Exploratory Data Analysis (EDA)

Week 3 – Part 1 – Motivation for EDA

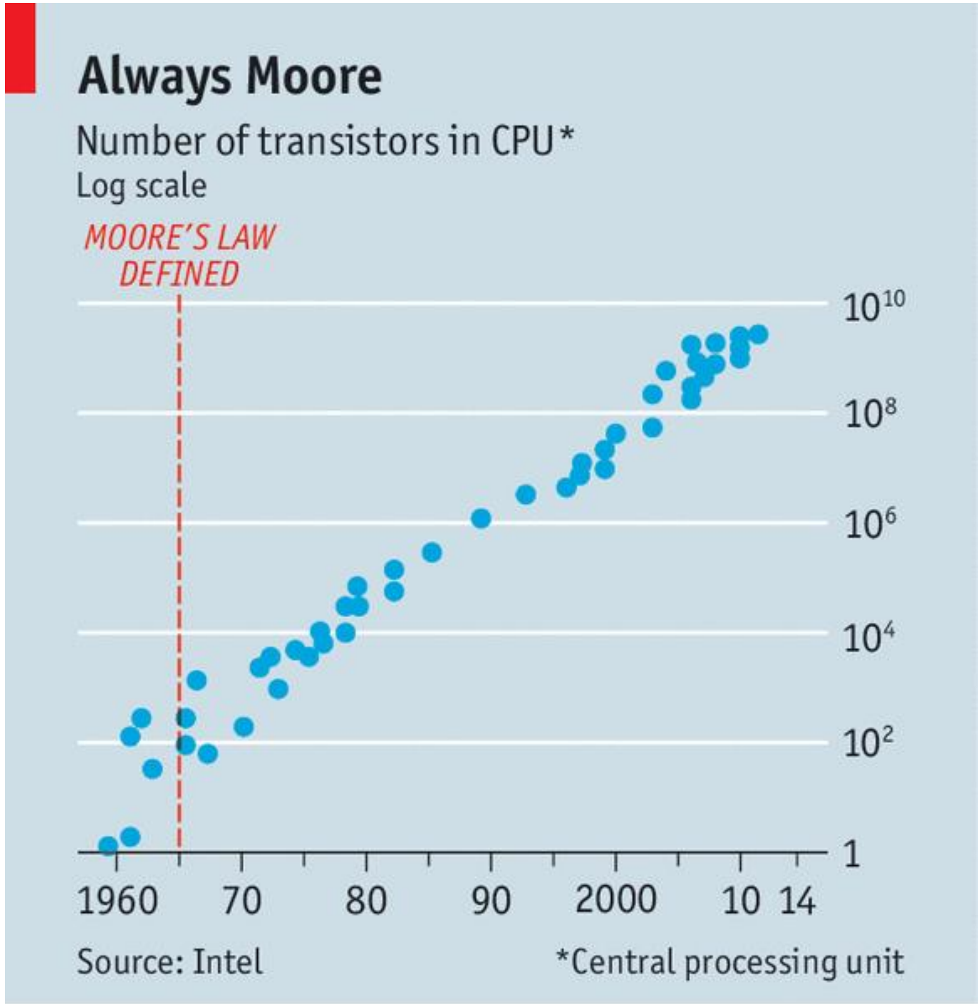
CS 457 - L1 Data Science

Zeehasham Rasheed

Data Trend

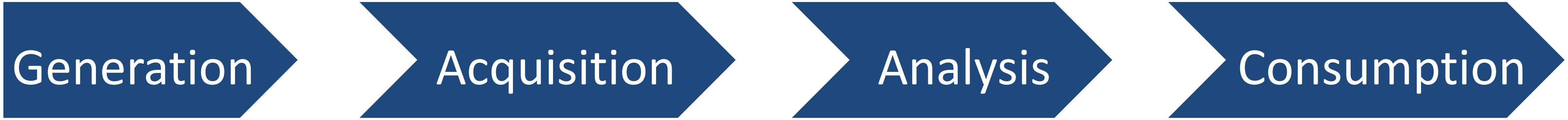
Moore’s Law

“The number of transistors in a dense integrated circuit (IC) doubles* about every two* years.”



Economist.com

Life of Data



Generation → Acquisition → Analysis → Consumption

Who creates data?

1. Nature
2. People
3. Machines

Who collects data?

1. Individuals
2. Organizations

Who crunches data?

1. Analysts
2. (Data) Scientists
3. Business Consultants
4. Bankers
5. Doctors
6. Whoever you find on LinkedIn with the word "data" in their job title.

Who consumes analysis?

Everyone!

What *Data* Means in This Course

1. A ***dataset*** is a table containing measurements of objects of the *same type*.
2. Every row is an object.
3. Every column is one attribute of an object.
4. Every cell is the measurement of the corresponding object and attribute.

Sample Data

Attributes of each flower

Samples of iris flowers

sepal_length	sepal_width	petal_length	petal_width	species
5.8	4	1.2	0.2	setosa
5.6	2.8	4.9	2	virginica
6.2	2.2	4.5	1.5	versicolor
6.3	3.4	5.6	2.4	virginica
6.3	2.5	5	1.9	virginica
5	3	1.6	0.2	setosa
4.7	3.2	1.3	0.2	setosa
5.8	2.7	5.1	1.9	virginica
4.9	2.5	4.5	1.7	virginica
6.1	3	4.6	1.4	versicolor

https://en.wikipedia.org/wiki/Iris_flower_data_set

Data Storytelling - Critical Skill

Storytelling has a 30X Return on Investment

Rob Walker and Joshua Glenn auctioned common items like mugs, golf balls, toys, etc. The item descriptions were **stories** purpose-written by 200+ contributing writers.

Items that were bought for $250 sold for over $8,000 – a return of over 3,000% for storytelling!

- **Stories are memorable and viral**
- People remember stories. They'll act on them.
- People share stories. That enables collective action.

Motivation for EDA

But analysts present their work, not their message

Data scientists present their analysis – what they did, and what they found. That’s not what the audience needs.

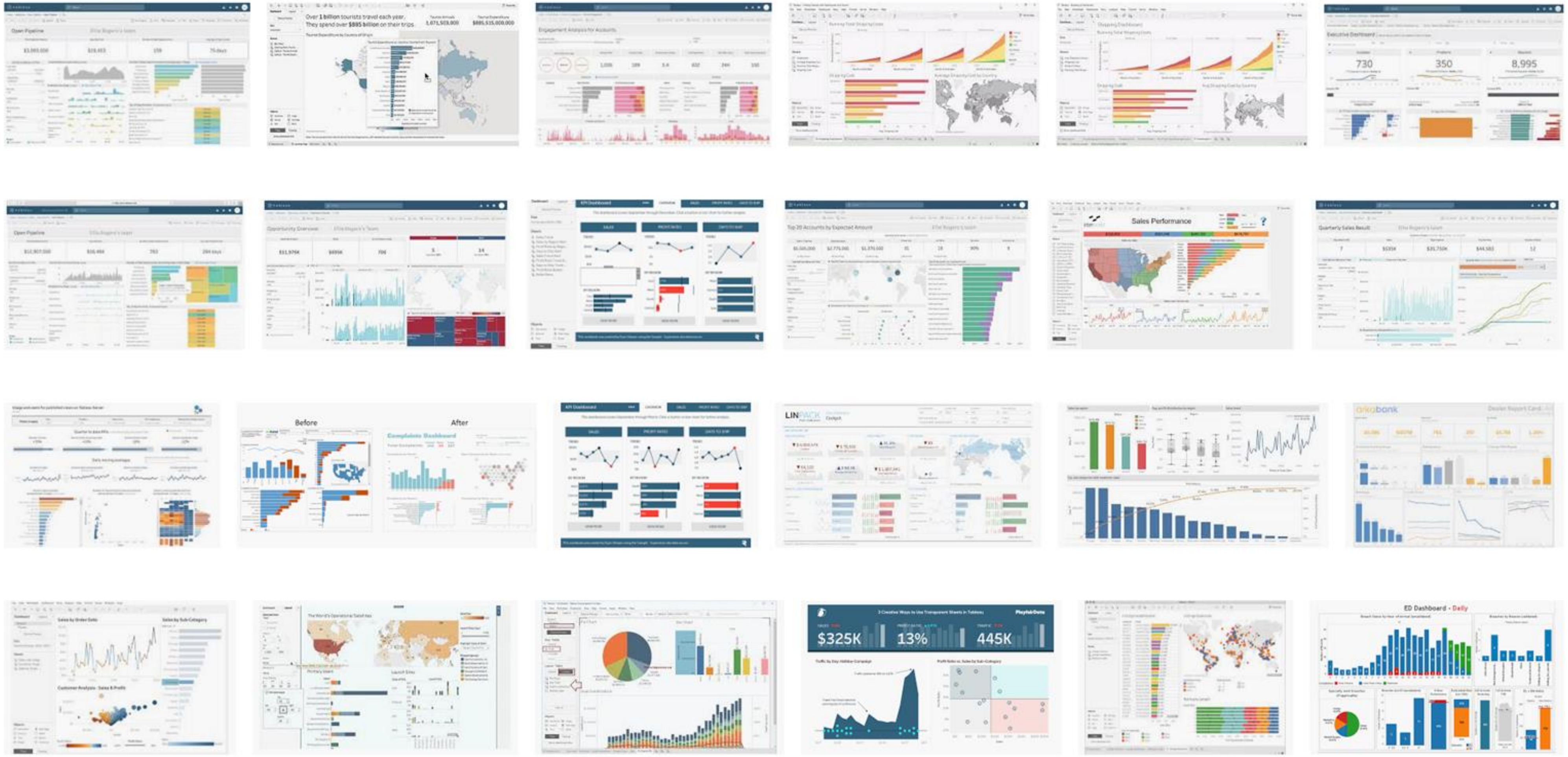
Audiences need a message that tells them what to do, and why. Told in an engaging way. As a story.

Share your data & analysis as data stories

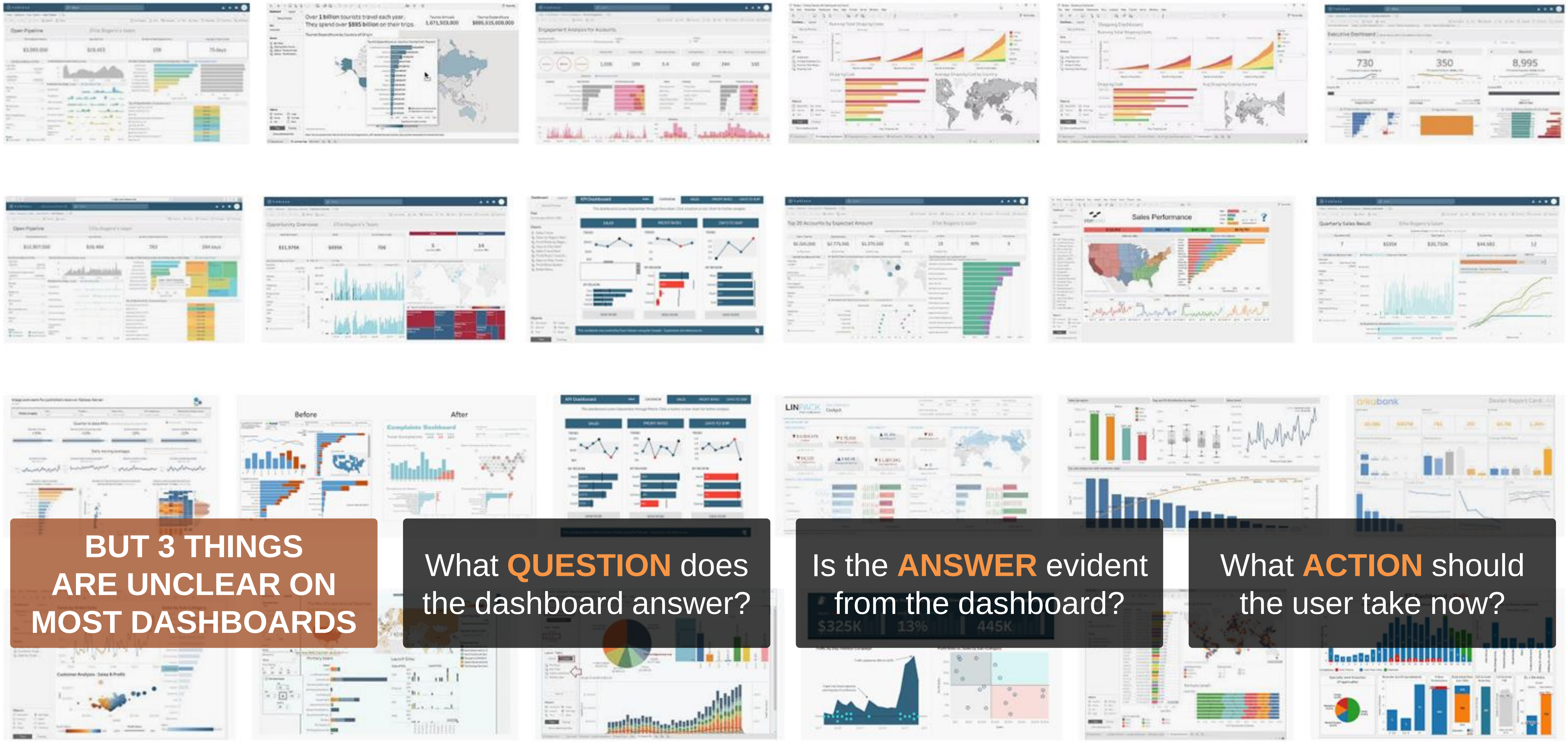
Whenever you share inferences from data – whether it’s as a presentation, or an email or document with your analysis, or as a dashboard – craft it as a story.

