National Aeronautics and Space Administration

An Introduction to Data Visualization

Chris Heinich

Table of Contents

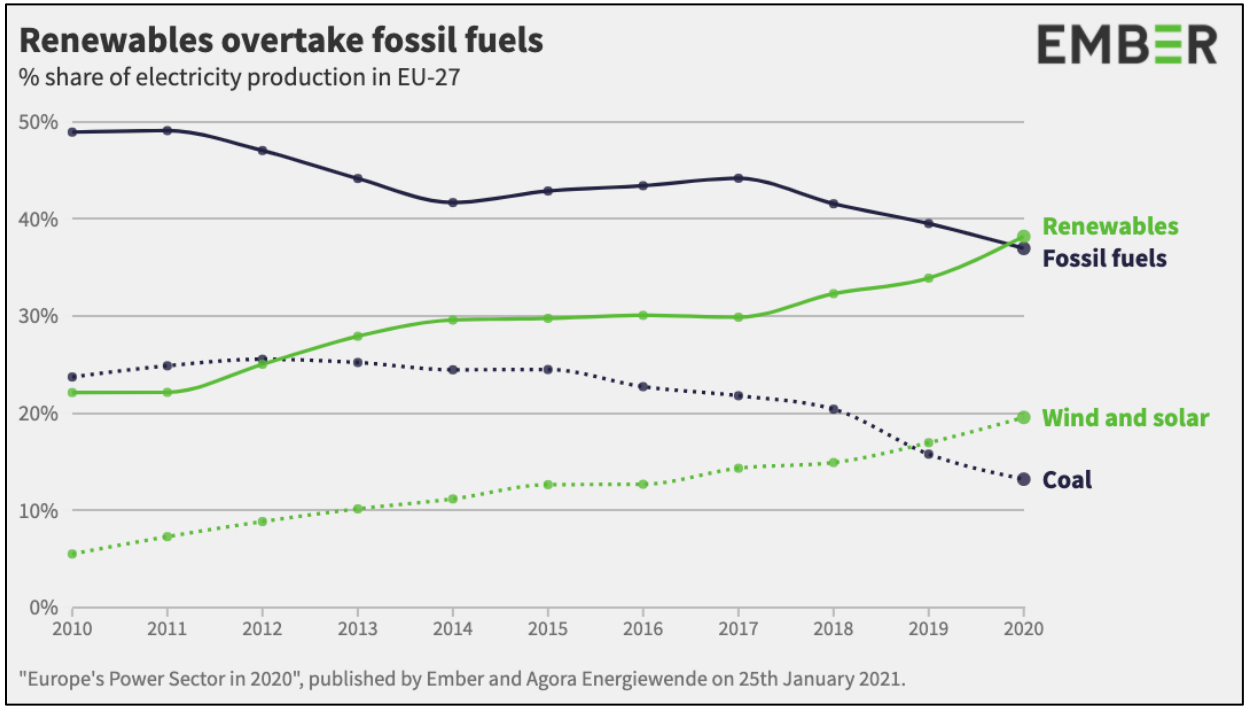
- Intro (What's Data Vis? Good? Bad?)
- Building Blocks
- Example Chart Types
- Design Tips for Charts and Dashboards
- Dashboard Example
- Sources/Further Reading

What is Data Vis? And Why?

- **Representing data and numbers with graphics**
- **Cross between mathematics, art, and communication**
- **Uses:**
  - Communicate findings
  - Explore data
  - Aid in analysis and discovery
- **Why not just show the numbers, or give to a computer?**
  - People are great at finding patterns in images. Less so with raw numbers
  - Computers are great at comparing numbers at once. But they need strict requirements.

What's Good Data Vis?

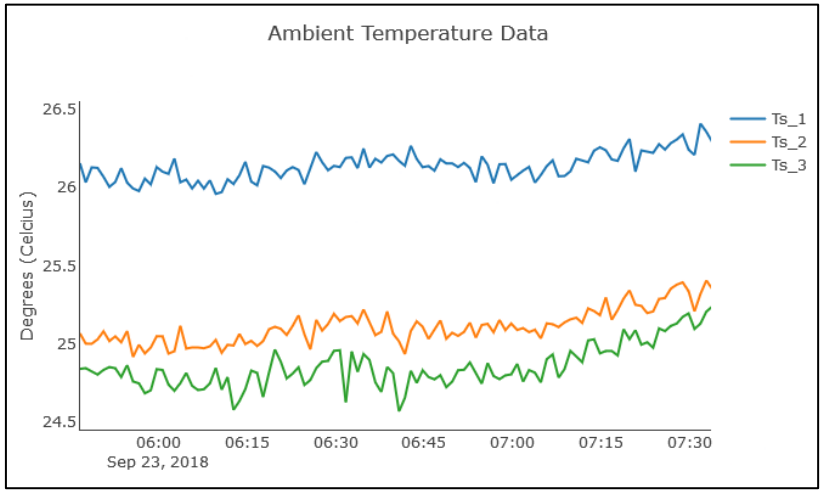
- **Built on correct data and calculations**
- **Provides an answer to a question your audience has**
- **Designed in a language the audience will understand**
  - Choose charts you don't have to explain
- **Summarizes data in an intuitive way**
  - Make the data seem simpler than it is
- **Brings attention to what's important**
  - Highlight the main point; hide the irrelevant
  - Gray can hide less important details
  - "10 Second Rule"



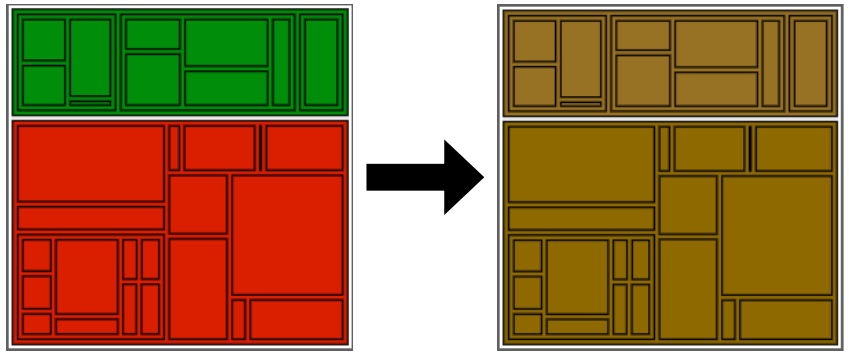
Jones, D. (2021, January 2021). EU power sector in 2020. Retrieved April 19, 2021, from https://ember-climate.org/project/eu-power-sector-2020/. (CC BY-SA 4.0 © Ember) https://creativecommons.org/licenses/by-sa/4.0/

What's Bad Data Vis?

- **Misleads or takes an answer out of context**
  - Careful on zooming and summarization
  - Ensure an "apples to apples" comparison
- **Overly complex**
  - Avoid making the users do math
  - Don't try to fit too many answers in one chart
- **Not accessible**
  - Hard to interact with; not color blind friendly; doesn't work on XYZ device
- **Interactivity not properly planned**
  - Need to ensure interactivity helps not hinders



This chart shows temperature readings of three sensors at NASA Langley Research Center. If we zoom in too much, the blue sensor looks significantly warmer than the others…even though the difference is only one degree.



Both tree maps show the same random data, generated by the Viz Palette project. They compare how two different people (one with no color deficiencies, and one with Deurteranopia) would see the same chart. Try it out yourself at https://projects.susielu.com/viz-palette?colors=["#008d0a","#da1e00"]

