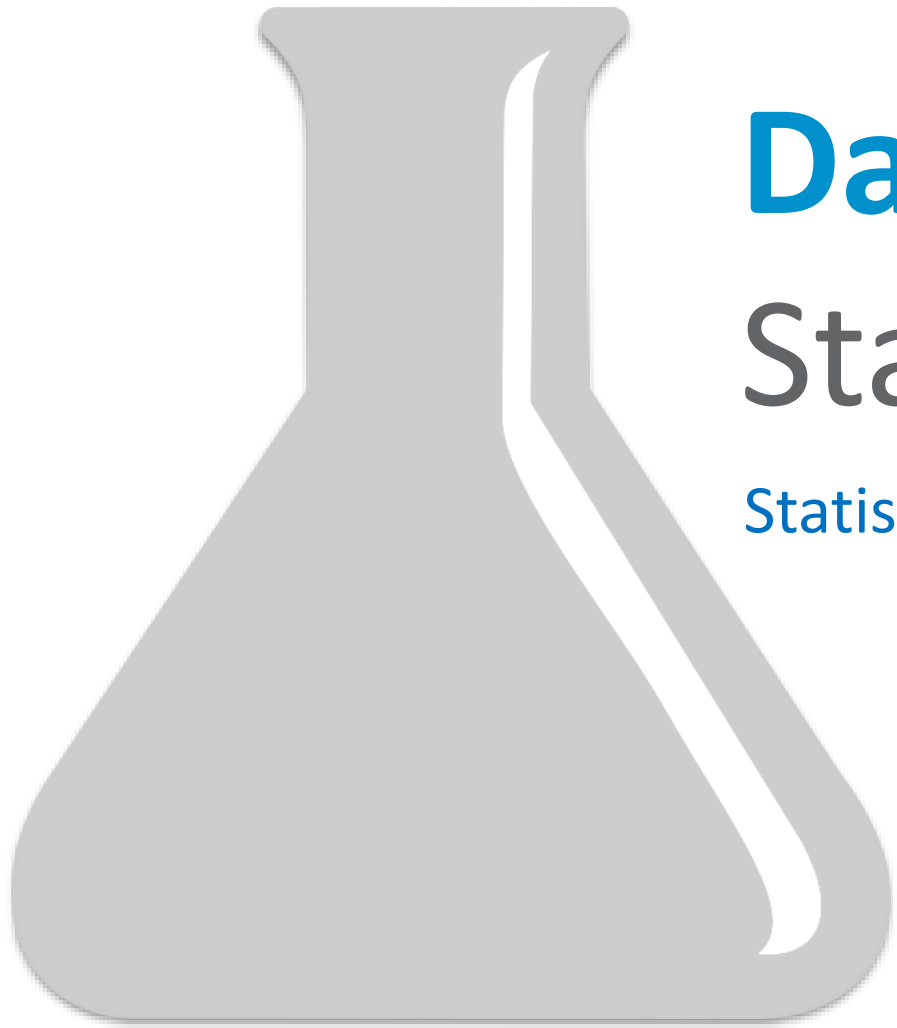




MOVING THE FUTURE FORWARD

BEOP.CTO.TP4
Owner: OCTO
Revision: 0001
Approved by: JAT
Effective: 01/30/2018

Buchanan & Edwards Proprietary:
Printed copies of this document are
UNCONTROLLED. Verify that this is
the correct version before use.



Data Science *On-Ramp*

Statistical Data Analysis

Statistical Methods that Drive AI & Machine Learning

Jon Tupitza

Practice Director, Data Platform & Predictive Analytics

Take-Aways

- Define “Data Science”
- Understand the basics of statistical data analysis
- Identify some data science tools



Agenda

- Intro to Data Science
- Statistical Data Analysis
 - Fundamental Statistics
 - Parametric Analysis
 - Non-Parametric Analysis
 - Categorical Analysis



What is Data Science?

Data Science is the exploration and quantitative analysis of all available structured and unstructured data to develop understanding, extract knowledge, and formulate actionable results.