Variety of Dashboards

- With the growth of self-service BI, most companies have lost track of how many dashboards they generated



3 Important Things that Matter



BUT 3 THINGS ARE UNCLEAR ON MOST DASHBOARDS

What **QUESTION** does the dashboard answer?

Is the **ANSWER** evident from the dashboard?

What **ACTION** should the user take now?

End of Part 1

Exploratory Data Analysis (EDA)

Week 3 – Part 2 – Types of Statistical Analysis

CS 457 - L1 Data Science

Zeehasham Rasheed

NOIR Summary

- **Nominal (Categorical)**
  - categories: qualitative, no implied order or size, *discrete*
    - color, gender, State, Country, …
- **Ordinal   (can contains categories)**
  - rank order: *discrete*
    - 1: dislike < 2: neutral < 3: like
    - Only relational operators

- **Numerical (Continuous)** contains numbers
- **Interval**
  - *distance/difference* measures have no meaning; *continuous, integer, floating point*
  - contains zero point on origin. However, the origin does not imply a true absence
    - 0^o Celsius does not mean absence of temperature
    - Date
- **Ratio**
  - *Size comparisons* have meaning; *continuous, integer, floating point*
    - 80kg = 2 x 40kg
    - can be 0, differences, ratios provides meaning. For example length, mass etc.

NOIR Facts

- **IMPORTANT**:
  - Data type determines what computations and statistical tests are appropriate or inappropriate!
  - e.g., can’t calculate mean or average Country or Gender

Types of Variables

A **variable**, **feature** or **dimension** is a column in the dataset.

A **numerical** or **continuous** variable contains numbers that have a meaning.

Numerical variables support arithmetic operations (+, -, *, /)

Poll: Q: *Which are the numerical variables in this dataset?*

Duration (in seconds)

Duration (in seconds)	Age	Gender	Country
510	22-24	Male	France
423	40-44	Male	India
83	55-59	Female	Germany
391	40-44	Male	Australia
392	22-24	Male	India
470	50-54	Male	France
529	22-24	Male	India
624	22-24	Female	United States of America
214	22-24	Male	United States of America

Types of Variables (2)

- Elements of an **ordinal** variable can be ordered.
- Ordinal variables support **order comparisons** (>, <, ==, <=, >=)
- Numerical variables are also ordinal variables. **Not** vice versa.
- **Poll: Q:** *Which are the ordinal variables in this dataset?*
  - *Age*

Duration (in seconds)	Age	Gender	Country
510	22-24	Male	France
423	40-44	Male	India
83	55-59	Female	Germany
391	40-44	Male	Australia
392	22-24	Male	India
470	50-54	Male	France
529	22-24	Male	India
624	22-24	Female	United States of America
214	22-24	Male	United States of America

Types of Variables: Numerical, Ordinal & Categorical

- Elements of a **categorical** or **nominal** variable are **independent** categories or classes.
  - AKA **discrete** variables
- Categorical variables only support **equality** and **inequality** (==, !=)
- They can only be **counted** or **grouped**
- **Poll: Q:** *Which are the categorical variables in this dataset?*
  - *Gender, Country*

Duration (in seconds)	Age	Gender	Country
510	22-24	Male	France
423	40-44	Male	India
83	55-59	Female	Germany
391	40-44	Male	Australia
392	22-24	Male	India
470	50-54	Male	France
529	22-24	Male	India
624	22-24	Female	United States of America
214	22-24	Male	United States of America

Statistics Types

- **Univariate, Bivariate & Multivariate Statistics**

- Univariate Statistics
  - Types of Variables: numerical, ordinal, ratio & categorical
  - Descriptive / Summary Statistics
  - Histograms & Bar Charts
  - Probability Distributions
- Bivariate Statistics
  - Correlation & Covariance
  - Groups & Aggregations
- Multivariate Statistics
  - Covariance Matrices
  - Regression
  - Principal Component Analysis (PCA)

End of Part 2

Exploratory Data Analysis (EDA)

Week 3 – Part 3 – Univariate Statistics

CS 457 - L1 Data Science

Zeehasham Rasheed

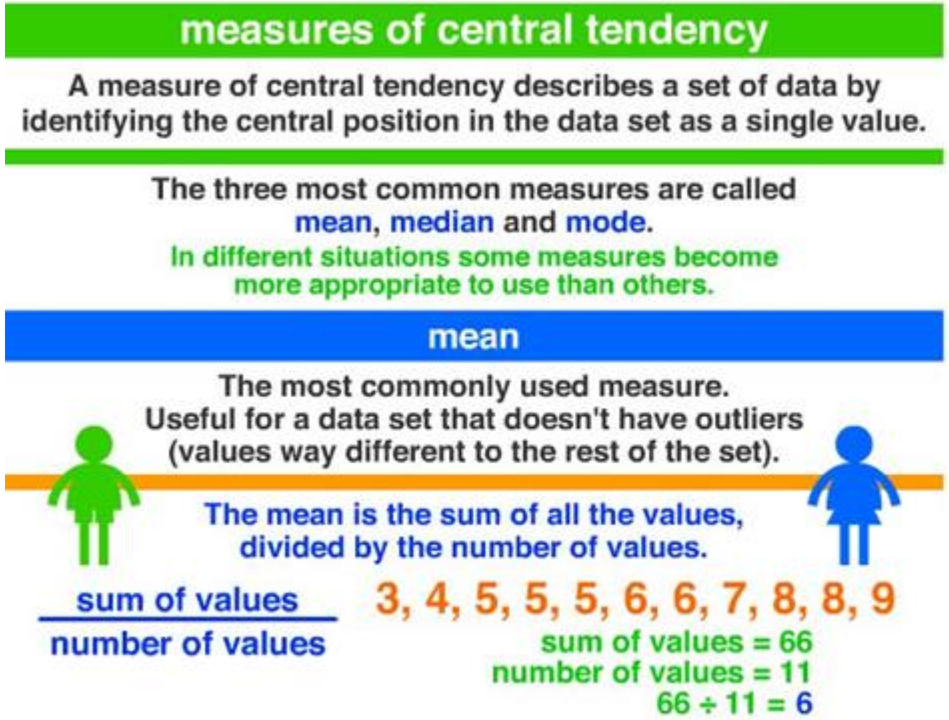
Descriptive Statistics

- Capture *general properties* of a given dataset or sample.
  - ***Central tendency*** measures describe the "*center"* of the distribution
    - Includes **mean** (**arithmetic**, geometric, harmonic, etc.), **median, mode**
  - ***Variation*** or *variability* measures describe data *spread*
    - *How far* the measurements lie from the "center".
    - Includes range, quartiles, **variance, standard deviation**

- **Fundamental Idea:**
  - learn and use stats appropriate for data types

Descriptive Statistics (2)

- Three most common measures
  - Mean, Median and Mode



measures of central tendency

A measure of central tendency describes a set of data by identifying the central position in the data set as a single value.

The three most common measures are called mean, median and mode.

In different situations some measures become more appropriate to use than others.

mean

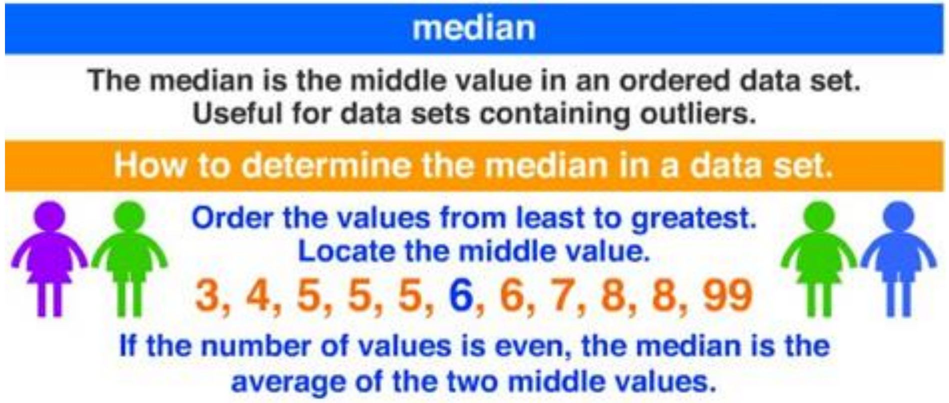
The most commonly used measure.
Useful for a data set that doesn't have outliers (values way different to the rest of the set).

The mean is the sum of all the values, divided by the number of values.

$$\frac{\text{sum of values}}{\text{number of values}}$$

3, 4, 5, 5, 5, 6, 6, 7, 8, 8, 9

sum of values = 66
number of values = 11
66 ÷ 11 = 6



median

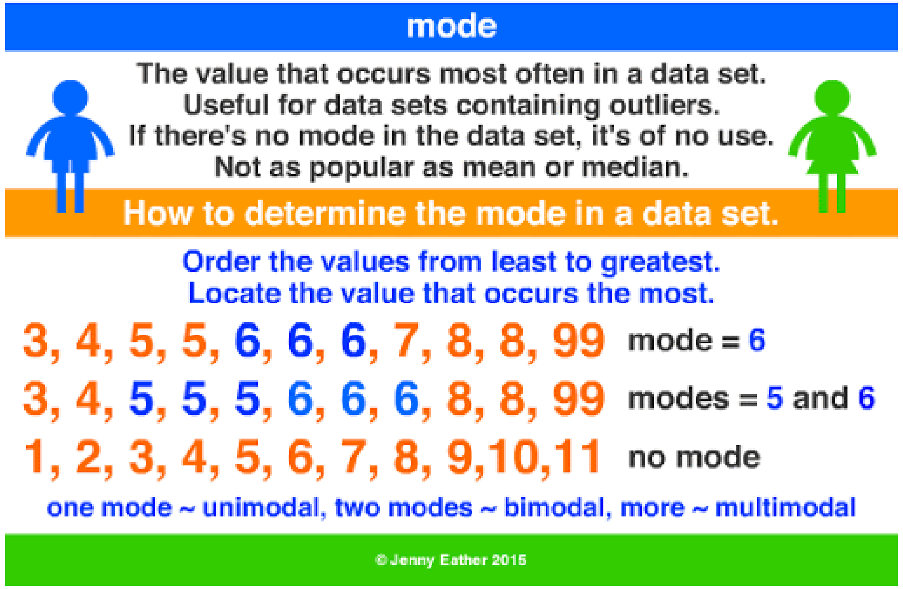
The median is the middle value in an ordered data set.
Useful for data sets containing outliers.

How to determine the median in a data set.

Order the values from least to greatest.
Locate the middle value.

3, 4, 5, 5, 5, 6, 6, 7, 8, 8, 99

If the number of values is even, the median is the average of the two middle values.



mode

The value that occurs most often in a data set.
Useful for data sets containing outliers.
If there's no mode in the data set, it's of no use.
Not as popular as mean or median.

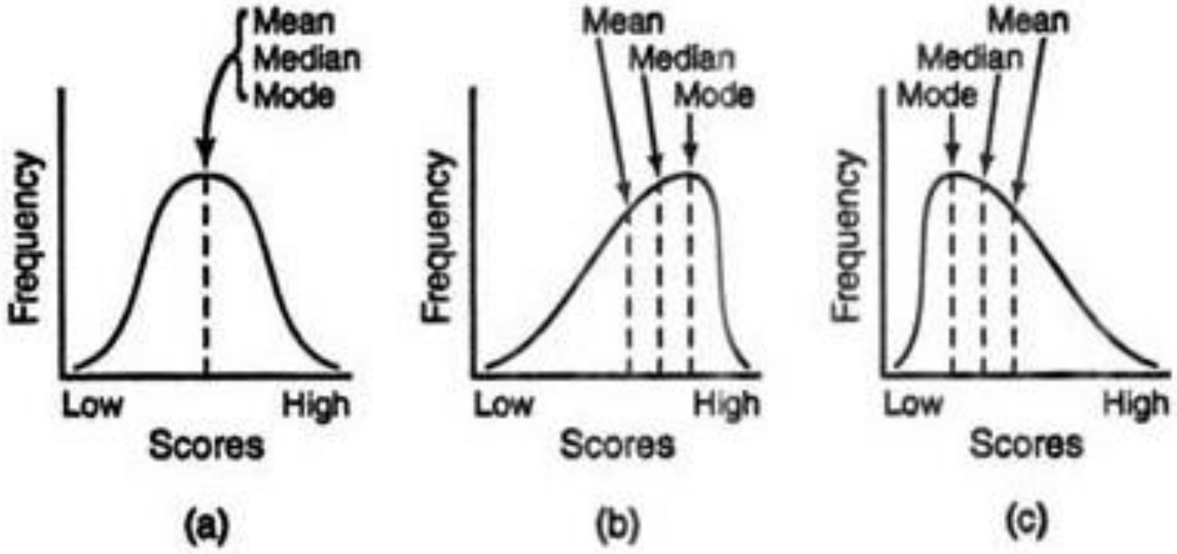
How to determine the mode in a data set.

