UDACITY Google spark

Palestine Launchpad Program

Data Analysis NanoDegree

Our agenda for today

1. Recap about the last session
2. Project 2 Statistics - Updated
3. QA from Google form
4. Selenium web driver example
5. Knowledge Sharing (image processing)
6. Case Study: Twjeehe 2024 dataset
7. QA

Palestine Launchpad
منصة فلسطين التقنية
Google

Recap

What did we talk about last session

UDACITY

Last Session we covered the following:

- Project 2 Statistics
- QA from Google form:
  - Facet plots
- Bar chart vs count chart
- Solve the line chart notebook

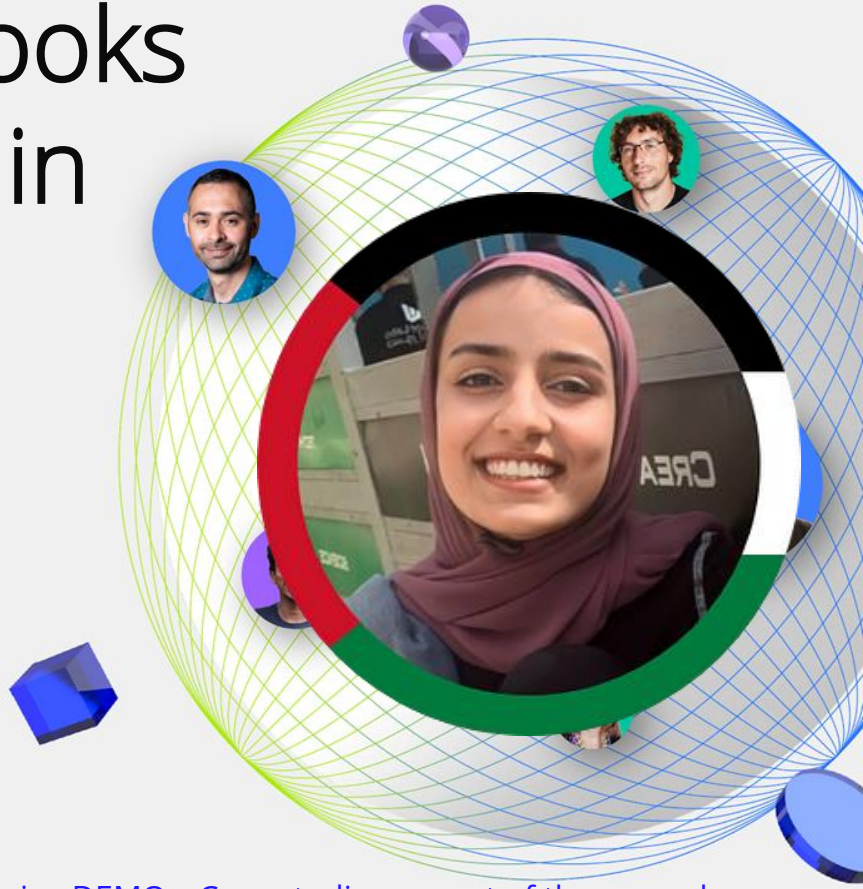
Last Session we agreed on the following:

To ensure you're keeping up, please:

- ❑ We discussed diving deeper into the FacetGrid x attribute
- ❑ I’ve uploaded the Twjeehe 2024 dataset to the Day12 folder on Google Drive. Analyze it and visualize the trends—let’s see your data insights shine! 📊

UDACITY

All the notebooks are gathered in one Place



[eng-aomar/Data-Analyst-nd-shared-notebooks: This repository contains DEMOs, Case studies as part of the second Palestine Launchpad with Google, Data Analyst Nanodegree. Session Lead by Alaa' (github.com)]