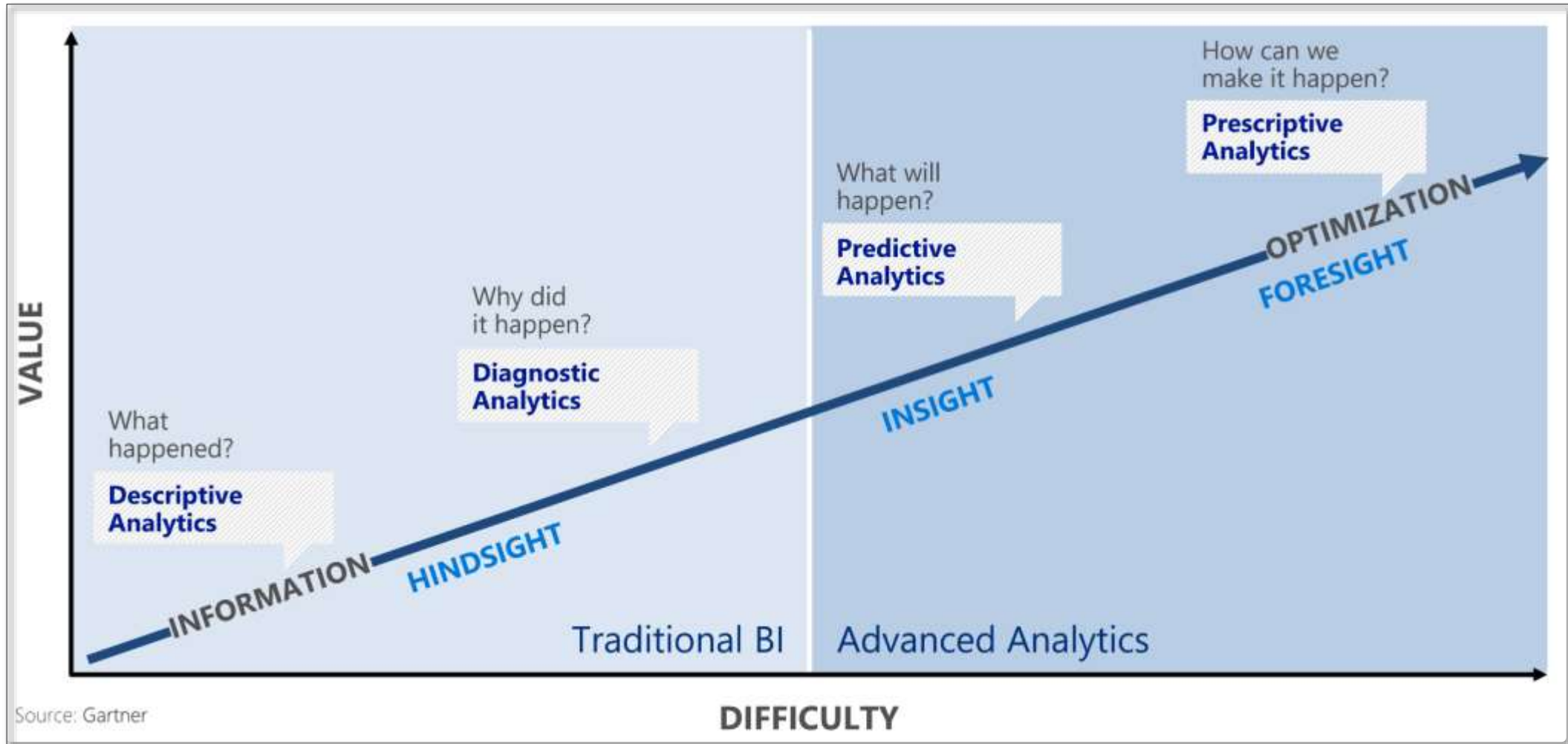
Transform raw data into a valuable asset

Replace intuition with data-driven analytical decisions

Use Data to Make Decisions that Drive Actions

Doesn't This Sound Familiar? How does it differ from Business Intelligence?

