Crash Course:
Data Analysis & Presentation

Week 2

Jordan Krell
jkrell@flintsciencefair.org

Tools

Elevate your analysis

Advanced Tools

- Petri dish samples
  - Color
  - # of Colonies
  - Size of Colonies
  - Growth Rate

- ImageJ / Fiji
  - Scientific image analysis
  - Fiji = imageJ + GUI and plugins
  - https://fiji.sc/
  - https://imagej.net/Welcome

Excel

Initial analysis + making graphs

Excel in Action

- Install the Analysis ToolPak

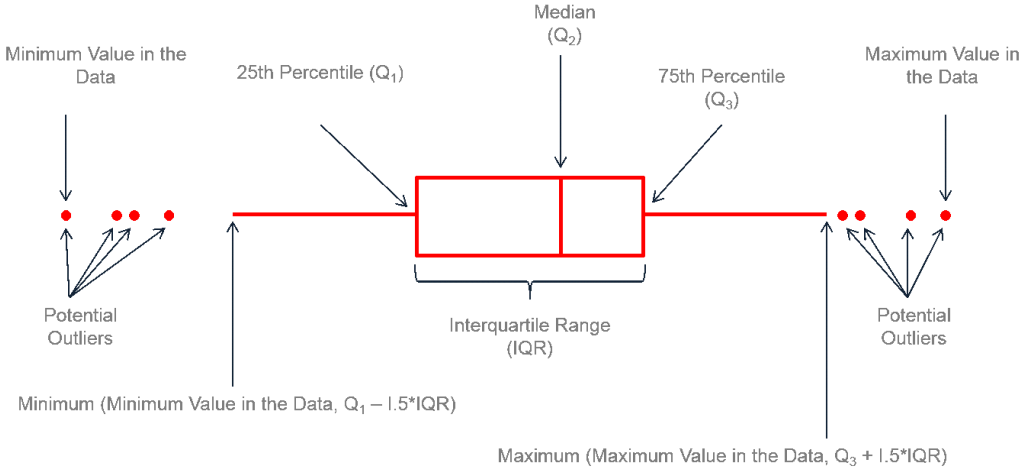
Excel in Action

- Make visuals
  - Scatter Plot
  - Bar Chart
  - Line Graph
- Statistics
  - Average
  - Standard Deviation
- Other
  - Tables
  - Color By
  - If Statements

Excel in Action

- Add error bars to chart
  - Standard Deviation

- Make a box plot
  - Median (middle value)
  - Q1, First Quartile (25%)
  - Q3, Third Quartile (75%)
  - Min = Q1 – 1.5*IQR
  - Max = Q3 + 1.5*IQR
    - IQR = Q3 - Q1



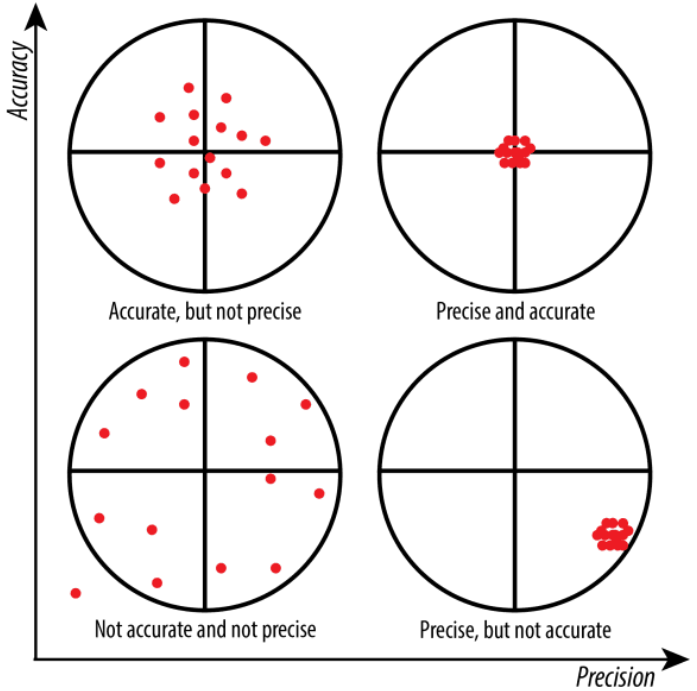
https://support.microsoft.com/en-us/office/create-a-box-plot-10204530-8cdf-40fe-a711-2eb9785e510f
https://www.contextures.com/excelboxplotchart.html

Analysis

Before using a tool, what are we doing?

Accuracy / Precision

- **Accuracy:** how close a measure value is to the true value.
- **Precision:** how close measure values are to each other.



https://wp.stolaf.edu/it/gis-precision-accuracy/

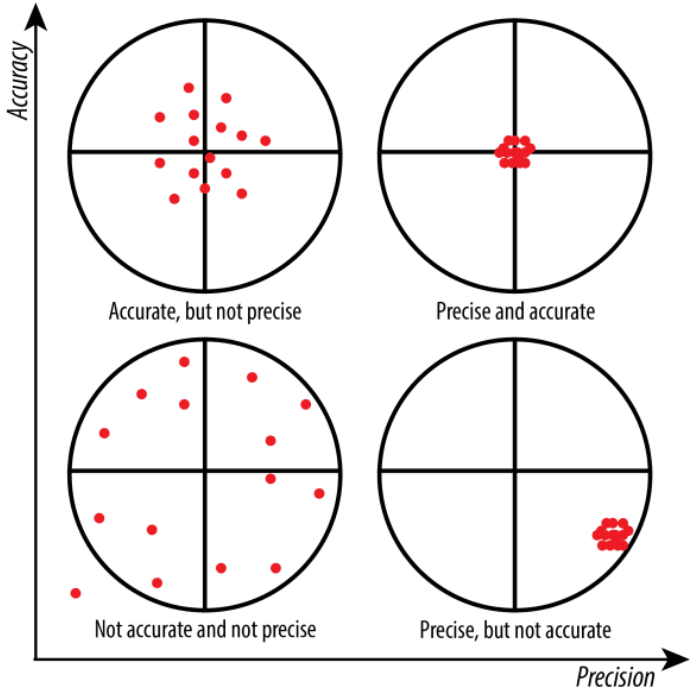
Accuracy / Precision & Errors

- **Accuracy:** Systematic / Biased Errors
  - How well equipment was used.
  - How well experiment was controlled.

- **Precision:** Random / Chance Errors
  - Precision of the equipment
  - Cannot or are unable to control these factors
  - Standard Deviation is a measure of Random Error

- **Uncertainties:** measurement of random errors
  - Errors incurred as a results of imperfect tools that only have a certain degree of precision

https://courses.lumenlearning.com/boundless-statistics/chapter/measurement-error/



Data Summaries

- Average: an estimate of the "true" value of the measurement

- Standard Deviation: a measure of the "spread" in the data
  - You can be reasonably sure (~70%) that if you repeat the ***same measurement*** one more time, that ***next measurement*** will be less than one standard deviation away from the average.

- Standard Error: an estimate in the uncertainty in the average of the measurements
  - You can be reasonably sure (~70%) that if you do the ***entire experiment*** again with the same number of repetitions, the ***average value from the new experiment*** will be less than one standard error away from the average value from this experiment.

https://www2.southeastern.edu/Academics/Faculty/rallain/plab194/error.html

Data Summaries – In Action

- Average = average
- Standard Deviation: use excel *stdev* function
  - N = number of samples
  - μ = mean (average of our samples)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- Standard Error: = Standard Deviation / sqrt (# of samples)

- In calculations
  - Measure within +- 0.1 meters
  - Box that we measure to be 1 x 1 m
  - To calculate our area: 0.9 x 0.9m and 1.1 x 1.1m
    - 0.81m2 and 1.21 m2
    - 1.01m2 +- 0.2m2

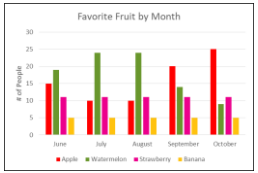
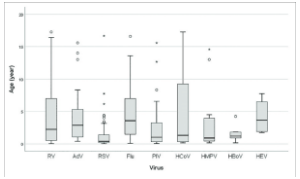
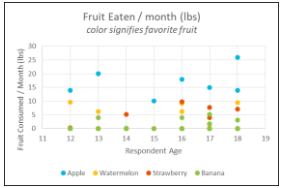
What Type of Chart & Analysis to Use

- Depends!

- What are our independent and dependent variables?
  - Independent: What we are changing / controlling
  - Dependent: The outcome / result from the experiment.
- Are our variables categorical or numeric?
  - Categorical: Represents characteristics
    - Sorted into groups w/ names or labels
    - Ex: A person's gender
  - Numeric: Data has meaning as a measurement
    - Expressed in terms of numbers
    - Ex: A person's height, weight

https://www.dummies.com/education/math/statistics/types-of-statistical-data-numerical-categorical-and-ordinal/

What Type of Chart & Analysis to Use

Independent Variable: Categorical Dependent Variable: Categorical Graphs: Bar Charts Analysis: Chi-squared test of independence 	Independent Variable: Categorical Dependent Variable: Numeric Graphs: Box Plots / Bar Charts Analysis: T-tests / ANOVA 
Independent Variable: Numeric Dependent Variable: Categorical Graphs: ?? Analysis: Regression, Classification	Independent Variable: Numeric Dependent Variable: Numeric Graphs: Scatter Plots Analysis: Regression 

Null Hypothesis

- Null Hypothesis: A hypothesis that we can reject (is false)
  - Alternate hypothesis: The hypothesis that we think is true

- How to reject a null hypothesis
  - State the null hypothesis and the alternate hypothesis
    - May be easier to do alternate first
  - Support or reject the null hypothesis (use a test that generates a p-value)

- With our analysis, we are often only able to reject a null hypothesis

https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/support-or-reject-null-hypothesis/

Analysis: Chi-squared tests

- Chi-Square Statistic χ^2
  - Measures how a model compares to observed data
  - Data must be random, raw, mutually exclusive, from independent variables
- Types
  - Chi-square goodness of fit: How well a sample of data matches a larger population
  - **Chi-square test for independence:** Is there a relationship between the variables

C = degrees of freedom

O = observed value

E = expected value

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

https://www.statisticshowto.com/probability-and-statistics/chi-square/

https://www.investopedia.com/terms/c/chi-square-statistic.asp#:~:text=A%20chi%2Dsquare%20(%CF%872,from%20a%20large%20enough%20sample.&text=Chi%2Dsquare%20tests%20are%20often%20used%20in%20hypothesis%20testing

In-Action: Chi-squared test of independence

- Alternate Hypothesis: Boys are more likely than girls to take German as a second language
- Null Hypothesis: Boys are not more likely than girls to take German as a second language

- Output of test = “p” value (probability that the variables are independent)
  - $p < 0.05$ = dependence (common value)
  - $p > 0.05$ = independent (common value)

https://www.real-statistics.com/chi-square-and-f-distributions/independence-testing/

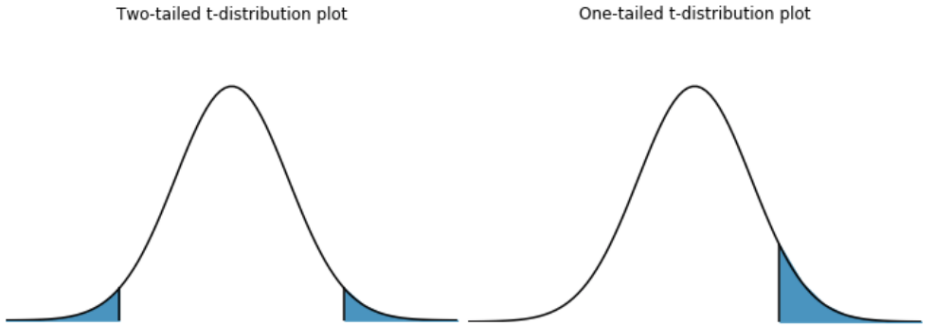
Analysis: T-test

- Is there a statistical difference between two groups
  - Assume that the means of the two distributions are equal
  - Able to reject (groups are highly probably different) or fail to reject a null hypothesis, never accepts
  - 20-30 samples in a group
- P-value (probability value): *is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct*
- Critical Value: *the boundaries of the acceptance region of the test*

