WELLCOME

STATISTIC WITH PYTHON

EMBARKING ON A JOURNEY INTO DATA SCIENCE

YA MANON

WHY STATISTICS?

WHY STATISTICS FOR BUSINESS INTELLIGENCE?

In this section we'll discuss the **role of statistics** in the context of **business intelligence** and the decision-making process, review key terms, and introduce the statistics workflow

TOPICS WE'LL COVER:

Why Statistics

Populations

Statistics Workflow

GOALS FOR THIS SECTION:

- *Identify scenarios when statistics helps use data to make smart decisions, and when it's not needed*
- *Understand the concepts of populations & samples*
- *Review the statistics workflow and the concepts that will be covered throughout the course*

WHY STATISTICS?

Business intelligence is about using data to make **smart decisions**

Statistics is about *evaluating* those decisions under *uncertain* circumstances

When do you need statistics?

1) You don't have all the data you're interested in

- You can only analyze some of the data you need to make your decision
- There's **uncertainty** involved

2) The decision you're making is important

- You don't want to make the wrong one based on your limited data
- There's something specific to evaluate

What difference between BI and Data Science?

POPULATION & SAMPLES

Why Statistics

Populations

Statistics Workflow

A **population** contains all the data you're interested in to make your decision

- It's the data you wish you had, but are unlikely to get
- Any figure that summarizes a population is called a **parameter**

A **sample** contains some of the data from the population

- It's the data you have (which should ideally represent the population)
- Any figure that summarizes a sample is called a **statistic**

Statistics lets you make reasonable estimates about **parameters** using **statistics**

HEY THIS IS IMPORTANT!

Statistics can't create certainty out of uncertainty, it just helps you make controlled decisions under it!

THE STATISTICS WORKFLOW

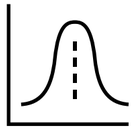
Why Statistics

Populations

Statistics Workflow

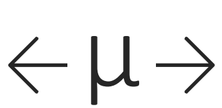
Descriptive Statistics

Understand what your sample data looks like



Probability Distribution

If the sample data fits a probability distribution, use it as a **model** for the entire population



Confidence Intervals

If the sample doesn't fit a distribution, use the central limit theorem to make **estimates** about population parameters

Hypothesis Testing

Continue to leverage the central limit theorem to draw **conclusions** about what a population looks like based on a sample

Regression Analysis

Use additional variables to increase the accuracy of your estimates and make **predictions** based on their relationships

HEY THIS IS IMPORTANT!

If you have all the population data, or simply need a bit of inspiration to make an "unimportant" decision, then descriptive statistics may be all you need!

DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS

In this section we'll cover understanding data with **descriptive statistics**, including frequency distributions, measures of central tendency, and measures of variability

TOPICS WE'LL COVER:

- **Statistics Basics**
- **Distributions**
- **Central Tendency**
- **Variability**

GOALS FOR THIS SECTION:

- *Identify the different types of variables in a dataset, along with their use cases*
- *Create frequency tables and plot the distributions of numerical variables using histograms*
- *Calculate the mean, median, mode, and standard deviation of a numerical variable*
- *Visualize the key descriptive statistics of a numerical variable using a box plot*

