Delete rows: `df=df.dropna(axis=0, how='all')` or `df=df.dropna(axis=0, how='any')`

Impute a specific value: `df['pop']=df['pop'].fillna(0)`

Impute mean: `df['pop']=df['pop'].fillna(np.mean(df['pop']))`

Interpolate: `df["pop"] = df["pop"].interpolate(method="quadratic")`

Remove Duplicate rows: `df = df.drop_duplicates()`

Analyzing Distributions

Frequency distribution

* qualitative data

Summary of the number of occurrences of a **given value** → Bar chart

* quantitative data

Summary of the number of occurrences in a **given bin** → Histogram

Relative frequency

↓

Ratio of frequency of a value to that of the total

Car
Tesla
Ford
Toyota
Toyota
Tesla
Toyota

Car	Frequency	Relative frequency
Tesla	2	2/6 = 0.33
Ford	1	1/6 = 0.16
Toyota	3	3/6 = 0.5

Methods of Central Tendency

Mean ($\bar{x}$): Average value for a variable

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$$

Median: Middle value of data when arranged in ascending order

Mode: Most frequently occurring value

Geometric Mean: *n*th root of the product of *n* values

$$\sqrt[n]{(x_1)(x_2)\ldots(x_n)}$$

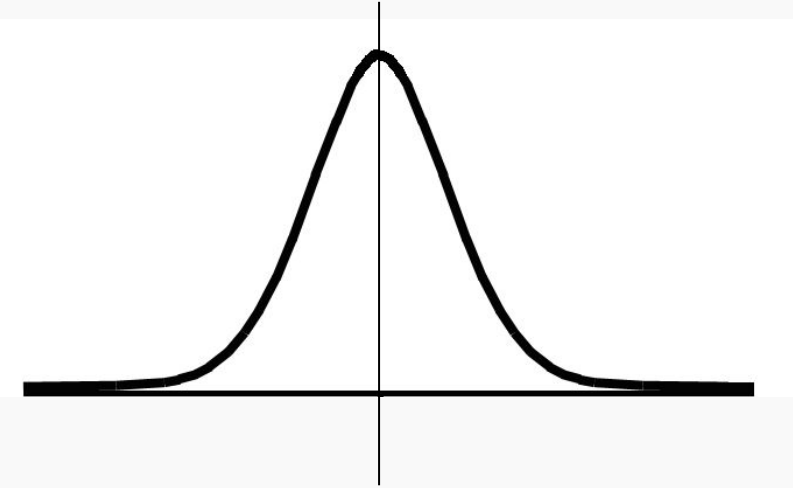
Profit
$20
$10
$10
$10
$100

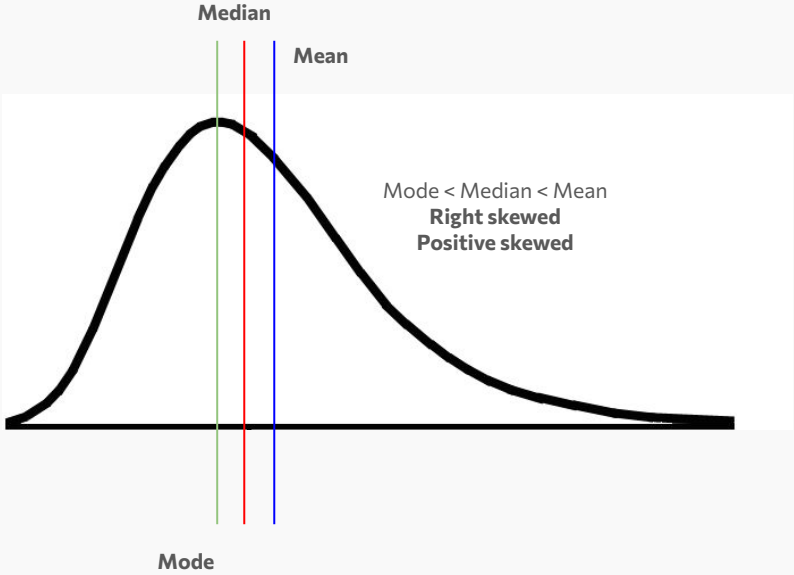
Mean = 30, median = 10, mode = 10, geometric mean = 18