Palestine Launchpad
منصة فلسطين التقنية
Google

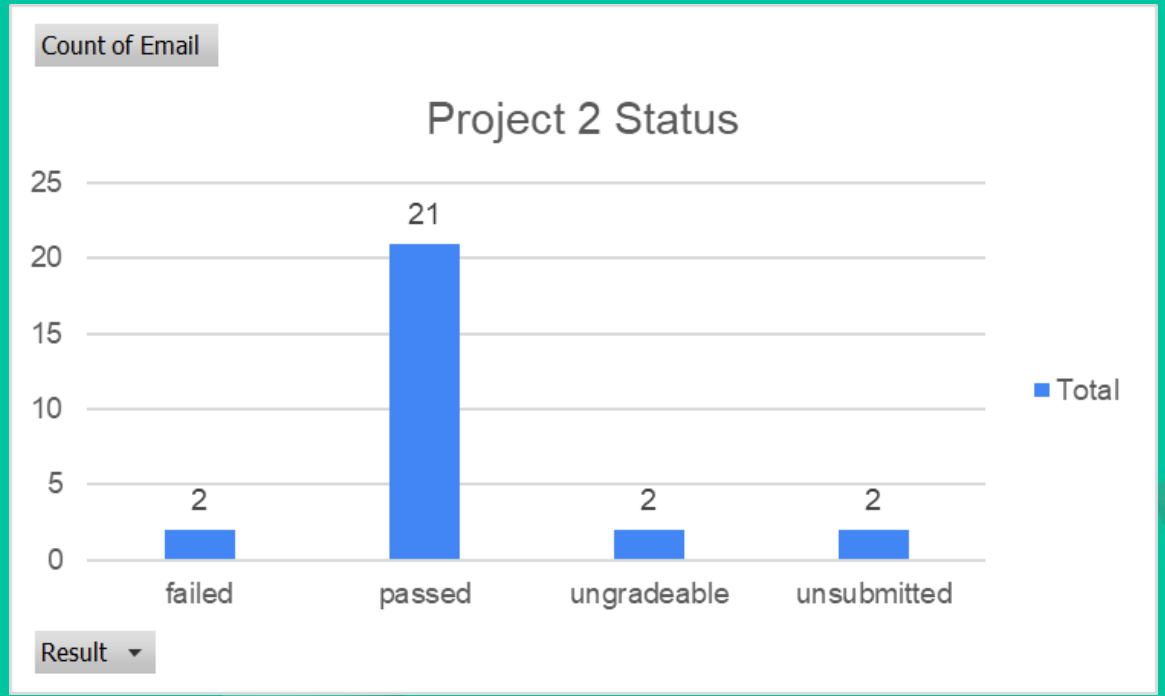
Project Two

Statisitcs about project tow results

UDACITY

UDACITY
Part of Accenture

SUBMIT



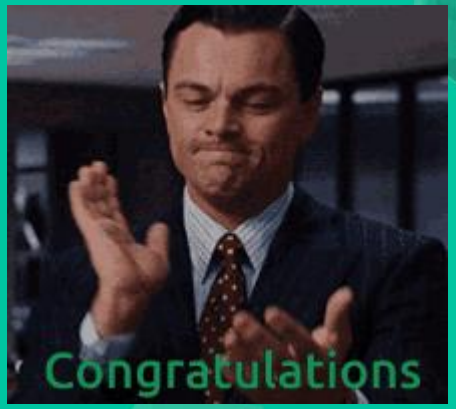
Your project successfully passed review!

Congratulations on passing your Real World Data Wrangling with Python project. Your project review is ready. Be sure to read through your full review and rate your reviewer.

Read Review

Then, keep the momentum going and continue learning!

Project two Statistics



		Attempts	Avg attempt
Passed	21	50	2.4
Total	**27**	**54**	2.0
Interventions	6	4	0.7
Pass rate	77.78%		