DESCRIPTIVE STATISTICS

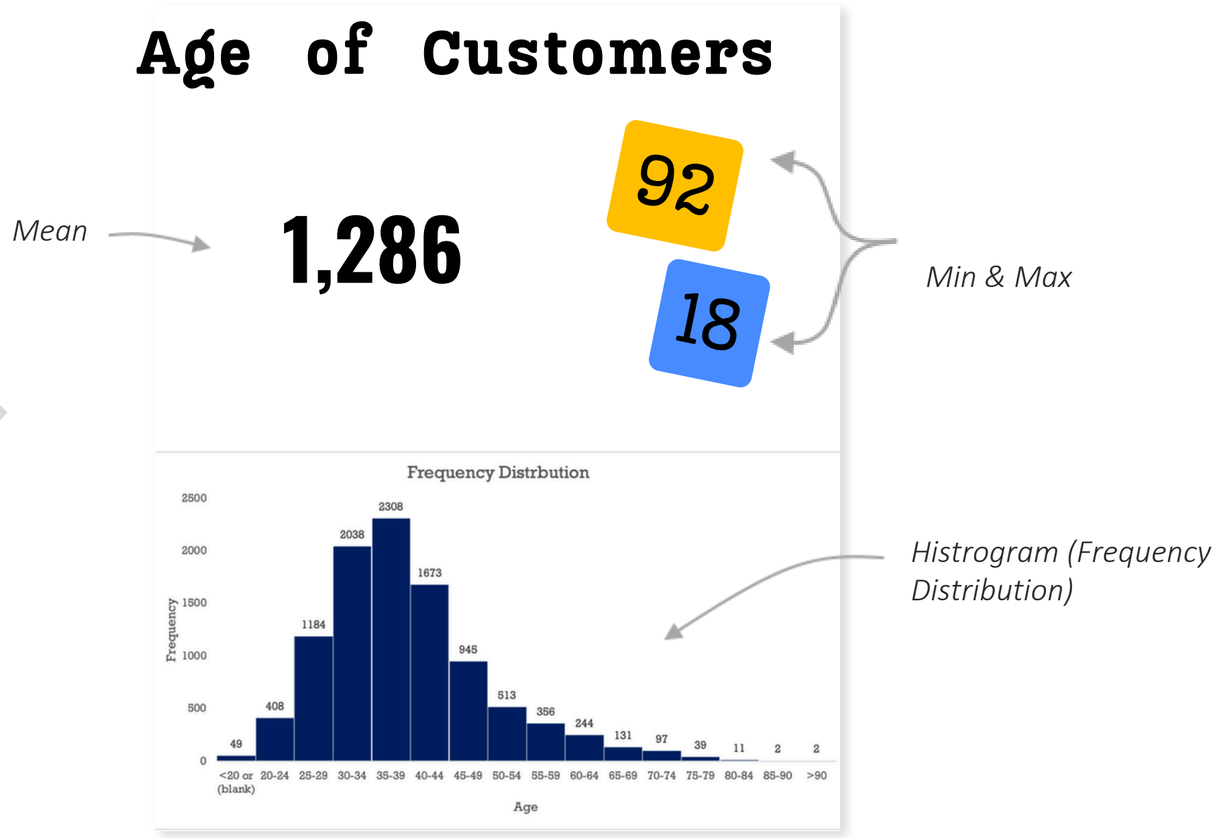
Descriptive statistics consists of the collection, organization, summarization and presentation of data.

They reduce a large array of numbers into a handful of figures that describe it accurately

- Statistics Basics
- Distributions
- Central Tendency
- Variability

customer_id	age
15634602	42
15647311	41
15619304	42
15701354	39
15737888	43
15574012	44
15592531	50
15656148	29
15792365	44
15592389	27
15767821	31
15737173	24
15632264	34
15691483	25
15600882	35
15643966	45
15737452	58
15788218	24
15661507	45
15568982	24

n=10000



TYPES OF VARIABLES

There are two main **types of variables** in a dataset: Numerical & Categorical

- **Numerical** or **Quantitative variables**
- **Categorical** or **Qualitative variables**

NUMERICAL:

customer_id	age	tenure	balance	products_nur	credit_card	active_meml	estimated_sa	churn
15634602	42	2	0	1	1	1	101348.88	1
15647311	41	1	83807.86	1	0	1	112542.58	0
15619304	42	8	159660.8	3	1	0	113931.57	1
15701354	39	1	0	2	0	0	93826.63	0
15737888	43	2	125510.82	1	1	1	79084.1	0
15574012	44	8	113755.78	2	1	0	149756.71	1
15592531	50	7	0	2	1	1	10062.8	0
15656148	29	4	115046.74	4	1	0	119346.88	1
15792365	44	4	142051.07	2	0	1	74940.5	0
15592389	27	2	134603.88	1	1	1	71725.73	0
15767821	31	6	102016.72	2	0	0	80181.12	0
15737173	24	3	0	2	1	0	76390.01	0
15632264	34	10	0	2	1	0	26260.98	0
15691483	25	5	0	2	0	0	190857.79	0
15600882	35	7	0	2	1	1	65951.65	0
15643966	45	3	143129.41	2	0	1	64327.26	0
15737452	58	1	132602.88	1	1	0	5097.67	1
15788218	24	9	0	2	1	1	14406.41	0
15661507	45	6	0	1	0	0	158684.81	0
15568982	24	6	0	2	1	1	54724.03	0
15577657	41	8	0	2	1	1	170886.17	0
15597945	32	8	0	2	1	0	138555.46	0
15699309	38	4	0	1	1	0	118913.53	1
15723737	46	3	0	2	0	1	8487.75	0