- **p_value >** α **(Critical value)**: Fail to reject the null hypothesis of the statistical test.
- **p_value ≤** α **(Critical value)**: Reject the null hypothesis of the statistical test.
- Common critical value = 0.05
  - This 0.05 means that, if we run the experiment 100 times, 5% of the times we will be able to reject the null hypothesis and 95% we will not.
  - p_value > 0.1: No evidence
  - p_value between **0.05** and 0.1: Weak evidence
  - p_value between **0.01** and **0.05**: Evidence
  - p_value between 0.001 and **0.01**: Strong evidence
  - p_value < 0.001: Very strong evidence

Start Here: https://towardsdatascience.com/the-statistical-analysis-t-test-explained-for-beginners-and-experts-fd0e358bbb62

Analysis: T-test



- Distribution Types
  - 1 Tailed or 2 Tailed
    - 1 tailed when differences are in a specific directio
    - 2 tailed is most common
- Paired or Unpaired?
  - Does the data come from the same participants?
    - Yes = Paired
- Variance: Equal or Unequal
  - Is the variance between our samples equal?
    - Equal: equal # of data points, numbers are similar
    - Conservative Choice: use unequal variances

1 vs 2 tailed: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-the-differences-between-one-tailed-and-two-tailed-tests/

Equal or Unequal: https://www.ruf.rice.edu/~bioslabs/Stats_tutorial/ttest17.html

The math: https://www.real-statistics.com/students-t-distribution/one-sample-t-test/

https://www.statisticshowto.com/probability-and-statistics/t-test/

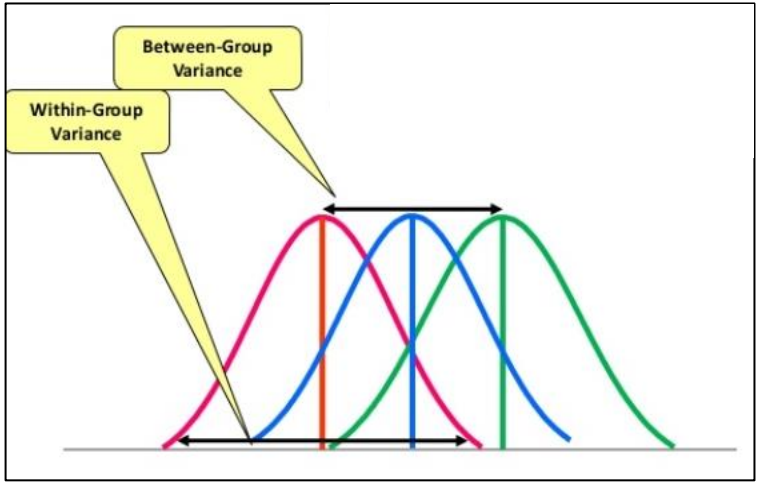
Analysis: ANOVA

- ANOVA = Analysis of variance
  - Outputs: variation within and between groups
- Determine if the results are significant
  - Do you reject the null hypothesis?
- T-test limited to 2 groups, ANOVA allows for more

- One-way or two-way
  - One-way: 1 independent variable
  - Two-way: 2 independent variables

https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/
https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/

Analysis: ANOVA

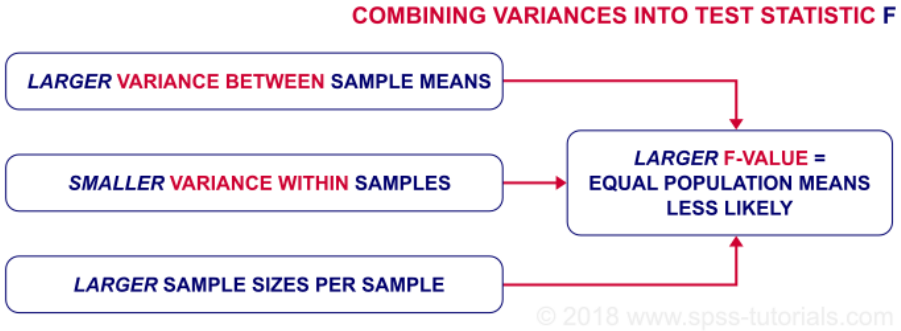


- ANOVA compares two things
  - Variation within groups
  - Variation between groups
- ANOVA Null Hypothesis
  - All population means are equal

- One-way ANOVA
  - Tells us that at least two groups are different from each other
  - But not which one!

In-Action: ANOVA

- F: measures if the means of different samples are significantly different or not
  - F = Between group variability / Within group variability
  - Fstatistic > Fcritical = we reject the null hypothesis
    - Fcritical is calculated for the desired significance level (α)



https://www.spss-tutorials.com/anova-what-is-it/#test-statistic

Analysis: Regression

- Method for finding trends in data
- Linear Regression: relationship between variables can be described with a straight line
- Non-linear Regression: relationship between variables can be described with a curved line

- Slope
- Y-Intercept
- R-Squared (correlation co-efficient)
  - How close are the variables related
  - 1.0 = perfect fit
  - Human behavior: > 0.5
  - Physical process: > 0.9
  - How High does R-squared need to be: https://statisticsbyjim.com/regression/how-high-r-squared/

https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/#definition

https://statisticsbyjim.com/regression/interpret-r-squared-regression/

Analysis: Margin of Error

- Margin of Error: Range of values above and below a *Confidence Interval*

- A poll says Candidate A will win an election with 52% of the vote with a confidence interval of 95% and Margin of Error of 4%
  - If we run the election 100 times, 95 of 100 times the candidate will receive 48 – 56% of the vote
    - 52 – 4 = 48
    - 52 + 4 = 56

https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/margin-of-error/

Analysis: Margin of Error

- How to calculate
  - Margin of error = Critical value x Standard Error of the sample.
    - Most common by far
  - Margin of error = Critical value x Standard Deviation for the population.

- Standard Error: Standard Deviation / sqrt (sample size)
- Critical Value:
  - Depends on your chosen confidence value
  - Look it up with a t or z-score table: Table
  - https://www.youtube.com/watch?v=RAnFyF_6zHk

https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/margin-of-error/

Analysis: Confidence Intervals

- Does not reflect the accuracy of a data set
- What is it saying: If the survey were repeated over and over, the results would match the results from the actual population 95% of the time
  - 0% = no confidence at all that if you repeated the survey you would get the same results
  - 100% = no doubt that if you repeated the survey that you would get the same results

Tips for Communication: Limits of Data & Study

- Some of the first questions you will be asked regarding data is:
  - What analysis did you conduct
  - Why did you conduct this analysis
  - What are the limitations on your data and experiment design
    - Sources of error in your measurement
    - External variables not accounted for

Data Presentation

Share your work so it can be understood.

Communications: What is our goal + how will it be seen / viewed?
Science and Engineering Fair

- What is our goal?
  - Share the results of our research.
    - Hypothesis is supported or not supported
    - Design solves a problem and fits our design criteria
    - How external factors have / do not have an effect

- How will it be viewed?
  - Preliminary Judging (5-6 judges)
    - 10-15 minutes for judges to learn about your project and score.
  - Finalist Judging (5-6 judges)
    - 12 minute Zoom interviews

Communications: What is our goal + how will it be seen / viewed?
Internal Design Reviews

- What is our goal?
  - Approve a product for sale to customers
    - Product is safe for all customer scenarios
    - Product meets all design requirements

- How will it be viewed?
  - In-depth design review meeting
    - 2-3 hour meeting to review with 5-10 colleagues
  - Report review
    - Report emailed and on drive for many to view on their own time

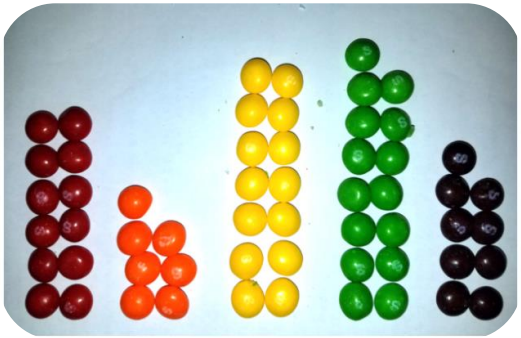
Communicating - Visuals

Process: Sorting Skittles by color

1) Open bag of Skittles and pour onto a table
2) Group Skittles by color
3) Count the number of each color

Step 1: Open bag of Skittles

Step 2: Group Skittles by color.

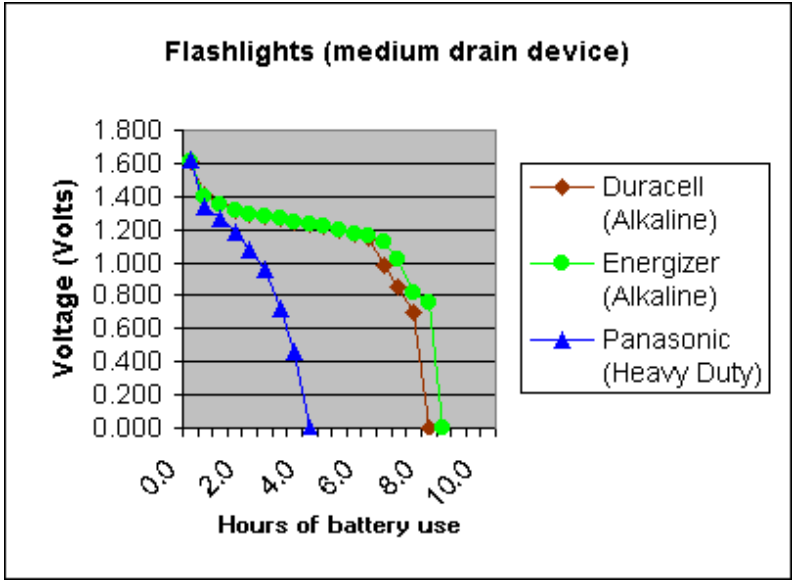


Step 3: Count the number of each color.

What is our goal?

- Communicate our results: what tells us more?

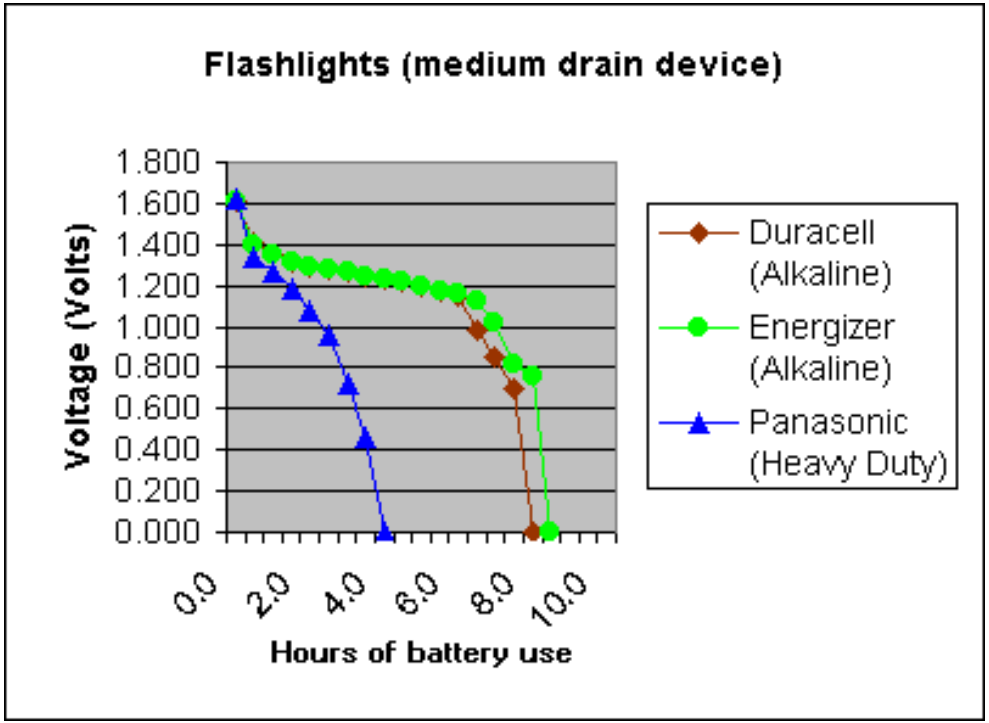
After conducting an analysis of the outcome with 3 trials, the Energizer Alkaline lasted an average of 9.5 hours, the Duracell Alkaline an average of 9.2 hours and the Panasonic Heavy Duty an average of 5.1 hours.