Advanced Analytics: *Understand & Control the Future!*

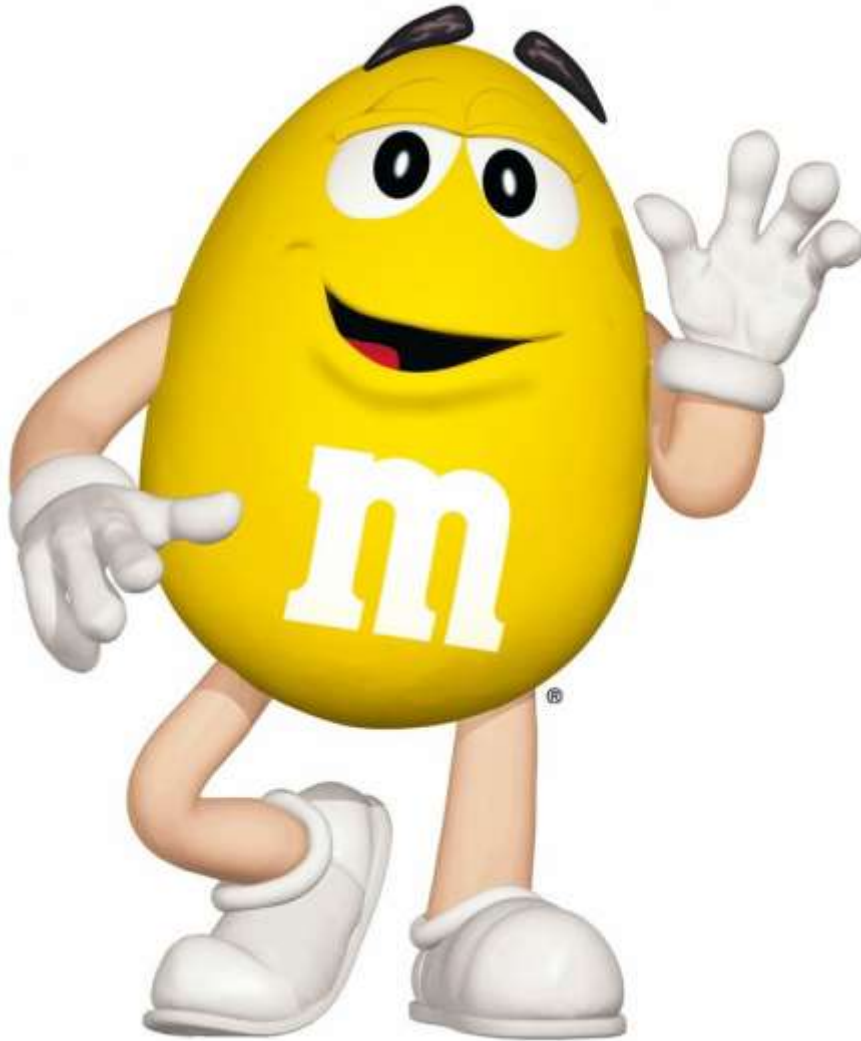


Agenda

- Intro to Data Science
- Statistical Data Analysis
 - Fundamental Statistics
 - Parametric Analysis
 - Non-Parametric Analysis
 - Categorical Analysis



Descriptive Statistics: Qualitative Data



Just **One** M&M...

- How Tall ?
- How Wide ?
- How Heavy ?
- What Color ?
- What Type ?

...Describe Each **Observation**

Descriptive Statistics: Quantitative Data – Populations

Populations have Parameters



All of the M&M's...

- How Many ?
- How Tall ?
- How Wide ?
- How Heavy ?
- What Color ?
- What Type ?

*It's often **impractical** or **impossible** to obtain data about an entire **population***

Descriptive Statistics: Quantitative Data – Samples

Samples have Statistics



Some of the M&M's...

- How Many ?
- How Tall ?
- How Wide ?
- How Heavy ?
- What Color ?
- What Type ?

...Therefore, **outcomes** must often be **inferred** from **samples**

Descriptive Statistics: Quantitative Data – Measures

Measures of Center:

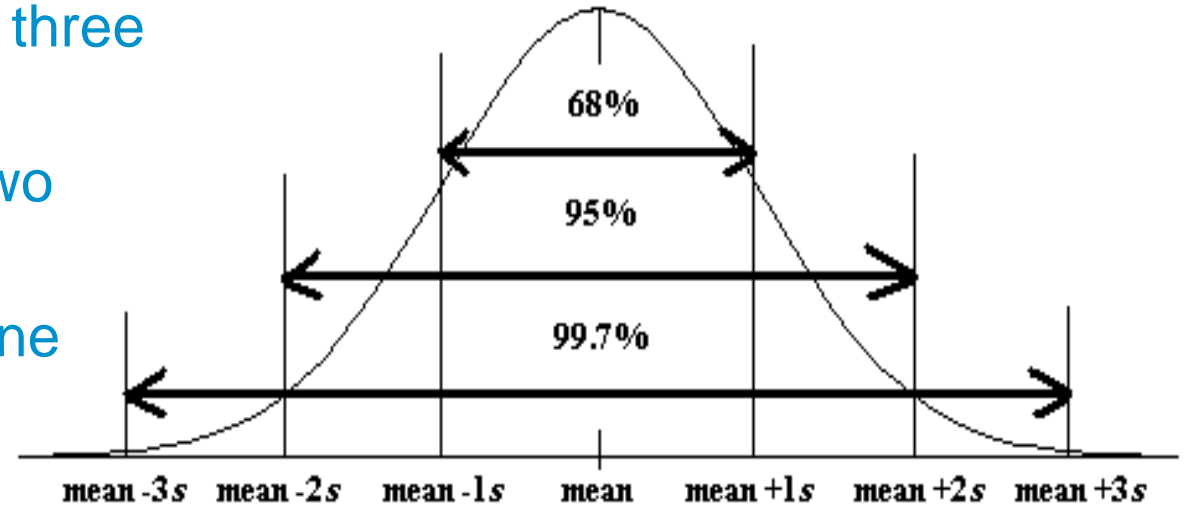
- **Mean**: helps summarize large data sets by determining their centrality
- **Median**: better representation of center for data sets containing outliers
- **Mode**: highlights the most frequently occurring value in a collection

Measures of Spread:

- **Minimum**: smallest value in a data set
- **Maximum**: largest value in a data set
- **Range**: difference between the largest and smallest ($\text{Range} = \text{Max} - \text{Min}$)
- **Mean Absolute Deviation**: The average **absolute** distance from the mean
- **Variance** (σ^2): squares the distance rather than using absolute value
- **Standard Deviation** (σ): square root of the variance ($\sqrt{\sigma^2}$)

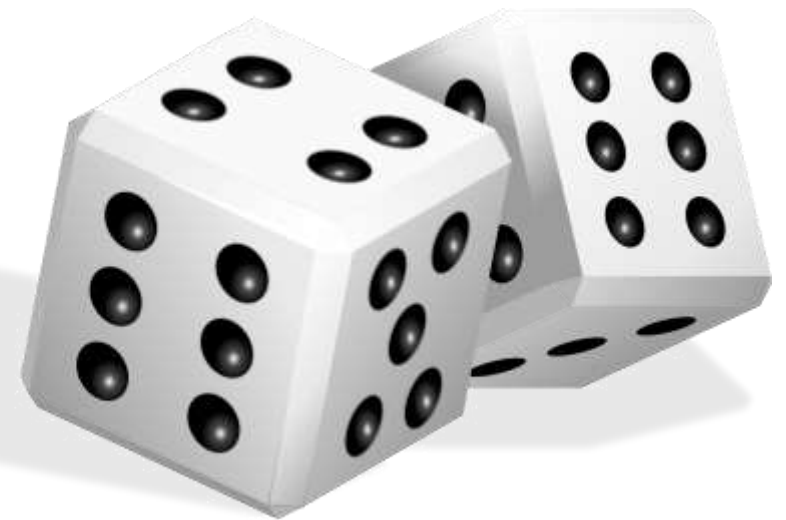
Descriptive Statistics: the Empirical Rule and Z-Scores

- **Normal** (Unimodal and Symmetric) distributions
 - Follow a definite pattern where mean, median and mode are equal
 - Useful for obtaining probabilities and interpreting outcomes
- **Empirical Rule:** states that for **any** unimodal and symmetric distribution
 - 99.7% of the observations fall within **three** standard deviations of the **mean**
 - 95% of the observations fall within **two** standard deviations of the **mean**
 - 68% of the observations fall within **one** standard deviation of the **mean**
- **Z-Score** ($x - \mu / \sigma$): a signed number
 - Reflects the **number** of standard deviations a value is **above** or **below** the **mean**
 - Useful for determining whether a data point is an outlier: **Higher Score = Unusual**