Possible question:

CATEGORICAL:

country	gender
France	Female
Spain	Female
France	Female
France	Female
Spain	Female
Spain	Male
France	Male
Germany	Female
France	Male
France	Male
France	Male
Spain	Male
France	Female
France	Female
Spain	Female

TYPES OF DESCRIPTIVE STATISTICS

There are 3 main **types of descriptive statistics** that can be applied to a variable:

- **Statistics Basics**
- Distributions
- Central Tendency
- Variability

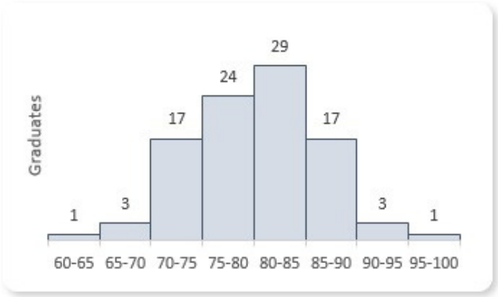
Distribution

*Represents the **frequency** of each value*

Examples:

- Frequency Tables
- Histograms

Grade Distribution



Central Tendency

*Represents the **middle** of the values*

Examples:

- Mean, Median, and Mode
- Skew

Class Average

80.17

Variability

*Represents the **dispersion** of the values*

Examples:

- Min, Max, and Range
- Quartiles & Interquartile Range
- Box & Whisker Plots
- Variance & Standard Deviation

Highest Grade

96.1

Lowest Grade

62.6

HEY THIS IS IMPORTANT!

Most measures of central tendency and variability can only be applied to numerical variables

FREQUENCY DISTRIBUTIONS

Statistics Basics

Distributions

Central Tendency

Variability

A **frequency distribution** counts the observations of each possible value in a variable

They are commonly depicted using frequency tables

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

n=17

FREQUENCY TABLE:

Undergrad Degree	Frequency	Relative Frequency
Art	1	6%
Business	8	47%
Computer Science	2	12%
Engineering	4	24%
Finance	2	12%

The relative frequency shows the count of each value as a % of the total

TIP: Use a PivotTable or the COUNTIFS() function to calculate frequencies for categorical variables in Excel

FREQUENCY DISTRIBUTIONS

For numerical variables, a frequency distribution typically counts the number of observations that fall into defined ranges or "bins" *(1-5, 6-10, etc.)*

They are commonly depicted using grouped frequency tables or histograms

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

FREQUENCY TABLE:

Undergrad Grade	Frequency
63.3	1
66	1
67.5	1
67.7	1
68.1	1
68.7	1
70.1	1
74	1
74.6	1
75.3	1
75.6	1
76	1
78.9	1
79.3	1
82.9	1
88.8	1
93.6	1

This isn't a meaningful representation of the distribution of the data

FREQUENCY DISTRIBUTIONS

For numerical variables, a frequency distribution typically counts the number of observations that fall into defined ranges or "bins" *(1-5, 6-10, etc.)*

- They are commonly depicted using grouped frequency tables or histograms

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

GROUPED FREQUENCY TABLE:

Undergrad Grade	Frequency	Cumulative Relative Frequency
60-65	1	6%
65-70	5	35%
70-75	3	53%
75-80	5	82%
80-85	1	88%
85-90	1	94%
90-95	1	100%
Grand Total	**17**	

The cumulative relative frequency shows the running total of the relative frequencies

TIP: Group the numerical values in a PivotTable or use the FREQUENCY() function with the upper limits to calculate frequencies for each bin in Excel

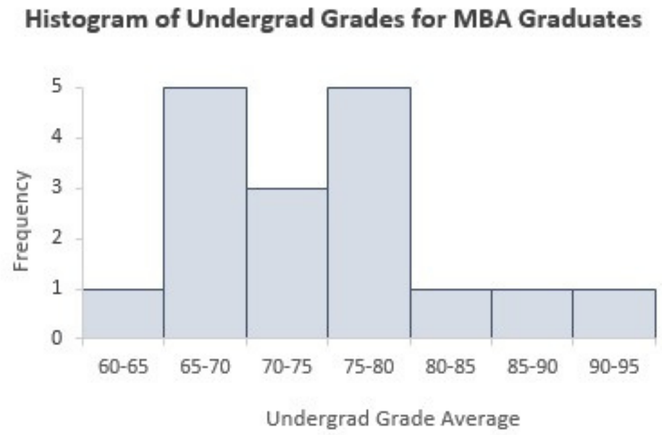
HISTOGRAMS

Histograms are used to visualize the distribution of a numerical variable

They also provide a glimpse of the variable's central tendency and variability

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

n=17



TIP: Create a histogram by using a column chart to plot the variable's frequency table, instead of using Excel's native histogram chart type (not as customizable)