National Aeronautics and Space Administration

Chart Building Blocks/Terminology

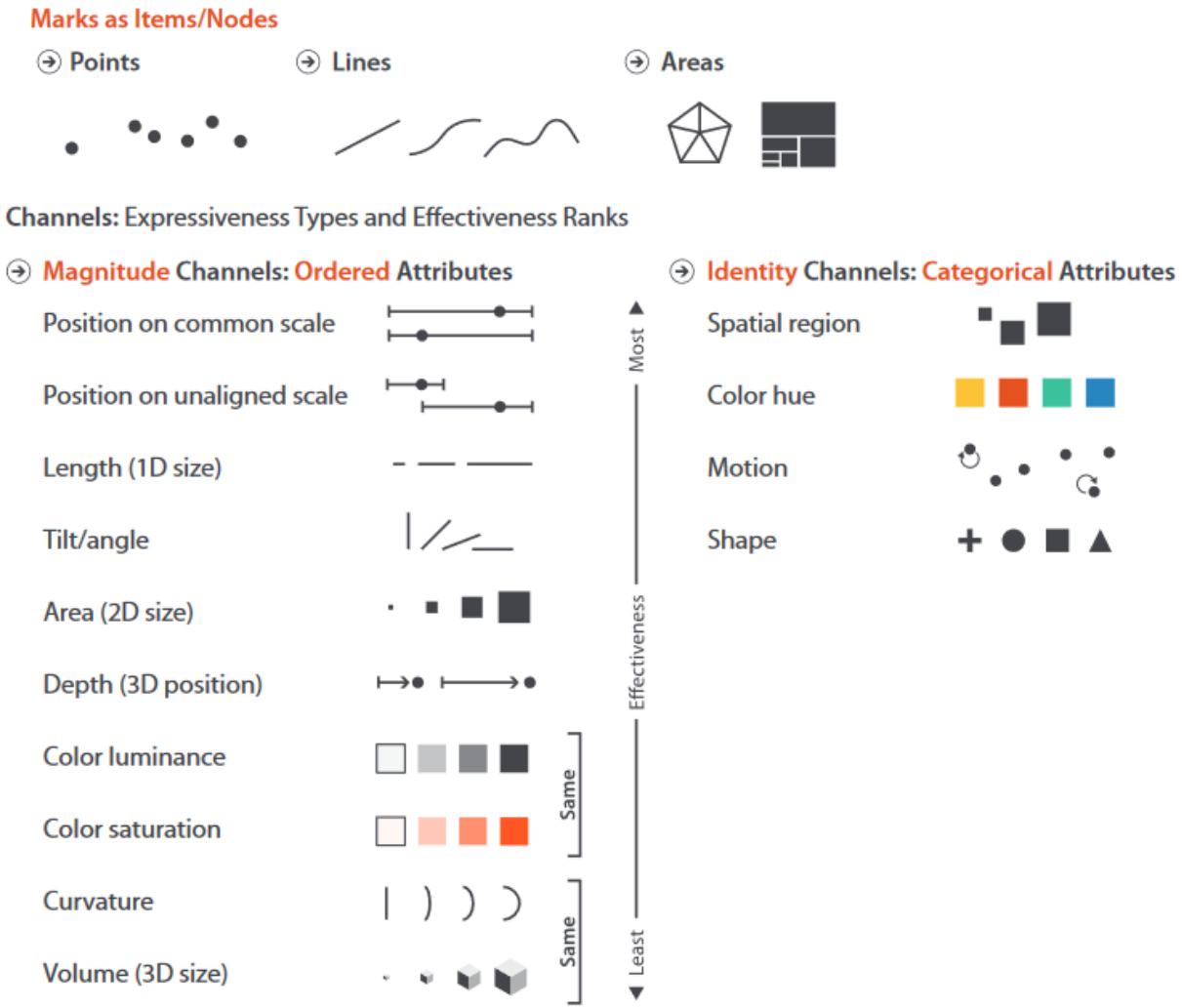
Thinking About Your Data

- Concept of “Items” (records) and “Attributes” (describes the record)
- Quantitative versus Qualitative
- Ordered? Sequential versus Diverging versus Cyclic
- Tabular? network? Hierarchal?
- Static versus Dynamic

Record ID	Date	Temperature (F)	Color	Location
1	12-2-2022	62	Green	Austin > Travis > Texas
2	12-3-2022	75	Yellow	San Marcos > Hays > Texas
3	12-4-2022	0	White	Kyle > Hays > Texas
4	12-5-2022	-12	White	Cocoa Beach > Brevard > Florida

Marks and Channels

- **Mark:** the representation of the item.
  - Example: a dot on a scatter plot
- **Channel:** the representation of an attribute value.
  - Example: location along the x/y coordinates; size of dot, color of dot
- Can use multiple channels for multiple attributes….but be careful!
- Not all channels are equal, and some will clobber others.

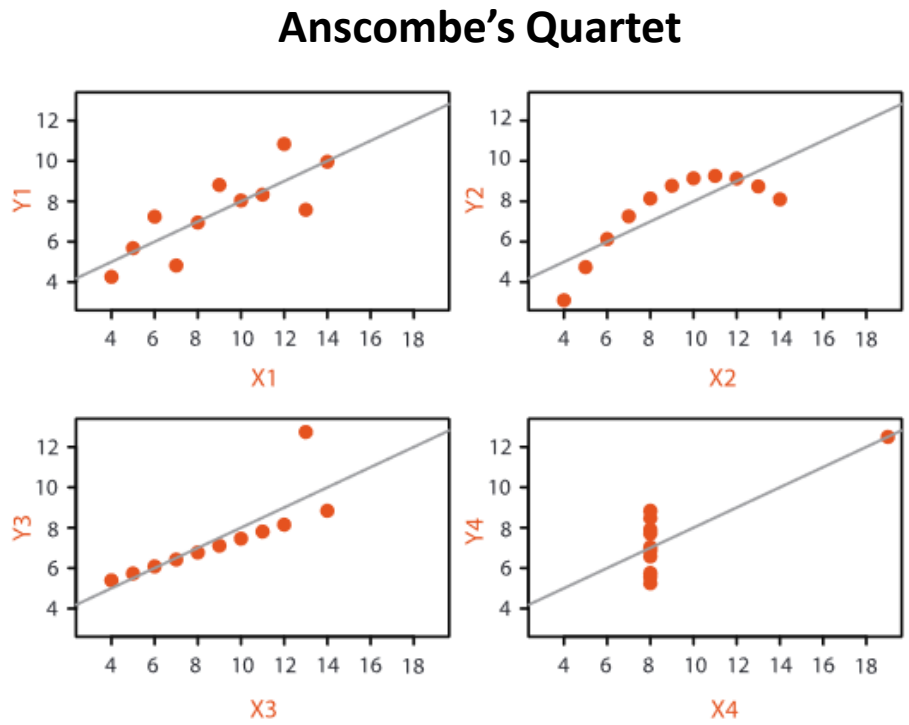


Visualization Analysis and Design. Tamara Munzner, with illustrations by Eamonn Maguire. A K Peters Visualization Series, CRC Press, 2014. (CC BY 4.0 © 2015 by Taylor & Francis Group, LLC) [1]
https://creativecommons.org/licenses/by/4.0/

Data Simplification Methods

- Consider ways to simplify your data for ease of communication.
  - Aggregation: averaging, adding, binning
  - Filter: remove unimportant items or attributes
  - Embed: hover text/annotations for interesting additional attributes

- Careful of simplifying too much; you may miss out on interesting features.



Plot from "Visualization Analysis and Design". Tamara Munzner. [1]

Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 17–21.