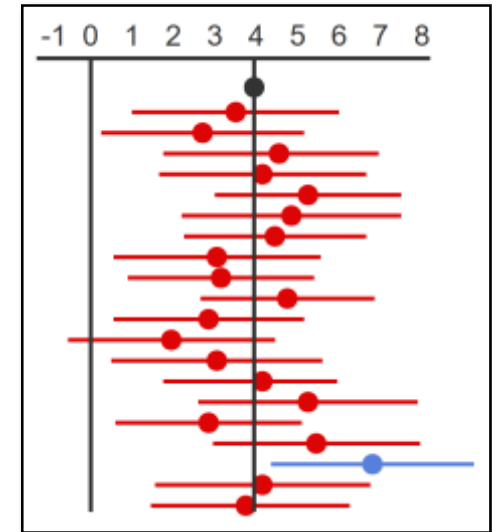
Probability: **Experimentation** *is what makes it Scientific*

- **Experiment:** A test that results in one of many possible outcomes
 - Roll a single die | Roll a pair of dice | Flip a coin | Flip a coin twice
- **Outcome:** The results of an experiment
 - Roll a 3 | Roll snake-eyes | Tails | Heads - Heads
- **Sample Space:** A collection encompassing all possible outcomes
 - 1, 2, 3, 4, 5, **or** 6 | 2, 3... 12 | Heads **or** Tails | HH, HT, TH, TT
- **Event:** A subset of the sample space
 - **Simple Event:** A subset consisting of a single outcome; e.g., roll a die and get a 5
 - **Compound Event:** A subset consisting of more than one outcome; e.g., roll an even number



Probability: Confidence Intervals

- Confidence Intervals provide a method that complements hypothesis testing
- Since the mean and standard deviation of the population are rarely known...
- A sample **statistic** is used to estimate a population **parameter**; e.g., mean
 - Each sample statistic represents only one possible outcome; i.e., many samples could potentially be drawn from a population with each one having its own value
 - Creates uncertainty regarding how well the **sample statistic** represents the corresponding **population parameter**; i.e., how accurate is an inference
- **Confidence Interval**: a range of values...
 - That describes the uncertainty surrounding an estimate
 - That expresses the accuracy of a given estimate
 - Wide intervals increase confidence the estimate lies in the interval
 - Believed to include a population parameter at a stated level of confidence



Inferential Statistics: Hypothesis Testing

Null Hypothesis (H_0)

Asserts there is no correlation between two measured phenomena; i.e., chance alone is responsible for the results.

Alternate Hypothesis (H_a)

Occurs when sufficient evidence exists that causes us to **fail to confirm** the null hypothesis.

Question: at what **threshold** are statistics derived from **samples** extreme enough to **infer** a conclusion regarding the **population**?

Inferential Statistics: **Statistical Significance**

If less than **significance level** (alpha), the null hypothesis is rejected.

Insignificant

If the samples are too small relative to the size of the population, or if the differences in outcomes are too small, then a conclusion cannot be inferred.

Significant

If the samples are large enough relative to the size of the population, and/or if the differences in outcomes are large enough, then a conclusion may be inferred.

P-Value: determines the probability of obtaining data at least as extreme as the data already observed; assuming the null hypothesis is true

Agenda

- Intro to Data Science
- Statistical Data Analysis
 - Fundamental Statistics
 - Parametric Analysis
 - Non-Parametric Analysis
 - Categorical Analysis



Parametric Analysis: Tests for Normally Distributed Data

Student's T-Test

- Determines if the hypothesized mean is equal to the true population mean

Two-Sample T-Test (Paired or Dependent)

- Determines if a statistically significant difference exists between two samples exposed to two different treatments

Two-Sample T-Test (Unpaired or Independent)

- Determines if a statistically significant difference exists between two samples exposed to the same treatment

Analysis of Variance (ANOVA):

- Determines if a statistically significant difference exists between the means of three or more populations

Agenda

- Data Science
- Statistical Data Analysis
 - Fundamental Statistics
 - Parametric Analysis
 - Non-Parametric Analysis
 - Categorical Analysis



Non-Parametric Data: **Non-Normal Distributions**

- **Non-Normal** (**Asymmetric** and/or **Multimodal**) distributions
 - Where mean, median and mode are **not** equal
 - Where distribution is [left or right] skewed; i.e., not evenly distributed
 - Where population or sample is too small; e.g., fewer than 30 observations
- **Resampling**:
 - Can be used to alter the data; making it more normal
 - Often aimed at enabling the analysis of rare events
 - Enables the use of parametric methods for data analysis
 - **Randomization** (permutation): Selects only once; without replacement
 - **Bootstrapping** (combination): Selects more than once: with replacement

Non-Parametric Analysis: Tests for Non-Normal Data

- Requires fewer assumptions regarding the distribution of the data
- Based on ranking observations by their frequency (count)

Wilcoxon Rank Sum

- Analogous to an independent two-sample t-test
- Ignores the values of the original data and compares the sum of the two groups' ranks.