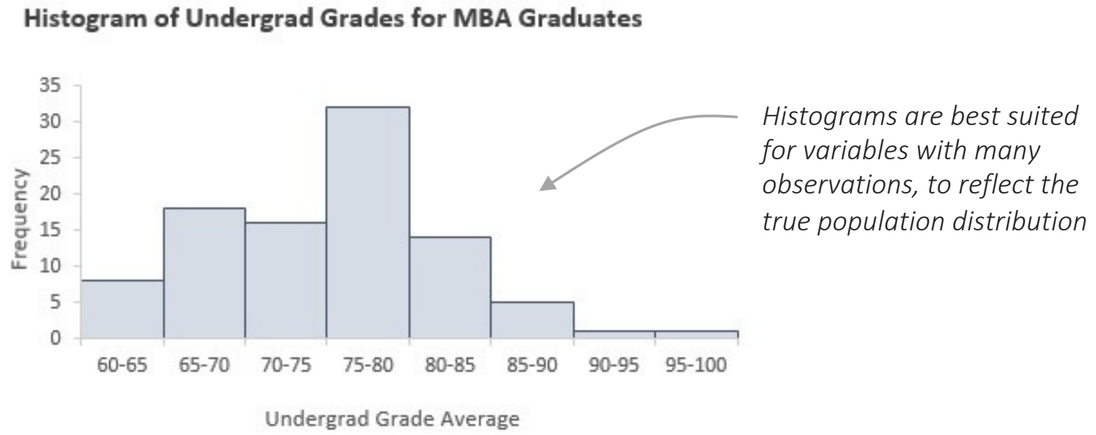
HISTOGRAMS

Histograms are used to visualize the distribution of a numerical variable

- They also provide a glimpse of the variable's central tendency and variability

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

n=95



TIP: Bin size can significantly change the shape and "smoothness" of a histogram, so select a bin width that accurately shows the data distribution

MEAN

Statistics Basics

Distributions

Central Tendency

Variability

The **mean** is the calculated “average” value in a set on numbers

- It is calculated by dividing the sum of all values by the count of all observations
- It can only be applied to numerical variables *(not categorical)*

customer_id	credit_score	age
15634602	619	42
15647311	608	41
15619304	502	42
15701354	699	39
15737888	850	43
15574012	645	44
15592531	822	50
15656148	376	29
15792365	501	44
15592389	684	27
15767821	528	31
15737173	497	24
15632264	476	34
15691483	549	25
15600882	635	35
15643966	616	45
15737452	653	58
15788218	549	24
15661507	587	45
15568982	726	24
15577657	732	41
15597945	636	32
15699309	510	38
15725737	669	46
15625047	846	38
15738191	577	25
15736816	756	36
15700772	571	44
15728693	574	43

Mean = 37.92

*(average **37.9***

TIP: Use the **AVERAGEIFS**() function if you want to calculate the mean for values that meet a specified criteria *(i.e., Mean by Undergrad Degree)*

LIMITATIONS OF THE MEAN

The main **limitation of the mean** is that it is sensitive to outliers *(extreme values)*

"The average income in America is not the income of the average American"

$35,000 $50,000 $65,000

*mean = **$50,000***

$100,000,000

*mean = **$25,037,500***

HEY THIS IS IMPORTANT!