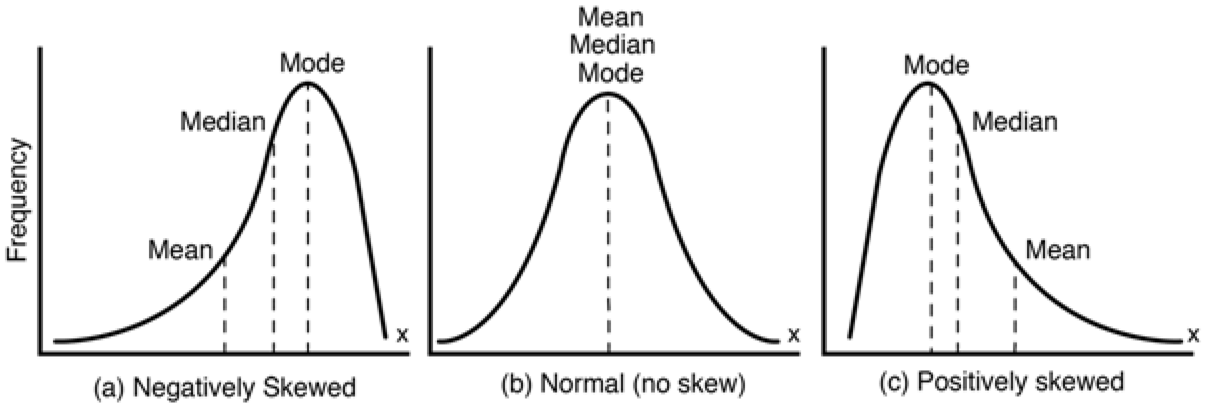
Order the values from least to greatest.
Locate the value that occurs the most.

3, 4, 5, 5, 6, 6, 6, 7, 8, 8, 99 mode = 6
3, 4, 5, 5, 5, 6, 6, 6, 8, 8, 99 modes = 5 and 6
1, 2, 3, 4, 5, 6, 7, 8, 9,10,11 no mode

one mode ~ unimodal, two modes ~ bimodal, more ~ multimodal

Central Tendency Example





(a) Negatively Skewed

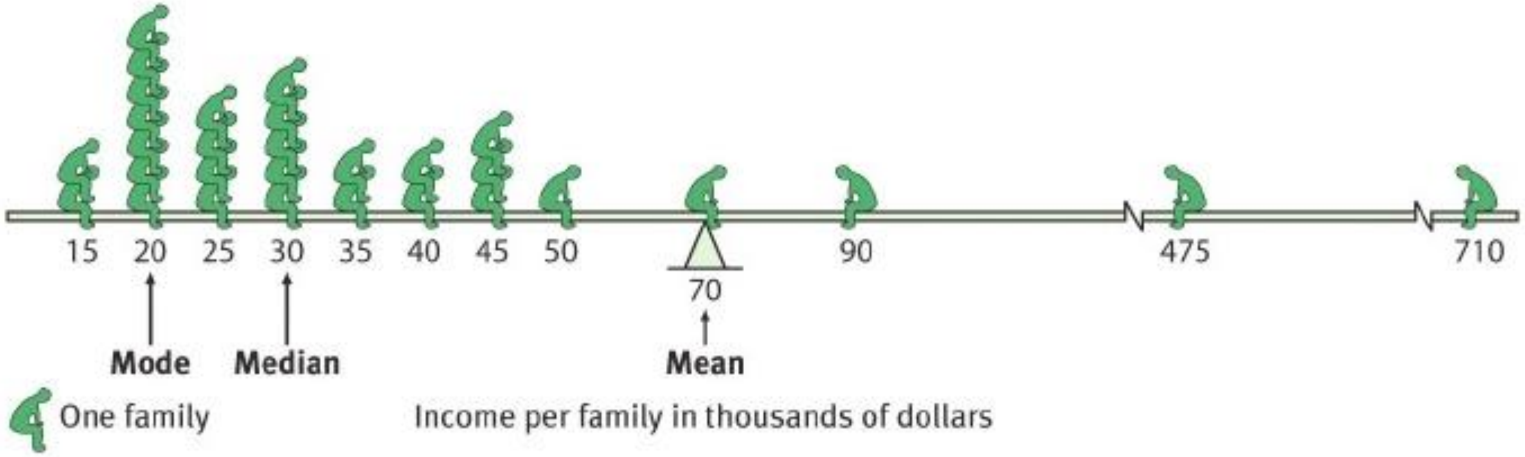
(b) Normal (no skew)

(c) Positively skewed

Detecting Outliers

- Detecting Outliers/Noise using Descriptive Statistics

- 475 and 710 can be considered as outliers.
- Mean was unable to control outliers
- Median was successfully able to detect outliers



End of Part 3

Exploratory Data Analysis (EDA)

Week 3 – Part 4 – Define and Aggregating Errors

CS 457 - L1 Data Science

Zeehasham Rasheed

Summarizing Information

Scenario:

1. You are stranded on a planet, waiting to be rescued.
2. A rescue mission is being arranged, but:

   1. What is the length of the day?
   2. How is the weather – cloudy, hot or cold?
   3. What is your heart rate?

Your transmitter has a limited bandwidth of sending only one row – and you can use it only once a week.

- How do you proceed?

- Which information would you transmit?

Day	Weather	Day Length	Heartbeats/ min
Day 1	Hot	9	85
Day 2	Cloudy	9	93
Day 3	Cold	9	84
Day 4	Cloudy	11	71
Day 5	Hot	11	73
Day 6	Hot	9	97
Day 7	Cold	10	79
Day 8	Hot	10	68

How to Summarize?

Day	Weather	Day Length	Heartbeats/ min
Day 1	Hot	9	85
Day 2	Cloudy	9	93
Day 3	Cold	9	84
Day 4	Cloudy	11	71
Day 5	Hot	11	73
Day 6	Hot	9	97
Day 7	Cold	10	79
Day 8	Hot	10	68

- How do you summarize this data so that you are **least wrong**?
- How to **measure error**?
- How to **minimise** that error?

- Every column is summarized (using mean, median or mode) as a single number, S_i
- *'S_i'* is a good summary if the discrepancy between *'S_i'* and **each value** of the i^{th} column is **small**.
- **Errors** and their **aggregation.**

Defining & Aggregating Errors

Important Question: Which descriptive statistics (mean, median, mode) is applied to which column?

- Each column *i* is summarised by s_i
- Each value in column *i* is x_{ij}
- Each x_{ij} creates its own error with s_i.

Intuitively:

- This error is small if $x_{ij} \approx s_i$ and
- Large if $x_{ij} >> s_i$ or $x_{ij} << s_i$

Weather	Day Length	Heartbeats/ min
Hot	9	85
Cloudy	9	93
Cold	9	84
Cloudy	11	71
Hot	11	73
Hot	9	97
Cold	10	79
Hot	10	68

Appropriate Descriptors

Some useful information:

- Weather is a **categorical** variable.
  - **Mode**
- Day length is **numerical**, but changes **slowly**.
  - **Mean**
- Heartbeat is also numerical, but may change **very drastically**.
  - **Median**

Weather	Day Length	Heartbeats/ min
Hot	9	85
Cloudy	9	93
Cold	9	84
Cloudy	11	71
Hot	11	73
Hot	9	97
Cold	10	79
Hot	10	68

Mathematical Notation

$$E_i^0 = \sum_j |x_{ij} - s_i|^0$$

Mode

$$s_i^0 = \arg min_{s_i} E_i^0$$

$$E_i^1 = \sum_j |x_{ij} - s_i|^1$$

Median

$$s_i^1 = \arg min_{s_i} E_i^1$$

$$E_i^2 = \sum_j |x_{ij} - s_i|^2$$

Mean
(μ)

$$s_i^2 = \arg min_{s_i} E_i^2$$

End of Part 4

Exploratory Data Analysis (EDA)

Week 3 – Part 5 – Advance Statistics

CS 457 - L1   Data Science

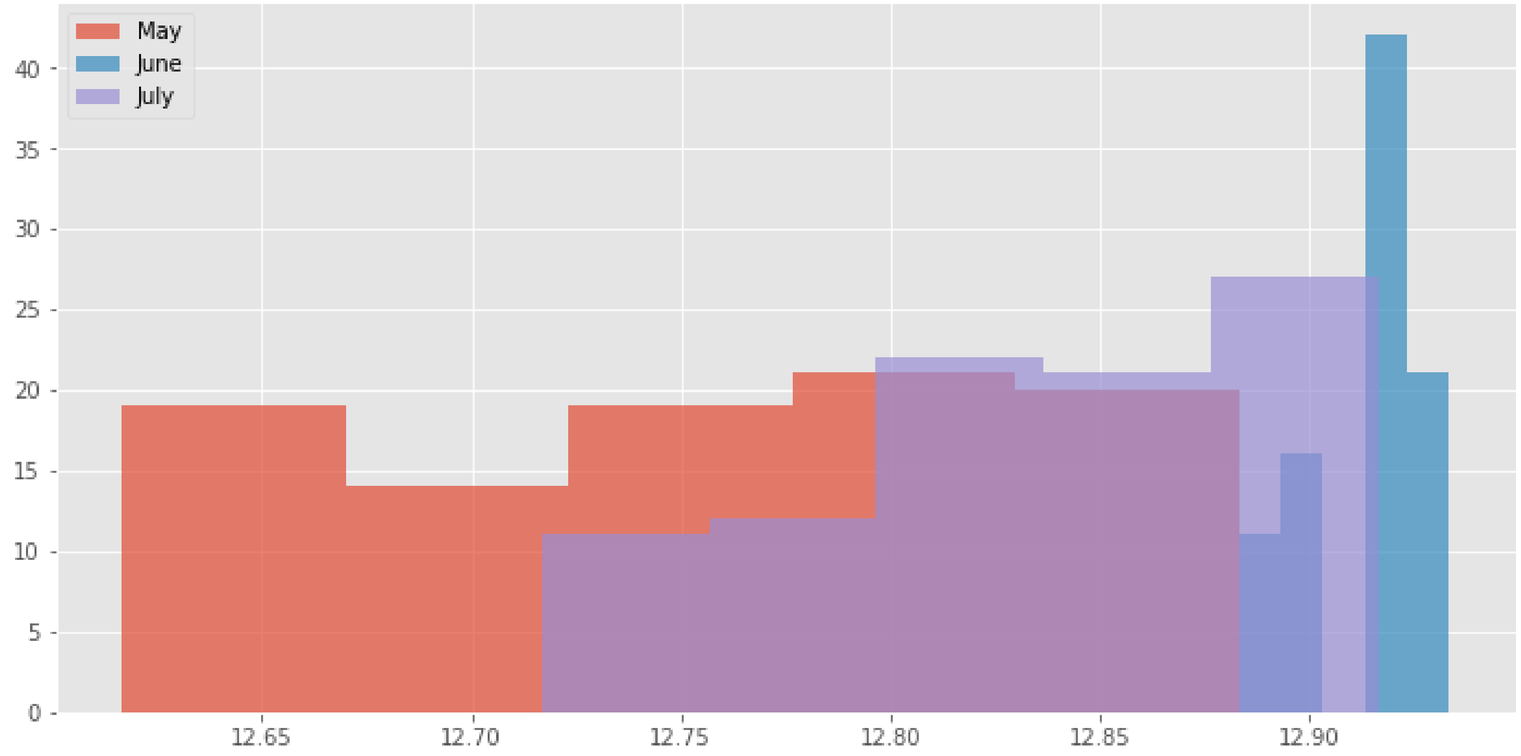
Zeehasham Rasheed

More Descriptive Statistics

- Variance
- Standard Deviation
- Frequency and Counts
- Probability Distributions