National Aeronautics and Space Administration

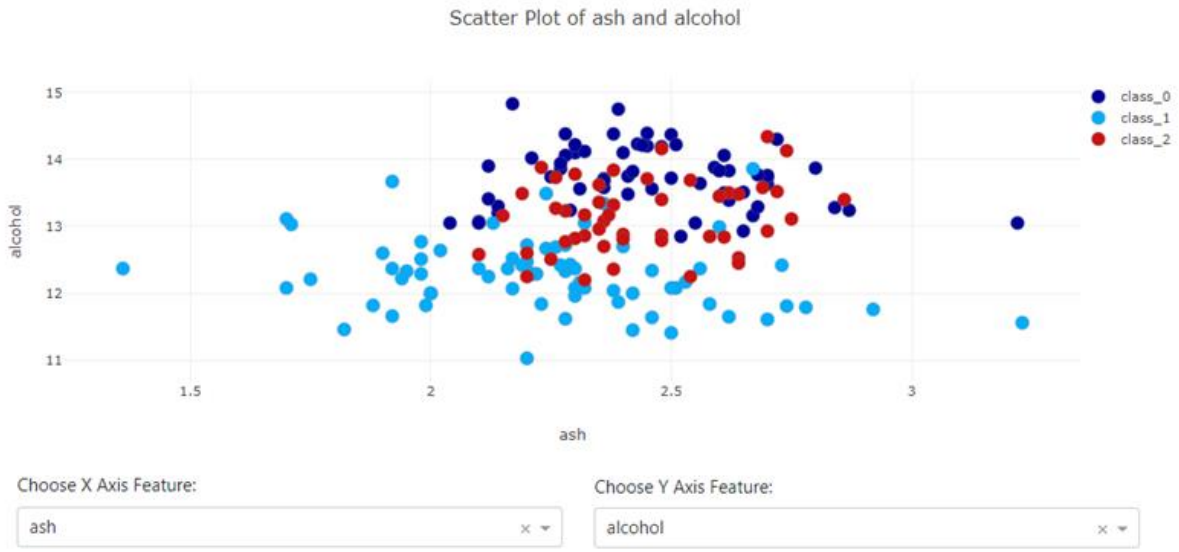
Chart Types

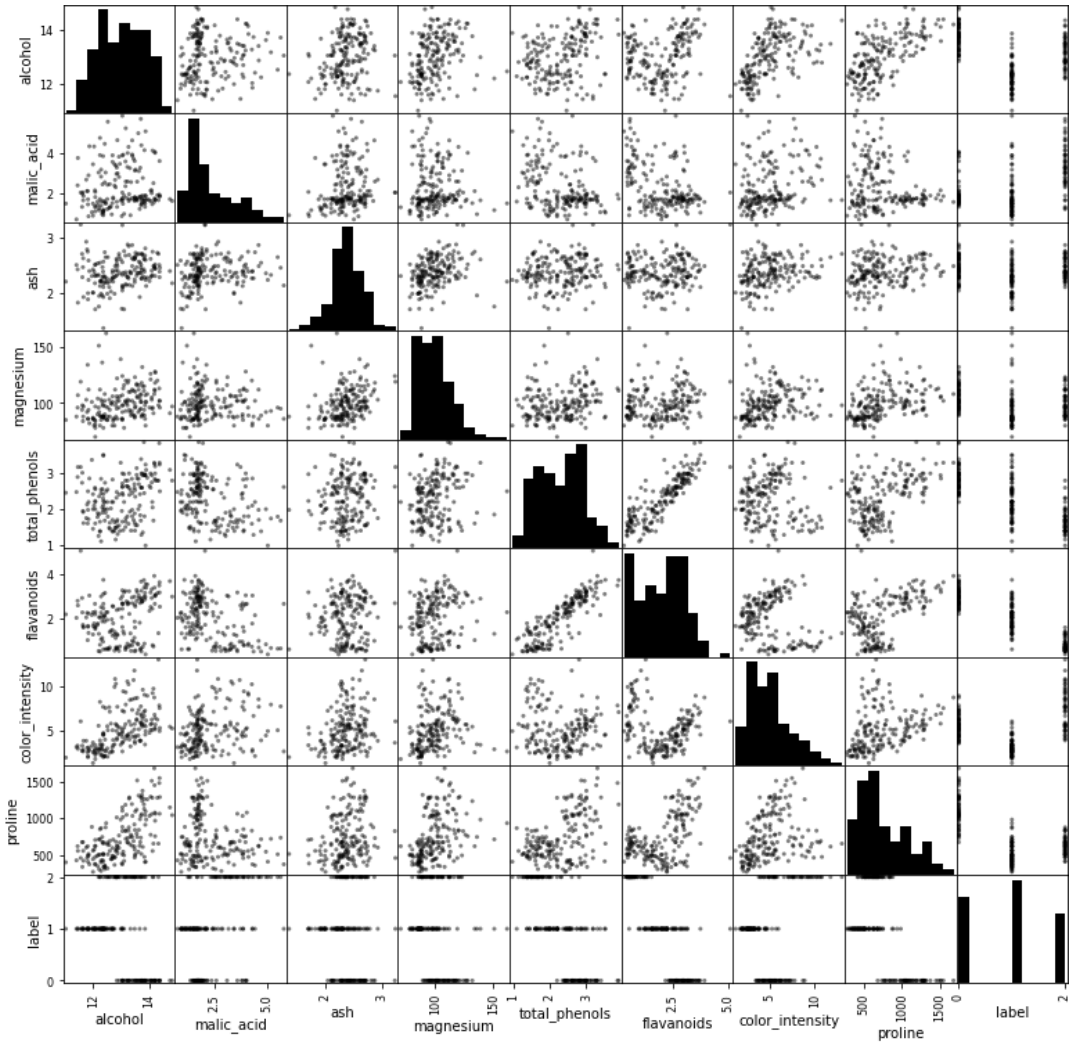
www.nasa.gov

Office of the Chief Information Officer

Example: Scatter Plots

Mark	The dot
Channels	X and Y position
Uses	Comparing quantitative data where the points don't necessarily have an order to them (ex non-temporal)
Optional Additions	Size channel for ordered quantitative Color channel for qualitative
Design Ideas	For many dots, consider adding dot border and alpha
Other Notes	Have a lot of attributes and need some EDA? Why not a scatter plot matrix (SPLOM)?

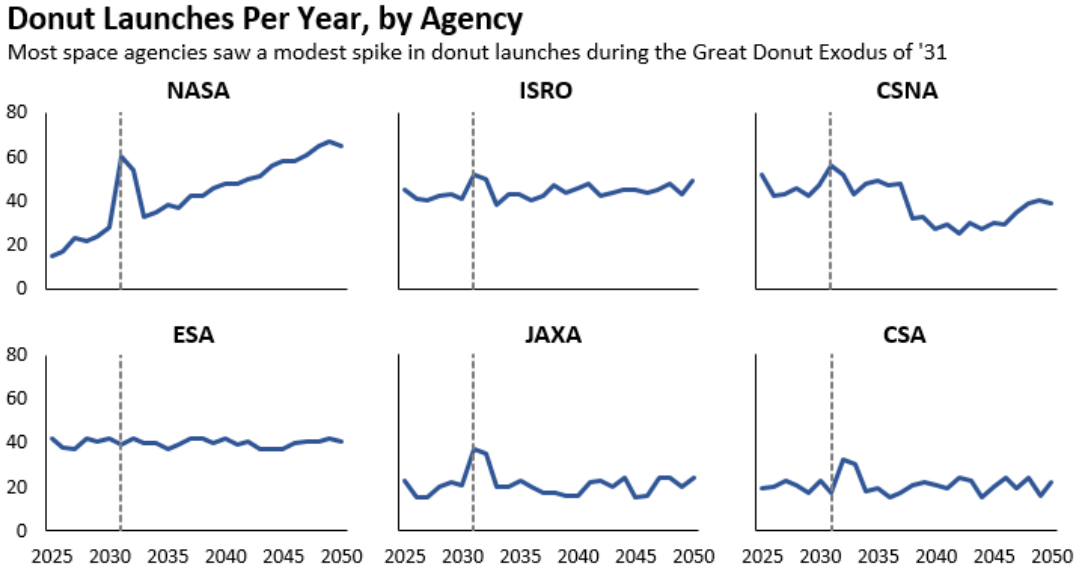


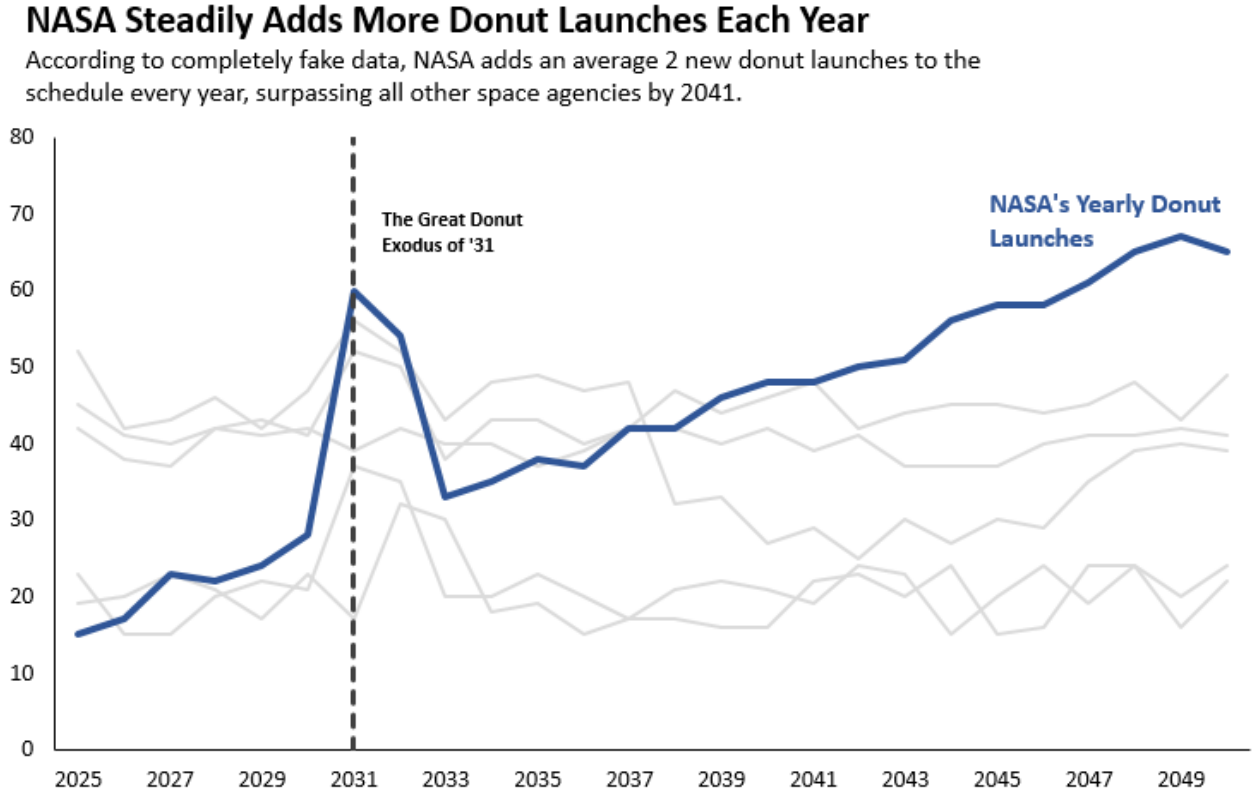


Charts exploring the Wine dataset provided by the UCI Machine Learning Repository [2]. On the right, an interactive scatter plot made that allows users to choose the features to view. Above, a scatter plot matrix that can view 9 features at once.