Mean is susceptible to **outliers***

Methods of Central Tendency

Mean = Median = Mode





Methods of Dispersion

Range: Difference b/w **min** and **max** values

Variance: Measure of variability (based on deviation about the mean)

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$$

Standard Deviation: Square root of variance

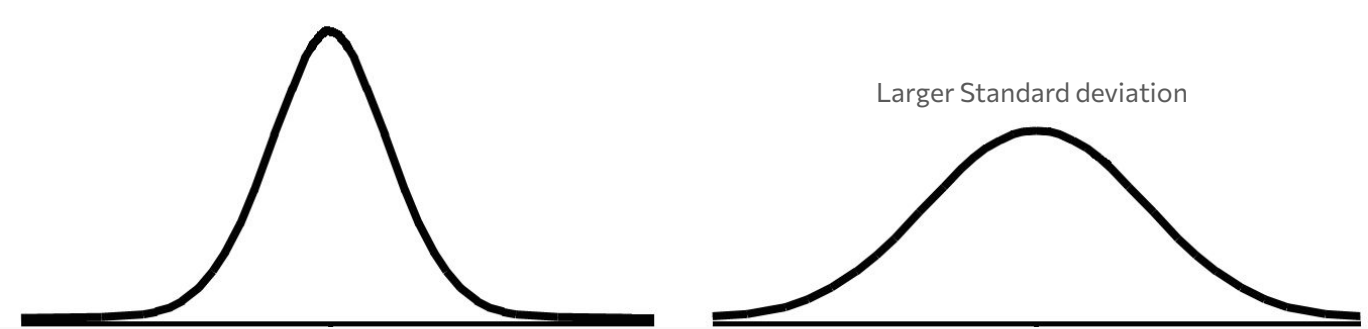
$$S = \sqrt{S^2}$$ sample

$$\sigma = \sqrt{\sigma^2}$$ population

Coefficient of Variation: How large **standard deviation** is *compared to* ***mean***