Variance

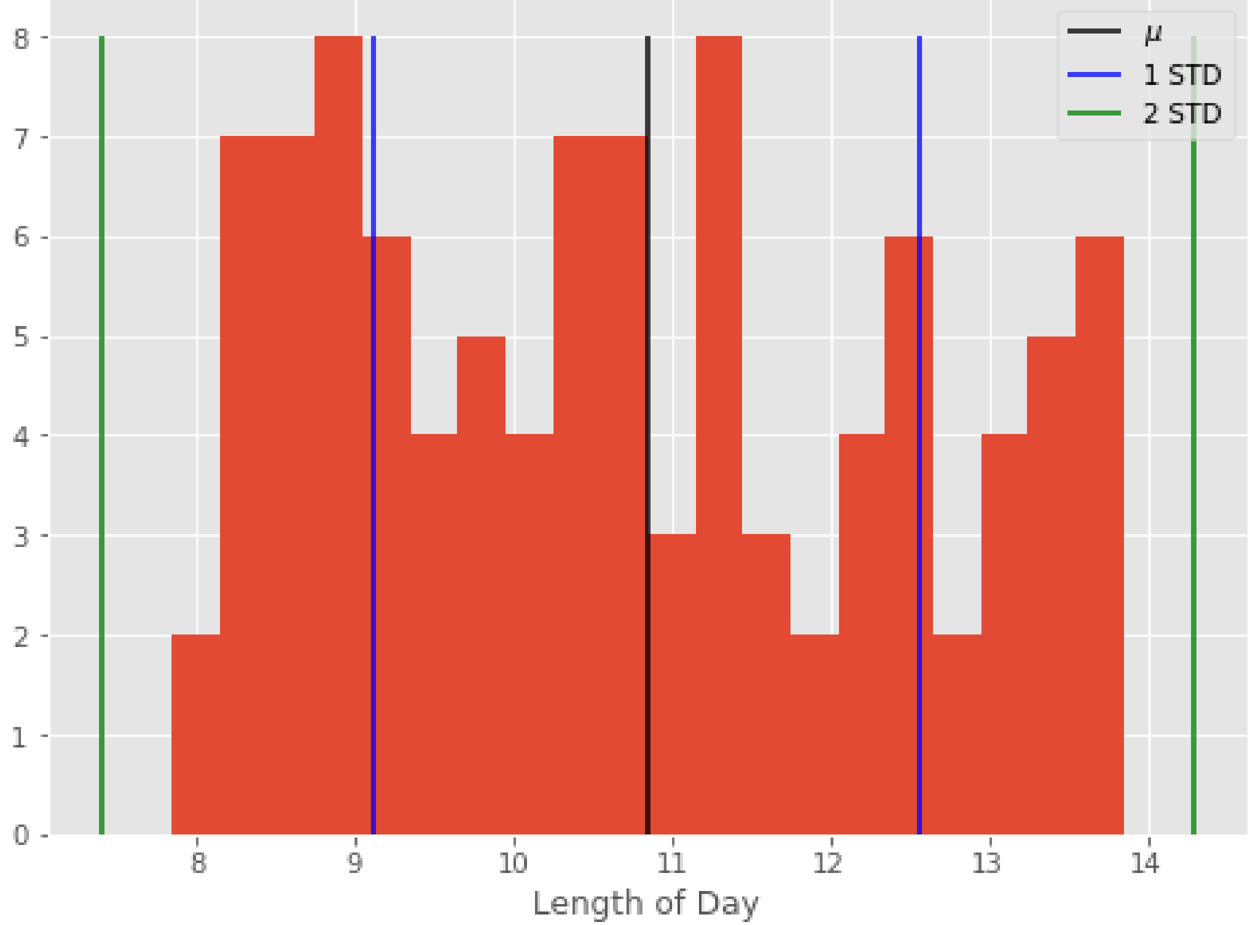
- **Variance:** How do we quantify the spread of the data?



Length of day in hours with sunlight

$$v_i = \frac{1}{N}\sum_{j=1}^{N}\left(x_{ij} - \mu_i\right)^2$$

- Standard Deviation (STD) measures deviation of data from a mean



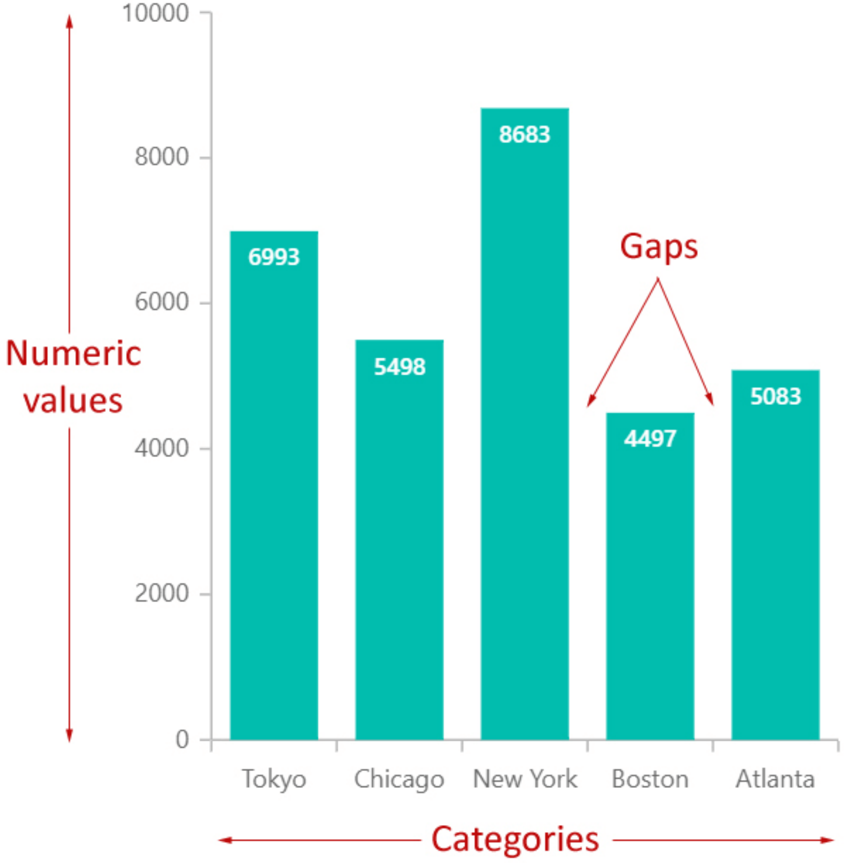
$$v_i = \frac{1}{N}\sum_{j=1}^{N}(x_{ij} - \mu_i)^2$$

$$\sigma_i = \sqrt{v_i}$$

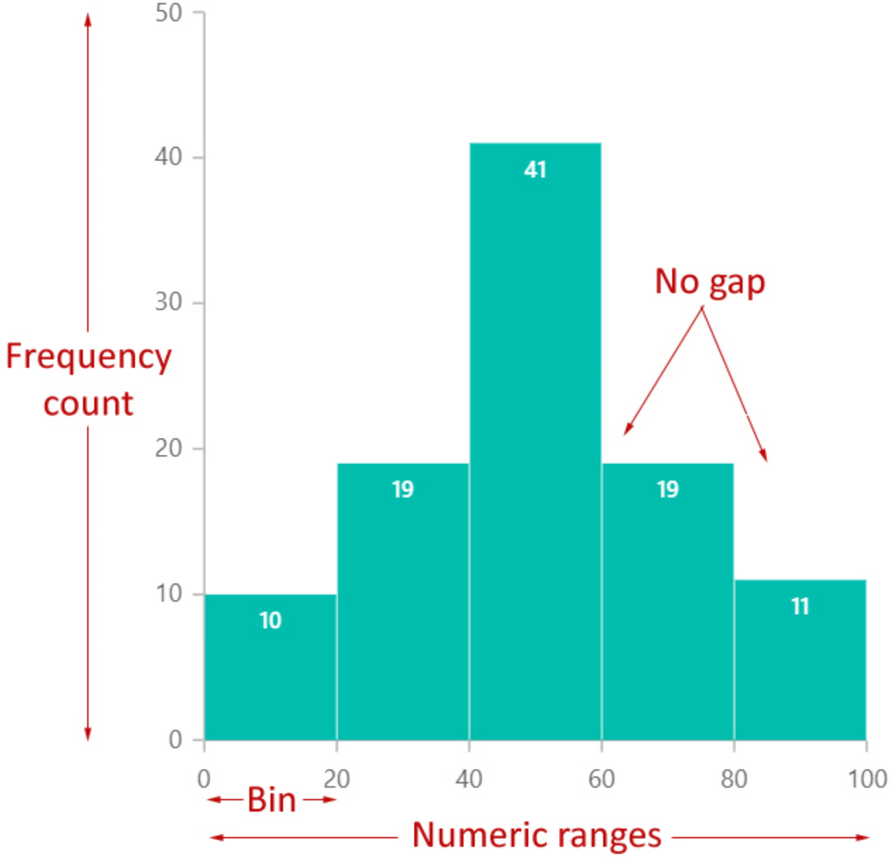
- STD has the same units as the data.
- STD corresponds to risk in business problems

Histograms & Bar Charts

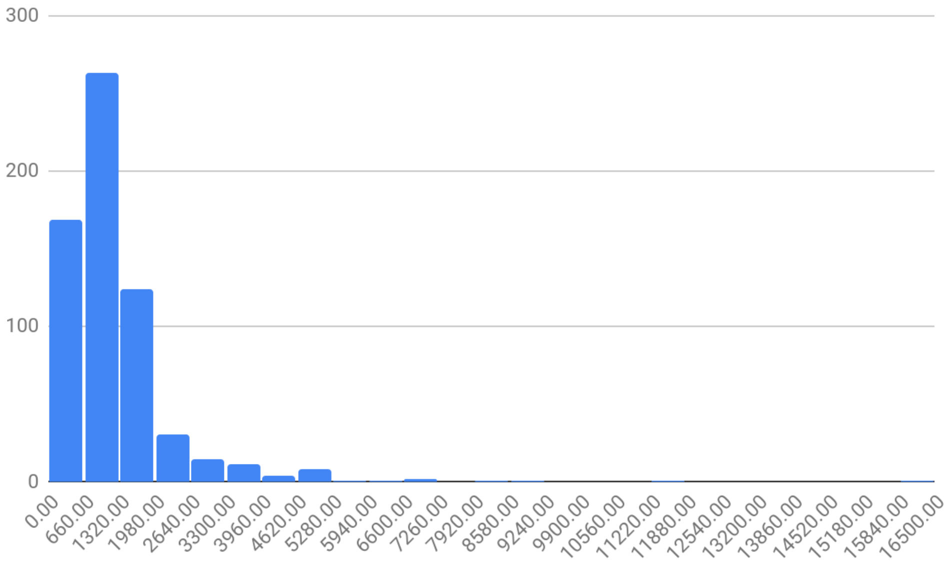
Bar Chart

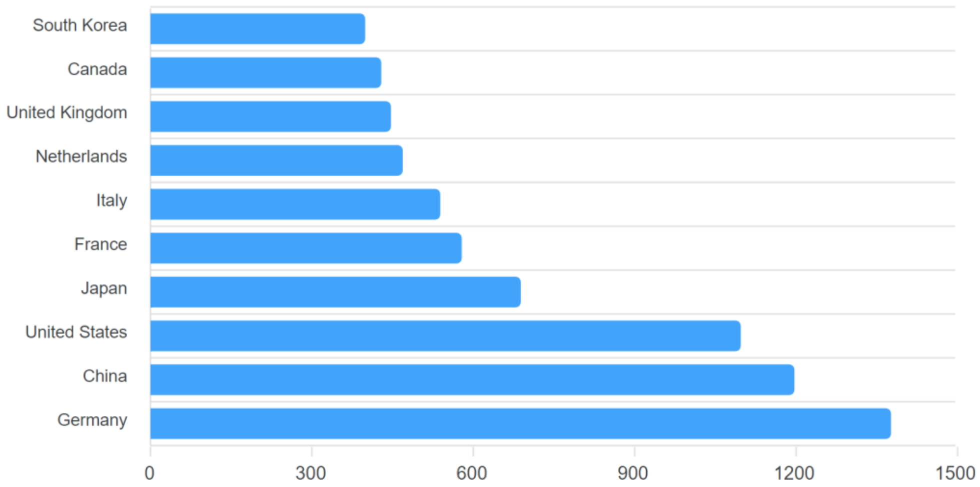


Histogram Chart



Histograms & Bar Charts (2)

