

Data Visualization

Visual tools for understanding your data

ICME Fundamentals of Data Science – Summer Workshop Series

Put in the chat where you're joining from today!



A decorative horizontal bar at the top of the slide, consisting of an orange segment on the left and a blue segment on the right, both with a slight gradient.

hello!

I'm Kaleigh Mentzer

ICME Summer Workshop Instructor

I'm Thanawat Sornwanee

ICME Summer Workshop Assistant

Introductions

Your Instructor: Kaleigh Mentzer



granica

ICME



Your Workshop Assistant: Thanawat Sorwanee



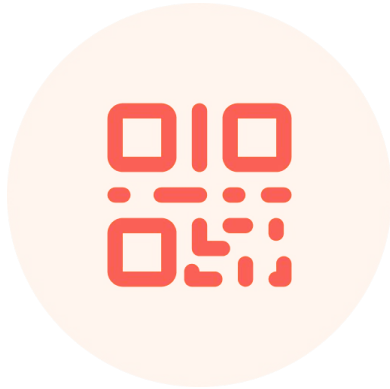
ICME

Course Website and Resources



bit.ly/icme-vis

slido



Join at slido.com
#1315187

① Click **Present with Slido** or install our [Chrome extension](#) to display joining instructions for participants while presenting.

slido



What industry are you in?

ⓘ Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

slido



How much experience do you have with Python?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

slido



How much experience do you have with data visualization?

① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

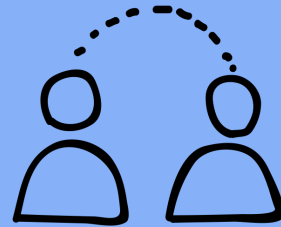
Course Plan

Workshop Plan:

- Data visualization in Python
- Taking you through the experience of starting with a new data set to communicating your insights
- Primarily focusing on tabular data



Day 1: Why Data Visualization and
Exploratory Data Analysis



Day 2: Data Visualization for
Communication

Workshop Plan: Day 1 – Why Data Visualization and Exploratory Data Analysis

- Intros and Course Plan
- Python as a tool for data visualization
- Exploratory Data Analysis (EDA)
 - What is it? Why is it important?
 - Nominal, Ordinal, and Quantitative data
 - Python data profiling tools
 - EXERCISE: Applying these tools to sample data
 - Missing values and outliers
 - Exploring correlations with pair plots
- Plotting basics in Python

Workshop Plan: Day 2 – Data Visualization for Communication

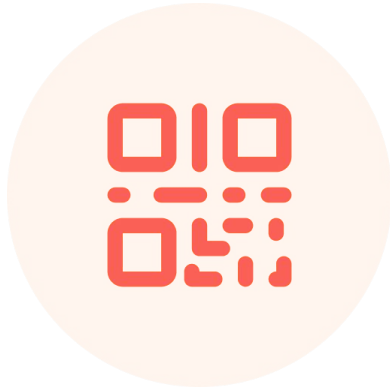
- Human perception and data visualization
- Improving on basic plots in Python
 - EXERCISE: Develop day 1 visualization
- Data visualization evaluation
 - Good and bad data visualization examples
 - EXERCISE: Peer feedback on visualizations
- Wrap up and additional resources

Demo Code and Data + Exercise Code and Data



- Link is on course website
- Demo code is available for reuse
- You'll be applying what I show to a new dataset

slido



Join at slido.com
#1315187

- ① Click **Present with Slido** or install our [Chrome extension](#) to display joining instructions for participants while presenting.

slido



Audience Q&A Session

- ① Click **Present with Slido** or install our [Chrome extension](#) to show live Q&A while presenting.

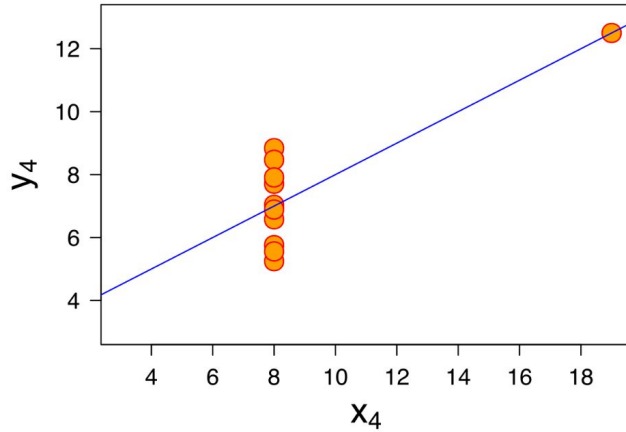
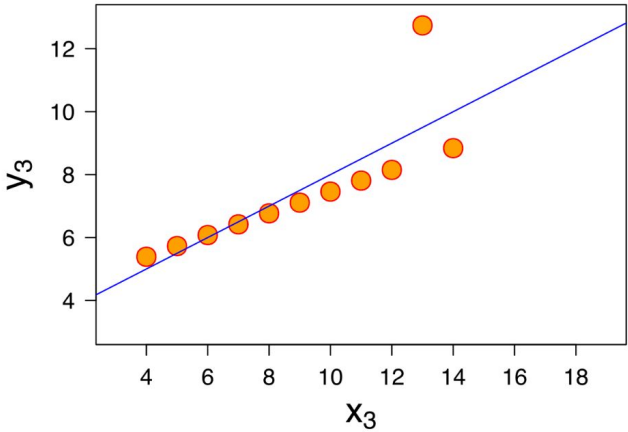
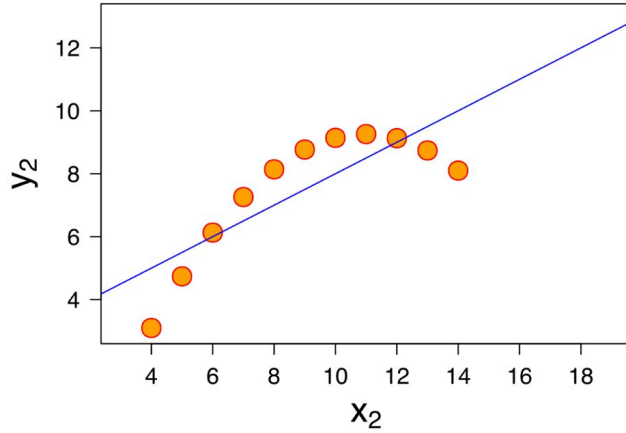
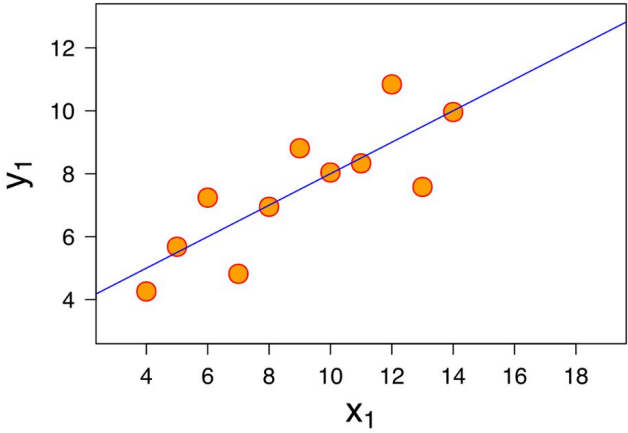
Why Data Visualization?