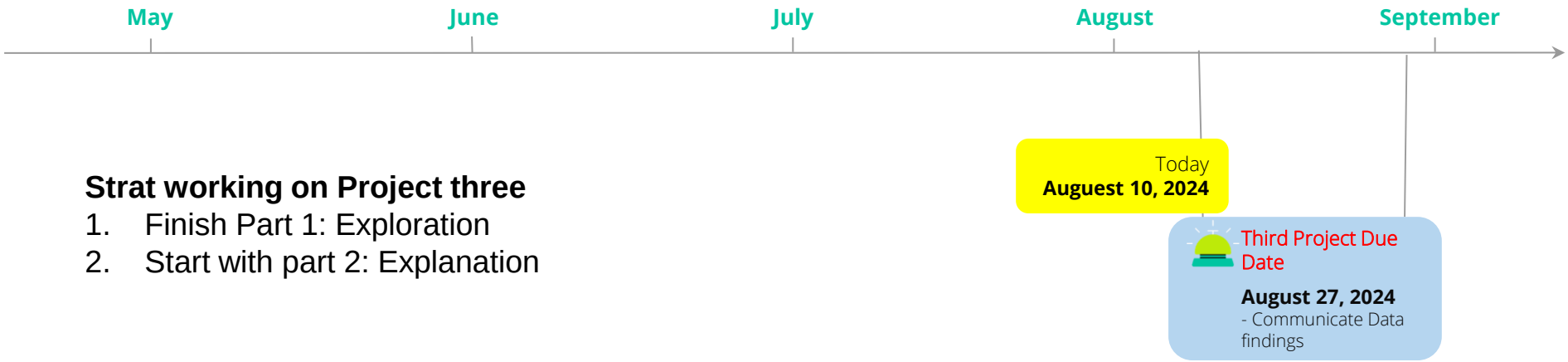
Palestine Launchpad
منصة فلسطين التقنية
Google

Project 3

Due Date

UDACITY

Third Project August 27, 2024



Strat working on Project three

1. Finish Part 1: Exploration
2. Start with part 2: Explanation

Program Period
May 9, 2024 - September 14, 2024

Palestine Launchpad
منصة فلسطين التقنية
Google.org

Data Gathering

Selenium WebDriver

UDACITY

Selenium WebDriver notebook

web_scraping_using_selenium.ipynb - Colab (google.com)

Palestine Launchpad
منصة فلسطين التقنية
Google

Bivariate Analysis

Charts Types

UDACITY

Bivariate Exploration

1. Line Chart
2. Scatterplots - required
3. Box Plots - required
4. Clustered Bar Chart
5. Heatmap

Google Form Feedback and QA

What Problems did you face? Please write details.

- - I'm facing a little trouble with Faceting maybe we could talk about it to get used to it since it's a very new topic to me, such as when do I use it and how to interpret it.

  - I think that I also have a hard time interpreting Multivariate visualization can we have examples of interpreting them, please?

  - Can we also have some examples for the violin plots (examples of interpreting them).

What is the difference between correlation and causation when analyzing the relationship between two variables? And what is the best plot to visualize each one of them ?

Do you have any comments, notes, please let me

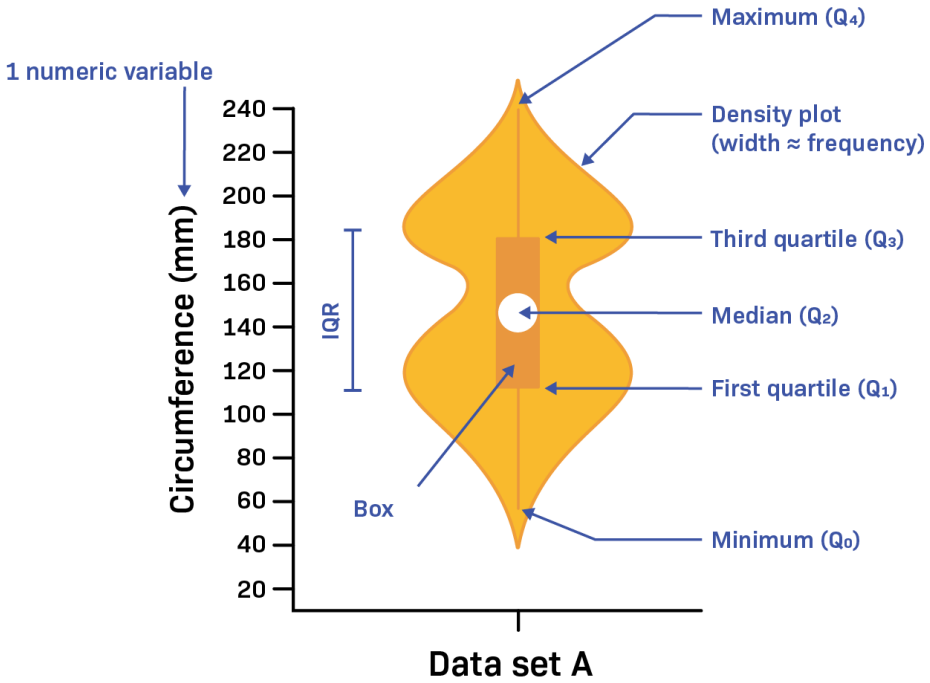
- About how to start work as data analyst, the starter point
- What is the difference between facet grid and pair grid and where to use each one through our analysis ?