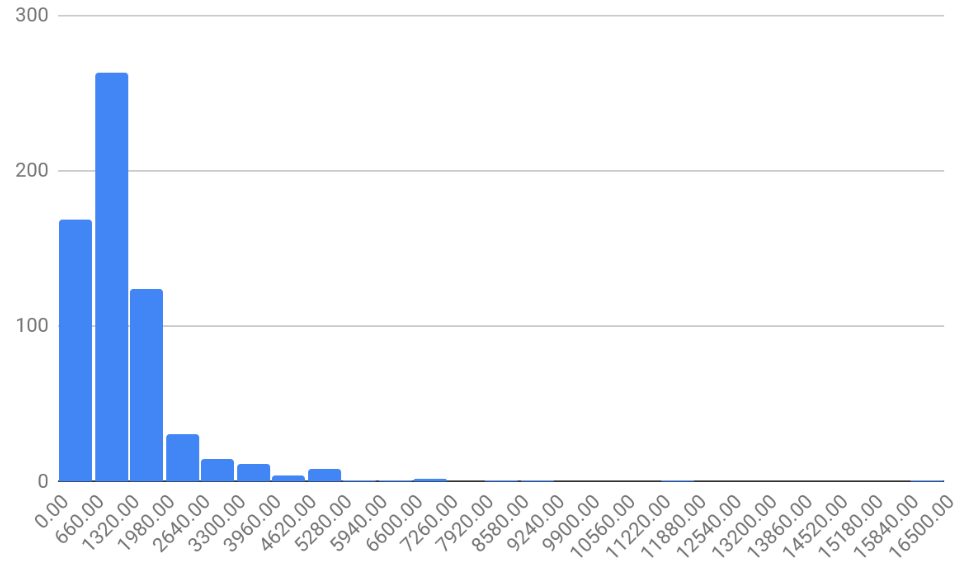


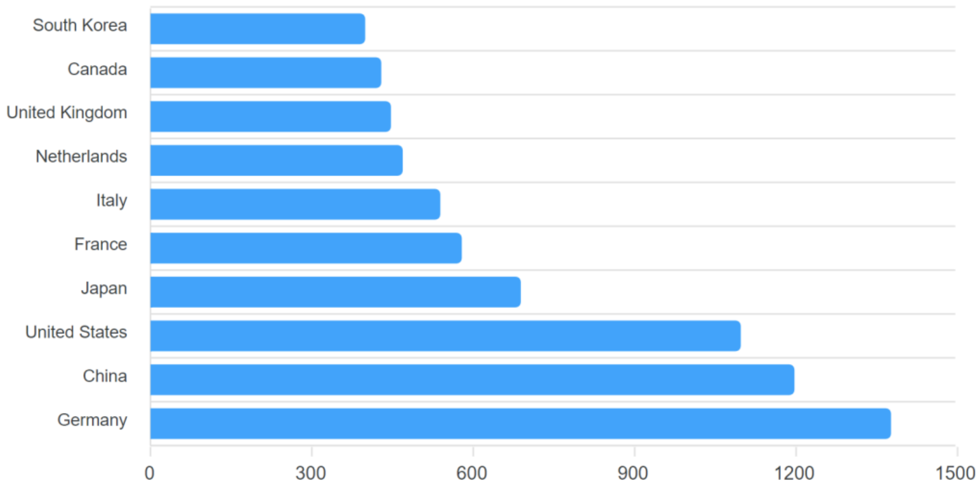


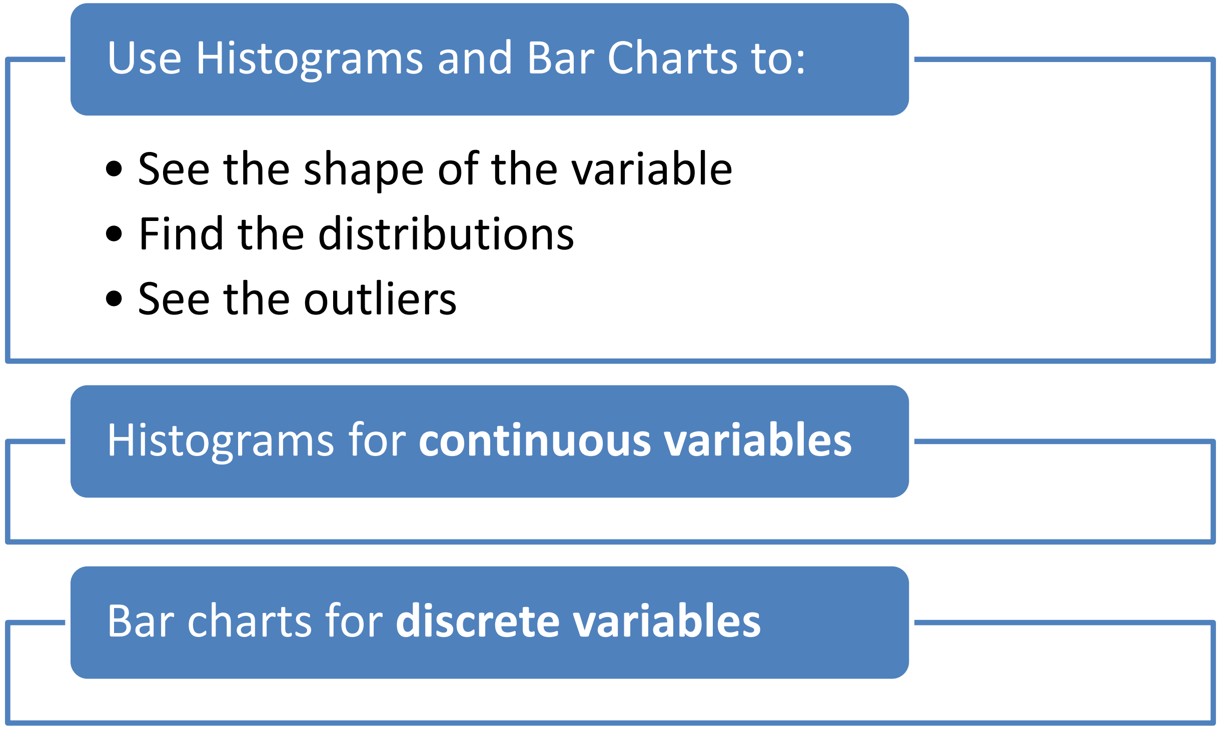
Poll: Which is the histogram and which is the bar chart?

Duration (in seconds)	Age	Gender	Country
510	22-24	Male	France
423	40-44	Male	India
83	55-59	Female	Germany
391	40-44	Male	Australia
392	22-24	Male	India
470	50-54	Male	France
529	22-24	Male	India
624	22-24	Female	United States of America
214	22-24	Male	United States of America

Benefits







Use Histograms and Bar Charts to:

- See the shape of the variable
- Find the distributions
- See the outliers

Histograms for **continuous variables**

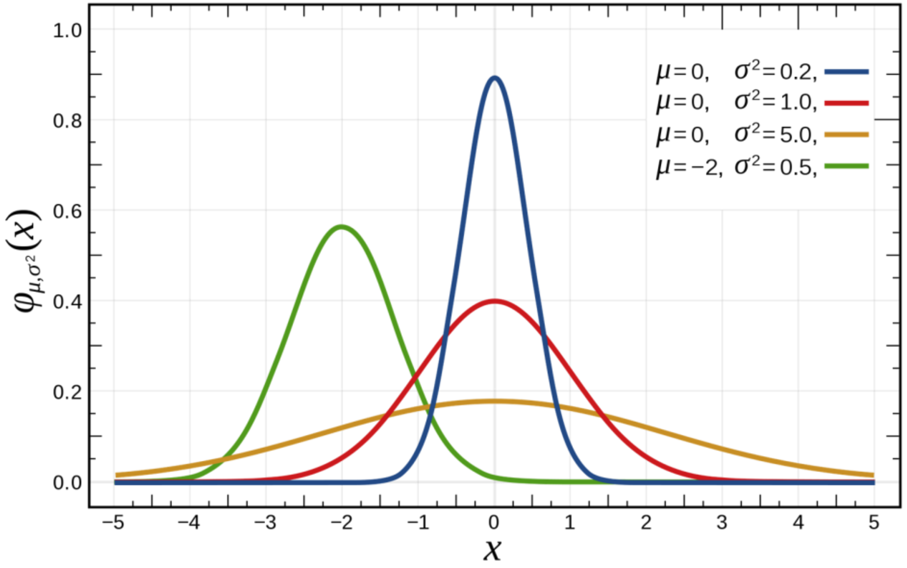
Bar charts for **discrete variables**

Why Probability Distribution

- Probability distributions help to model our world, enabling us to obtain estimates of the probability from our data that
  - a certain event may occur
  - or estimate the variability of occurrence for any event

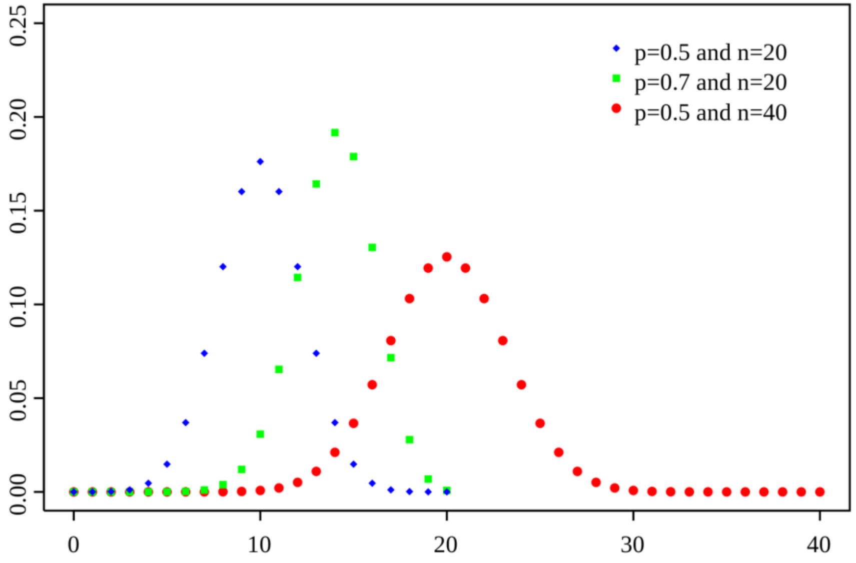
- Some practical uses of probability distributions are:
  - **Inferential Statistics** to draws conclusions using estimates that cannot be derived from descriptive statistics
  - To describe, and possibly predict, the probability of an event **(Machine Learning)**

Common Probability Distributions



Normal / Gaussian Distribution:

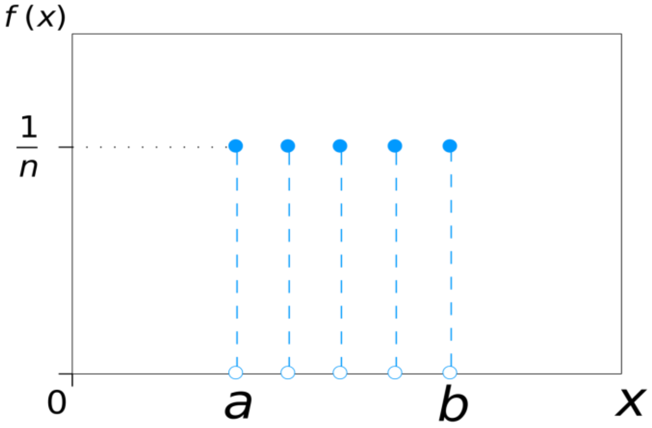
1. **Continuous valued** distribution.
2. Parameters: mean & standard deviation
3. Defines the *central tendency* and *spread* of any naturally occuring quantity.



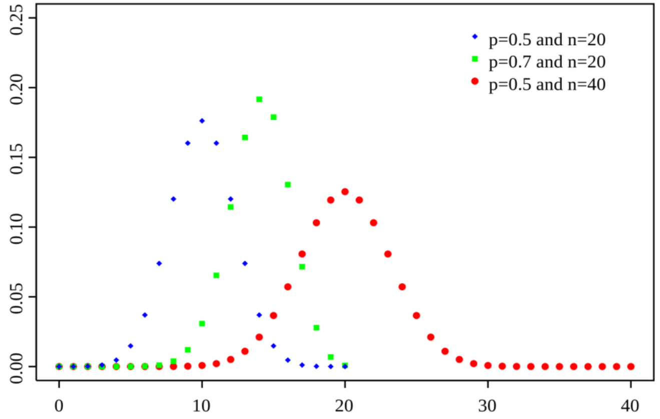
Binomial Distribution:

1. **Discrete valued** distribution.
2. Parameters: p & n
3. Defines the *number of success/failures in a sequence of binary events.*