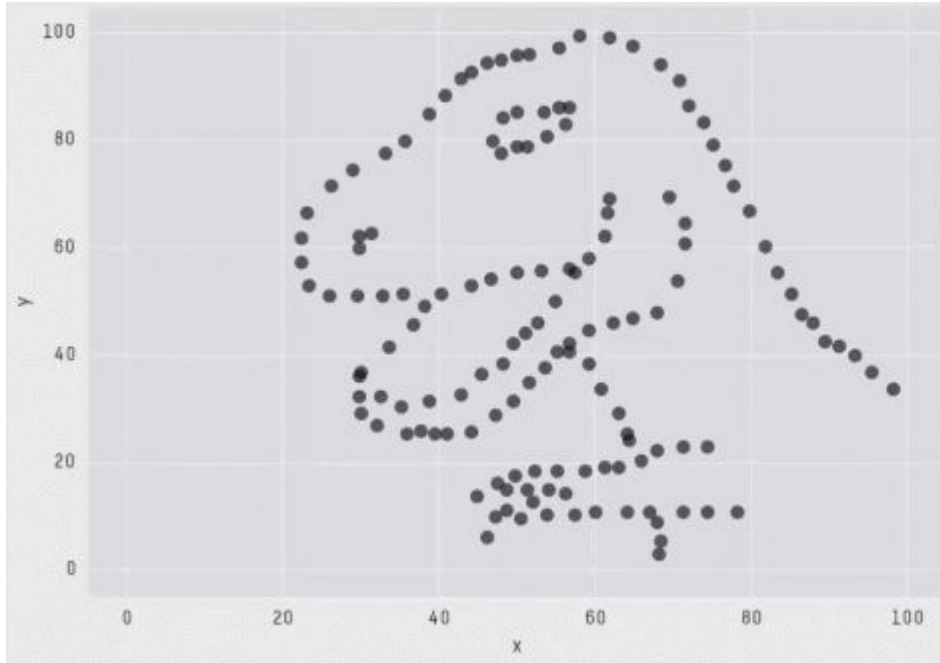
What do you notice about this data?

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	± 0.003
Correlation between x and y	0.816	to 3 decimal places

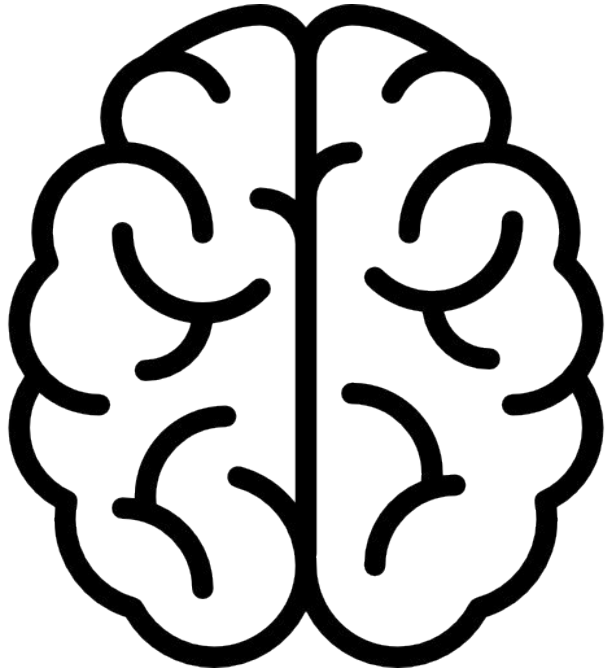
Anscombe's Quartet





```
X Mean: 54.2659224  
Y Mean: 47.8313999  
X SD   : 16.7649829  
Y SD   : 26.9342120  
Corr.  : -0.0642526
```

It's really hard to understand your data without visualization.



The human brain is exceptionally good at pulling out patterns visually – take advantage of that!



Key Use #1:
Analyzing Data

Visualization also plays an important role in communicating data insights.



Just like it's easier for you to understand data by seeing it, the same holds true for others.



Key Use #2:
Data Communication

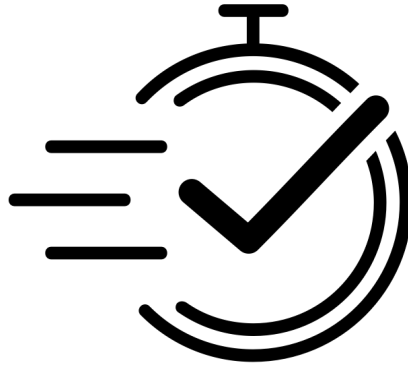
Additional considerations for designing for others – must be able to stand alone without the contextual understanding you gained from using the data

Python Data Visualization

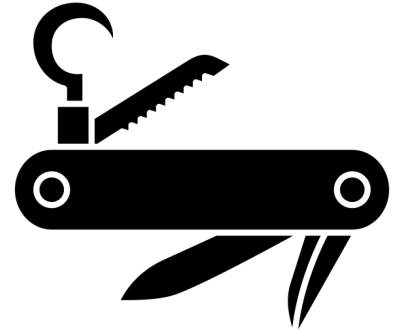
Why Python is good for data visualization



Convenience: You're likely already using Python for your data science needs

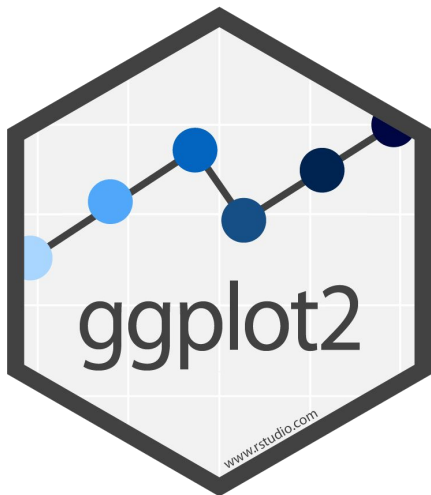


Speed: There are built-in functions that help plot data quickly

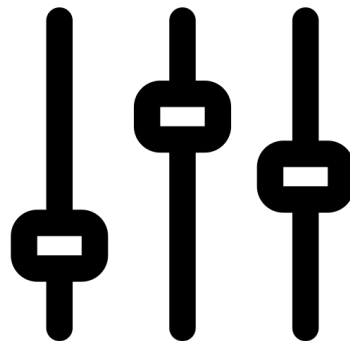


Versatility: There are packages for most of your data visualization needs

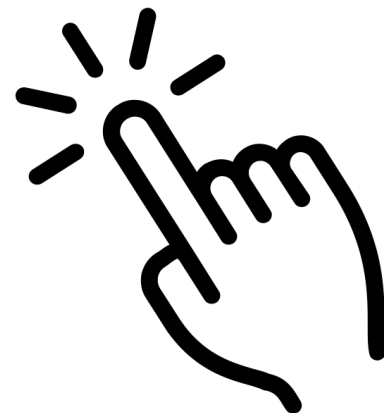
Python's data visualization shortcomings



Aesthetics: Less beautiful defaults than R



Customizability: Not as customizable as JavaScript/D3



Interactivity: Poor at sharable interactive plots

Our plotting packages: Matplotlib and Seaborn

matplotlib

The bread and butter of Python visualization (built-in package)

```
import matplotlib.pyplot as plt
```

 **seaborn**

Add on package for better aesthetics and more plotting functionality

```
import seaborn as sns  
sns.set_theme()
```

Our data handling package: Pandas



Data handling library in Python for tabular data

```
import pandas as pd
```

Has some built in data visualization tools!