$$\left(\frac{Standard\ deviation}{Mean} \times 100\right)\ \%$$

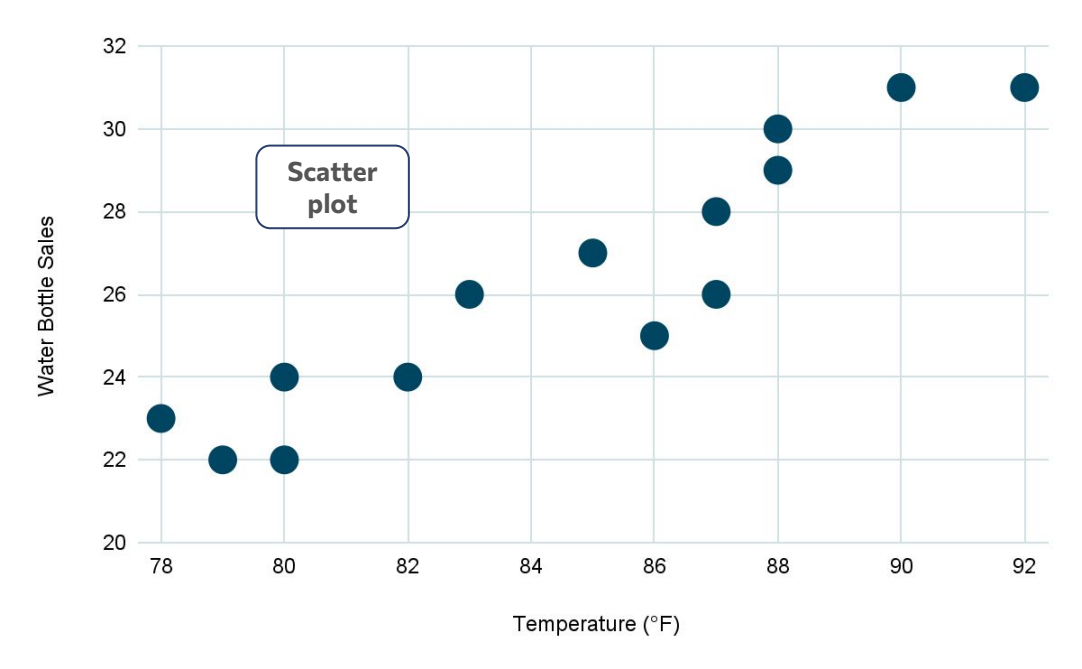
Methods of Dispersion



Measures of dispersion get larger as the spread of data items increases

Methods of Association

Temperature (° F)	Water Bottle Sales
78	23
79	22
80	24
80	22
82	24
83	26
85	27
86	25
87	28
87	26
88	29
88	30
90	31
92	31



Covariance

descriptive measure of the linear association between two variables

Sample $$Cov(x,y) = \frac{\sum(x_i - \bar{x}_x)(y_i - \bar{y}_y)}{n-1}$$

Population $$Cov(X,Y) = \frac{\sum(X_i - \mu_x)(Y_i - \mu_y)}{N}$$

Employees	**Sales**
3	5
4	10
10	15
15	22
Mean (32/4) = **8**	Mean (52/4) = **13**

Cov(x,y) = (3-8)(5-13) + (4-8)(10-13) + (10-8)(15-13) + (15-8)(22-13) **/ (4-1) = 38.66**

Review	**Inventory**
3	8
2	15
5	1
4	4
Mean (14/4) = **3.5**	Mean (28/4) = **7**

Cov(x,y) = (3-3.5)(8-7) + (2-3.5)(15-7) + (5-3.5)(1-7) + (4-3.5)(4-7) **/ (4-1) = -7.667**

Correlation

Correlation coefficient measures the relationship between two variables

Varies b/w **-1** and **+1**

$$r_{xy} = \frac{Cov_{xy}}{S_x S_y}$$

Employees	**Sales**
3	5
4	10
10	15
15	22
Mean (24/4) = **6**	Mean (52/4) = **13**
Stdev = **5.59**	Stdev = **7.25**

Corr(x,y) = 38.66/ (5.59 * 7.25) = 0.97

Strong **positive** correlation

Strong **negative** correlation

Corr(x,y) = -7.667 / (1.29*6.05) = -0.98

Review	**Inventory**
3	8
2	15
5	1
4	4
Mean (14/4) = **3.5**	Mean (28/4) = **7**
Stdev = **1.29**	Stdev = **6.05**