Violin Plot

- A violin plot plays a similar role as a box-and-whisker plot. It shows the distribution of data points after grouping by one (or more) variables. Unlike a box plot, each violin is drawn using a kernel density estimate of the underlying distribution.

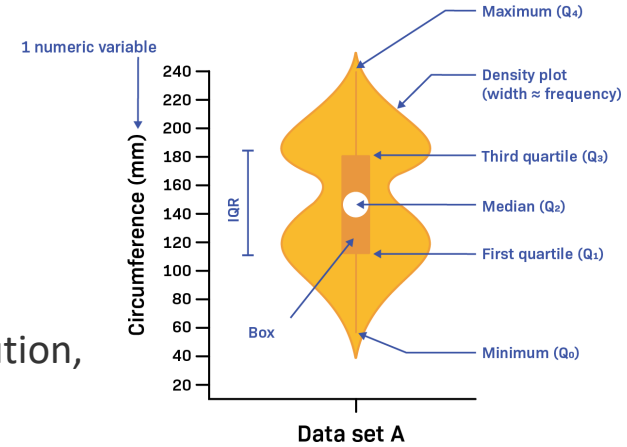
- seaborn.violinplot — seaborn 0.13.2 documentation (pydata.org)
- Visualizing categorical data — seaborn 0.13.2 documentation (pydata.org)
- https://colab.research.google.com/drive/18NCzIqzRfU9VjVpMKF2CWEEPZtbWZz0t?usp=sharing
- How to Interpret Violin Charts - LabXchange

Violin Plot


1 numeric variable
Circumference (mm)
240
220
200
180
160
140
120
100
80
60
40
20
IQR
Box
Maximum (Q4)
Density plot
(width ≈ frequency)
Third quartile (Q3)
Median (Q2)
First quartile (Q1)
Minimum (Q0)
Data set A

When looking at a violin plot, it's important to note the following:

- A violin plot shows how a data set varies along one variable by combining a boxplot with a PDF.
- The boxplot summarizes the center and spread:
  - The white dot in the center of the box represents the median.
  - The length of the box represents the interquartile range (IQR).
  - The length of the line that extends out of the box represents the range.
- A PDF of the data set, which shows the shape of the distribution, is centered and symmetrically arranged along the boxplot.
- A violin plot does not show sample size, so if you are comparing groups with very different sample sizes, it is important to explicitly note the sample sizes.



QUESTION(S) THIS TYPE OF DATA HELPS TO ANSWER:

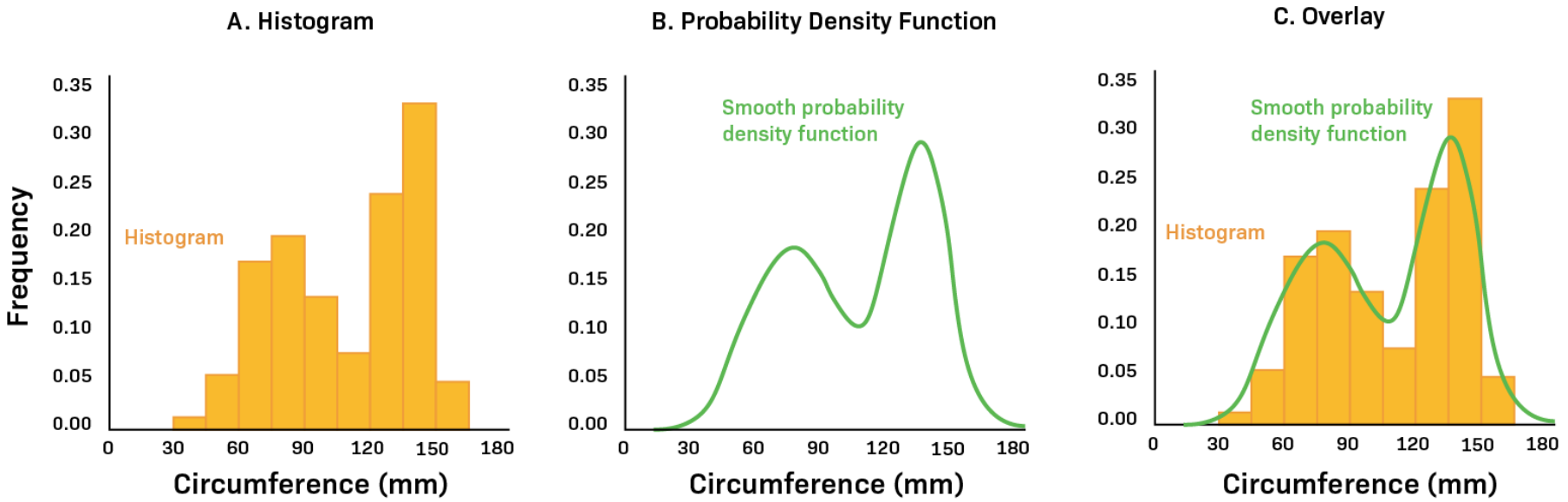
- What is the center of the data? Where is the median? How many modes does my data set have?
- What is the spread of the data? What are the range and IQR?
- What is the shape of my data? Is it symmetrical, skewed, uniform, or multimodal?
- How different are two (or more) data sets?