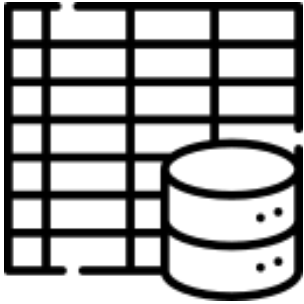
Exploratory Data Analysis

You're starting a new data-driven project.

- Ultimate goal might be:
 - Finding new business opportunities by leveraging patterns in data
 - Building a machine learning model
 - Developing a new research hypothesis
- You've just acquired a new data set.

What do you do first?

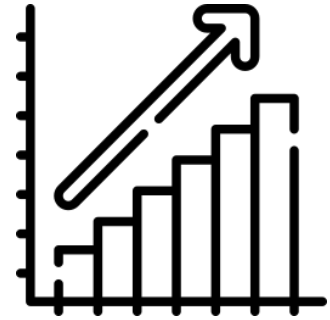
How do you begin to work with a new dataset?



Understand what data
you actually have



Ask and answer
questions with the
data



Find meaningful
patterns

Conclusions can be **misleading** or **wrong** if you do not understand your data well.

The exploratory data analysis process

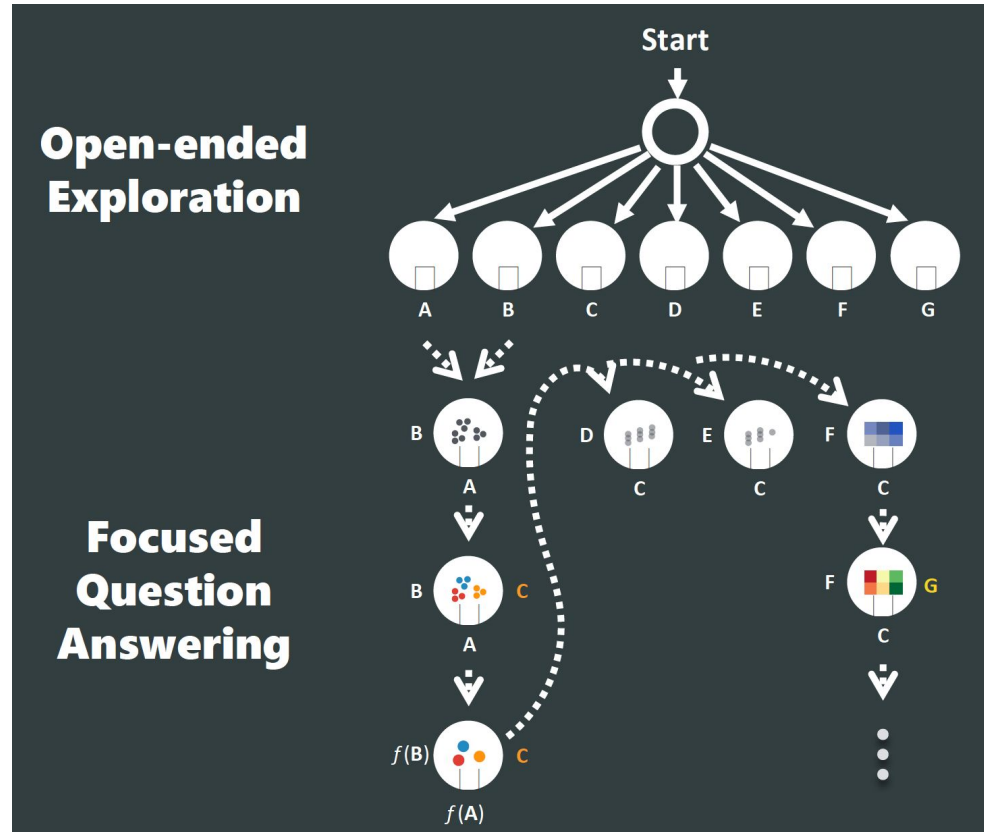
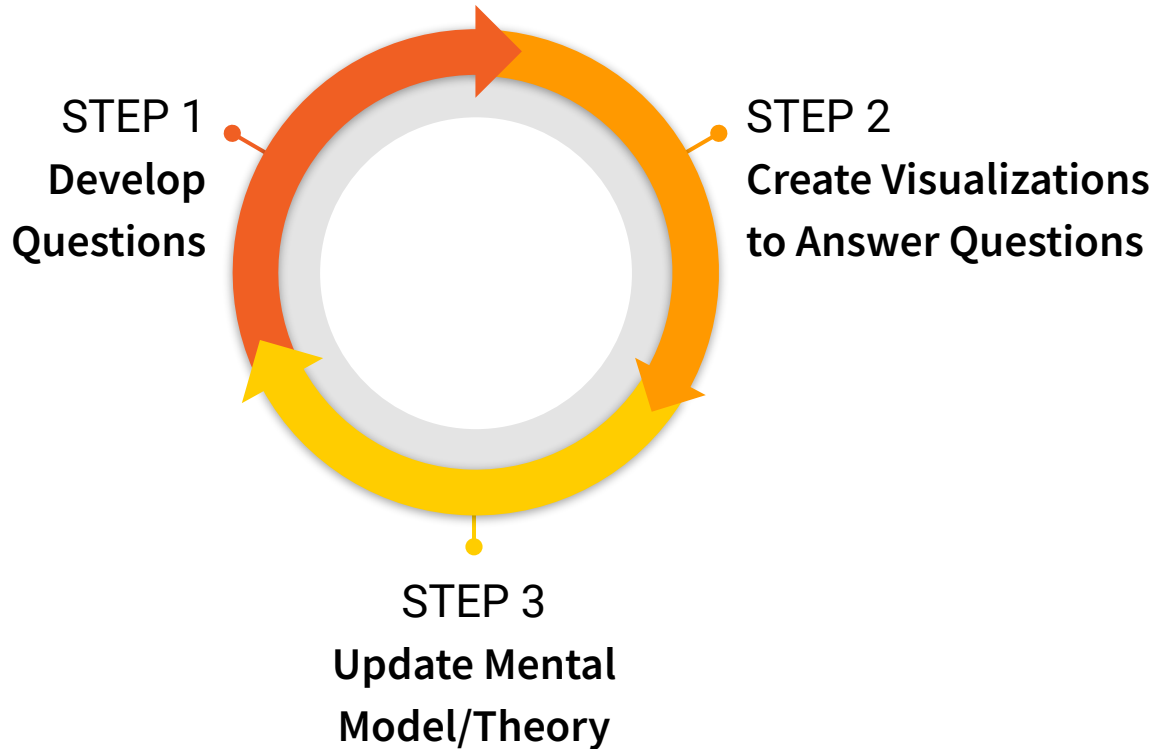