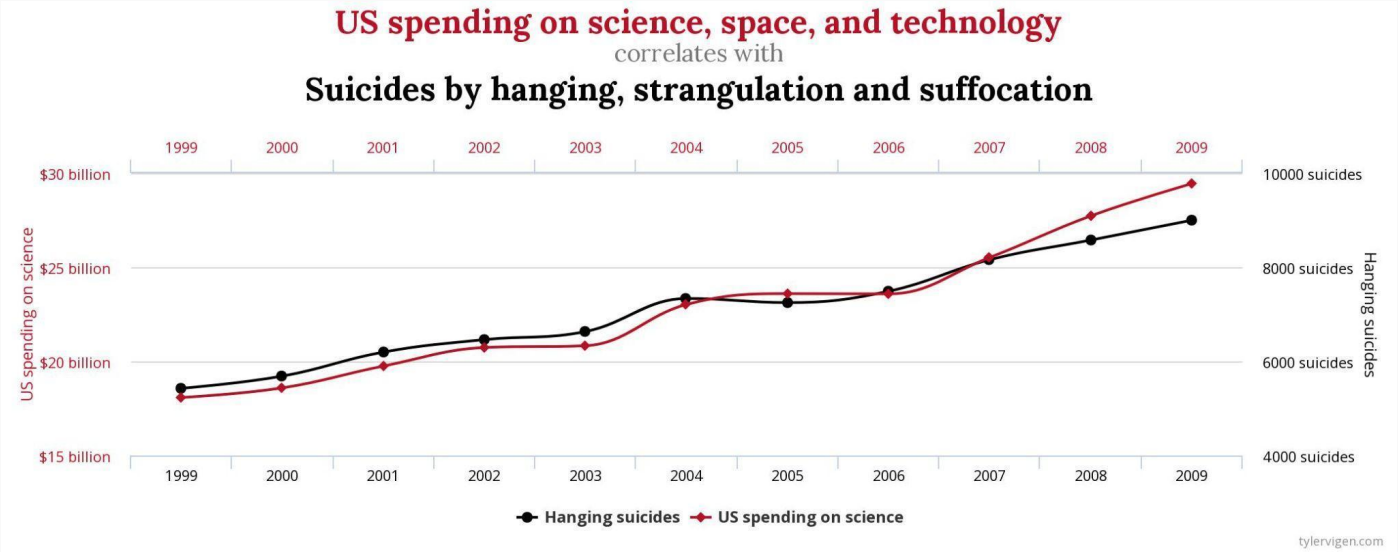
Regression

Temperature (° F)	Water Bottle Sales
78	23
79	22
80	24
80	22
82	24
83	26
85	27
86	25
87	28
87	26
88	29
88	30
90	31
92	31

An equation to estimate the relationship between variables

- Line - **linear regression**
- Polynomial function - **polynomial regression**
- More than 2 variables - **multi-regression**

Causation



https://tylervigen.com/spurious-correlations

Visualizations: Comparison

Visualization	Data	Questions
Bar Chart	1 categorical, 1 quantitative	Comparison
Stacked Bar Chart	2 categorical, 1 quantitative	Comparison across categories and sub-categories
Clustered Bar Chart	2 categorical, 1 quantitative	Comparison across categories and sub-categories
Column Chart	1 categorical, 1 quantitative	Comparison
Stacked Column Chart	2 categorical, 1 quantitative	Comparison across categories and sub-categories
Clustered Column Chart	2 categorical, 1 quantitative	Comparison across categories and sub-categories

Visualizations: Trend

Visualization	Data	Questions
Line Chart	Time, 1 quantitative	Trend over time
Multi-line Chart	Time, 1 categorical, 1 quantitative	Trend over time, for different categories
Multi-area Chart	Time, 1 categorical, 1 quantitative	Compre category values over time
Ribbon Chart	Time, 1 categorical, 1 quantitative	Change of category ranking over time

Visualizations

Composition		
Visualization	**Data**	**Questions**
Pie/Donut	1 categorical, 1 quantitative	Composition
Tree Map	1 categorical, 1 quantitative	Composition
Decomposition Tree	1 or more hierarchical categorical variables, 1 quantitative	Composition and Comparison across multiple levels

Category	Visualization	Data
Summary	Histogram/box plot	Evaluate the distribution of a quantitative variable
Geographic	Map	1 geographic, 1 quantitative
Correlation	Scatter/Bubble Chart	Evaluate correlation between the variables

Exercise

- Perform descriptive statistics on Happiness and summarize the findings. Visualize the distribution and compare the insights from the description and visualization.
- Perform descriptive statistics on Age and summarize the findings. Visualize the distribution and compare the insights from the description and visualization.
- Is there a correlation between the Happiness and income? Use a visualization.

Thank you!

Any questions?

gude.vinayaka@outlook.com