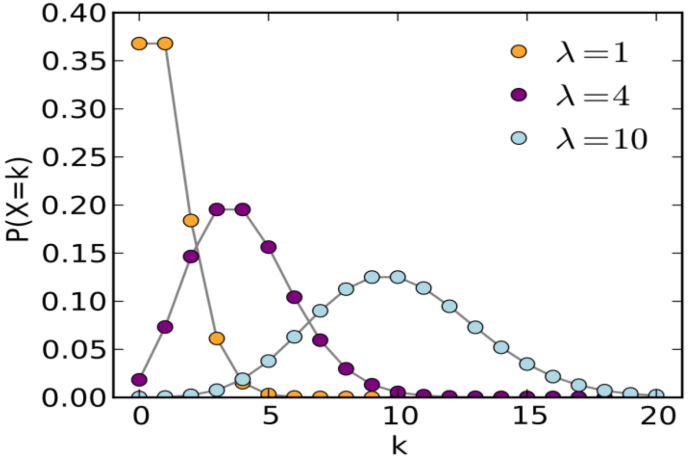
More Probability Distributions



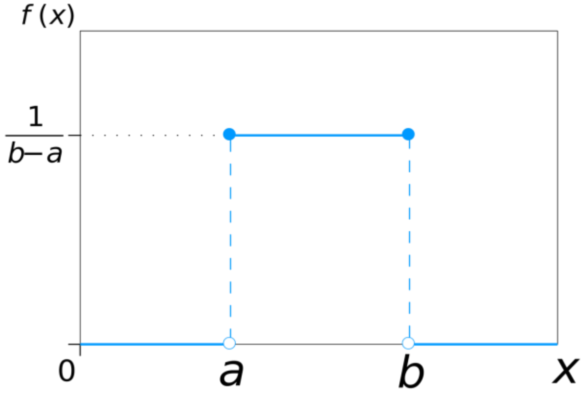
Discrete Distribution



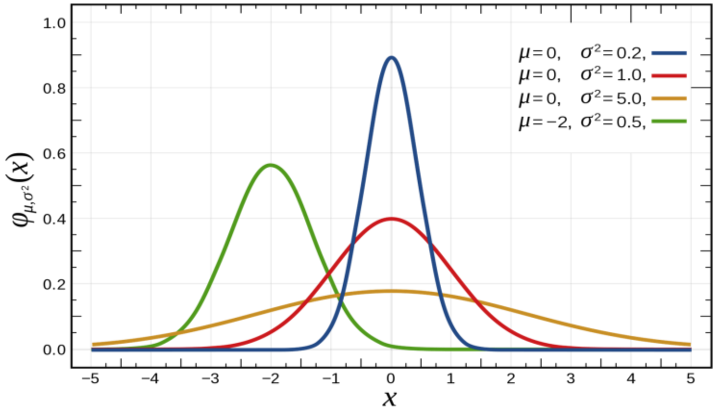
Binomial Distribution



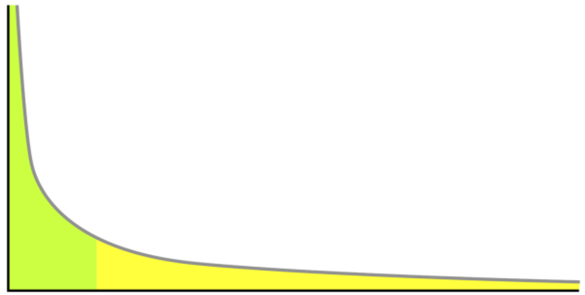
Poission Distribution



Discrete Distribution



Normal/Gaussian Distribution



Power Law Distribution

End of Part 5

Exploratory Data Analysis (EDA)

Week 3 – Part 6 – Bivariate Analysis

CS 457 - L1 Data Science

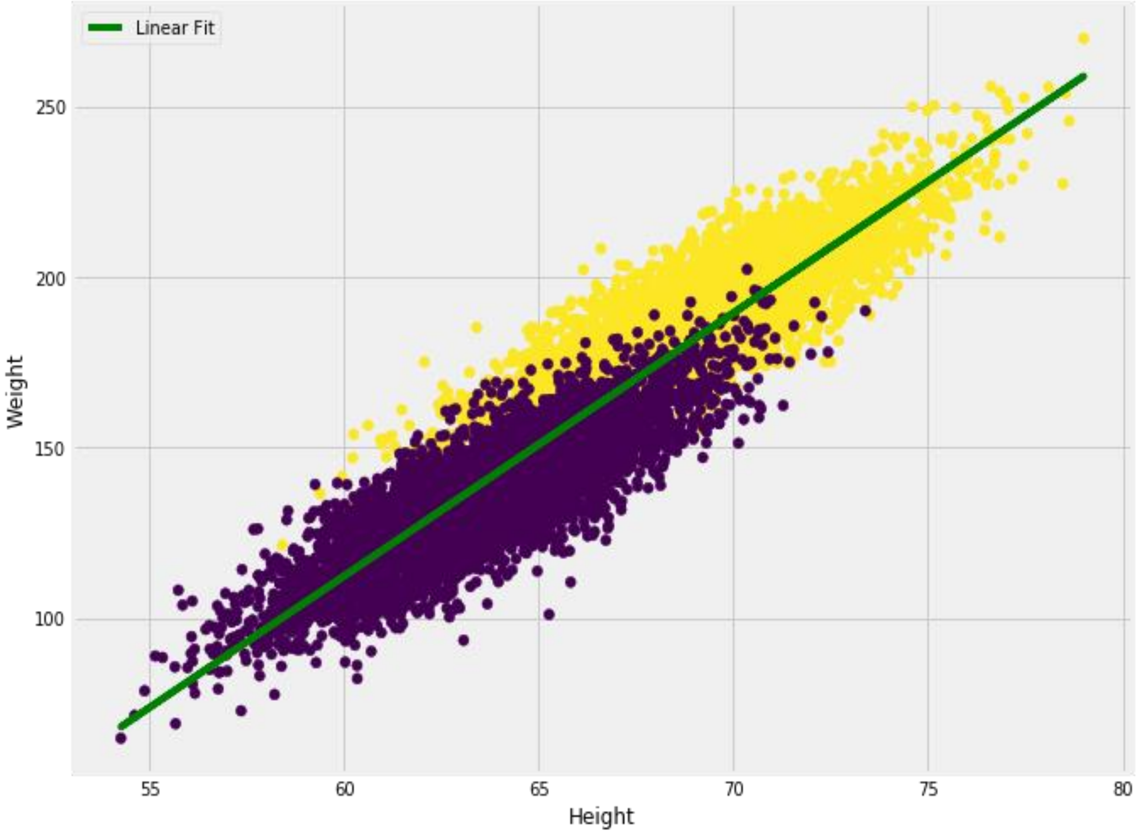
Zeehasham Rasheed

Bivariate Statistics

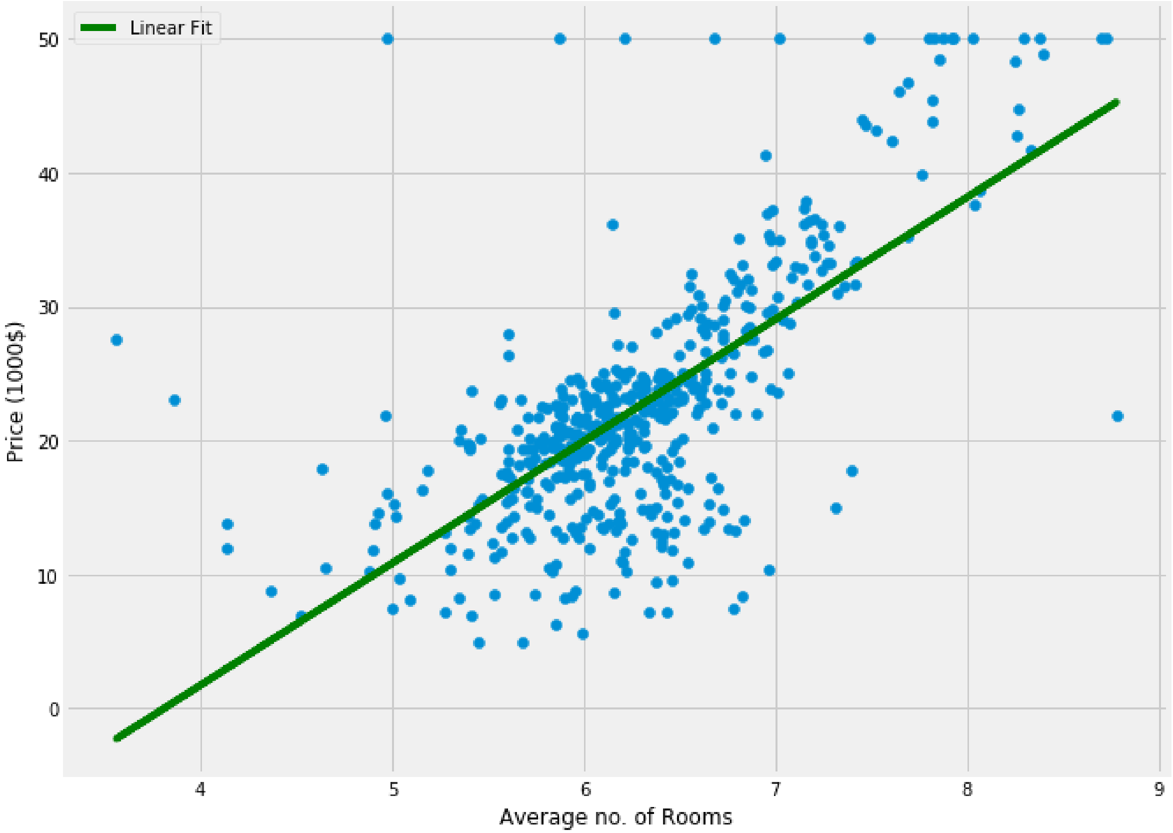
- Continuous vs Continuous
- Continuous vs Discrete
- Discrete vs Discrete

Scatter Plots Showing Relationship

- Continuous vs Continuous



Plot 1



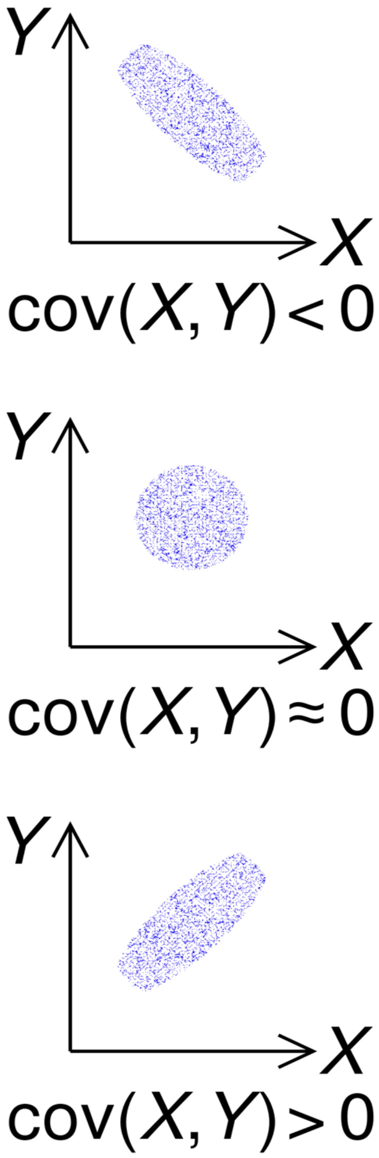
Plot 2

Covariance

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

Covariance is the measure of the **joint variability** between two random variables.

Covariance is a measure of how two variables change together, but its magnitude is unbounded, so it is difficult to interpret.



Correlation

Correlation or **statistical dependence** is any relationship between two random variables, or **bivariate** data.

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

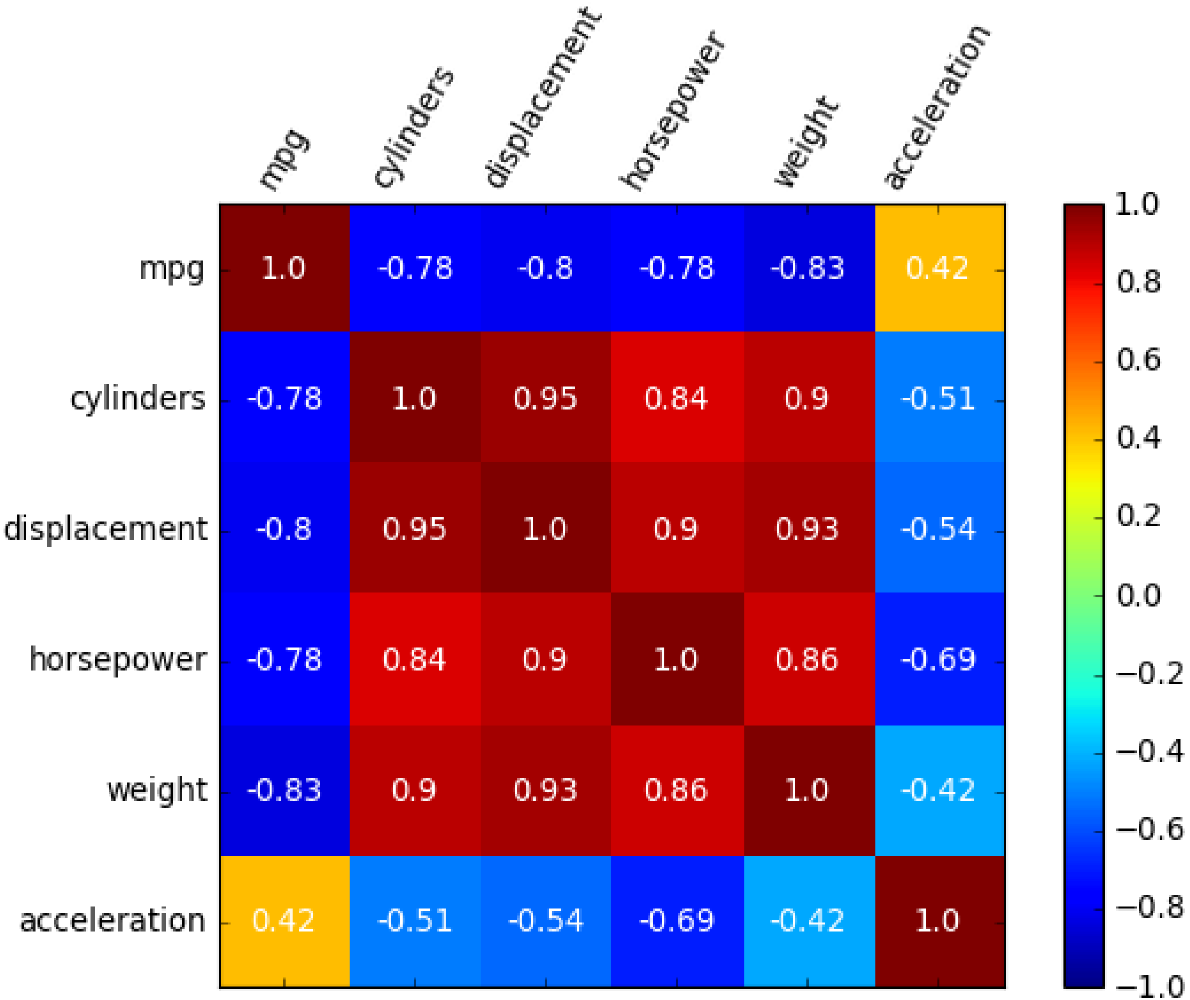
By dividing covariance by the product of the two standard deviations, one can calculate the normalized version of the statistic. This is called the correlation coefficient.