Figure from Hari Subramonyam, adapted from Hullman

The exploratory data analysis process

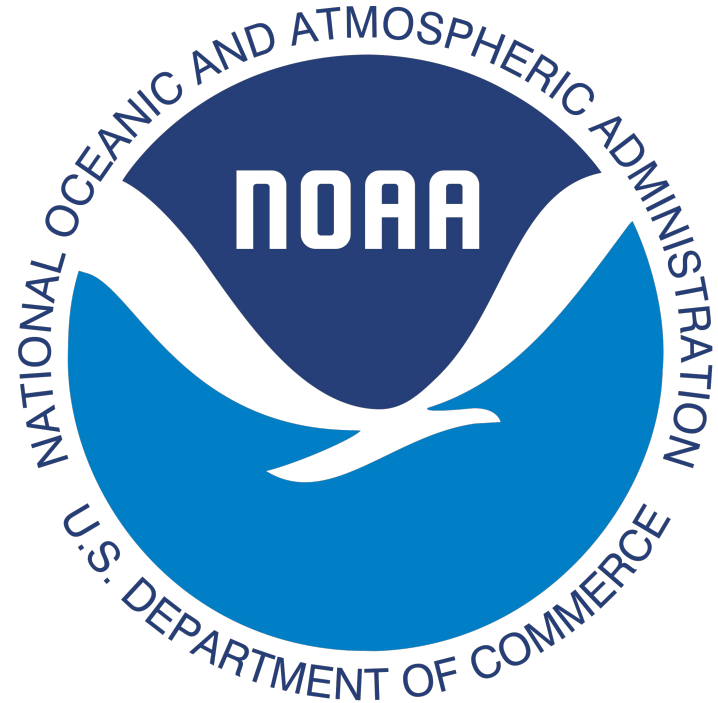


Data visualizations here don't have to be pretty or sophisticated – they're just for you to understand your data.

We're going to go through these exercises with some data on weather anomalies.

United States weather from 2013

Randomly sampled 5,000 points from throughout the year



What does the data look like?

	date	station_name	degrees_from_mean	longitude	latitude	max_temp	min_temp	type
0	2013-09-11	HARTFORD BRAINARD FLD	20.35	-72.6506	41.7361	34.4	21.7	Strong Hot
1	2013-07-16	BOISE LUCKY PEAK DAM	6.92	-116.0542	43.5253	37.8	21.1	Weak Hot
2	2013-10-04	WINTHROP UNIV	7.42	-81.0317	34.9381	30.0	13.9	Weak Hot
3	2013-11-28	WHITING FLD NAS	-12.15	-87.0167	30.7167	11.1	-3.2	Weak Cold
4	2013-06-30	TIMPANOGOS CAVE	10.43	-111.7075	40.4447	35.6	20.0	Weak Hot
5	2013-07-01	SEWARD AP	-4.90	-149.4167	60.1283	11.7	8.9	Weak Cold
6	2013-05-14	LOGAN UTAH ST UNIV	11.26	-111.8033	41.7456	30.6	13.3	Weak Hot

Data Dictionary – What do the columns mean?

Column	Field Description
date	The date of the weather anomaly. (Date)
degrees_from_mean	The number of degrees that the temperature was above or below the monthly mean temperature. (Float)
longitude	The longitude of the weather station where the anomaly was recorded. (Float)
latitude	The latitude of the weather station where the anomaly was recorded. (Float)
max_temp	The maximum temperature (C) recorded at the weather station on the date of the anomaly. (Float)
min_temp	The minimum temperature (C) recorded at the weather station on the date of the anomaly. (Float)
station_name	The name of the weather station where the anomaly was recorded. (String)
type	The type of anomaly, either high or low temperature. (String)

Data Attributes Types

Type	Definition	Example
Nominal	Labels or categories	Cat, Dog, Hippo
Ordinal	Ordered values that don't tell you more than order	Agree, Somewhat Agree, Somewhat Disagree, Disagree
Quantitative	A number that gives you information about differences (interval) or proportions (ratio)	5 miles, 75% correct

QUIZ: What types are the columns of our weather anomaly data?

slido



What data type is "max_temp", the maximum temperature recorded at the weather station on the date of the anomaly?

- ① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

slido



What data type is "station_name", the name of the weather station where the anomaly was recorded?

- ① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

slido



What data type is "type", the type of anomaly, either high or low temperature?

- ① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

slido



What data type is "date", the date of the weather anomaly?

- ① Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.

What types are the columns of our weather anomaly data?

Column	Data Attribute Type
date	Quantitative (interval)
degrees_from_mean	Quantitative (interval)
longitude	Quantitative (interval)
latitude	Quantitative (interval)
max_temp	Quantitative (interval)
min_temp	Quantitative (interval)
station_name	Nominal
type	Ordinal

Why do we care?

The types of visualizations we can create depend on the type of the data.

Coffee Break

15:00



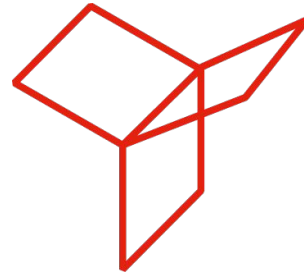
A Pandas Crash Course

- `df.columns` - to see list of columns
- `df["column_A"]` - to select a single column
- `df[["column_A", "column_B", "column_A"]]` - to select multiple columns
- `df.loc[df["column_A"] < 100]` - to filter by a column value

< Code Demo >

A quick overview of your data – profiling packages

- Provide a quick overview of your data with minimal effort
- Identifies:
 - Missing/invalid data
 - Duplicates
 - Data quality alerts
 - Univariate patterns
 - Correlations



YData

Formerly



**PANDAS
PROFILING**

< Code Demo >

Try it yourself!