Example: Line Charts

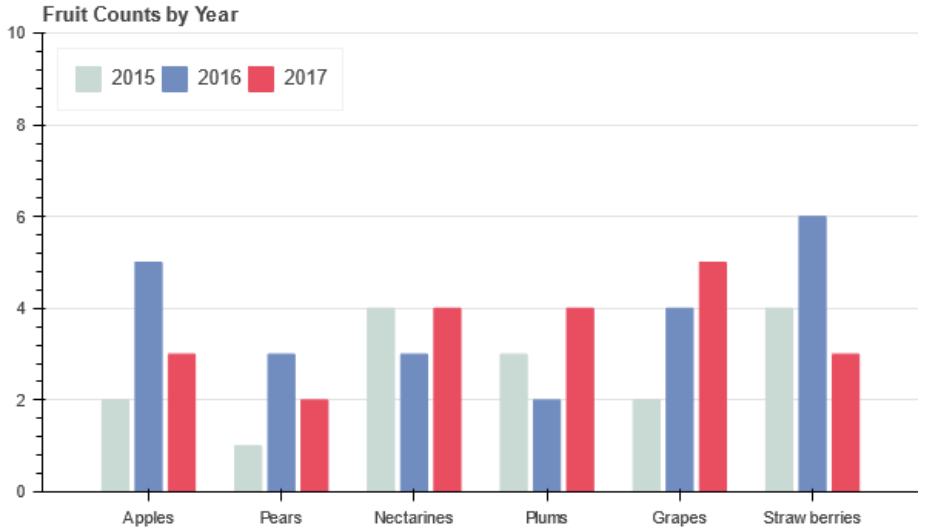
Mark	The dot and line
Channels	X and Y position
Uses	Show one quantitative attribute and one ordered attribute (ex: signal over time)
Optional Additions	For multiple trends, show multiple lines. For many trends, consider a small multiples
Design Ideas	Highlight one trend from many with bold color and make the rest gray
Other Notes	Can add on top of other charts like bar charts or histograms to show raw data versus cleaned



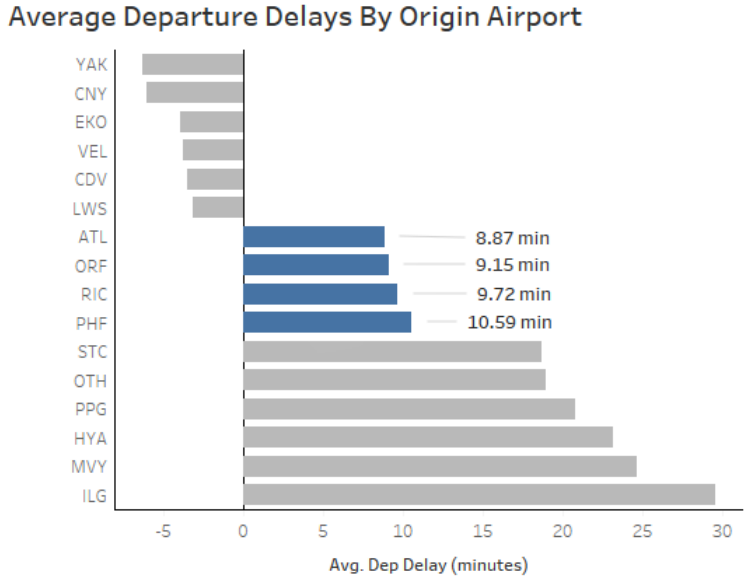


Example: Bar Charts

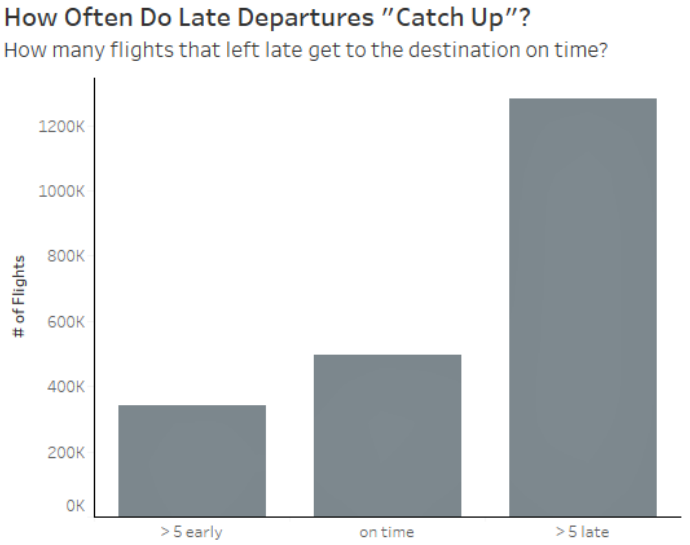
Mark	The bar
Channels	Length and position
Uses	Comparing categorical data
Optional Additions	Group several bars into categories Use stacked bars to show parts of a whole
Design Ideas	Try a horizontal layout for long labels Sort by bar height for easier comparisons Color only for grouped/stacked bars or highlights



Bokeh Contributors. Bar_dodged.py, retrieved April 5, 2022 from https://docs.bokeh.org/en/latest/docs/gallery/bar_dodged.html. (BSD 3 Clause) https://github.com/bokeh/demo.bokeh.org/blob/main/LICENSE.txt



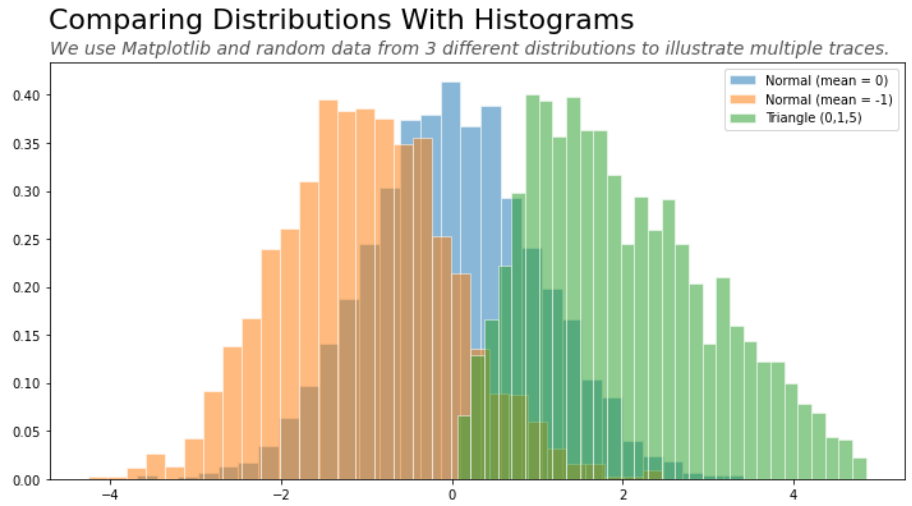
These two charts were built using data compiled and hosted by Google Cloud Platform for training purposes. The dataset stores information about US flights in 2015. [5]



Examples: Histograms & Boxplots

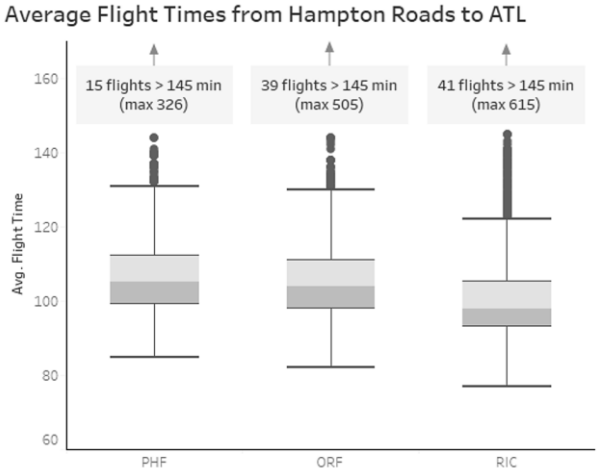
Histogram

Mark	Bar: represents a "bin" of data (bin is a range, ex 1-2)
Channels	Length = # of values within the bin Position on x axis = value of the bin
Uses	Showing/comparing distributions of quantitative data
Optional Additions	For 2-3 distributions, graph them all on same axes and have different color for each distribution
Design Ideas	For multiple distributions, add an alpha to see all them Adjust bin size as needed (too big and you miss the interesting; too small and noise takes over) For far outliers, consider removing and adding annotations.