I. Violin plots allow for quickly approximating where the data is centered and how it is spread.

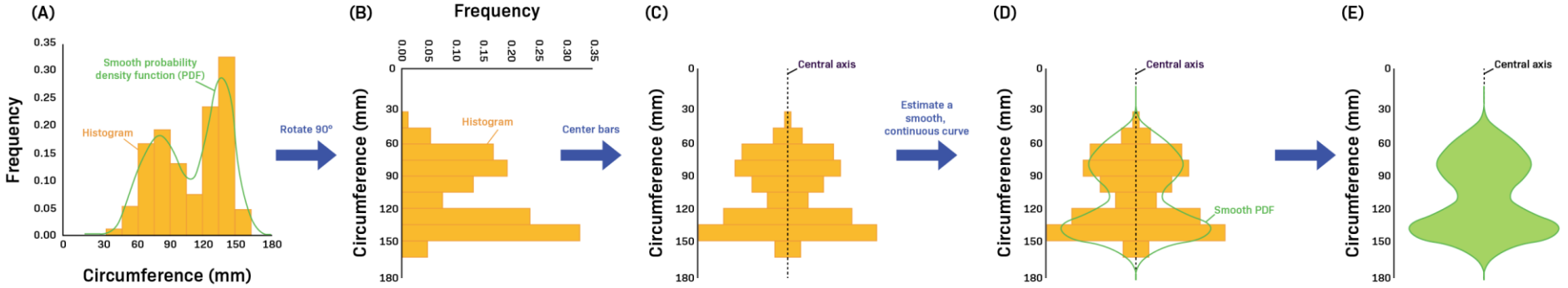
Since a violin plot includes a boxplot, the center and spread can be read just as they would be using a boxplot.

II. Violin plots show the shape of the distribution by including a probability density function.

A violin plot is a boxplot with a probability density function (PDF) added on top of it. A PDF is essentially a smoothed histogram indicating how frequently each value occurs, as shown in Figure 2. Compared to a histogram, a PDF provides a smoother distribution by smoothing out the noise.



An example of a probability density function. A histogram is presented in panel A. A probability density function (PDF) of the same data set is presented in panel B. A PDF provides a smoother distribution than a histogram. In panel C, the PDF is overlaid on top of the histogram to show how it smooths out and captures the general shape of the distribution described by the histogram.



The violin shape of a violin plot can be thought of as a rotated, centered, and smoothed histogram. Violin plots use a probability density function (PDF) to show the shape of a data set. The violin shape comes from starting with a histogram, rotating it 90°, centering the bars along a central axis, and smoothing the shape by estimating a PDF. In this example, the violin shape in panel E is proportional to the histogram in panel A. The last step in making a violin plot is to add a boxplot along the central axis.