Correlation coefficient converts relationship to a number from -1 to 1

Try this: http://guessthecorrelation.com/

Correlation Plot

- Vehicle Features correlation
  - Cylinders vs MPG
  - Horsepower vs Cylinders
  - Cylinders vs Weight



	mpg	cylinders	displacement	horsepower	weight	acceleration
mpg	1.0	-0.78	-0.8	-0.78	-0.83	0.42
cylinders	-0.78	1.0	0.95	0.84	0.9	-0.51
displacement	-0.8	0.95	1.0	0.9	0.93	-0.54
horsepower	-0.78	0.84	0.9	1.0	0.86	-0.69
weight	-0.83	0.9	0.93	0.86	1.0	-0.42
acceleration	0.42	-0.51	-0.54	-0.69	-0.42	1.0

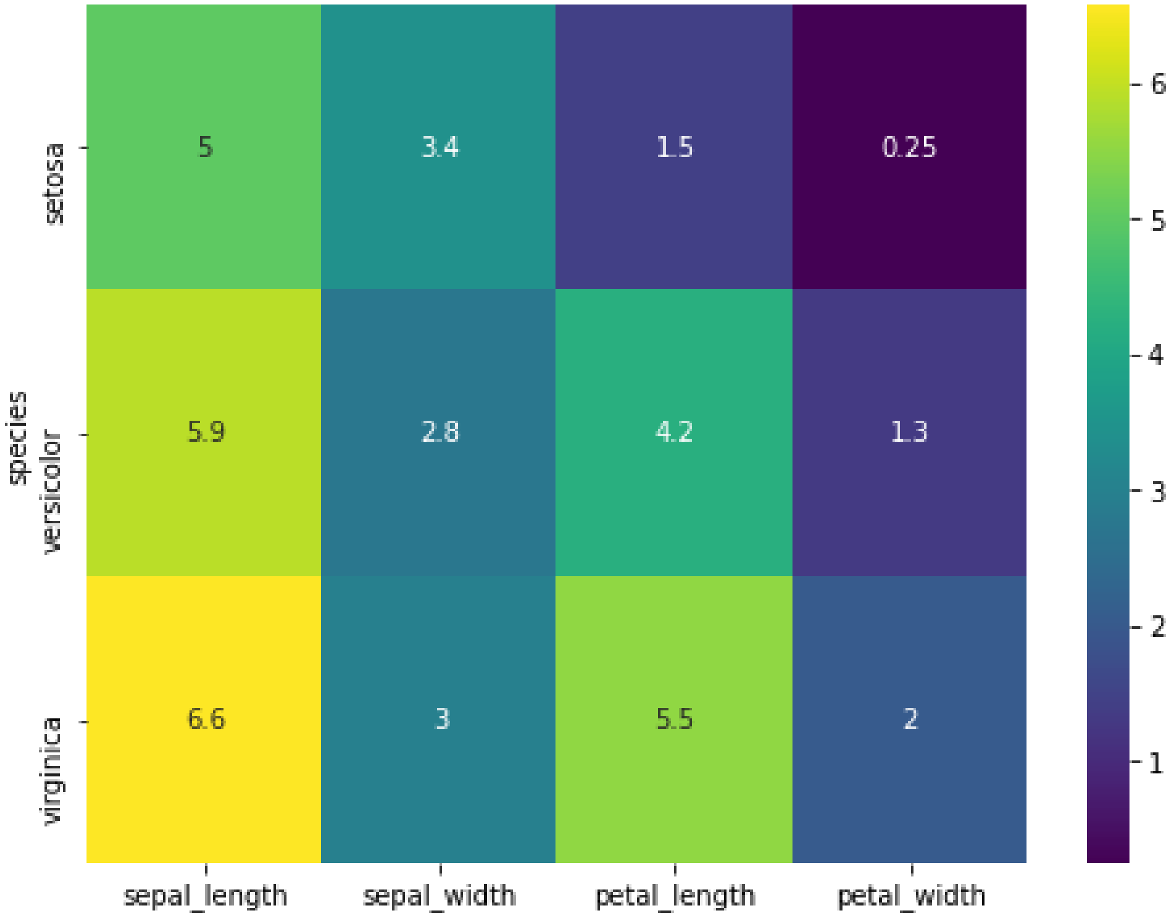
Correlation matrix of the Auto-MPG dataset

Grouping & Aggregation

- **Continuous vs Discrete**

- Pick a discrete feature - ***dimension***
- Pick a continuous feature - ***metric***

- Filter the data by a unique value in **dimension**, find some aggregation of the **metric:**
  - Sum
  - Average/Mean
  - Min, Max, etc.



Multivariate Statistics

- Multivariate Analysis of Variance (M AN O VA or ANOVA)
  - Generalization of Bivariate Methods
- Multiple Regression
- Principal Component Analysis (PCA)
  - All will be discussed in upcoming weeks

End of Part 6

Exploratory Data Analysis (EDA)

Week 3 – Part 7 – Finding Insights

CS 457 - L1 Data Science

Zeehasham Rasheed

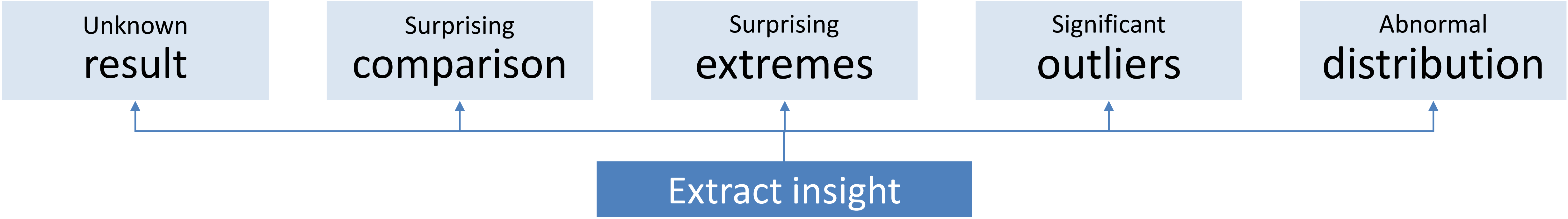
We Want to

- Create meaningful insights
- Present them effectively
- Make them easy to consume

- **Analysis ≠ Insights**

- **What separates insights from analysis?**
  - **Patterns of Insights**

Patterns of Insights



Unknown result

Surprising comparison

Surprising extremes

Significant outliers

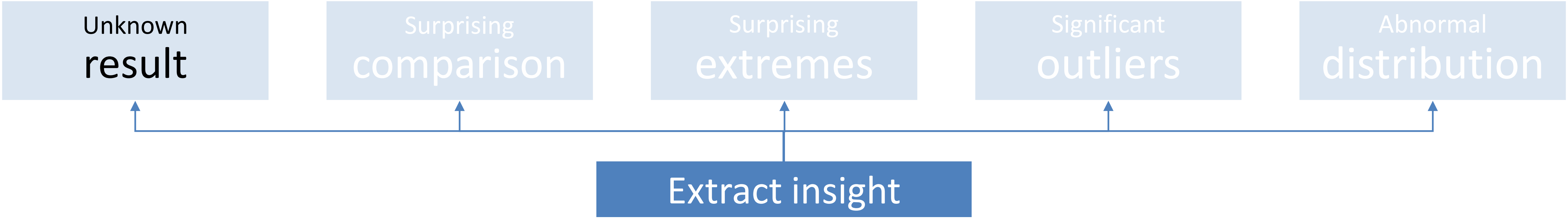
Abnormal distribution

Extract insight

Almost all significant insights fall into one of these patterns

Knowing these patterns beforehand helps frame questions and look for answers

Unknown Results



Unknown result

Surprising comparison

Surprising extremes

Significant outliers

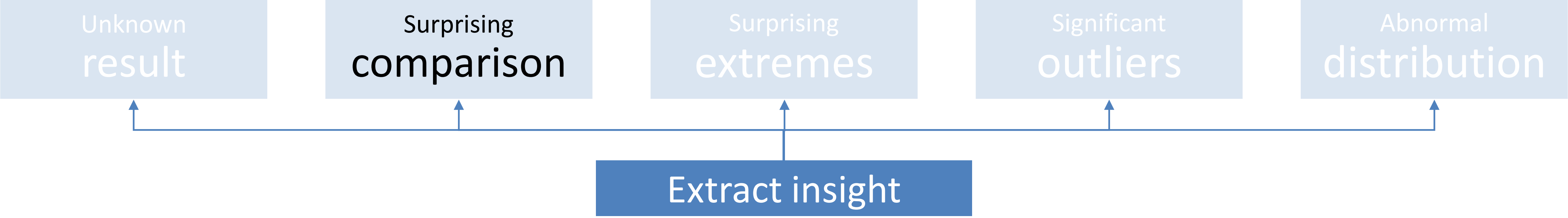
Abnormal distribution

Extract insight

Examples:

- The national animal of Scotland is a unicorn.
- Revenue has increased by x% from the last quarter.
- Sales have decreased by y% in this financial year.

Surprising Comparison



Unknown result

Surprising comparison

Surprising extremes

Significant outliers

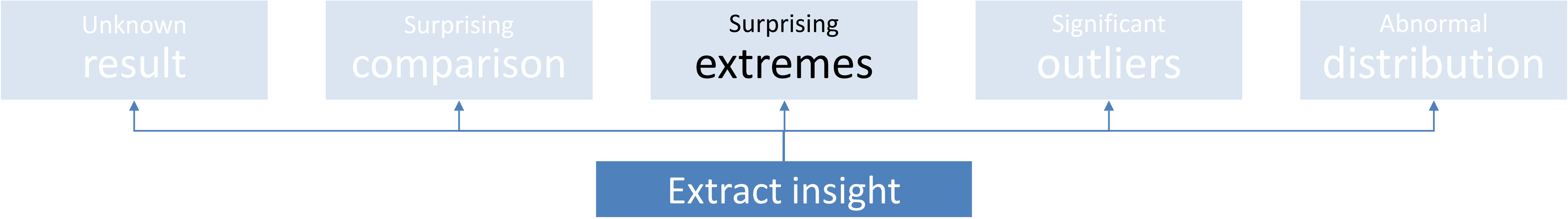
Abnormal distribution

Extract insight

Examples:

- Revenue has **doubled** since the last quarter! 😎
- Sales are only **half** as compared to our competitor. 😕

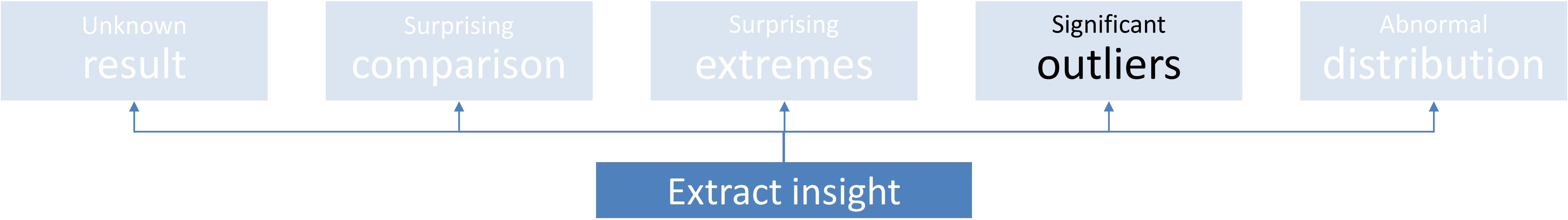
Surprising Extremes



Examples:

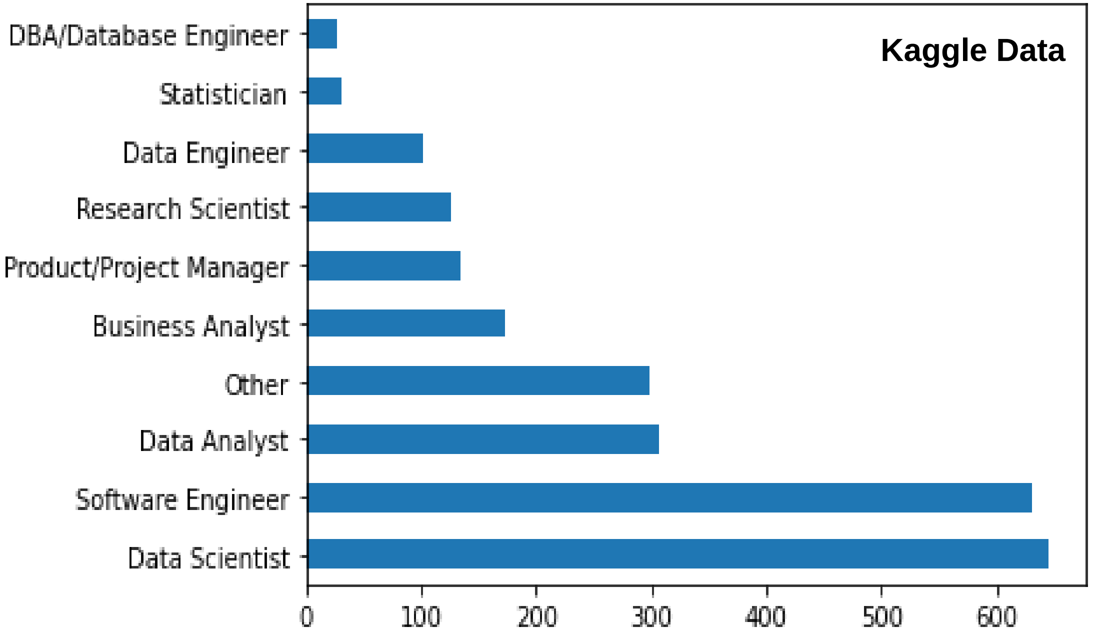
- Men who skip breakfast get more coronary heart disease. American men 45 to 82 who skip breakfast showed a **27 percent** higher risk of coronary heart disease over a 16-year period. (Harvard University medical researchers)
- Smart people like curly fries. Liking “Curly Fries” on Facebook is predictive of high intelligence. (Researchers at the University of Cambridge and Microsoft Research)

Significant Outliers

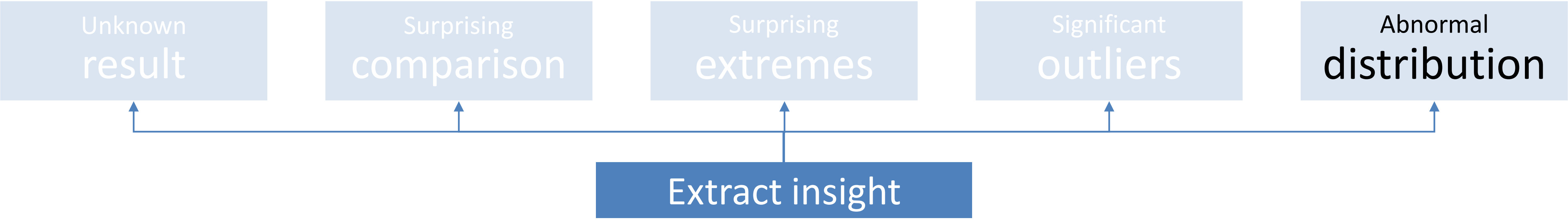


Examples:

- Lahore is a hub of startup activity taking 68% of total funding and 46% of all startup activity
  - followed by Karachi and Islamabad at 13% and 2.5% of total funding.
- There are twice as many software engineers on Kaggle than data analysts.



Abnormal Distributions



Unknown result

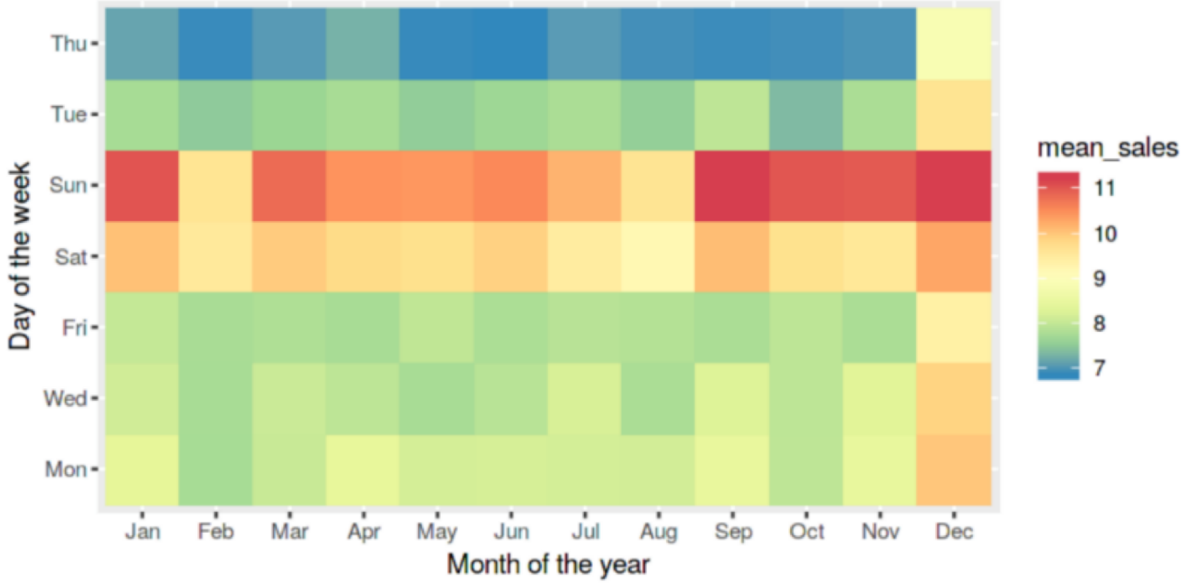
Surprising comparison

Surprising extremes

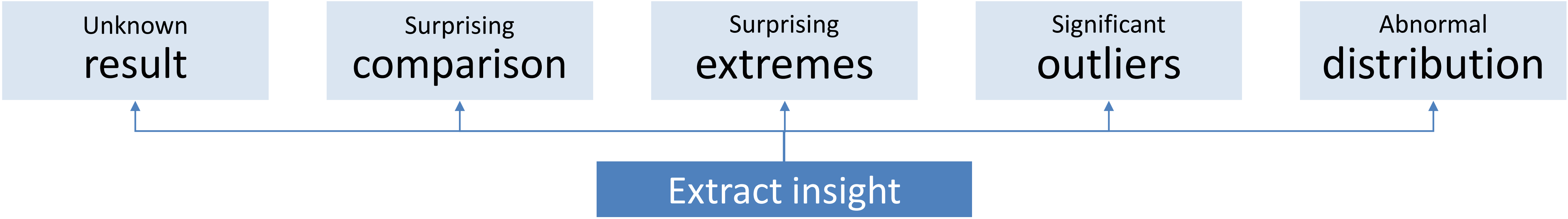
Significant outliers

Abnormal distribution

Extract insight



Patterns of Insights



Unknown result

Surprising comparison

Surprising extremes

Significant outliers

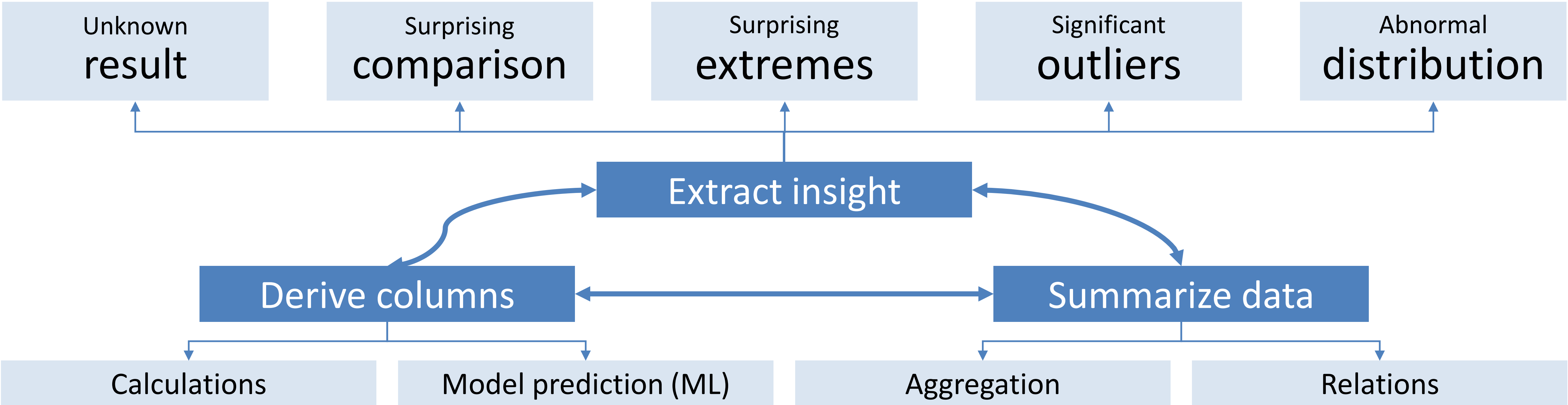
Abnormal distribution

Extract insight

USE THESE TO **CATEGORIZE** YOUR ANALYSIS

Summary of Insights

- Categorize your analysis



Unknown result

Surprising comparison

Surprising extremes

Significant outliers

Abnormal distribution

Extract insight

Derive columns

Summarize data

Calculations

Metadata lookup

Transformations
(e.g. Count, Frequency, Binning)

Model prediction (ML)

Cluster

Classify

Aggregation

Describe, Summary Stats

Group

Pivot

Relations

Hierarchies

Correlations

End of Part 7

Exploratory Data Analysis (EDA)

Week 3 – Part 8 – Framing Questions with Patterns of Insights

CS 457 - L1 Data Science

Zeehasham Rasheed

Unknown Results

Pattern	Question Template	Examples
Unknown result	What is **{{ metric }}** of **{{ value }}**?	
		• What is the average rainfall in Uganda?
		• What is the cheapest place to buy coffee in Karachi?
		• What is the income of the poorest person in the richest country?

Surprising Comparison

Pattern	Question Template	Examples
Surprising Comparison	Is the {{ metric }} of {{ value_X }} {{ greater or less than }} {{ same metric }} of {{ value_Y }}? If so, by how much? (In absolutes or percentage?)	
		• What % of urban population has access to PPE in the USA vs in Italy?
		• Are richer countries happier than poorer countries?
		• Between demonetization and the Covid-19 lockdown, where did the informal sector of the economy suffer more losses?

Surprising Extremes & Significant Outliers

Pattern	Question Template	Examples
Surprising Extremes	What is the {{ maximum / minimum }} of {{ value }}?	
		• How tall is the tallest person in the room?
		• Which is the most developed city? Is it also the richest?
		• Which batsman has the best strike rate? Does it match their batting average?
Significant Outliers	How much {{ greater or less than }} is the {{ highest or lowest }} value than the successor?	
		• How taller is the tallest person in the room than the second tallest person?
		• How much more developed is the most developed city than the second most developed city?

Abnormal Distributions

Pattern	Question Template	Examples
Abnormal Distribution	What is the expected distribution of a {{ value }}? Does the data match that distribution?	• In *Sholay*, is Amitabh Bachchan's coin toss really random?
		• Are dates of birth uniformly distributed across the calendar?
		• Do 9 of 10 startups fail?

Insights must be **BUS**: Big Useful and Surprising

IS THE INSIGHT
BIG

The analysis must, of course, be statistically significant.
But it should also be **numerically significant**.
We want a result that substantially changes the outcome.

IS THE INSIGHT
USEFUL

What should the audience do after hearing the insight?
Can they take an **action** that improves their objective?
Even if it’s informational, what should they do next?

IS THE INSIGHT
SURPRISING

Is this something they didn’t know? Is it non-obvious?
Does it overturn a domain-driven belief or a gut feel?
Or does it bring consensus to a group with divided opinion?

Marking each analysis as BUS (High, Medium, Low)

Insights	Big	Useful	Surprising
Project managers get paid 5X as much as data analysts	High	Medium	Low
Business analysts get paid twice as much as data analysts	High	High	High
Office supplies sell the least in the South. Sales in South are only 50% of sales in East or West.	Medium	High	Low
A startup in Bengaluru has a 6X more chance to get private equity funding than any other place.	High	High	Low
About 50% of American small businesses do not have a website	High	Medium	Low
The recommendation system influences about 80% of content streamed on Netflix	Big	Low	Low

Summary

1. Analysis is NOT insights.
2. Five Patterns of Insights:
   a. Unknown Results
   b. Surprising Comparisons
   c. Surprising Extremes
   d. Significant Outliers
   e. Abnormal Distributions
3. Insights have to be **BUS** (Big, Useful, Significant)

End of Part 8