Box and Whiskers/Boxplot

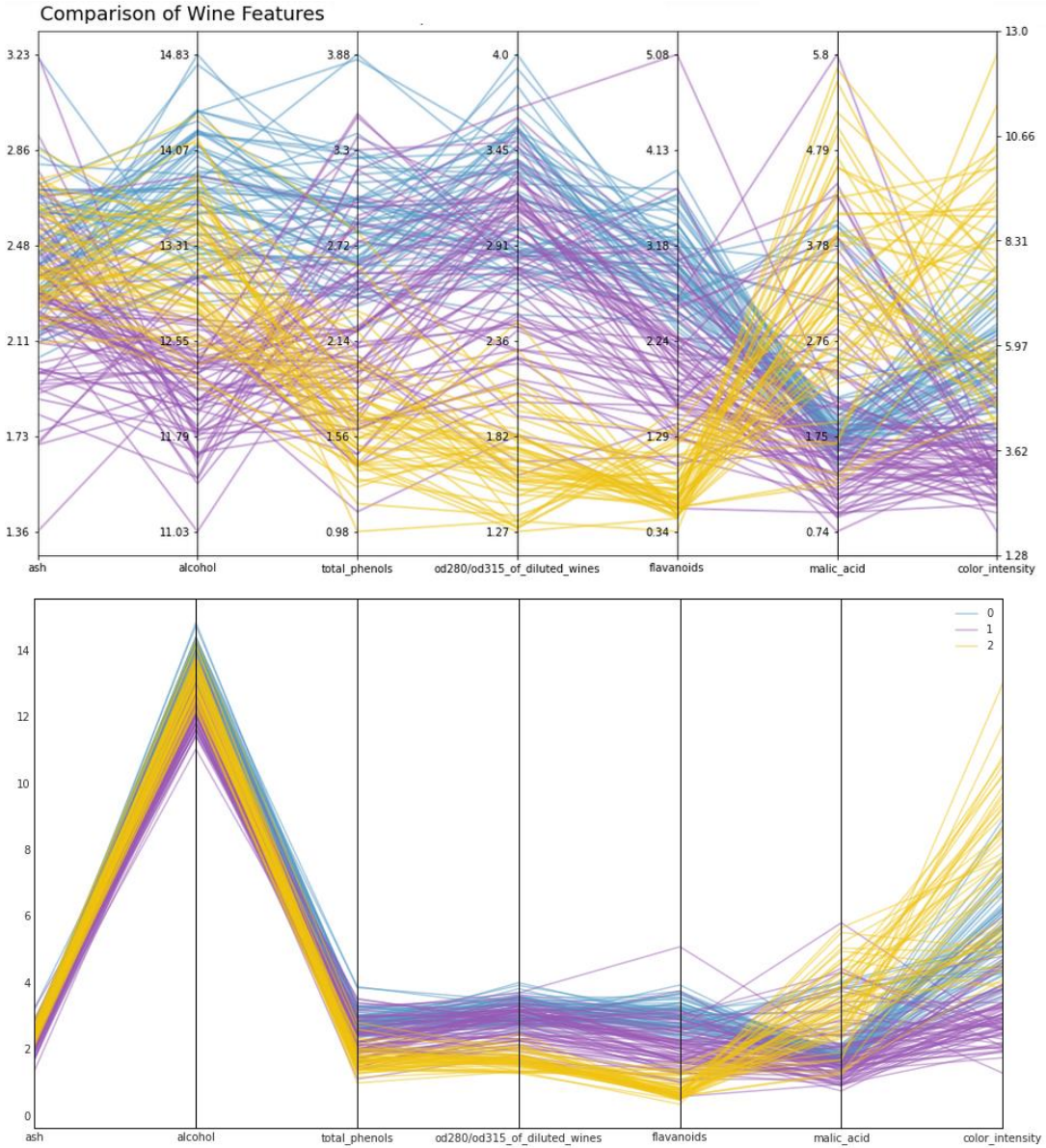
Marks	The box-whiskers-dots glyph Box = interquantile range. Midline is median; bottom and top are Q1 and Q3 (respectively) Whiskers/lines = rest of data excluding outliers Dots = outliers
Channels	height & location of box; whisker length; dot locations
Uses	Showing/comparing distributions of quantitative data
Optional Additions	For multiple distributions, chart side by side for easy comparison
Design Ideas	For far outliers, consider removing and adding annotation.



Lakshmanan et al., 2018 [5]

Example: Parallel Coordinates

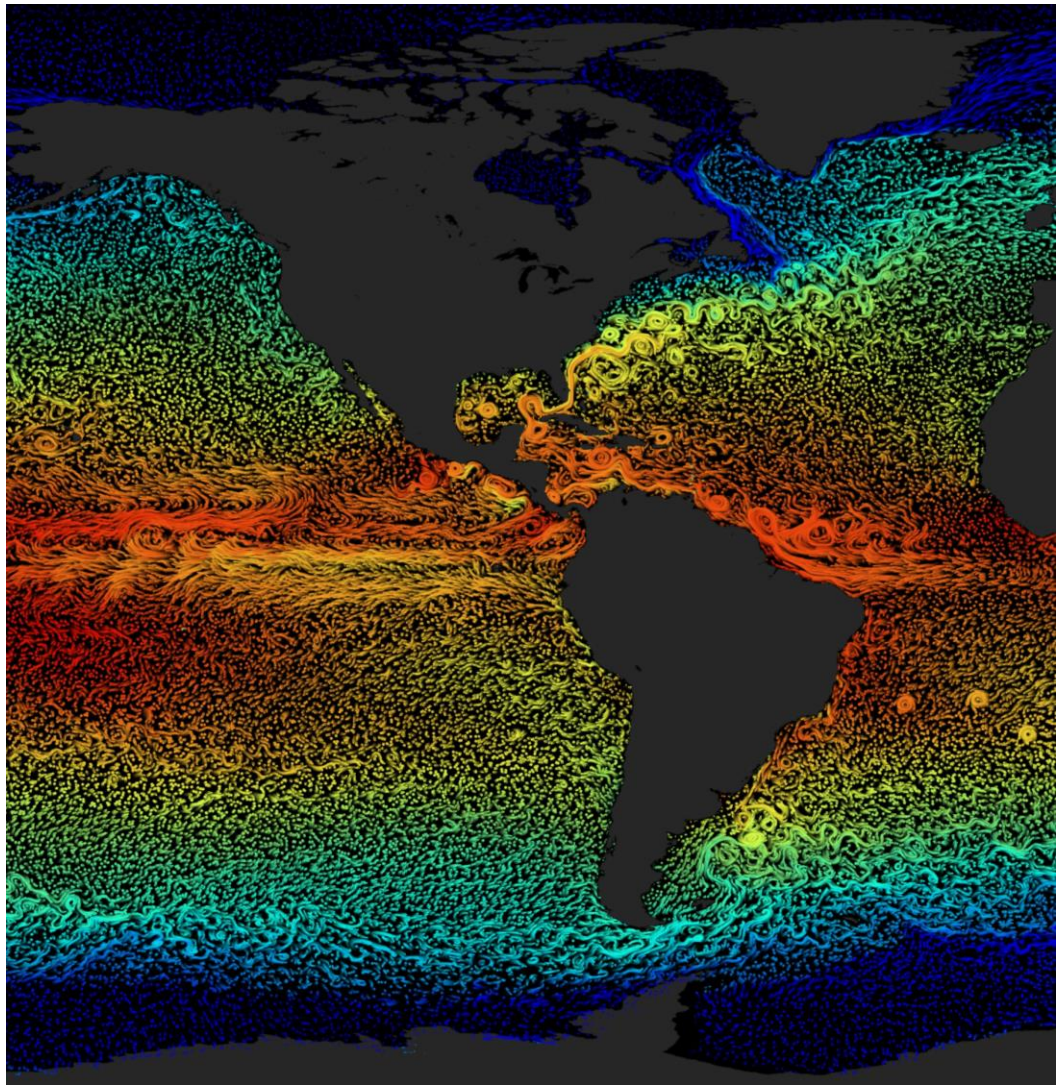
Mark	A line spanning the axes Line denotes a single record, and “zigs” along the parallel axes to denote the records value for each attribute
Channels	Location on “Y” = magnitude Location on “X” = attribute
Uses	EDA; visualizing many quantitative attributes at once Compare correlation between attributes Finding the range of attributes across the population
Optional Additions	For “groups” of data (ie multiple samples of a species), use color to categorize the lines
Design Ideas	Highlight and bold interesting lines Compare auto scaled axes or shared axes to see which is more legible



Example charts created to view features in the Wine Dataset [2]. Both show the same data, but the first scales each axes, while the second shares axes across features.

Example: Vector Fields

Mark	Arrow/directional glyph/line
Channels	Location in grid angle/curvature = direction length of glyph = magnitude
Uses	Often associated with fluid dynamics/velocity While complex charts, can be good at showing “sinks”, “sources” and other interesting features of complex datasets
Optional Additions	Color to denote additional features
Design Ideas	Try different glyphs to see which works best Add a sequential colormap to emphasize high/low Match “intuition”; ex “up/down” goes along y, and “side to side” goes along x

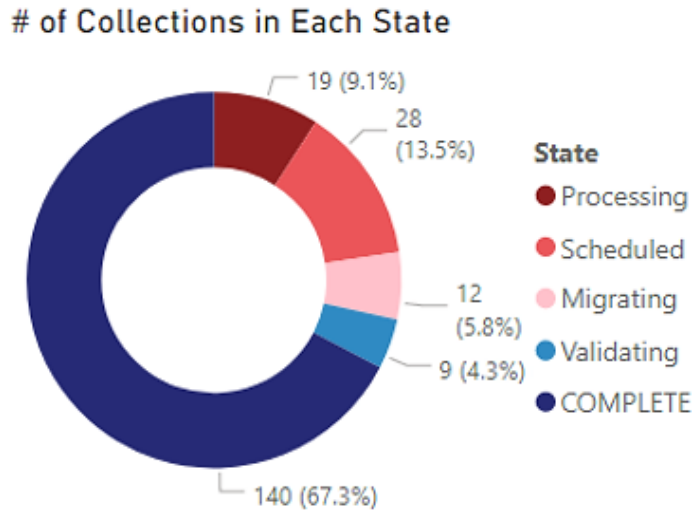


Sub view of the Flat Map Ocean Current Flows with Sea Surface Temperatures by NASA Goddard, created using model output from Estimating the Circulation and Climate of the Ocean, Phase II (ECCO2). The view shows both current and sea surface temperatures in the Americas for January 2005

NASA/Goddard Scientific Visualization Studio, MIT/JPL [6]

Pies and Donuts: The Controversy

Mark	Pie/Donut “slices”
Channels	Color = category Area/Angle/Arc Length of slice = percent of whole
Uses	Showing parts of a whole (categorical data)
Optional Additions	Hover text/text labels for specific numbers
Design Ideas	Keep to only a few categories An ordered color map can indicate sequence To highlight one category, use shades of gray for all other categories, and color for highlight. OR: only show the highlight (and all else is “other”)



An overview of collection status during a migration project. While there were 5 states, a diverging color map hints at “incomplete” and “complete/near completion”.