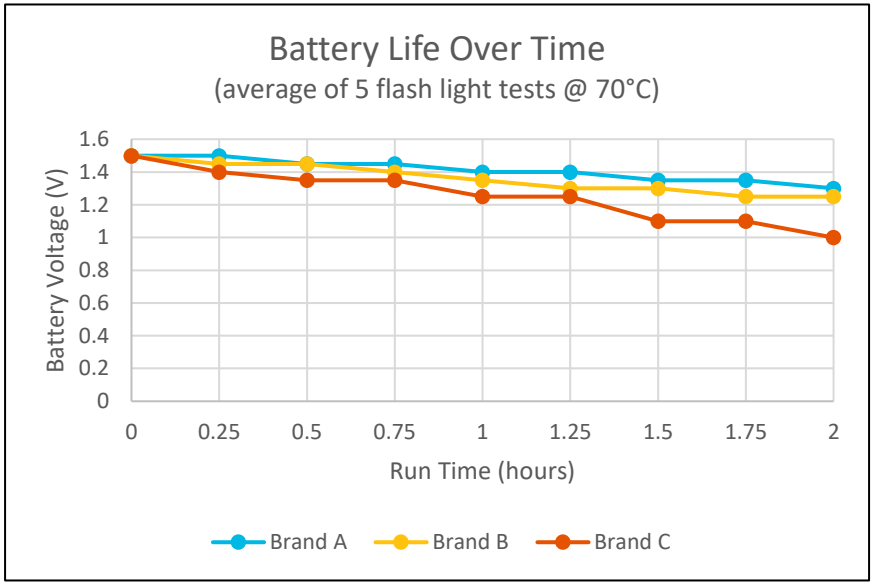
https://www.sciencebuddies.org/science-fair-projects/science-fair/data-analysis-graphs

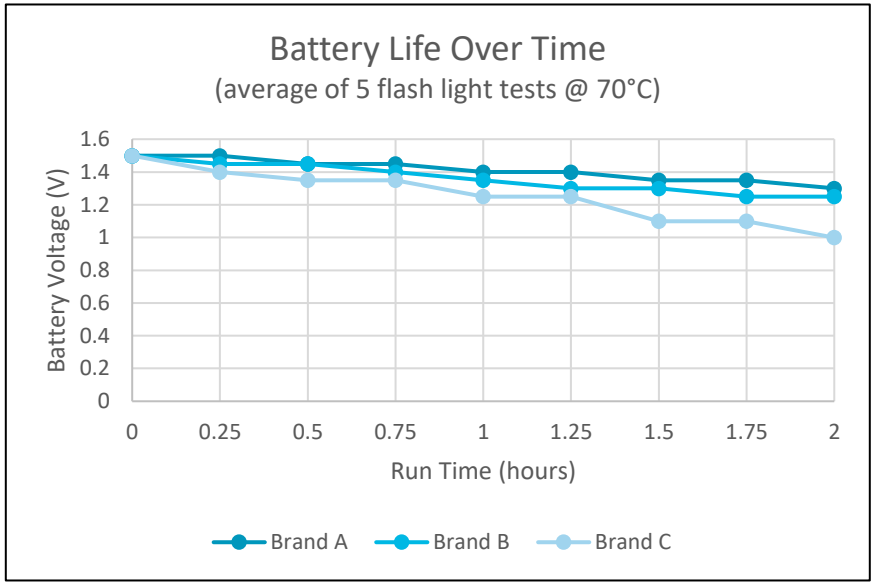
Graph Etiquette

- Title
- Axis Labels
  - Include units
- Legend / Key
- Readability
  - Background
  - Colors
  - Excess Precision