KEY TAKEAWAYS

- ✓ A violin plot combines the strengths of a boxplot with those of a histogram and uses a smoothed probability density function (PDF) rather than a histogram to avoid the subjectivity of binning.
  - ✓ Violin plots include a boxplot, and identifies the minimum, first quartile, median, third quartile, and maximum in the same way that a boxplot does.
- ✓ Violin plots are a great way of visualizing multimodal data.
- ✓ Violin plots allow for quick graphical examination and comparison of one or more data sets.

How to Interpret Violin Plots

Similar to histograms, violin plots show the shape of a data set. However, unlike histograms, **violin plots show the shape of a data set by using a Probability Density Function (PDF),** or a density plot, which is effectively a smoothed-over histogram. The width of the PDF describes how frequently that value occurs in the data set. The wider regions of the density plot indicate values that occur more frequently.

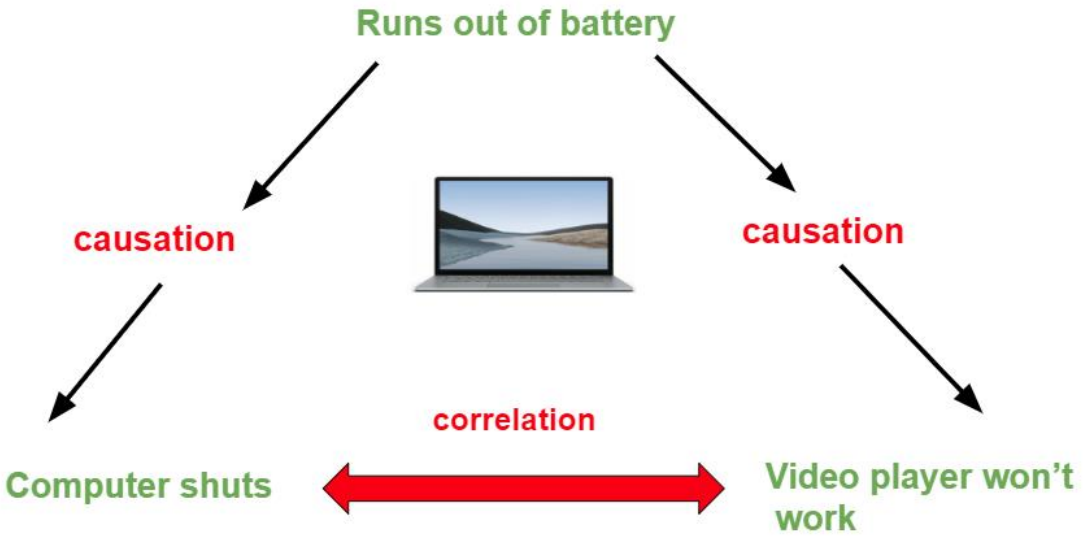
Narrower regions of the density plot indicate values that occur less frequently. In addition to showing the shape of a data set, a violin plot also summarizes a data set using five summary values: the minimum, first quartile, median, third quartile, and maximum. Violin plots combine the utility and versatility of a boxplot with a histogram's ability to show the underlying distribution.This allows violin plots to be useful when graphing multimodal data. Violin plots are also useful for visualizing multiple distributions at once for comparison.

Violin Plot [Simply explained] (youtube.com)

Violin Notebook

- https://colab.research.google.com/drive/18NCzIqzRfU9VjVpMKF2CWEEPZtbWZz0t?usp=sharing