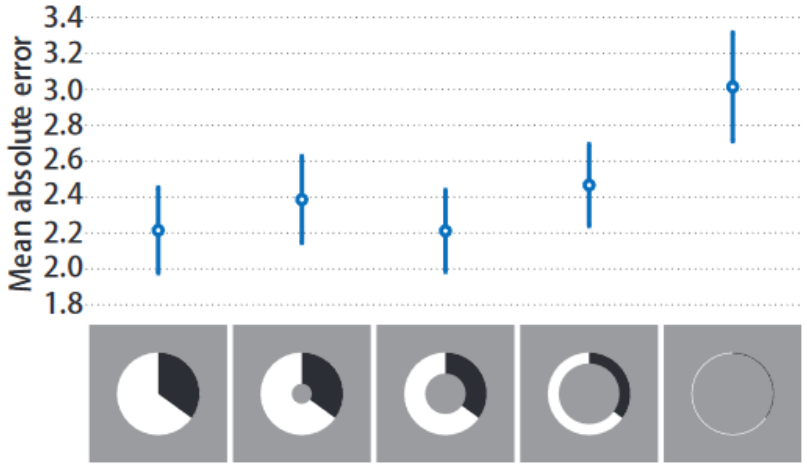


Figure 5: *Mean absolute error and 95% confidence intervals for conditions IR_0 to IR_4 in Experiment 2. Mean absolute error of IR_4 is significantly larger than the other conditions.*

“Simkin and Hastie studied the spontaneous response of 200 undergraduate students to different types of chart. ***The results showed that most people make comparisons when presented with bar charts and make proportion judgments when presented with pie charts****, indicating that people have certain expectations for the use of these charts and the information conveyed by them.” (Cai et al., 2018) [7]*

National Aeronautics and Space Administration

Dashboard Design

www.nasa.gov

Office of the Chief Information Officer

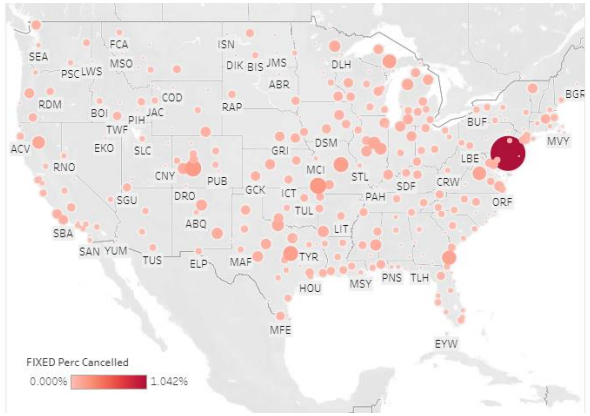
Designing Charts and Dashboards

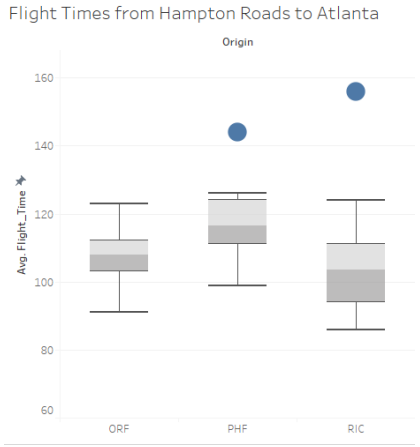
- **Get a use case/general set of questions.**
- **Find some data, a data dictionary, and any info you can.**
- **Make some specific questions/do Exploratory Data Analysis (EDA) to further refine.**
- **Sketch out some charts that could answer those questions.**
  - For dashboard: sketch out a layout using the chart sketches.
- **Make a first draft (in the tool of your choice).**
- **Validate calculations and visualizations.**
- **Reiterate as needed.**

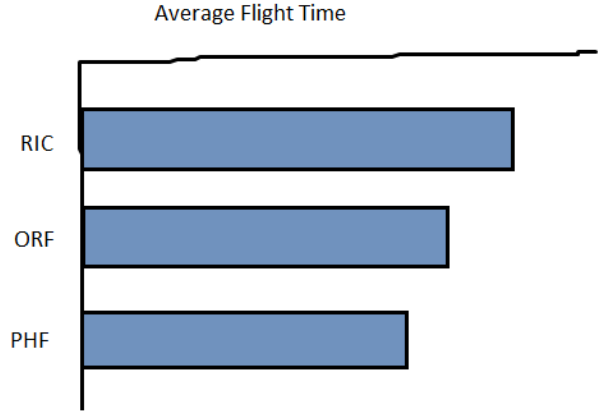
Example: The Sketching

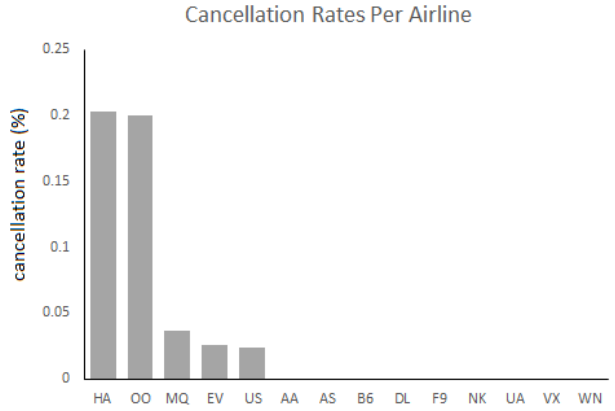
Choosing an Airport from Hampton Roads to Atlanta, Georgia

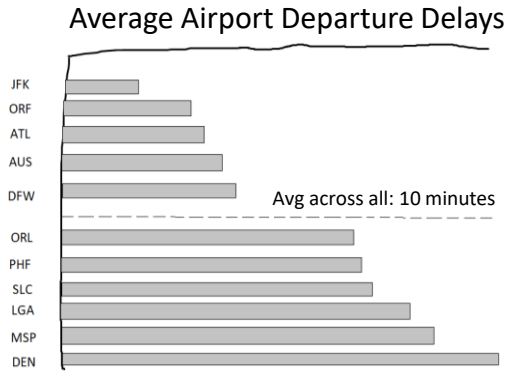
Our user lives in the Hampton Roads area of Virginia and often finds themselves flying to Atlanta, Georgia to visit family and friends. We'd like to provide this user some recommendations on which airports they should use to fly.

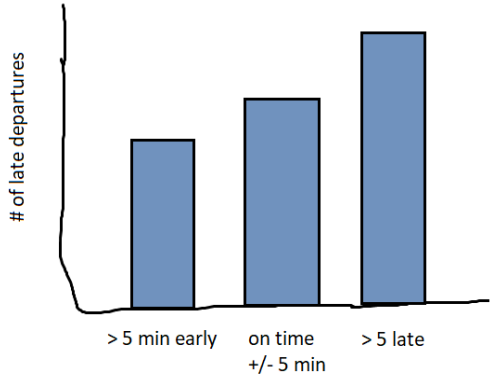










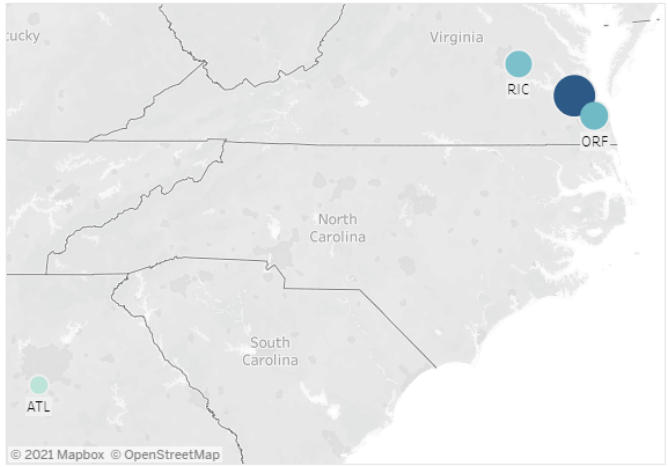


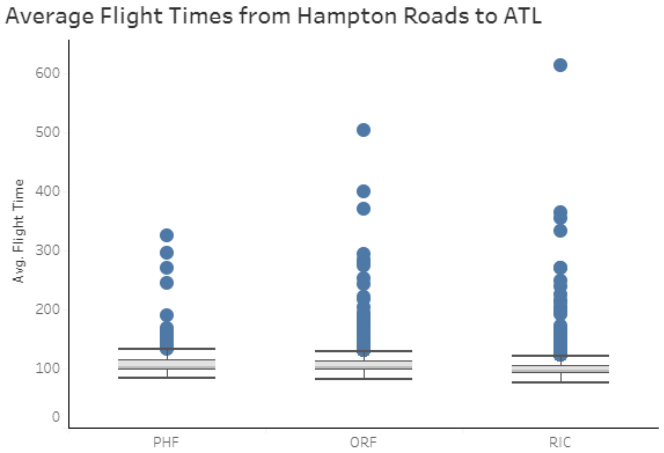
Lakshmanan et al., 2018 [5]