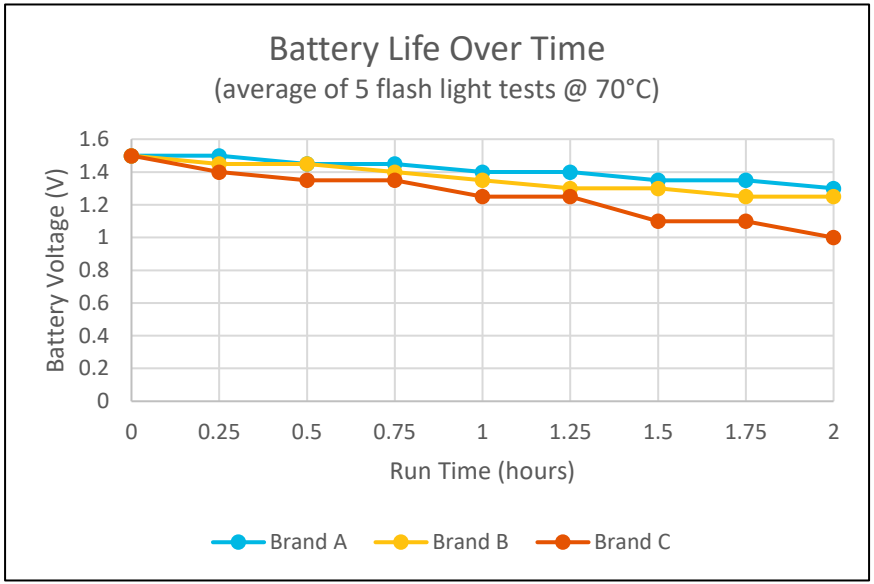


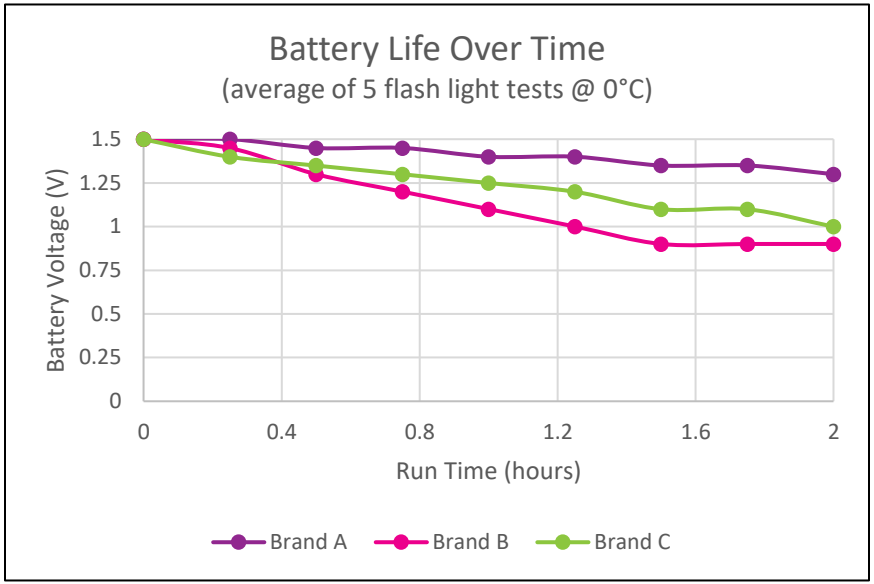
Graph Etiquette - Colors



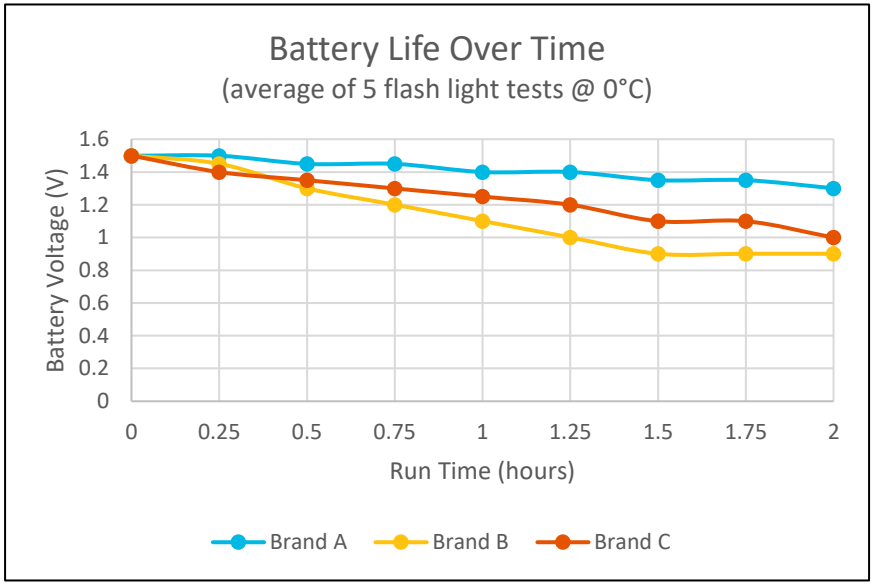


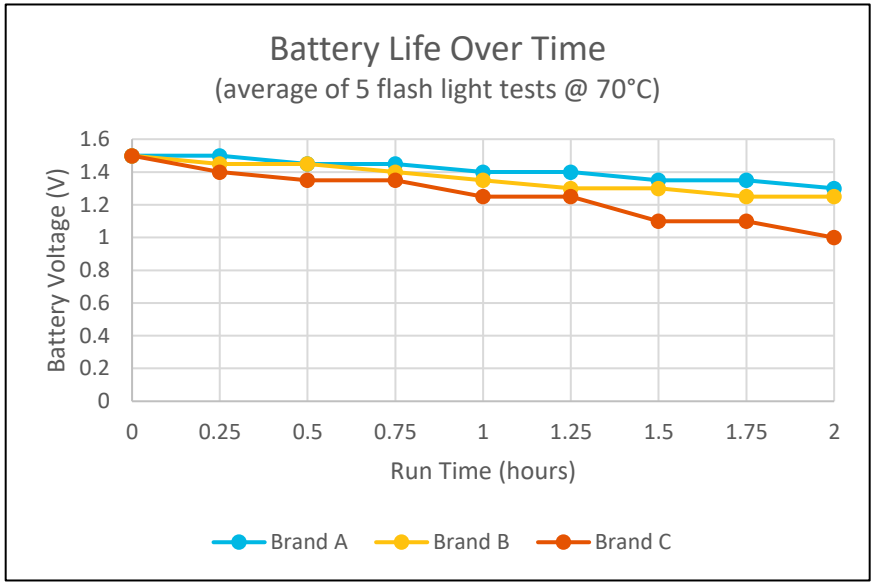
Graph Etiquette - Consistency





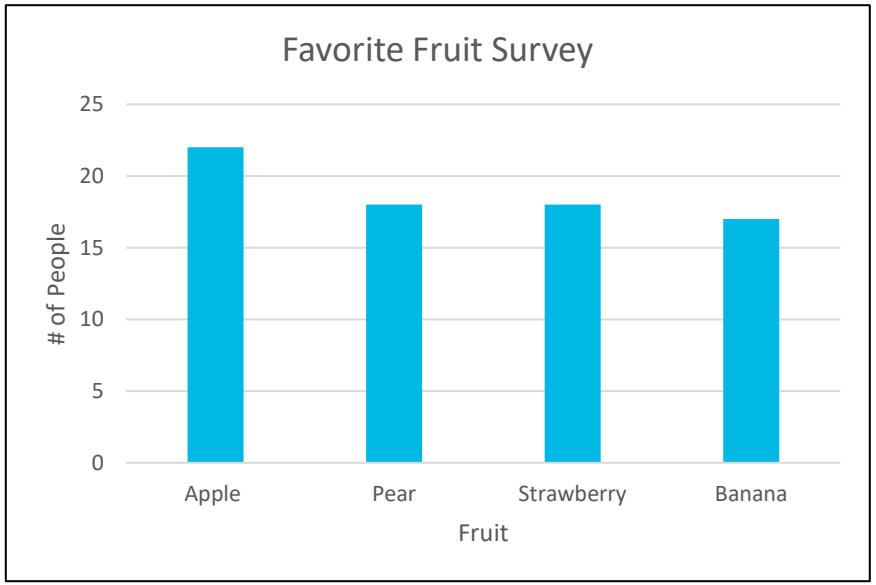
Graph Etiquette - Consistency

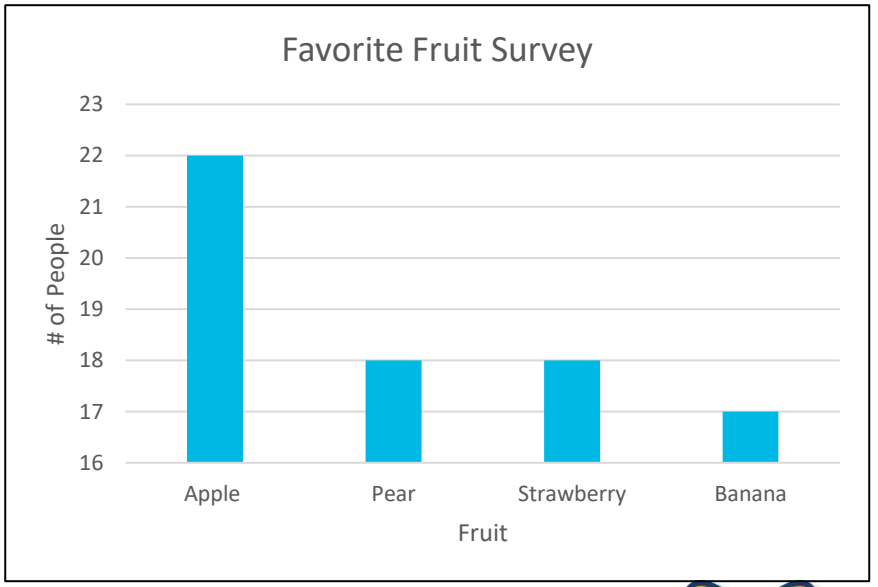




Graph Etiquette – “Cable News” Graphs

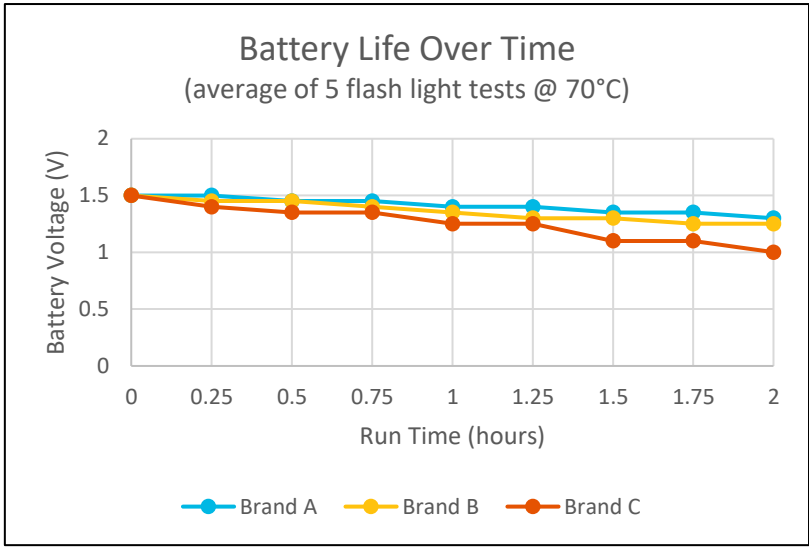
- For a bar chart our baseline always needs to be 0





Graph – Layout Rule of Thumb

- X-Axis: Independent Variable
  - What we are changing
- Y-Axis: Dependent Variable
  - The outcome
- Battery Experiment
  - Independent Variable: Run time of flashlight
  - Dependent Variable: Battery voltage



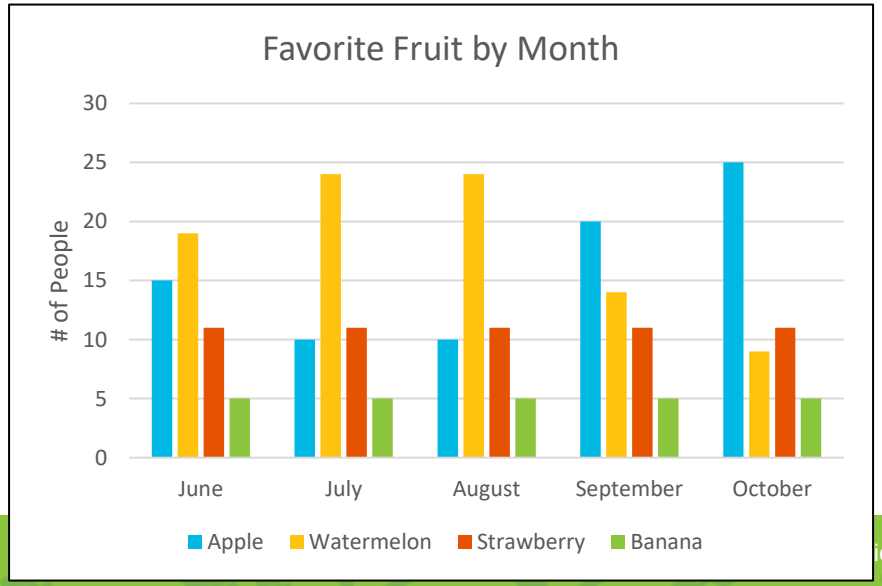
What is our goal?

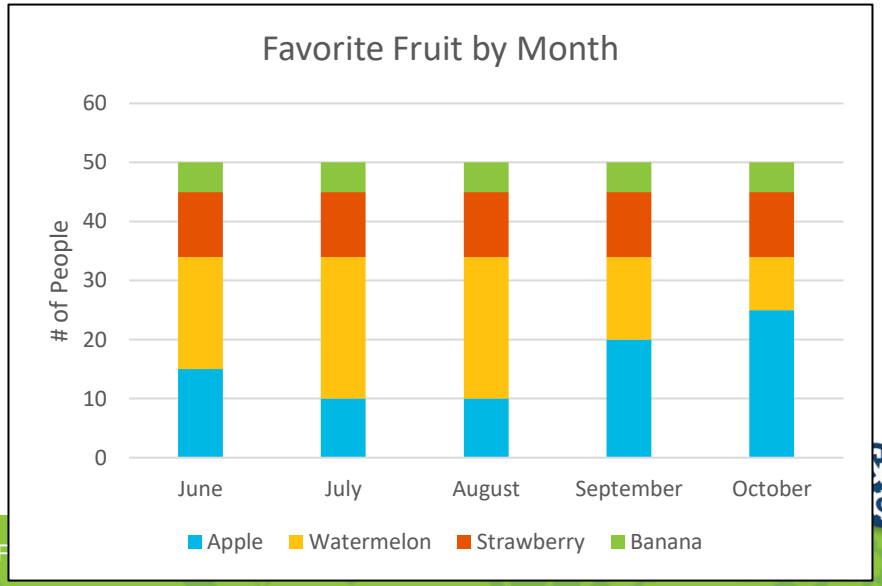
- Communicating: How is our data changing over time
  - Line Chart
  - Scatter Plot (with connecting lines, aka line chart w/ dots)
  - Bar Chart (each time period is a bar)
  - Box Plot (advanced, will show at end)

- Examples are data from a survey of 50 people
  - Survey sent each month
  - Ask participants to choose their favorite fruit
  - Goal: Does someone's favorite fruit change throughout the year

Change Over Time – Bar Chart

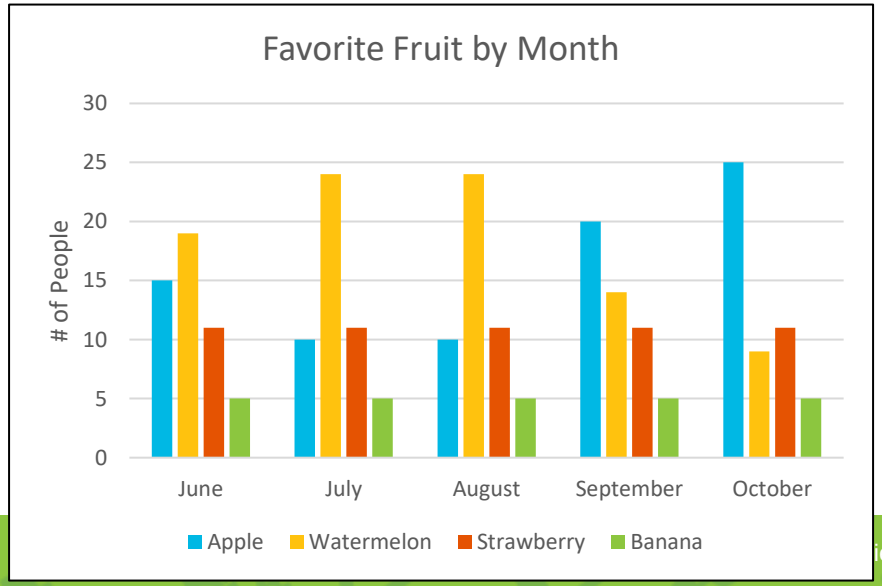
- Bar Chart (each time period is a bar) https://chartio.com/learn/charts/bar-chart-complete-guide/
  - Good choice if your independent variable is not numerical
    - Standard bar chart (left)
    - Stacked bar chart (right)

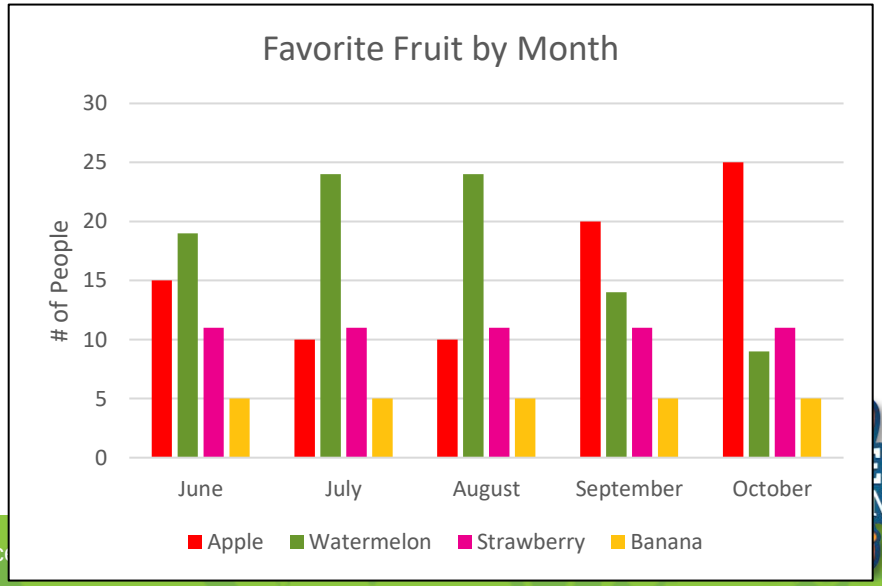




Hints – Color is Your Friend

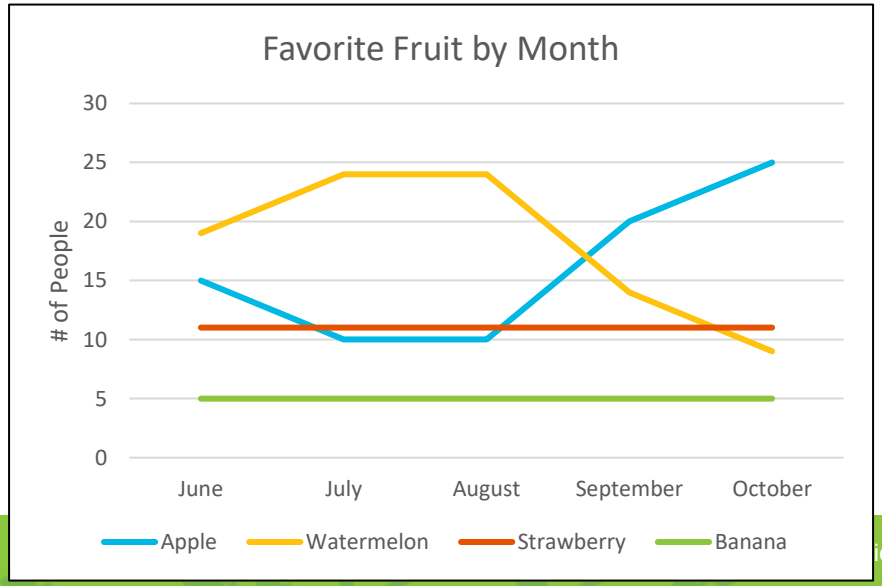
- If applicable: have your colors match the impression of the item
  - Apple: red
  - Banana: yellow