While the mean is typically great for making a "best-guess" estimate of a value, it's important to complement this value with other descriptive statistics like the distribution, median, and mode to see if the mean value is being distorted by outliers

MEDIAN

The **median** is the "middle value" in a sorted set of numbers

- Unlike the mean, the median is NOT sensitive to outliers
- When there are two middle-ranked values, the median is the average of the two

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

Undergrad Grade
63.3
66
67.5
67.7
68.1
68.7
70.1
74
74.6
75.3
75.6
76
78.9
79.3
82.9
88.8
93.6

n=17

Median = **74.6**

MEDIAN

The **median** is the "middle value" in a sorted set of numbers

- Unlike the mean, the median is NOT sensitive to outliers
- When there are two middle-ranked values, the median is the average of the two

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

Undergrad Grade
63.3
66
67.5
67.7
68.1
68.7
70.1
74
74.6
75.3
75.6
76
78.9
79.3
82.9
88.8

n=16

Median = **74.3**

*(average of **74** and **74.6**)*

MODE

The **mode** is the “most frequent” value in a variable

- It can be applied to both numerical and categorical variables

Mode = “**Business**”

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

Mode = **N/A**

MODE

The **modal class** is the group with the highest frequency

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	93.6
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Computer Science	67.7
Engineering	75.3
Engineering	68.1
Finance	63.3

GROUPED FREQUENCY TABLE:

Undergrad Grade	Frequency
60-65	1
65-70	5
70-75	3
75-80	5
80-85	1
85-90	1
90-95	1
Grand Total	**17**

Mode = **65-70**, **75-80**

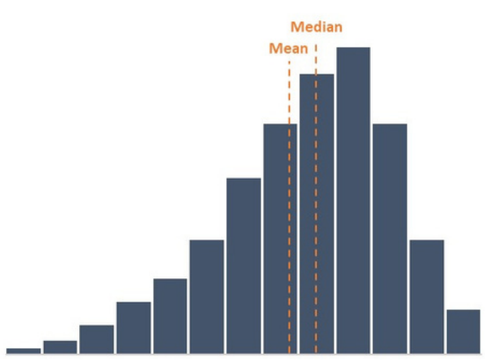
*This is a **multi-modal** distribution, which indicates that there may be another variable impacting the undergrad grades*

SKEW

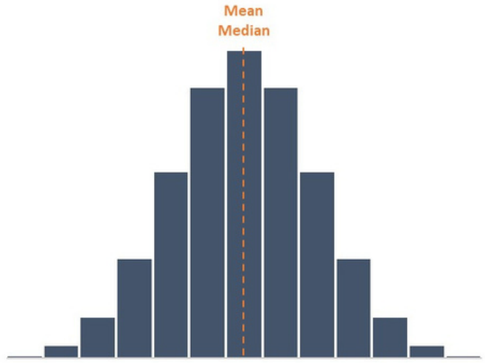
The **skew** represents the asymmetry of a distribution around its mean

- In a **zero-skewed** distribution, the mean and median are equal
- In a **right-skewed** *(or positive)* distribution, the mean is typically greater than the
- median In a **left-skewed** *(or negative)* distribution, the mean is typically smaller than the median

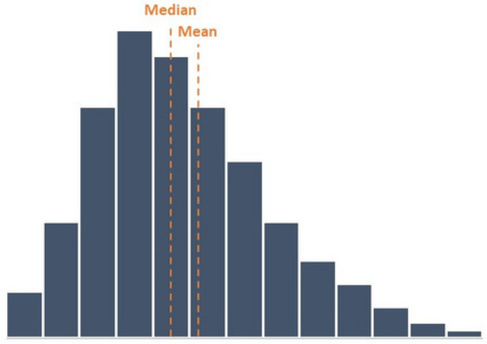
Left skew



Zero skew



Right skew



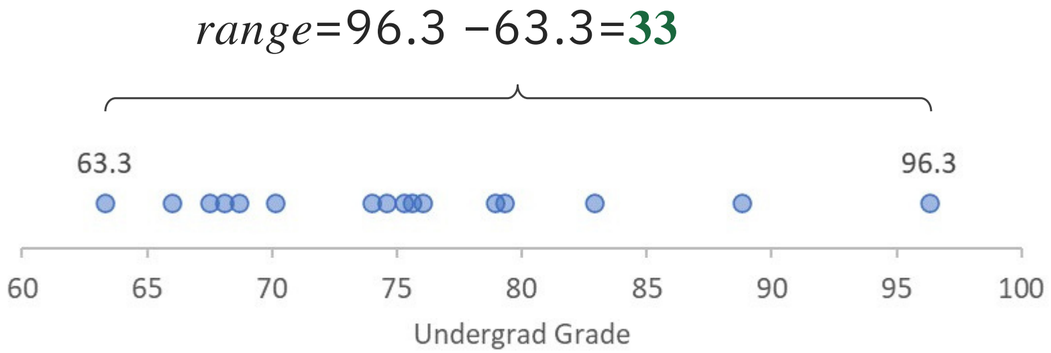
*This is one of the properties of a **normal distribution** (more on that later!)*

RANGE

The **range** is the spread from the lowest *(min)* to the highest *(max)* value in a variable

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=16



HEY THIS IS IMPORTANT!