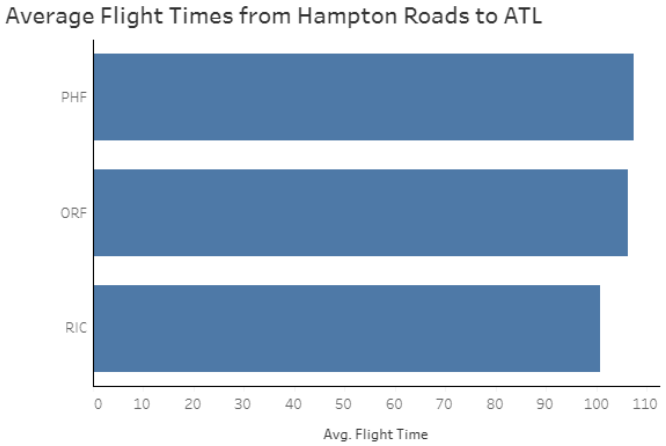
Example: First Draft

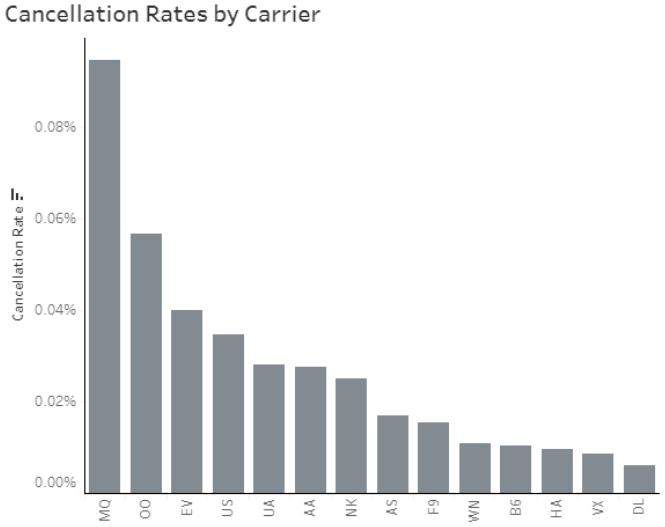
Choosing an Airport from Hampton Roads to Atlanta, Georgia

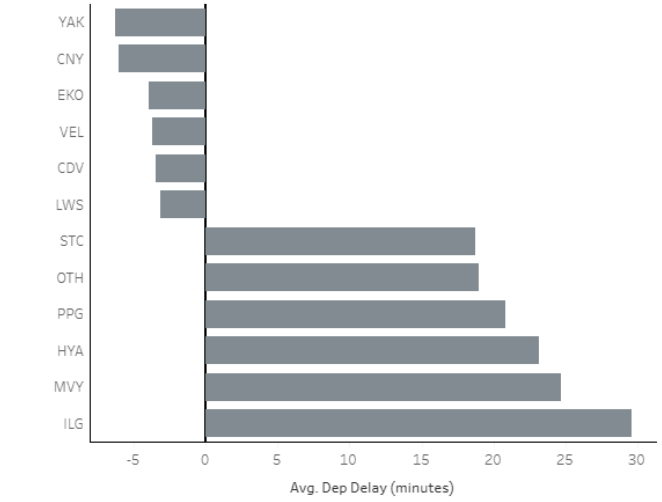
Our user lives in the Hampton Roads area of Virginia and often finds themselves flying to Atlanta, Georgia to visit family and friends. We'd like to provide this user some recommendations on which airports they should use to fly.

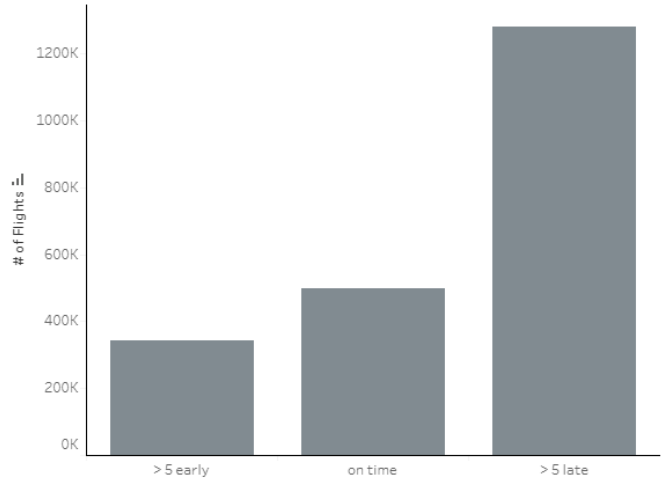










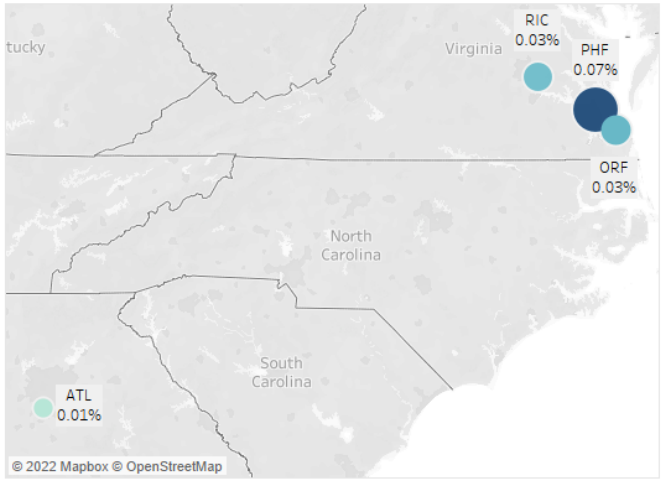


Example: Final Draft

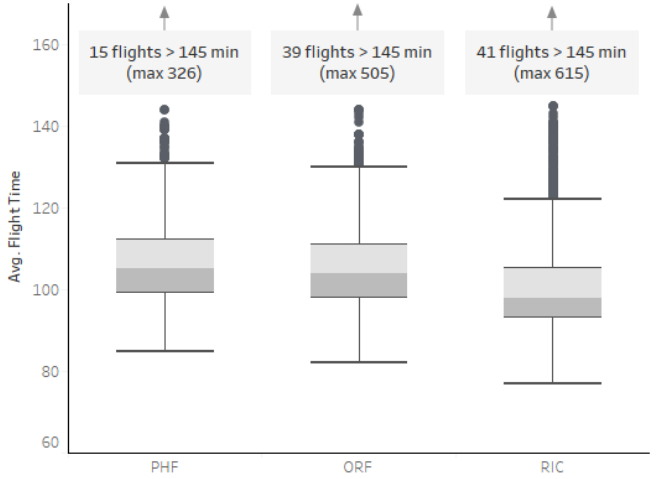
Choosing an Airport from Hampton Roads to Atlanta, Georgia

Our user lives in the Hampton Roads area of Virginia and often flies to Atlanta, Georgia to visit family and friends. We'd like to provide some recommendations on which airports they should use to fly.