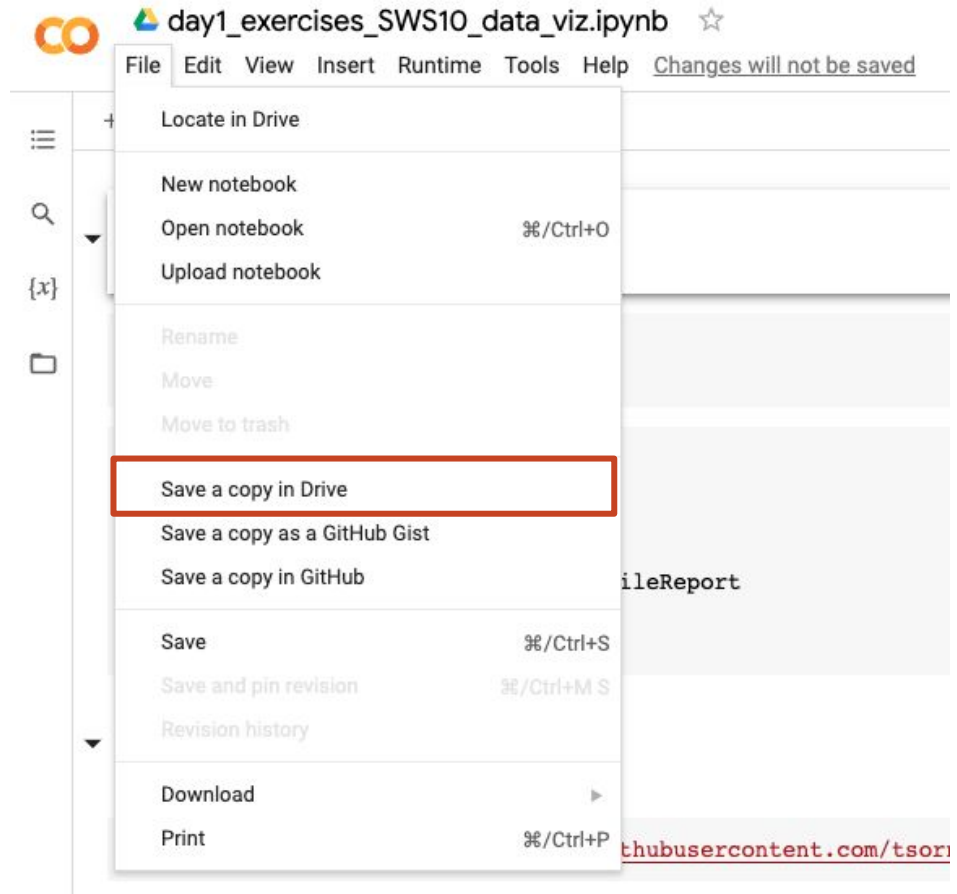


Airline Delays measured by the US Bureau of
Transportation Statistics

Exercise Notebook available on
course website or
bit.ly/icme-vis-exercise1

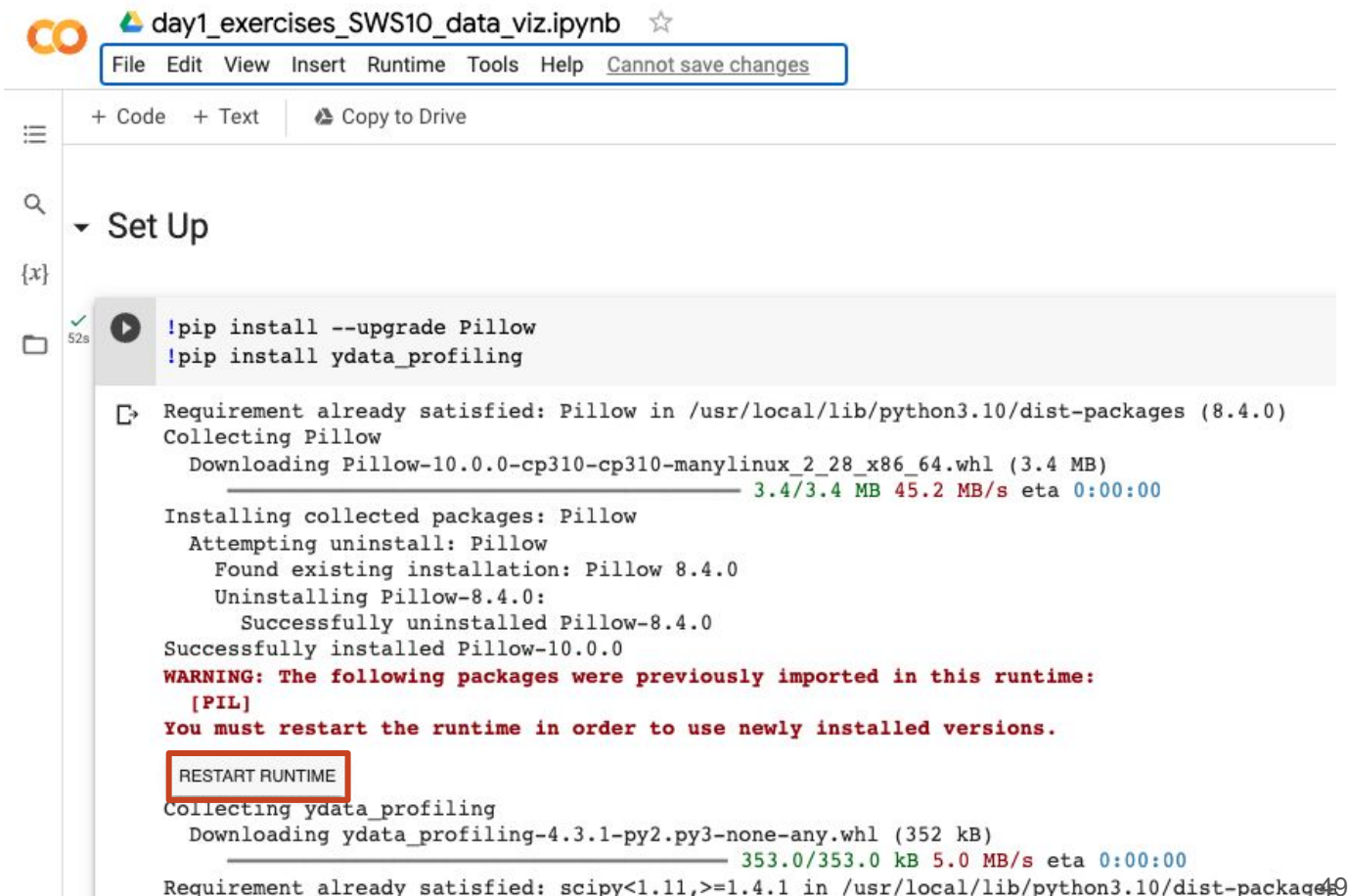
*Hint: The demo notebook is also
on the website - feel free to copy
the syntax from there!*

Creating your personal copy



The screenshot shows the Google Colab interface for a notebook titled "day1_exercises_SWS10_data_viz.ipynb". The "File" menu is open, displaying various options. The option "Save a copy in Drive" is highlighted with a red rectangular border. Other visible options include "Locate in Drive", "New notebook", "Open notebook" (with keyboard shortcut ⌘/Ctrl+O), "Upload notebook", "Rename", "Move", "Move to trash", "Save a copy as a GitHub Gist", "Save a copy in GitHub", "Save" (with keyboard shortcut ⌘/Ctrl+S), "Save and pin revision" (with keyboard shortcut ⌘/Ctrl+M S), "Revision history", "Download", and "Print" (with keyboard shortcut ⌘/Ctrl+P). The background shows a sidebar with navigation icons and a portion of the notebook content, including a "FileReport" section.