Kruskal-Wallis

- Analogous to a one-way analysis of variance (ANOVA) test
- Extends the Wilcoxon Rank Sum test to three or more groups of data

Non-Parametric Analysis: Wilcoxon Rank Sum Test


- Wilcoxon Rank Sum Test with Two Groups

Placebo	Painkiller
60	50
40	40
20	60
20	70
10	60

Determines if two independent samples come from the same population.

Placebo	1	2.5	2.5	4.5	8	18.5
Painkiller	4.5	6	8	8	10	36.5

$$W = 18.5$$



Value	10	20	20	40	40	50	60	60	60	70
Group	1	1	1	1	2	2	1	2	2	2
Rank	1	2.5	2.5	4.5	4.5	6	8	8	8	10

Non-Parametric Analysis: **Kruskal-Wallis Rank Sum Test**

- Kruskal-Wallis Test with Three Groups

Placebo	Painkiller	Ibuprofen
60	50	70
40	40	50
20	60	60
20	70	50
10	60	80

Determines the statistical difference between two or more groups of an independent variable

Value	10	20	20	40	40	50	50	50	60	60	60	60	70	70	80
Group	1	1	1	1	2	2	3	3	1	2	2	3	2	3	3
Rank	1	2.5	2.5	4.5	4.5	7	7	7	10.5	10.5	10.5	10.5	13.5	13.5	15

Agenda

- Data Science
- Statistical Data Analysis
 - Fundamental Statistics
 - Parametric Analysis
 - Non-Parametric Analysis
 - Categorical Analysis



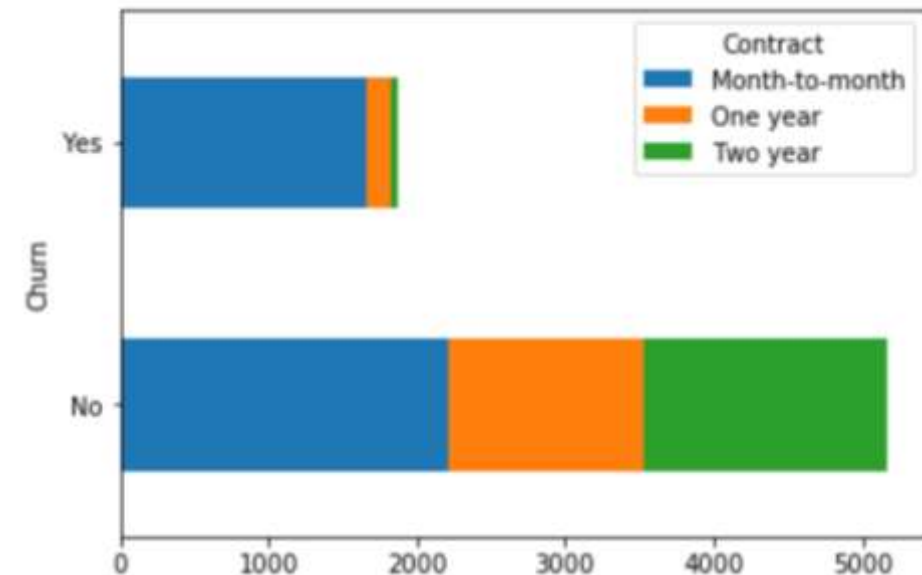
Categorical Analysis: Qualitative Data

- **Non-numeric:** Takes on the value of one of several categories
 - **Ordinal:** Having a natural order like small, medium and large
 - **Nominal:** Having no natural order like apples, oranges and bananas
- Analysis involves counting the instances of each category

Contingency Table

Contract	Month-to-month	One year	Two year
Churn			
No	2220	1307	1647
Yes	1655	166	48

Stacked Bar Chart



Questions