While the range is generally a good indicator of the variability in a numerical variable, a single outlier can cause it to change significantly

INTERQUARTILE RANGE

The **interquartile range** is the spread of the *middle half* of the values in a variable

- In other words, it's the spread from the **first quartile** to the **third quartile**

Undergrad Grade	
	63.3
25%	66
	67.5
	68.1
	68.7
25%	70.1
	74
	74.6
	75.3
25%	75.6
	76
	78.9
	79.3
25%	82.9
	88.8
	96.3

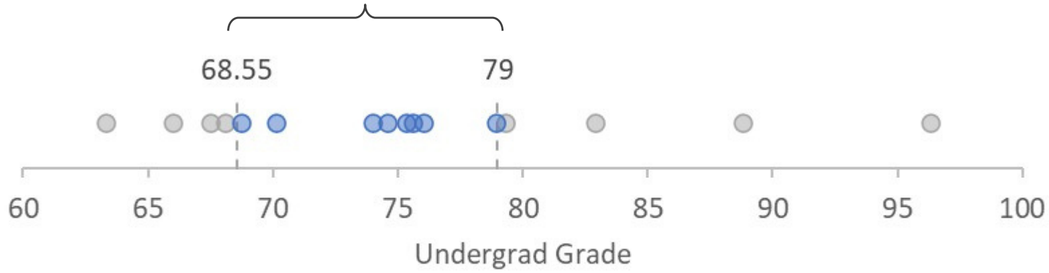
Q1 = 68.55

Median