One small workaround



The screenshot shows a JupyterLab notebook titled "day1_exercises_SWS10_data_viz.ipynb". The interface includes a menu bar with "File", "Edit", "View", "Insert", "Runtime", "Tools", "Help", and "Cannot save changes". Below the menu bar, there are buttons for "+ Code", "+ Text", and "Copy to Drive". The notebook content is organized into a "Set Up" section. A terminal cell is active, showing the execution of two pip commands: `!pip install --upgrade Pillow` and `!pip install ydata_profiling`. The output of the first command shows that Pillow 8.4.0 was already installed and is being upgraded to 10.0.0. The output of the second command shows that ydata_profiling 4.3.1 is being installed. A red box highlights the "RESTART RUNTIME" button at the bottom of the terminal cell.

```
!pip install --upgrade Pillow
!pip install ydata_profiling

Requirement already satisfied: Pillow in /usr/local/lib/python3.10/dist-packages (8.4.0)
Collecting Pillow
  Downloading Pillow-10.0.0-cp310-cp310-manylinux_2_28_x86_64.whl (3.4 MB)
    3.4/3.4 MB 45.2 MB/s eta 0:00:00
Installing collected packages: Pillow
  Attempting uninstall: Pillow
    Found existing installation: Pillow 8.4.0
    Uninstalling Pillow-8.4.0:
      Successfully uninstalled Pillow-8.4.0
  Successfully installed Pillow-10.0.0
WARNING: The following packages were previously imported in this runtime:
[PIL]
You must restart the runtime in order to use newly installed versions.
RESTART RUNTIME
Collecting ydata_profiling
  Downloading ydata_profiling-4.3.1-py2.py3-none-any.whl (352 kB)
    353.0/353.0 kB 5.0 MB/s eta 0:00:00
Requirement already satisfied: scipy<1.11,>=1.4.1 in /usr/local/lib/python3.10/dist-packages
```

Click on the play button
or click on the cell and
run Cmd+Enter (Mac)
or Ctrl+Enter
(Windows)

10:00

The word "colab" is written in a bold, lowercase, sans-serif font. The letters "c", "o", and "a" are orange, while "l", "a", and "b" are blue. A horizontal line with a gradient from orange to blue passes behind the text.

colab

Your Task

1. Get Google Colab notebook running.
2. Make a `ydata_profiling` report for the airline delay data.

slido



Audience Q&A Session

- ① Click **Present with Slido** or install our [Chrome extension](#) to show live Q&A while presenting.

What if we had missing values?

There's no universal answer here.

- Options:
 - Omission - Potential bias
 - Imputing values (filling with mean, median, etc.) – may not make sense
 - Visually representing missing values (e.g., in another color) – potentially distracting
- Remember to look for column values that could indicate missing values, like “NULL” or “NaN”

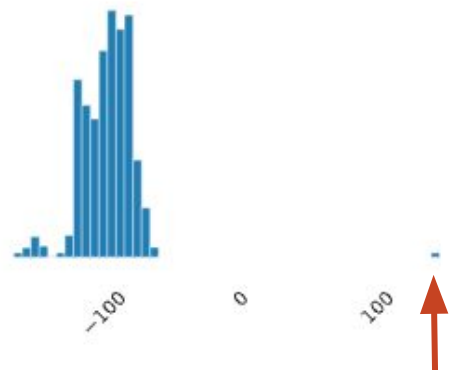
General Rule: Think about why the data is missing before you decide how to handle missing values!

What if we had outliers?

longitude

Real number (\mathbb{R})

Distinct	2128	Minimum	-166.5433
Distinct (%)	42.6%	Maximum	144.7961
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	4987
Infinite (%)	0.0%	Negative (%)	99.7%
Mean	-97.967102	Memory size	39.2 KiB



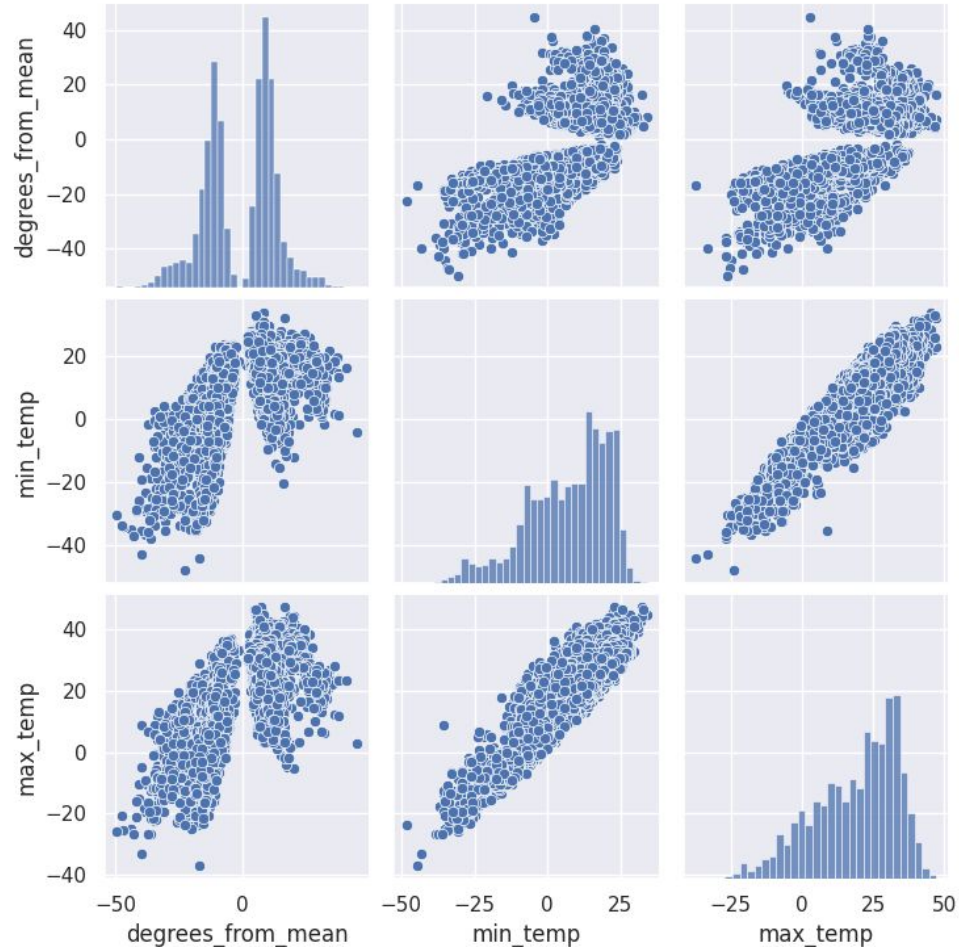
Dig in to weird features you see in the visualization.

```
df.loc[df.longitude > 0]
```