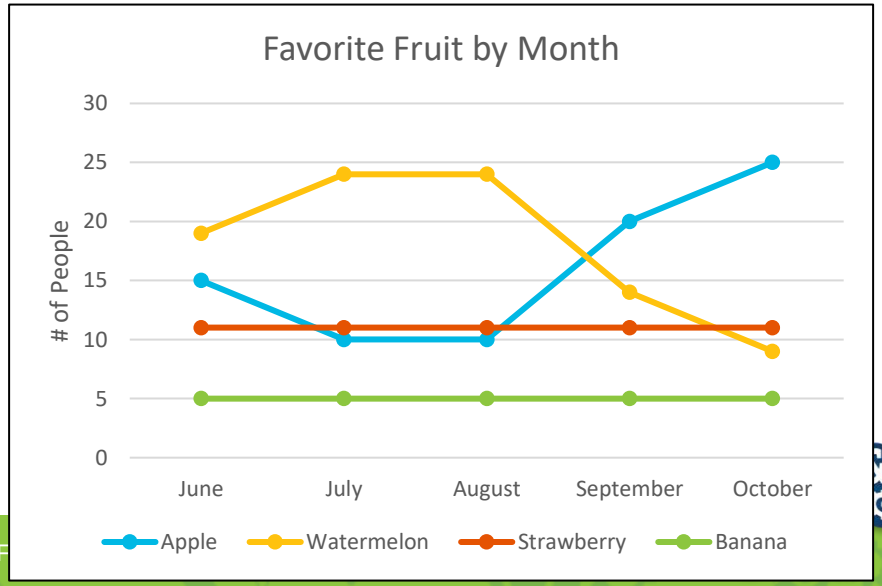




Change Over Time

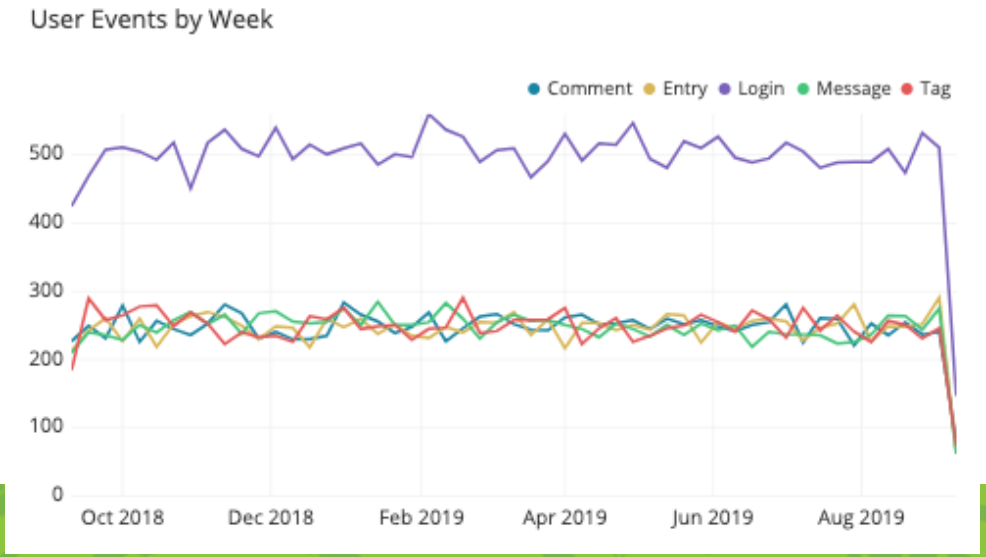
- Line Chart and Scatter Plot w/ lines
  - Good choice if your independent variable is numerical
    - Line chart (left)
    - Scatter plot w/ lines (right)





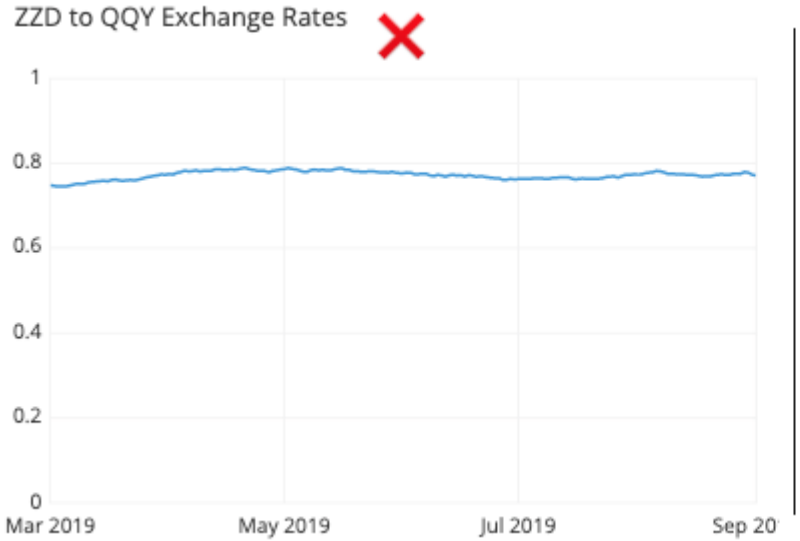
Hints – More is Not Always Better

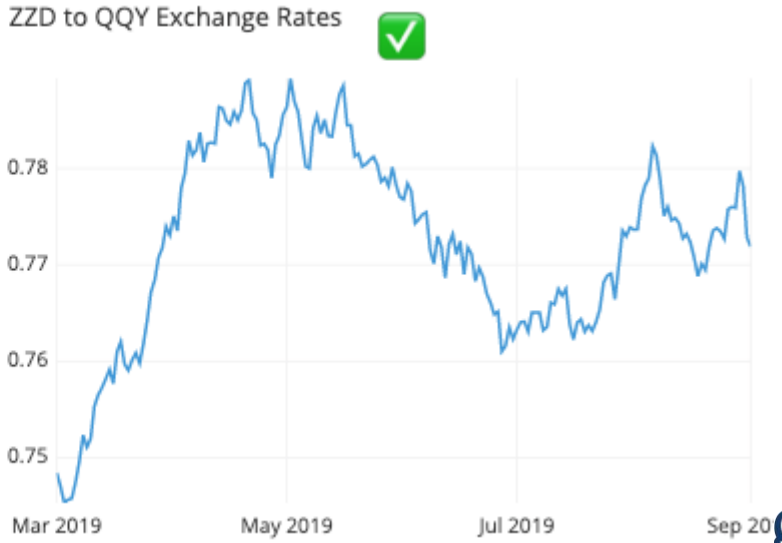
- Think about what you want to communicate
  - Good: clearly able to see logins are greater than other activities
  - Bad: Difficult to determine differences between Entry, Message, xxxx



Hints – Baseline Value

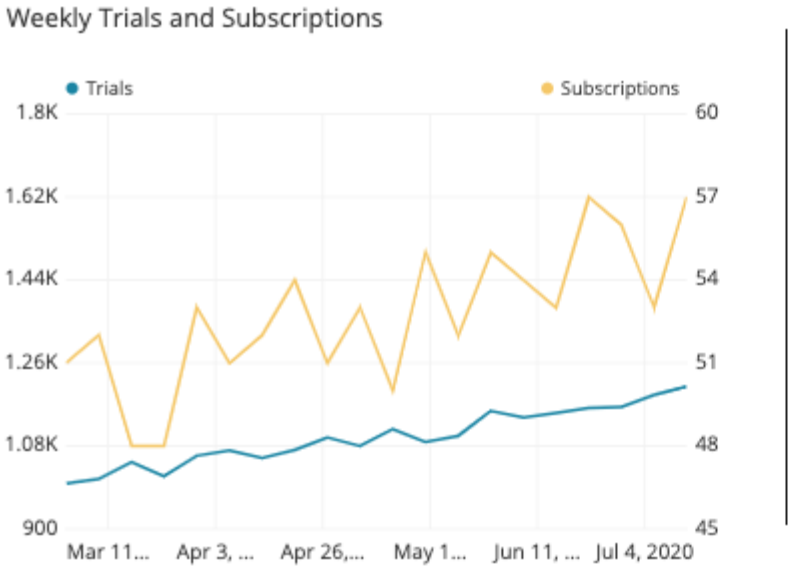
- Think about what you want to communicate
  - For Line Chart: Emphasize changes in value

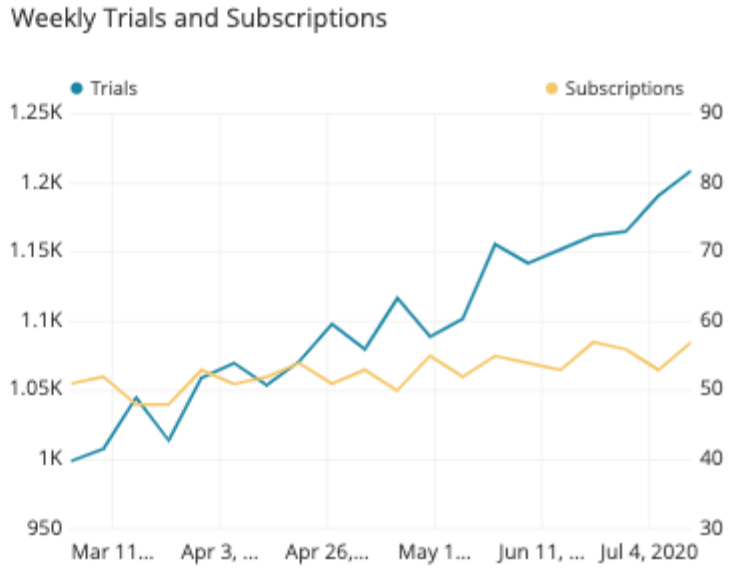




Hints – Dual Axis

- We can compare different dependent variable variables on a 1 graph
  - Be consistent with your axis scaling (don't place a large offset on the scaling)





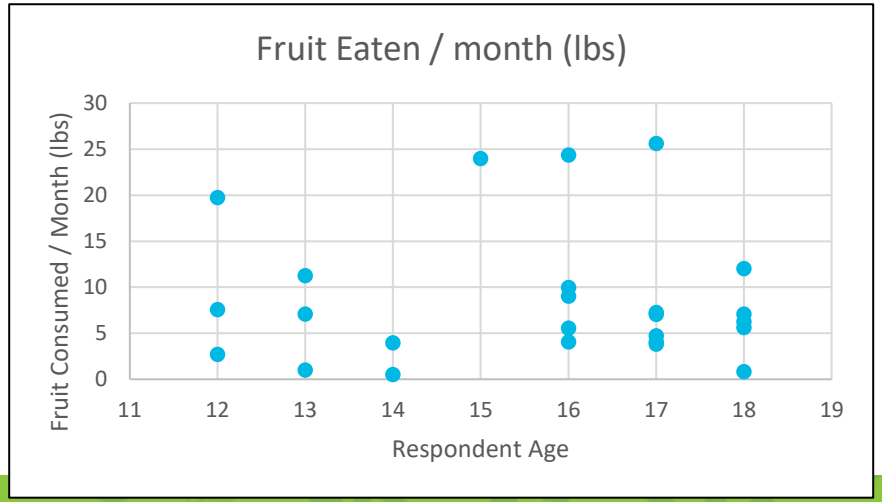
What is our goal?