Q3 = 79

n=16

$$IQR = 79 - 68.55 = \mathbf{10.45}$$



BOX & WHISKER PLOTS

Box & whisker plots are used to visualize key descriptive statistics

Statistics Basics

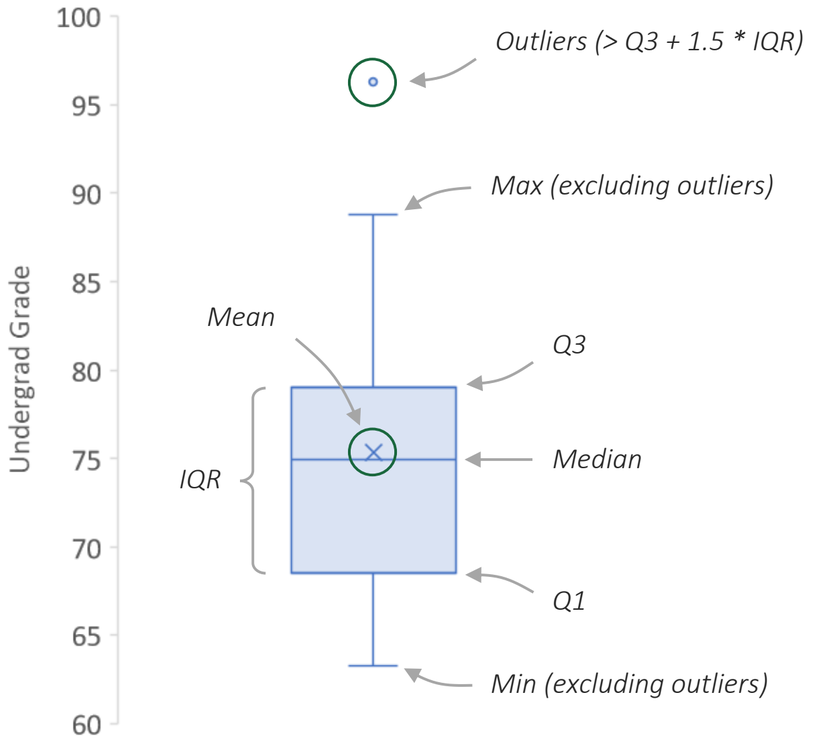
Distributions

Central Tendency

Variability

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=16



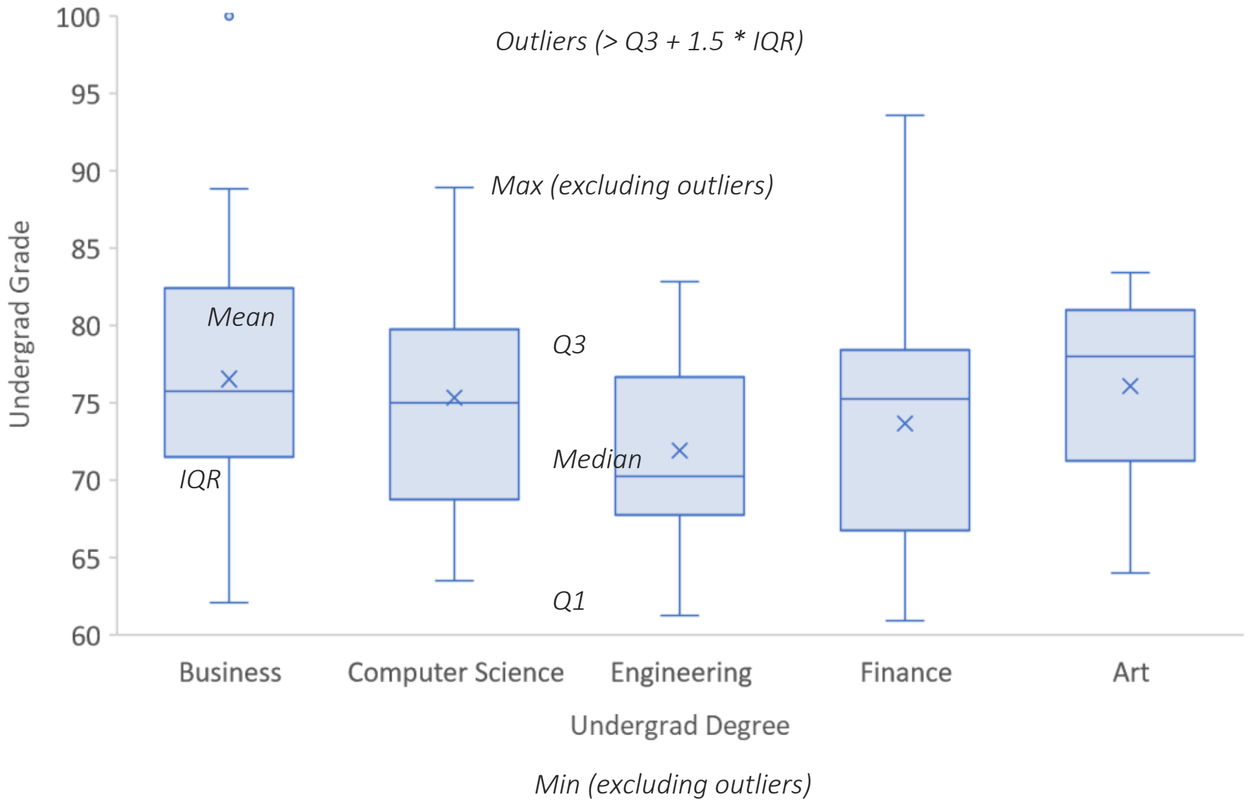
BOX & WHISKER PLOTS

Box & whisker plots are used to visualize key descriptive statistics

- They can be used to quickly compare statistical characteristics between categories