	date	station_name	degrees_from_mean	longitude	latitude	max_temp	min_temp	type
603	2013-04-17	GUAM INTL AP	1.79	144.7961	13.4836	32.2	25.0	Weak Hot
988	2013-03-30	GUAM INTL AP	3.38	144.7961	13.4836	31.7	26.1	Weak Hot
1027	2013-04-13	GUAM INTL AP	5.11	144.7961	13.4836	32.2	26.7	Strong Hot
1111	2013-04-16	GUAM INTL AP	2.39	144.7961	13.4836	32.8	25.0	Weak Hot
1412	2013-03-17	GUAM INTL AP	2.53	144.7961	13.4836	32.2	24.4	Weak Hot
1525	2013-05-14	GUAM INTL AP	4.12	144.7961	13.4836	32.2	27.8	Weak Hot
1833	2013-06-03	GUAM INTL AP	3.38	144.7961	13.4836	32.8	27.2	Weak Hot
2338	2013-06-29	GUAM INTL AP	3.38	144.7961	13.4836	32.2	27.2	Weak Hot
2752	2013-07-29	GUAM INTL AP	6.56	144.7961	13.4836	33.3	27.2	Strong Hot
3081	2013-05-15	GUAM INTL AP	4.12	144.7961	13.4836	32.8	27.8	Weak Hot
3215	2013-11-07	GUAM INTL AP	2.97	144.7961	13.4836	31.1	26.7	Weak Hot
3732	2013-07-25	GUAM INTL AP	4.35	144.7961	13.4836	32.8	27.8	Weak Hot
4460	2013-04-09	GUAM INTL AP	2.89	144.7961	13.4836	33.3	26.1	Weak Hot

Pair plots are useful for determining pairwise relationships.

```
sns.pairplot(df)
```



Once you have an idea about the structure of your variables, you can dig in to specific relationships.

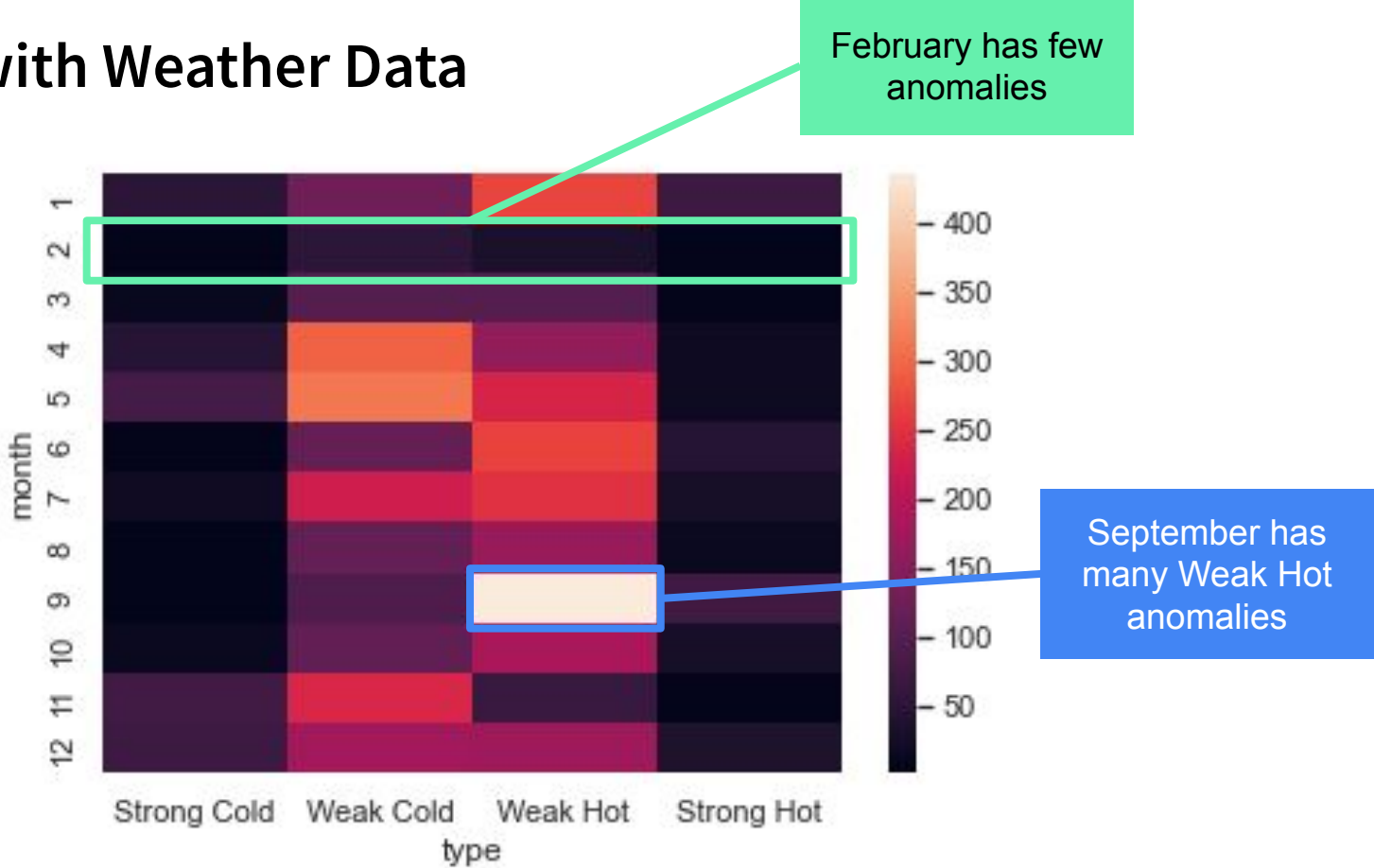
- Visualization type depends on (1) the number of variables and (2) the type of variable(s)
 - 1 Variable:
 - Quantitative: histogram
 - Nominal: bar chart
 - 2 Variable
 - Quantitative: scatter plot or line plot
 - Nominal: heatmap