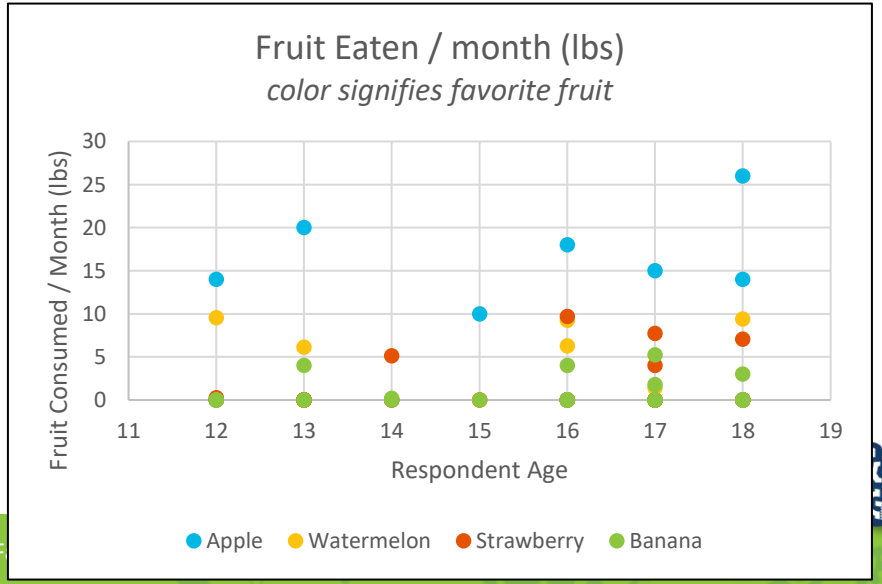
- Communicating: Observe relationships between groups
  - Scatter Plot
  - Bubble Chart
  - Grouped Bar Chart

- Examples are data from a survey of 50 people
  - Survey sent each month
  - Ask participants to choose their favorite fruit
  - Goal: Does someone's favorite fruit change throughout the year

Relationship Between Groups

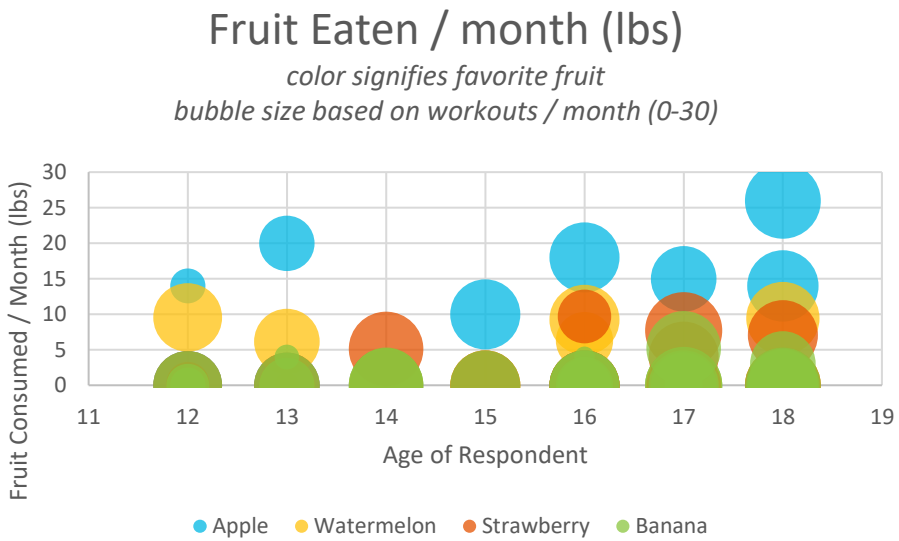
- Scatter Plot https://chartio.com/learn/charts/what-is-a-scatter-plot/
  - Good choice for determining correlation between groups, finding outliers, versatile
    - Scatter Plot (left)
    - Scatter plot w/ color for favorite fruit





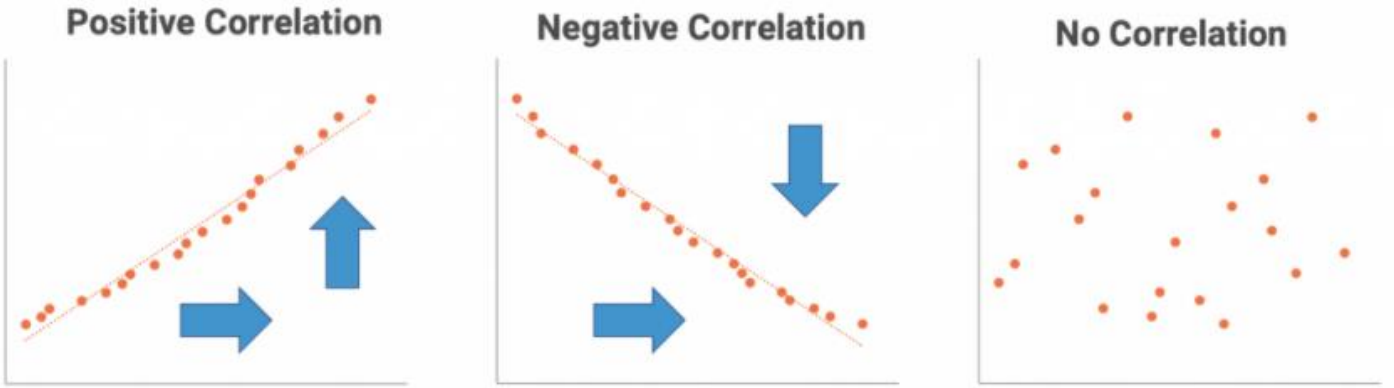
Relationship Between Groups

- Bubble Chart
  - Good choice for determining correlation between groups, finding outliers, versatile



Correlation

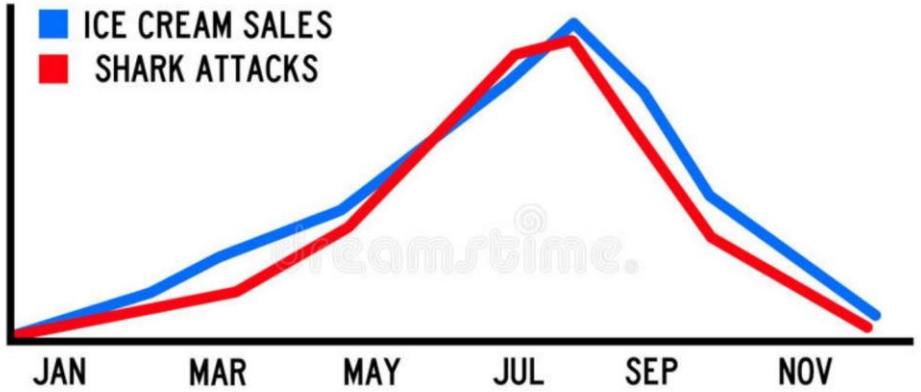
- Correlation: the relationship between the two variables
  - How much does one variable affect the other?
  - Positive correlation: Both variables move in the same direction
  - Negative correlation: Variables move in opposite directions
  - No correlation: No link between the two variables



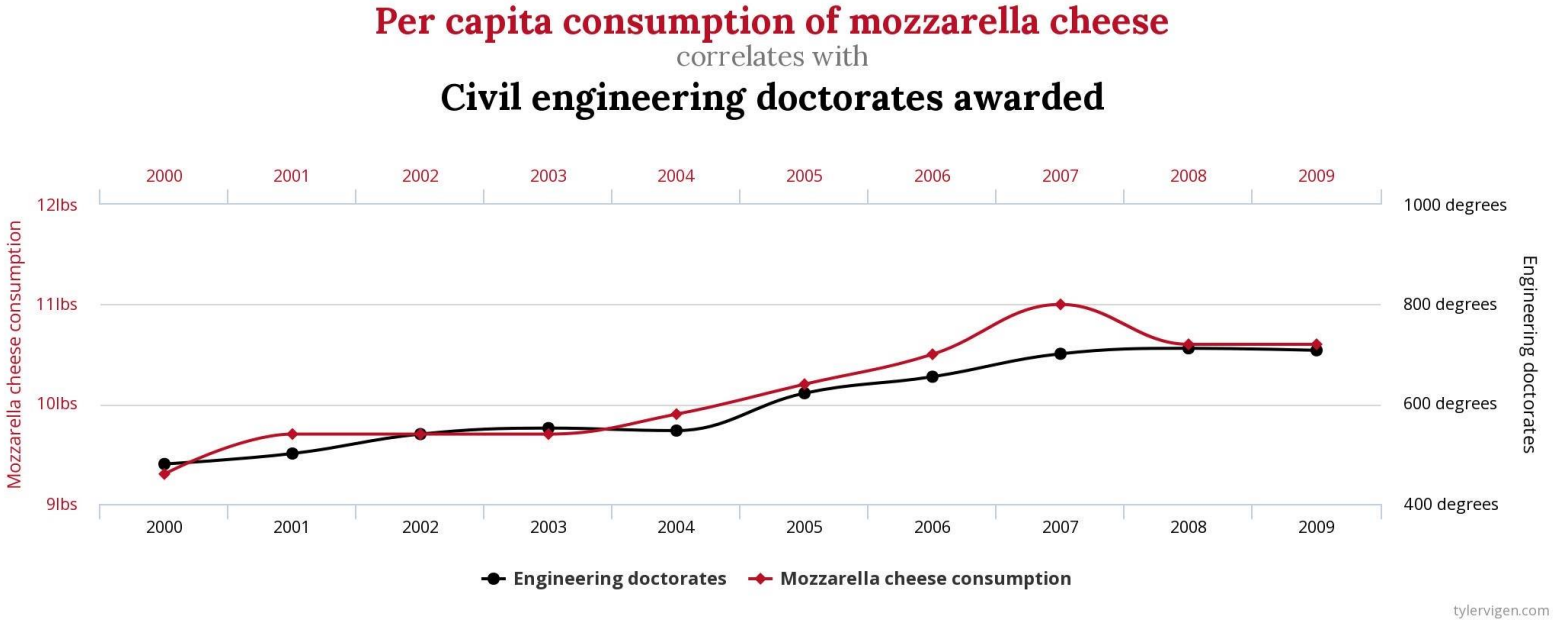
https://mylearningsinaiml.wordpress.com/2018/11/21/scatter-plots/
https://astutesolutions.com/blog/articles/causation-vs-correlation