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=1965



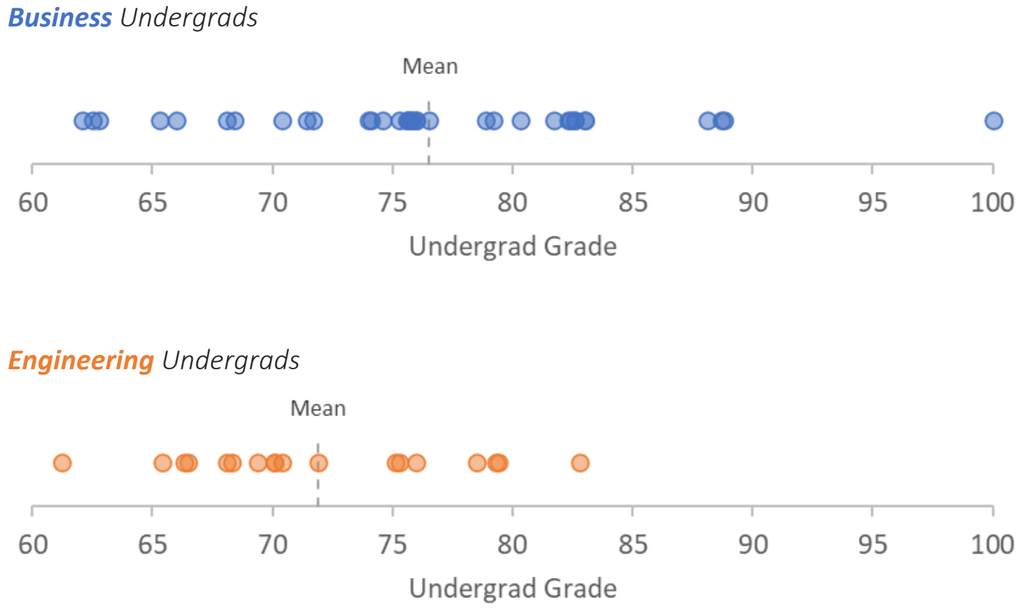
STANDARD DEVIATION

The **standard deviation** measures, on average, how far each value lies from the mean

The *higher* the standard deviation, the *wider* a distribution is *(and vice versa)*

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=95



VARIANCE

The **variance** is the square of the standard deviation

Statistics Basics

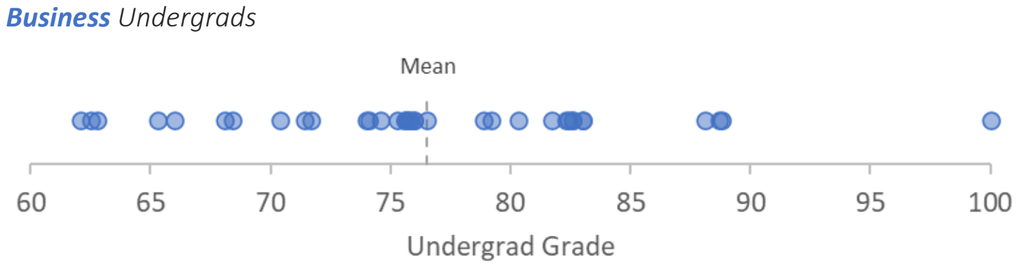
Distributions

Central Tendency

Variability

Undergrad Degree	Undergrad Grade
Business	78.9
Business	74
Business	74.6
Engineering	79.3
Engineering	70.1
Business	88.8
Business	66
Art	82.9
Business	96.3
Business	75.6
Finance	67.5
Computer Science	68.7
Business	76
Engineering	75.3
Engineering	68.1
Finance	63.3

n=95



HEY THIS IS IMPORTANT!

The variance does have its place in some statistical tests, so it shouldn't be discarded, but as a single numerical measure of a variable's dispersion the standard deviation is more effective

PRO TIP: COEFFICIENT OF VARIATION

The **coefficient of variation** measures the standard deviation relative to the mean

- It is used to compare the standard deviations of variables with significantly different means

Undergrad Grade	Employability (Before)
74	133
74.6	122
79.3	236
70.1	143
88.8	354
66	214
82.9	225
96.3	261
75.6	277
67.5	282
68.7	322
76	326
67.7	421
75.3	368
68.1	279
62.3	268

n=95

$$CV = \frac{standard\ deviation}{mean}$$

	Undergrad Grade	Employability (Before)
Standard Deviation:	7.42	85.94
Mean:	74.9	239.9
Coefficient of Variation:	**0.099**	**0.358**

*On average, undergrad grades differ from the mean by **~10%** of its value, while employability scores differ by **~ 35%***

KEY TAKEAWAYS: DESCRIPTIVE STATISTICS

- There are two main types of variables: **numerical** & **categorical**
  - *Numerical variables are meant to be aggregated, and categorical variables are used to create groups*
- The **distribution** represents the “shape” of a variable
  - *Histograms are a great way to visualize this “shape” by plotting the frequency of each value (or class)*
- The **mean** & **median** locate the “center” of a distribution
  - *Don’t focus on using one instead of the other, rather on using both to complement each other*
- The **standard deviation** measures the dispersion around the mean
  - *Use a box plot alongside the standard deviation to provide additional context on the variability and center*