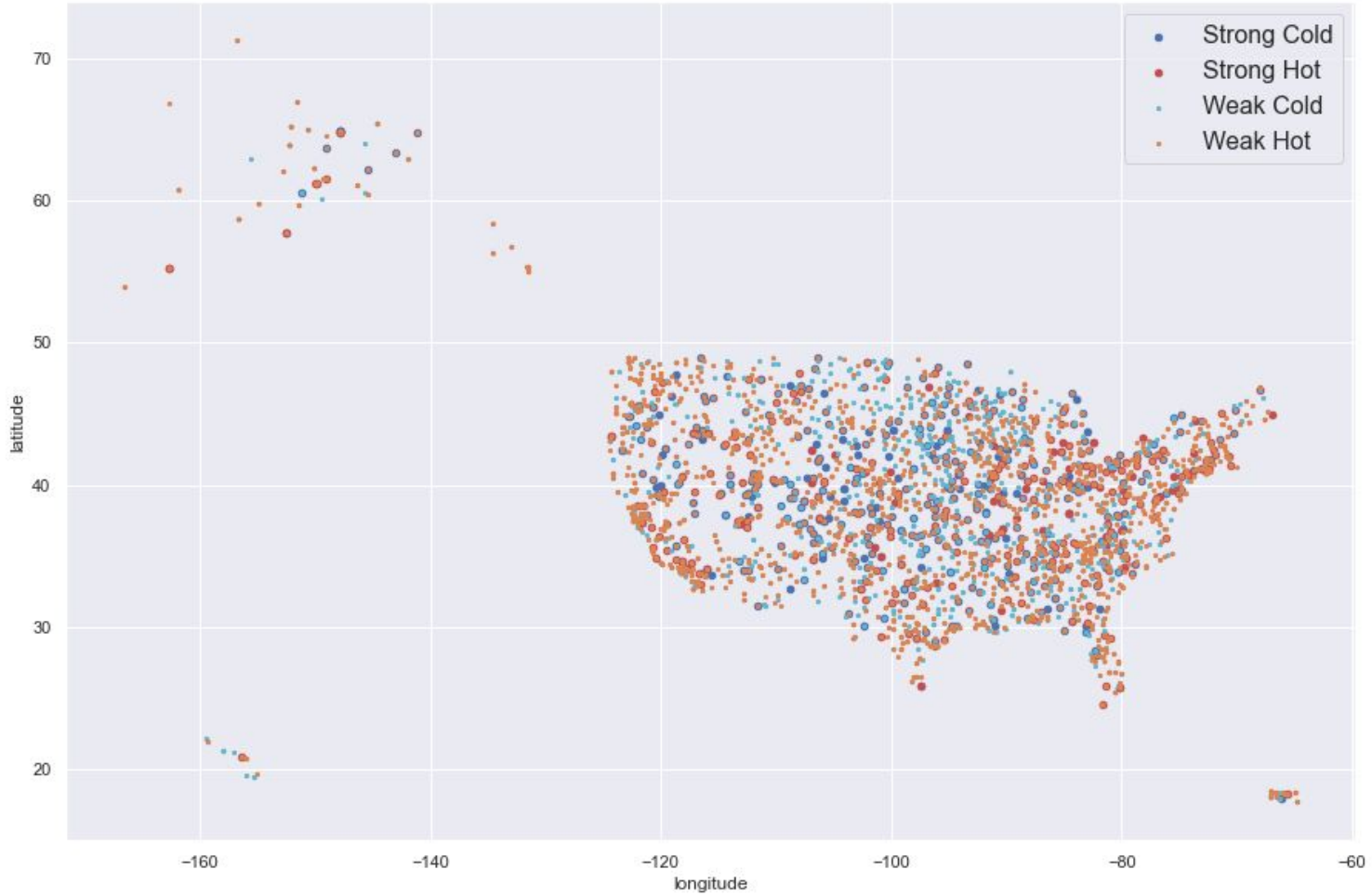
< Code Demo >

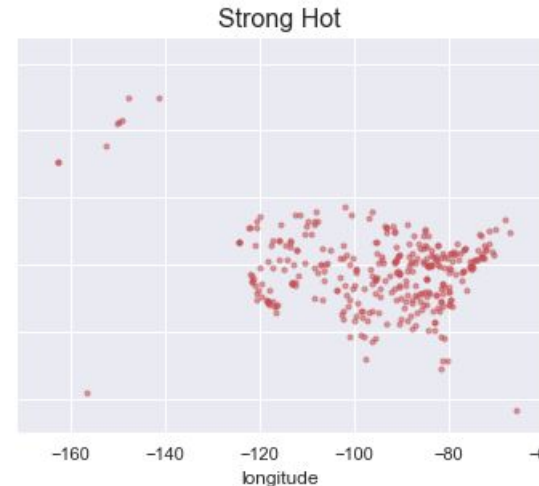
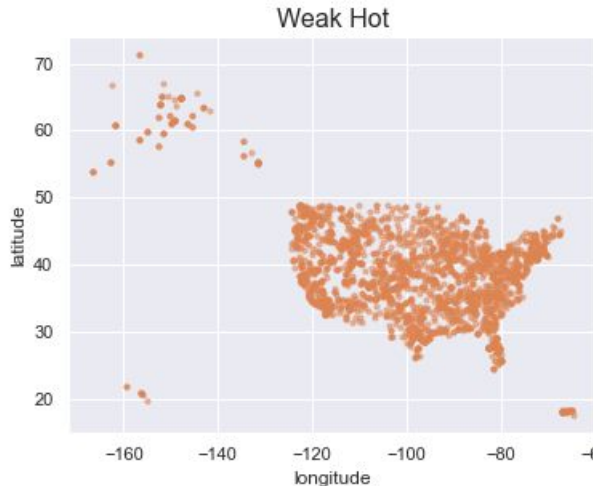
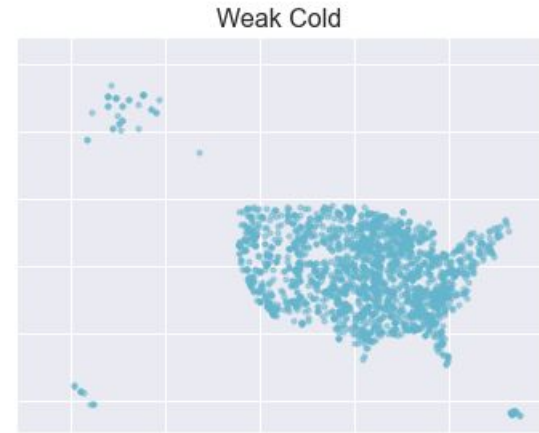
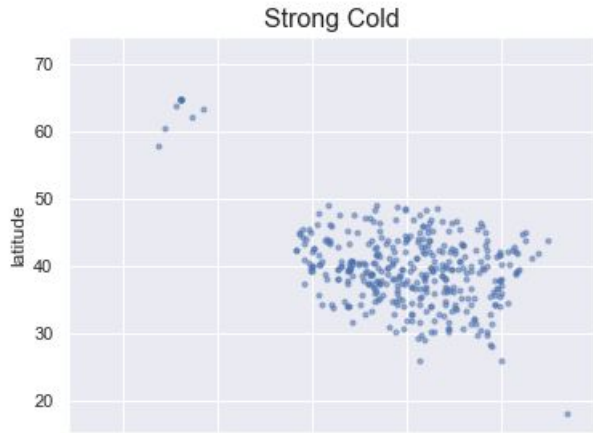
Your Homework

1. **Explore the Airline Delay dataset** using the tools you learned today.
2. **Create a scatter plot of departure delay vs arrival delay.** How correlated are the two? What does this suggest about why flights are delayed?
3. **Develop 1-3 candidate ideas** for a visualization to polish tomorrow.

Example with Weather Data







Your Homework

See you tomorrow at 1 pm PT!

Questions?
kmentzer@stanford.edu

1. **Explore the Airline Delay dataset** using the tools you learned today.
2. **Create a scatter plot of departure delay vs arrival delay.** How correlated are the two? What does this suggest about why flights are delayed?
3. **Develop 1-3 candidate ideas** for a visualization to polish tomorrow.