Correlation =/ Causation

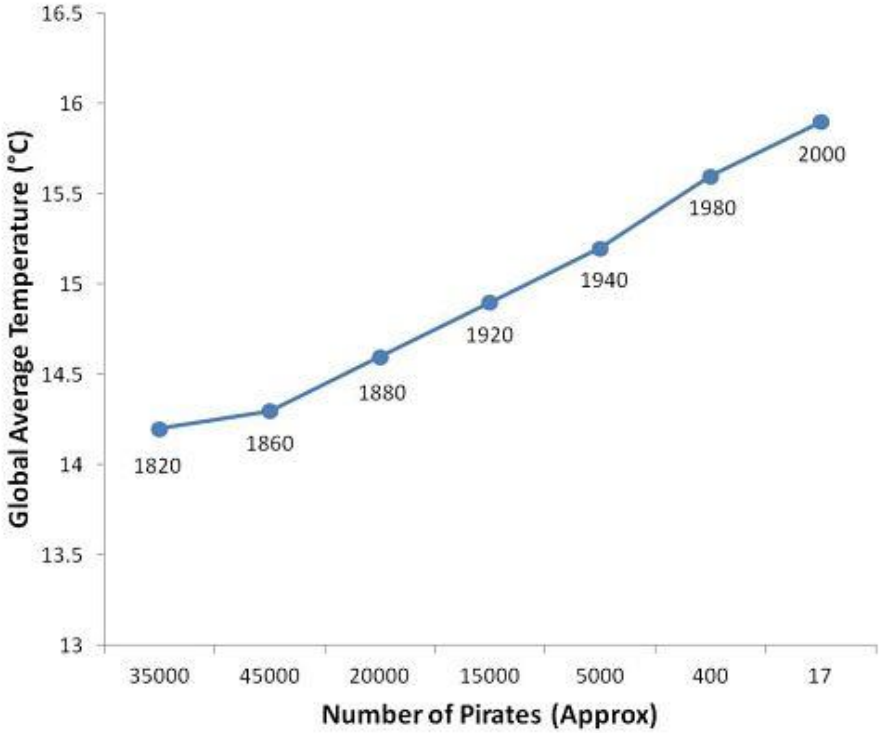
- Correlation: A change in one variable mirrored by a positive or negative change in the other.
  - Spurious Correlation: strong relationships between variables that are not caused by one another.
- Causation: One variable is changing as a result of the other variable.



Correlation =/ Causation



Correlation =/ Causation



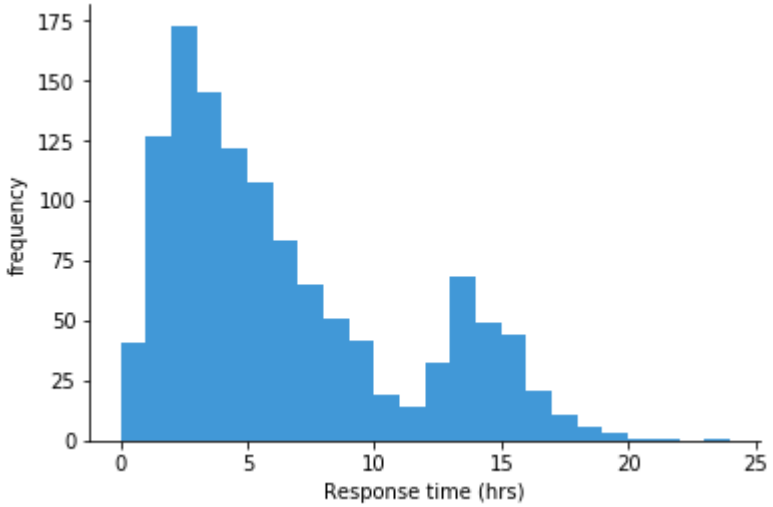
What is our goal?

- Communicating: How our data is distributed
  - Bar Chart
  - Histogram
  - Density Curve
  - Box Plot (advanced, will show at end)

- Examples are data from a survey of 50 people
  - Survey sent each month
  - Ask participants to choose their favorite fruit
  - Goal: Does someone's favorite fruit change throughout the year

Data Distribution

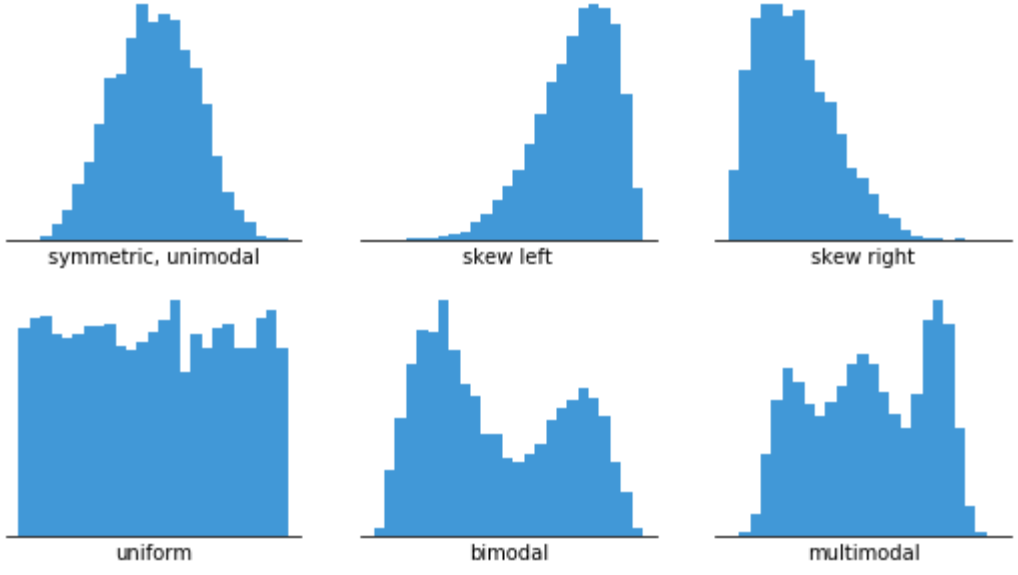
- Histogram
  - Plots the distribution of a numeric variable's values as a series of bars
  - The x-axis values are "binned" together (ex. each hour is binned together)



https://chartio.com/learn/charts/histogram-comp

Data Distribution

- Histogram
  - Very good at showing the distribution of our data



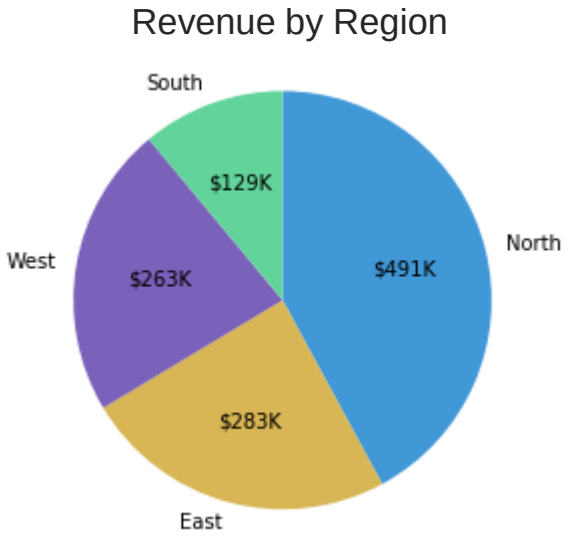
What is our goal?

- Communicating: Part to Whole Comparison (understanding the components that make up the total)
  - Pie Chart
  - Doughnut Chart (pie chart w/ the center missing)
  - Stacked Bar Chart
  - Stacked Area Chart

- Examples are data from a survey of 50 people
  - Survey sent each month
  - Ask participants to choose their favorite fruit
  - Goal: Does someone's favorite fruit change throughout the year

Part to Whole Comparison

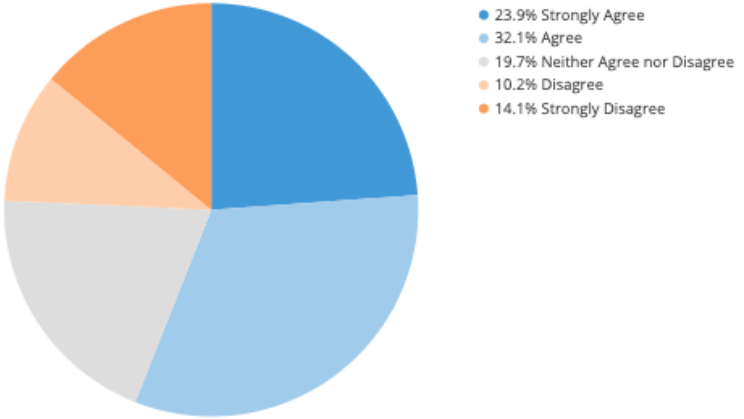
- Pie Chart
  - Comparing each variable relative to the whole data set
  - Only use for the above use case

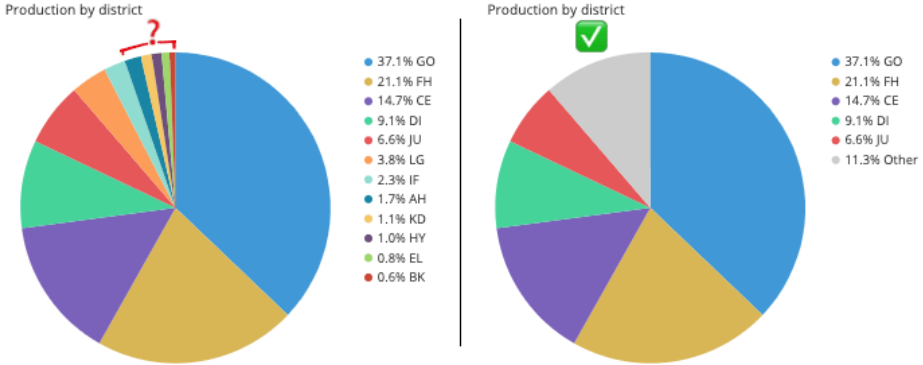


https://chartio.com/learn/charts/pie-chart-complete-guide/

Hints – Pie Chart

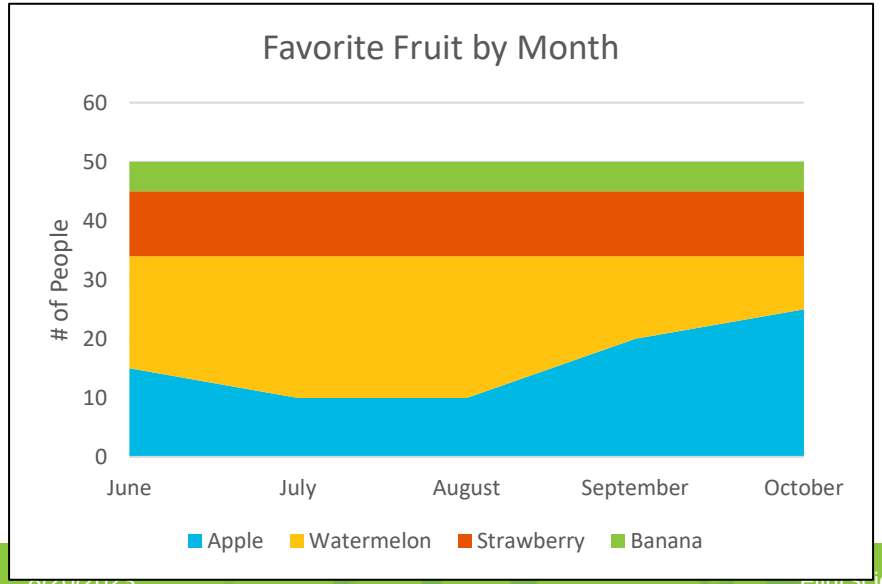
- Include annotations (%, value)
- Order slices by size
- Limit the number of slices
  - Group many “small” slices into “other”