Correlation vs Causation

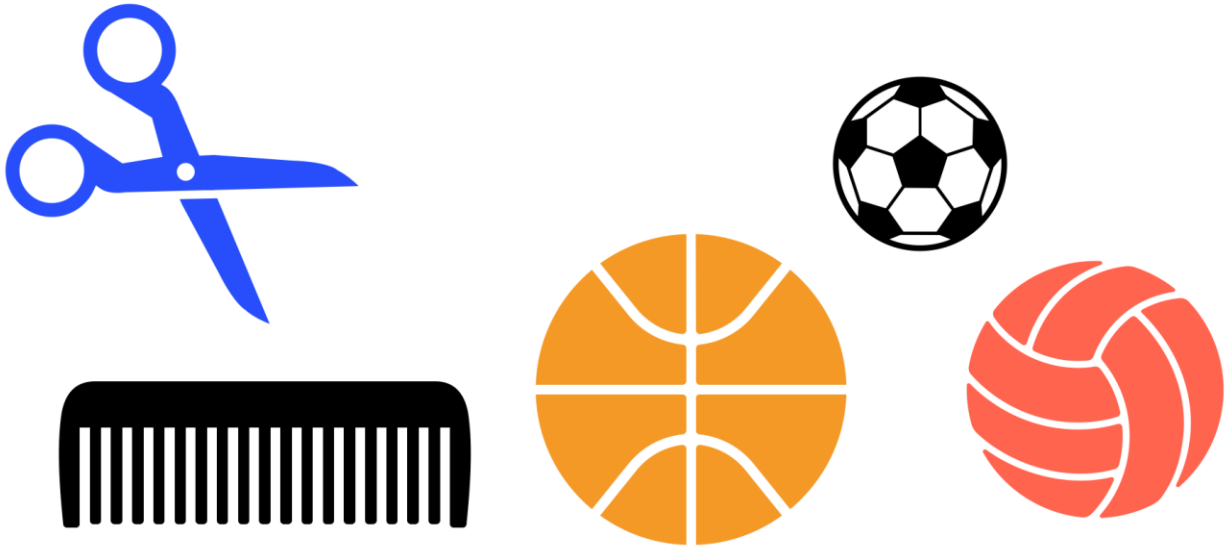


- Difference between correlation and causation (video) | Khan Academy

Correlation

A connection between two or more things that don't cause each other

For example, the amount of sports you play might have a connection or be similar to how often you get your hair cut



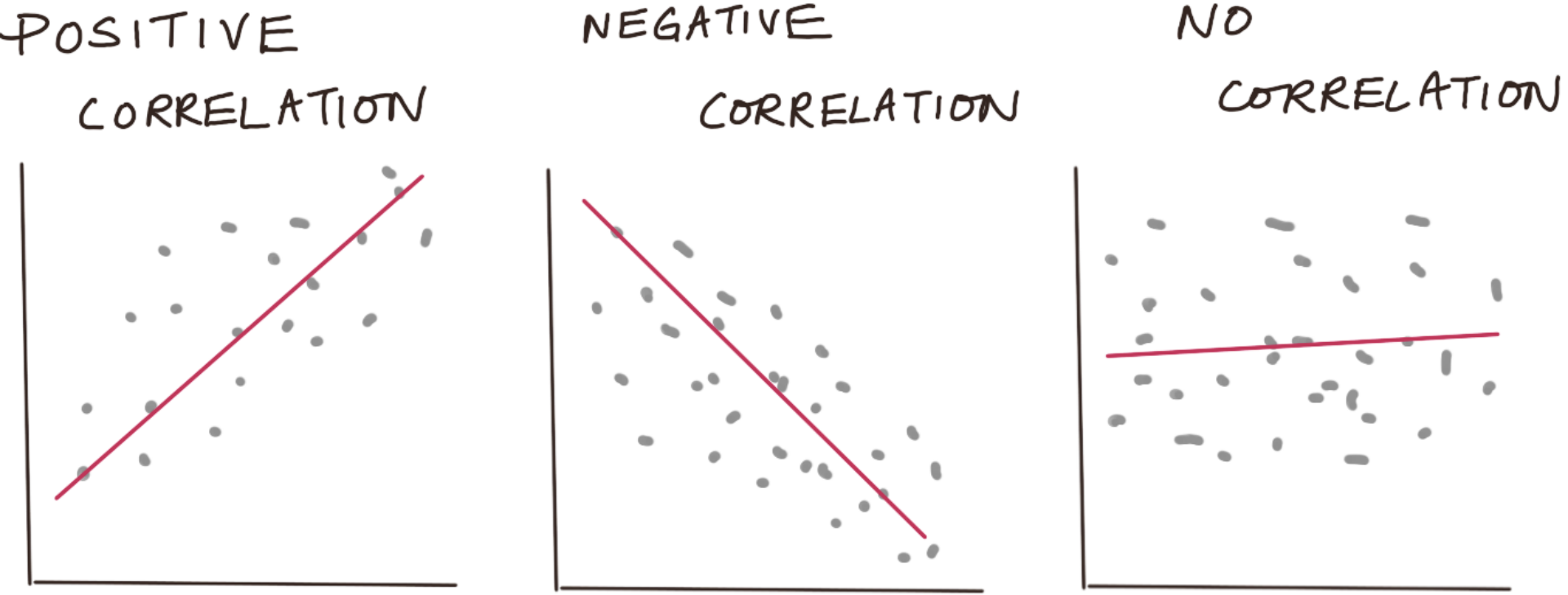
What is correlation?