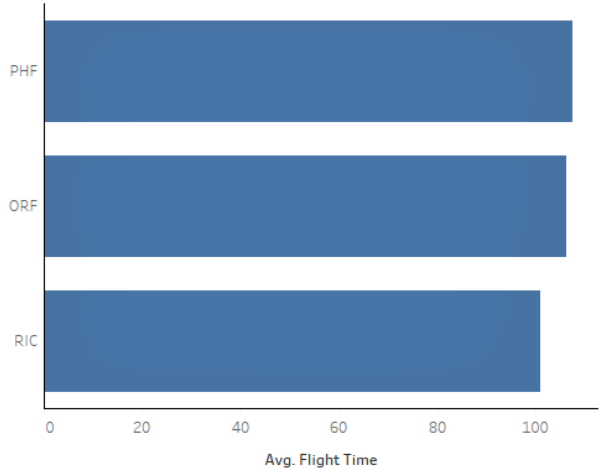
Cancellation Rates By Airport



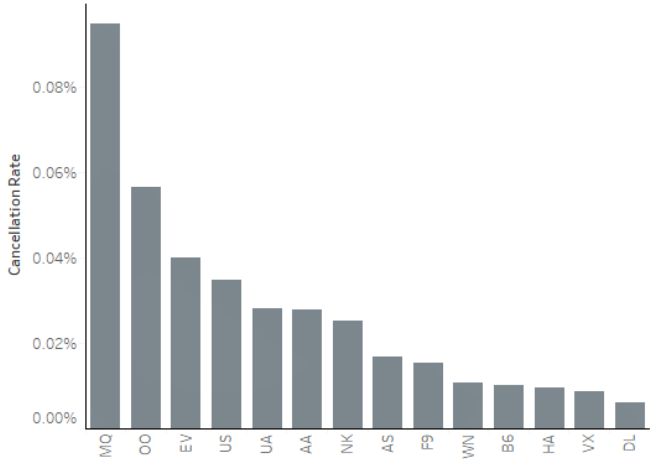
Average Flight Times from Hampton Roads to ATL



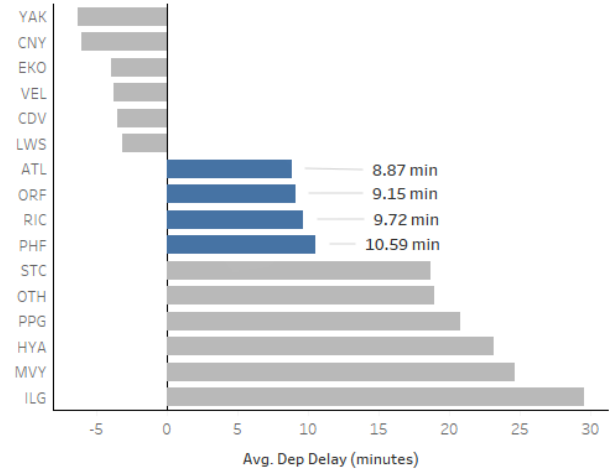
Average Flight Times from Hampton Roads to ATL



Cancellation Rates by Carrier

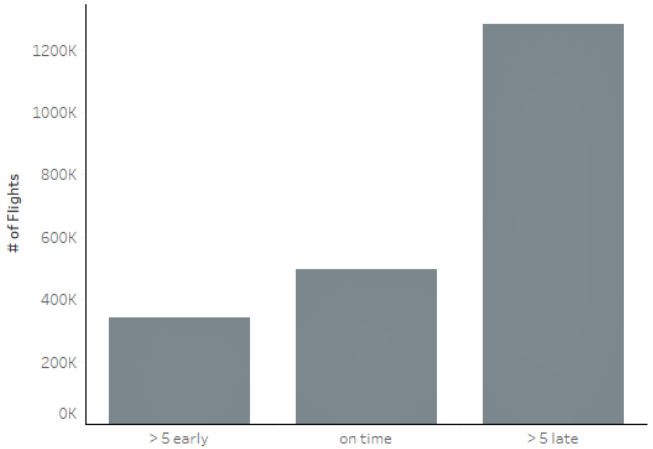


Average Departure Delays By Origin Airport

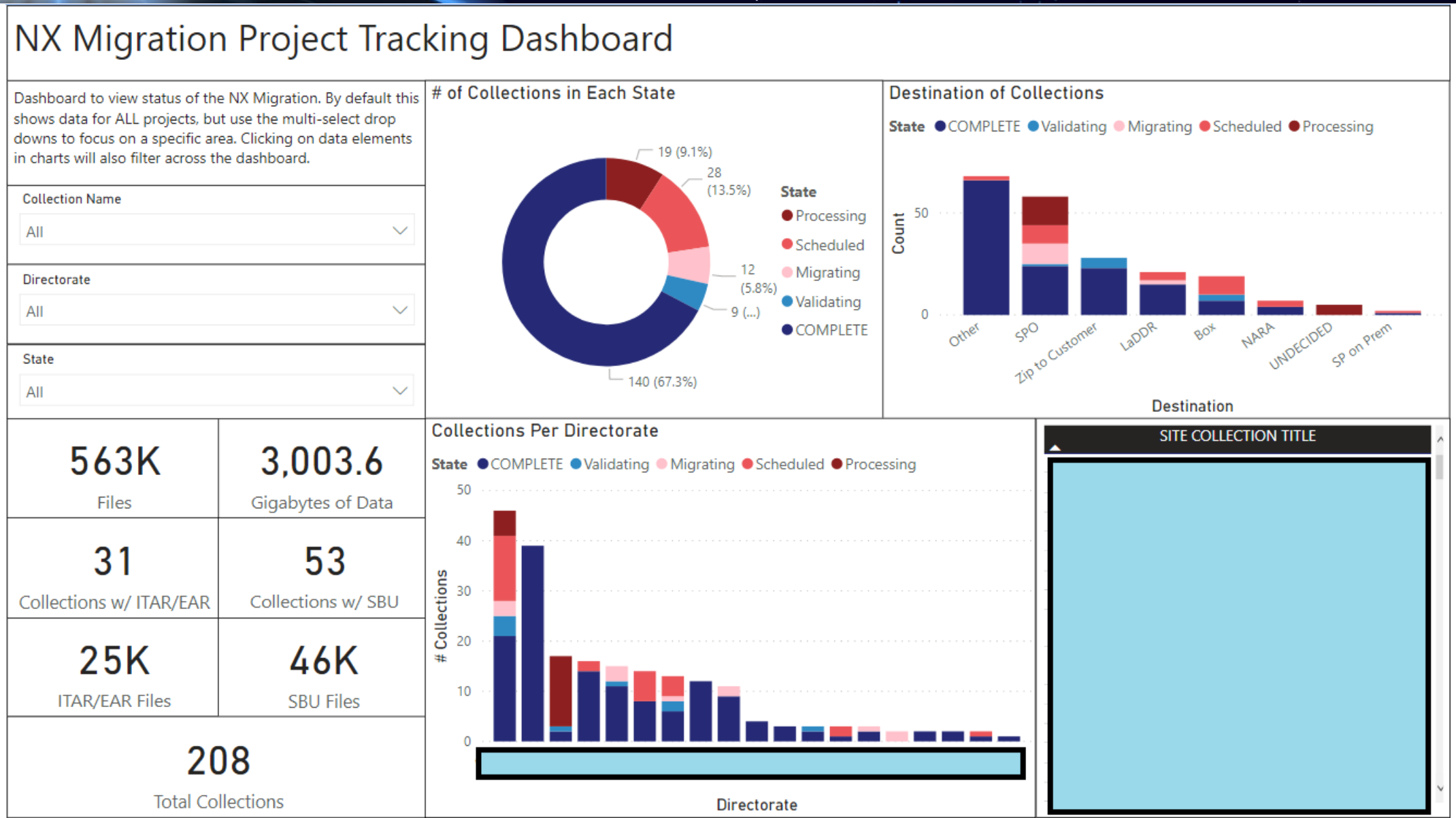


How Often Do Late Departures "Catch Up"?

How many flights that left late to get to the destination on time?



Example: NX Migration Dashboard



How to Learn More

- **Government Guidelines**
  - https://designsystem.digital.gov/components/data-visualizations/
  - https://cfpb.github.io/design-system/guidelines/data-visualization-guidelines
- **Explore different chart types**
  - https://datavizcatalogue.com
- **Look at sources that have good data visualizations**
- **Look at sources that have bad data visualizations** (and think about how you might improve them)
- **Ask for feedback (from a trusted source/stakeholders)**
  - Have a fresh set of eyes double check the charts make sense
- **Practice, Practice, Practice!**

Sources and Further Reading

[1] Munzner, T., & Maguire, E. (2015). *Visualization Analysis & Design*. CRC Press. Retrieved from https://www.cs.ubc.ca/~tmm/vadbook/#figures

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] Carter, S., & Quealy, K. (2014, August 26). Home Prices in 20 Cities. *The New York Times*. Retrieved from https://www.nytimes.com/interactive/2014/01/23/business/case-shiller-slider.html.

[4] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] Lakshmanan, V. (2018). Chapter 8. In *Data Science on the google cloud platform: Implementing end-to-end real-time data pipelines: From ingest to machine learning*. essay, O'Reilly Media.
Retrieved data using tutorial on the accompanied GitHub: https://github.com/GoogleCloudPlatform/data-science-on-gcp/tree/edition1_tf2 (Apache 2.0 license).

[6] NASA/Goddard Space Flight Center Scientific Visualization Studio (2011, February 10). “Flat Map Ocean Current Flows with Sea Surface Temperatures (SST)”. Retrieved from https://svs.gsfc.nasa.gov/3821.

[7] Cai, X., Efstathiou, K., Xie, X., Wu, Y., Shi, Y., & Yu, L. (2018). A study of the effect of doughnut chart parameters on proportion estimation accuracy. *Computer Graphics Forum*, *37*(6), 300–312. https://doi.org/10.1111/cgf.13325

[8] Bokeh Contributors. “Bar_dodged.py,” retrieved April 5, 2022 from https://docs.bokeh.org/en/latest/docs/gallery/bar_dodged.html. (BSD 2 Clause) *https://github.com/bokeh/demo.bokeh.org/blob/main/LICENSE.txt*

[9] The pandas development team. (2020). pandas-dev/pandas: Pandas 1.0.3 (v1.0.3). Zenodo. https://doi.org/10.5281/zenodo.3715232

[10] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.