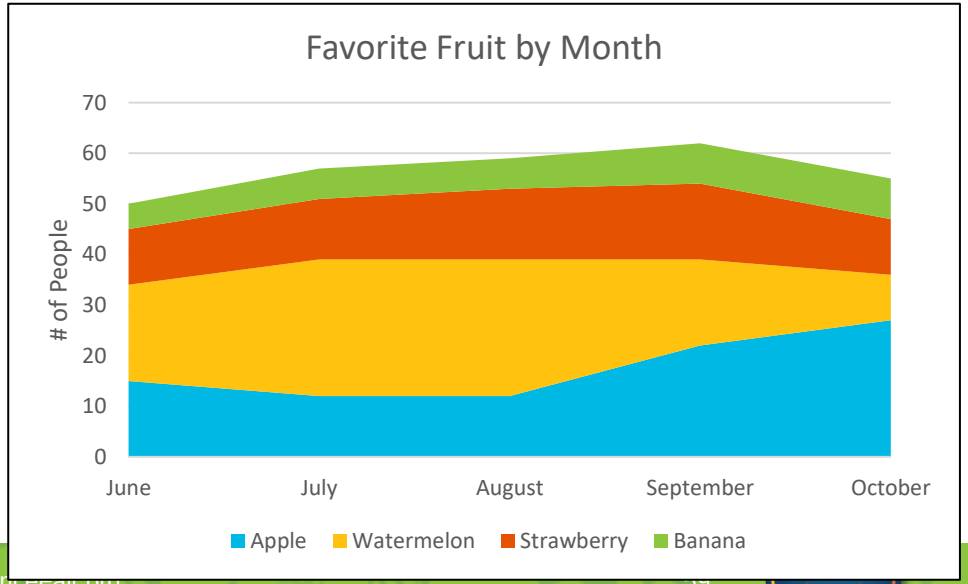




Part to Whole Comparison

- Area Chart https://chartio.com/learn/charts/area-chart-complete-guide/
  - Line Chart + Bar Chart
  - Very powerful if the whole is also changing (left: same # / month; right: different # / month)



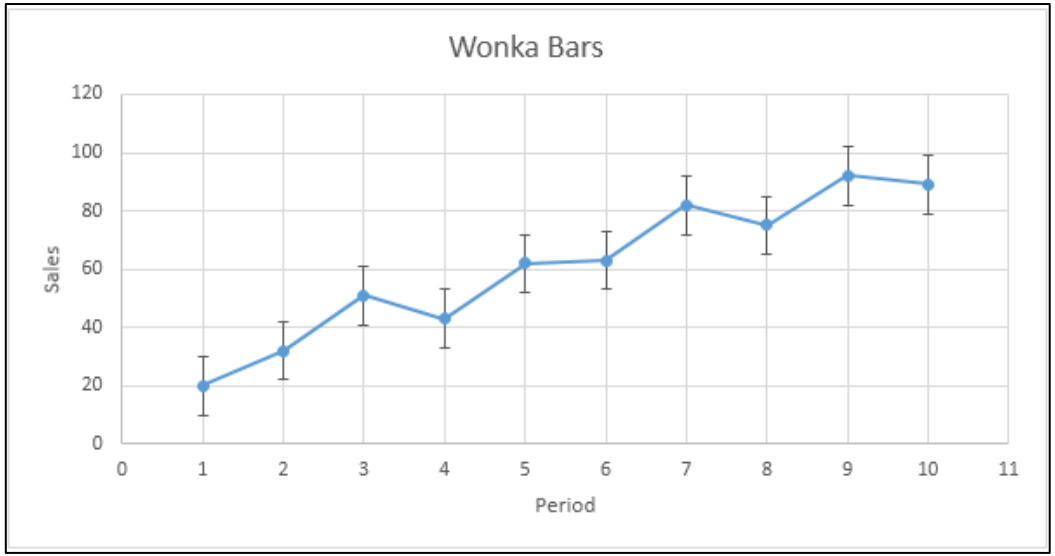


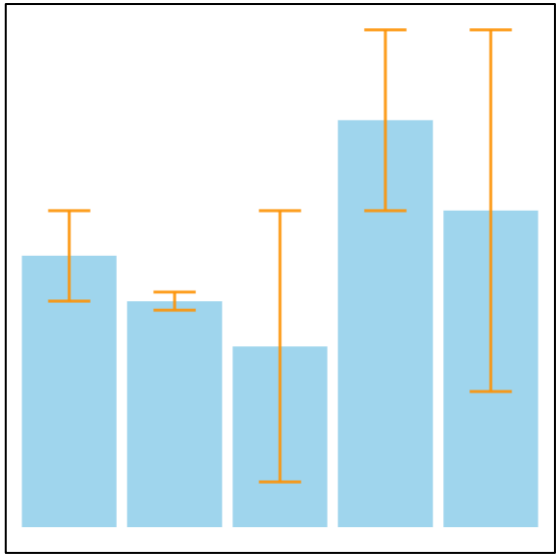
Uncertainty

- What is uncertainty?
  - The variability of our measurement.

- How do we communicate our uncertainty?
  - Uncertainty / Error bars
  - Uncertainty shading
  - Box plots

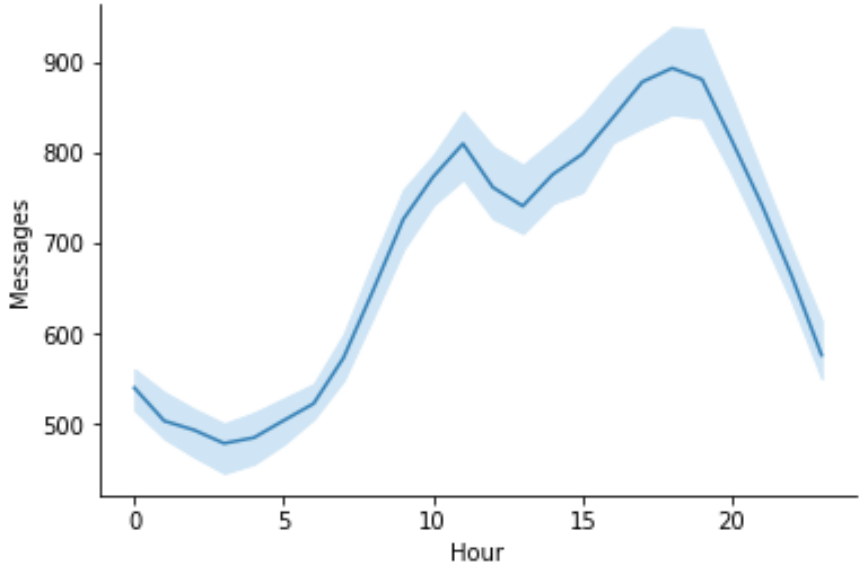
Error Bars





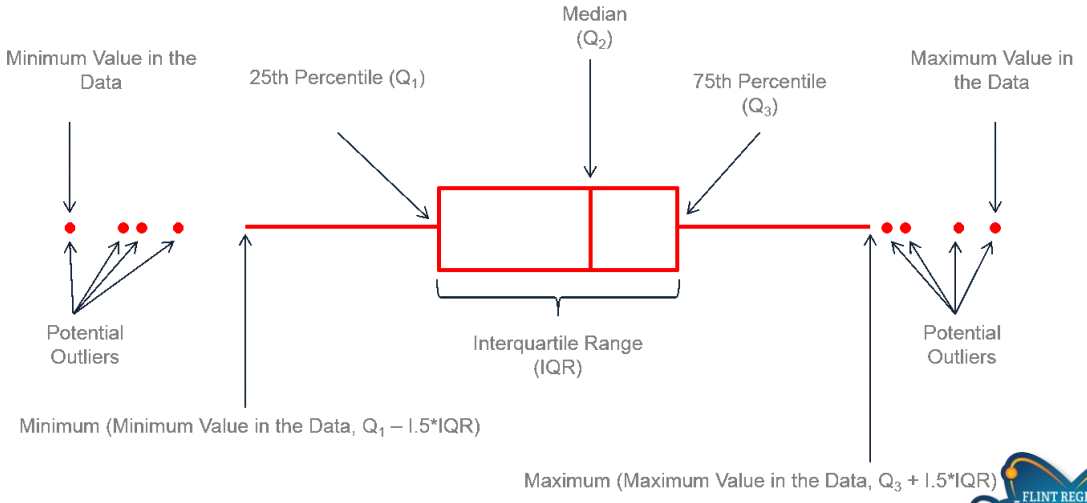
Shading for uncertainty

- An alternative to error bars is to add shading for uncertainty.

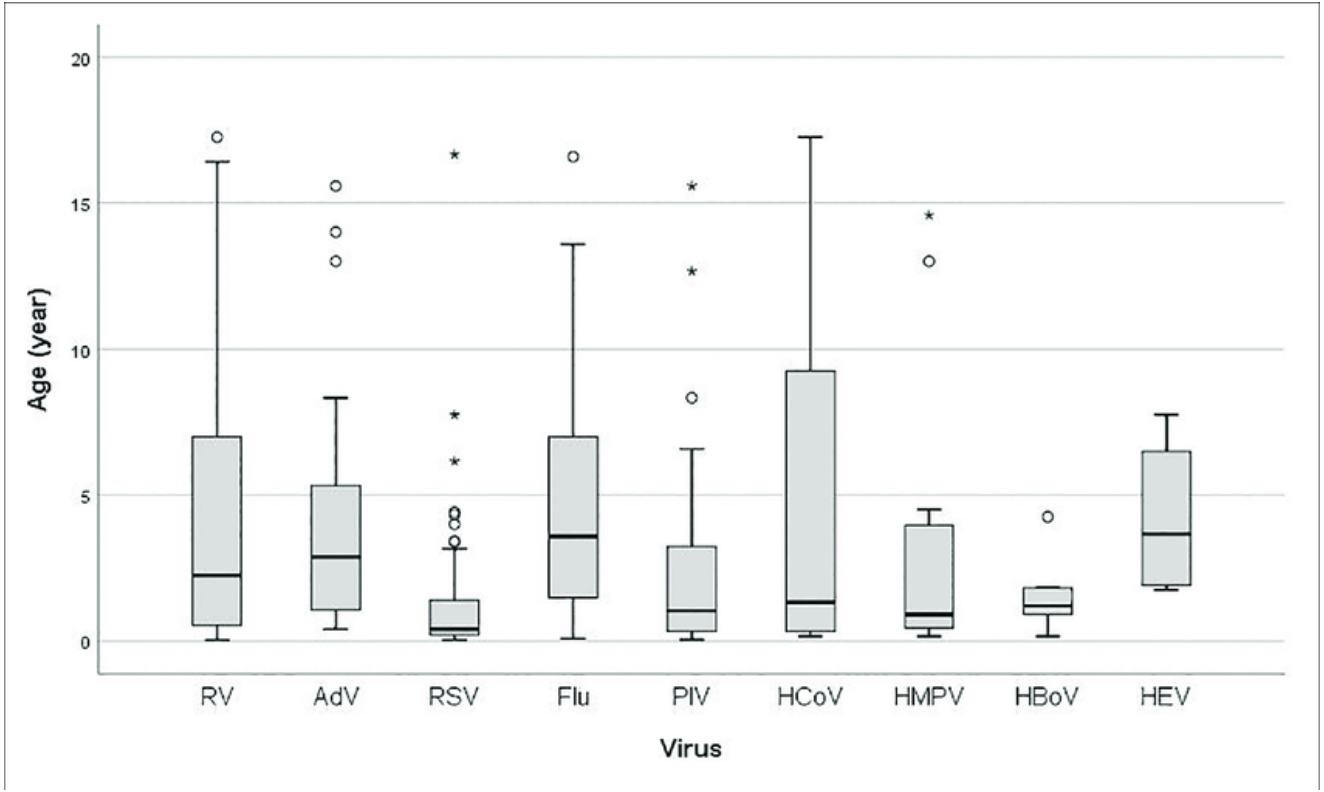


Box Plot

- Combines many concepts https://chartio.com/resources/tutorials/what-is-a-box-plot/
  - Provides a 5 number summary in 1 graph
    - Minimum
    - Maximum
    - Median (Average)
    - First Quartile (25%)
    - Third Quartile (75%)



Box Plot



Thank You!

Reach out anytime:

Jordan Krell
jkrell@flintsciencefair.org