- Correlation measures the linear relationship between variables. In a positive correlation, when the value of one variable goes up, the other does as well. When one variable goes down, the other variable descends, too.

Correlation Types

- You can represent the strength of the relationship between variables using a correlation coefficient ranging from -1 to +1, where the closer the linear relationship is to zero, the weaker the correlation is:
  - 1 = Perfect positive correlation
  - 0.5 = Weak positive correlation
  - **0 = Zero correlation**
  - -0.5 = Weak negative correlation
  - -1 = Perfect negative correlation

Correlation Types in Analysis



Correlation Types in Analysis

There are several statistics that you can use to quantify correlation. In this tutorial, you'll learn about three correlation coefficients:

- Pearson's r
- Spearman's rho
- Kendall's tau

Correlation Types in Analysis

Pearson’s coefficient measures linear correlation, while the Spearman and Kendall coefficients compare the ranks of data. There are several NumPy, SciPy, and pandas' correlation functions and methods that you can use to calculate these coefficients. You can also use Matplotlib to conveniently illustrate the results.

Demo

- NumPy, SciPy, and pandas: Correlation With Python – Real Python
- https://colab.research.google.com/drive/1vrn1myVRWtcILRgb2sEOx9bpo2j0sQbD?usp=sharing

Pair plot

- seaborn.pairplot — seaborn 0.13.2 documentation (pydata.org)

- https://colab.research.google.com/drive/1VMf1R-lMxHT_LSTVFfsEYvQj80o9sj2l?usp=sharing

Palestine Launchpad
منصة فلسطين التقنية
Google for

Knowledge Sahring

Udacity, Students and Session Leads

UDACITY

Image Processing

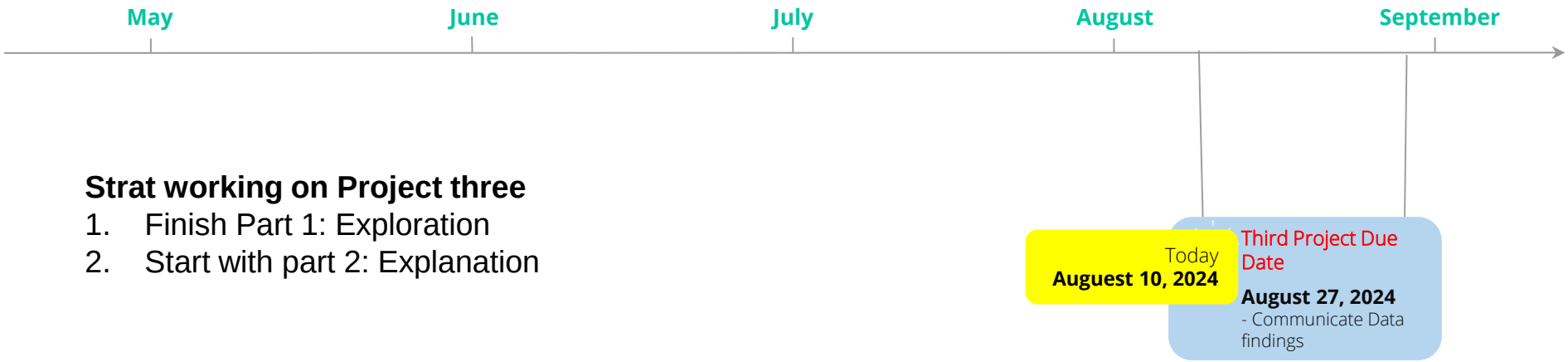
Palestine Launchpad
منصة فلسطين التقنية
Google

Next Steps

Let's get things started...

UDACITY

Second Project July 2, 2024



Strat working on Project three

1. Finish Part 1: Exploration
2. Start with part 2: Explanation

Palestine Launchpad
منصة فلسطين التقنية
Google

Thank You!

And Good Luck!

UDACITY