



Fundamentals of Data Visualization

A class at the
University of Applied Sciences Northwestern Switzerland

Nils Fahrni
BSc Data Science Student

December 7, 2023

Contents

1	Introduction	2
2	Visualization basics and chart types	3
3	Visual Perception	5
4	Design Principles vs. Data	7
5	Grammar of Graphics Tools	9
6	Evaluation	11

Chapter 1

Introduction

In our data-driven age, the ability to visualize complex datasets not only simplifies information but also tells a story, highlights patterns, trends, and anomalies that might otherwise remain hidden. Data visualization acts as the bridge between the raw, often impenetrable world of numbers and an accessible visual representation that can be easily understood, interpreted, and acted upon from any perspective.

In this report, we delve into the fundamentals of data visualization, elucidating its importance and methodologies. By applying these principles, we present a series of visualizations centered around wildfire data from California. Through these graphics, I aim to shed light on the frequency, intensity, and geographical spread of these fires over recent years.

Join me on this journey through the five learning objectives, as we navigate the terrain of data visualization and uncover the hidden tales of California's fiery landscape — a landscape where not only trees and vegetation are at risk, but also a realm of danger for its myriad forest animals.

The data utilized in this report has been sourced from California's Department of Forestry and Fire Protection, available at www.fire.ca.gov. This comprehensive dataset covers an eight-year span, from 2013 to 2020.

Chapter 2

Visualization basics and chart types

We aren't innately equipped with the knowledge to interpret any type of chart which displays data. As creators of data visualizations, we need to ask ourselves the question of "Who is the visualization at hand for?". It's crucial for us to comprehend our audience's perspective and recognize when an alternative graph might captivate them more effectively or be easier on their eyes, aiding in enhancing our audience's graphical literacy.

Scott Klein, deputy managing editor at *ProPublica* once wrote,

"There is no such thing as an innately intuitive graphic. None of us are born literate in reading visualizations."

This realization brings light to the fact that we cannot create visualizations that need absolutely no thinking. There will always be a subjective interpretation — but, as a data scientist, one can help to make the process of interpretation easier and more streamlined for those who have to interpret the insights.

Visualizing frequencies

A great way to visualize frequencies of nominal data is to use **bar charts**. In this example we visualize the wildfire season's frequency throughout the average year.

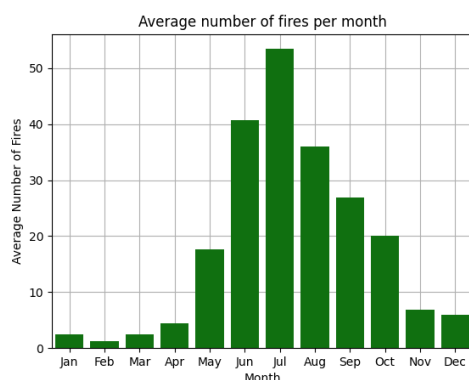


Figure 2.1: Bar Chart

When to use bar charts

Bar charts are among the simpler types of charts. They allow us to get an overview of how frequent a category is in a quick manner. It not only shows us a single category, but displays all the categories at once, allowing us to compare the magnitudes.

When to not use bar charts

Bar charts can have a few flaws. It is important to notice that bar charts work best if

there aren't too many categories present. Such a chart can also get over-cluttered fast if too much information is added, which may not be the result one wants.

Comparing densities

A violin plot is a method of plotting numeric data and can be understood as a combination of a box plot and a kernel density plot. It provides a visualization of the distribution of the data, its probability density, and its cumulative distribution, with the added ability to visualize the data's symmetrical nature and the presence of multiple modes.

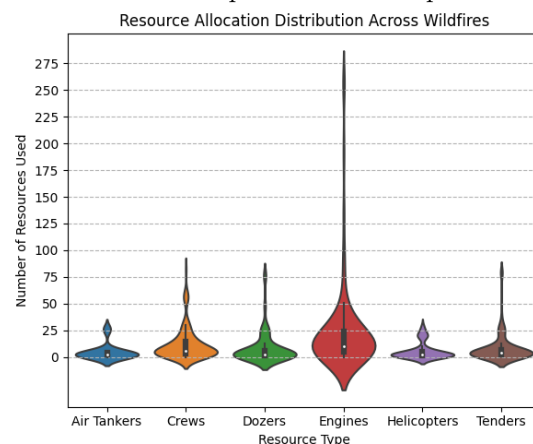


Figure 2.2: Violin Plot

When to use Violin Plots

Violin plots excel at visualizing data distributions across categories. For the wildfire dataset, they aptly showcase how resource deployments vary across different fire conditions, revealing both average usage and distribution nuances.

When to avoid using Violin Plots

Violin plots may not be the best choice when the dataset has limited data points or when the primary interest is in summary statistics rather than detailed distributions. For instance, in the wildfire context, if we're solely focused on the average number of helicopters used per year, a simple bar chart would be more straightforward and interpretable. Generally, Violin Plots are said to be on the harder side when it comes to interpretability.

Geospatial Visualization

In this example we have location based data of wildfires at our hand. If we want to know which Californian counties were most heavily affected, this view offers a great opportunity to visualize that. This following visualization gives us a location based overview on how many acres were burned in each Californian region.

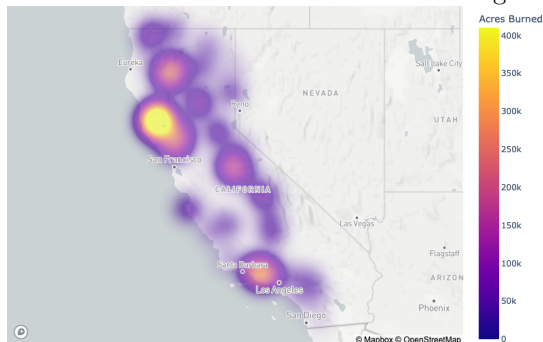


Figure 2.3: Heatmap

The problem with geospatial visualizations

Something we notice when looking at the heatmap is that we only get a rough visual estimate on how large the fires actually were. The cells on the heatmap don't represent the actual area that was burned, only their color does. The cells start to get bigger when other fires are in close proximity. This can lead to misinterpretation.

When to use geospatial heatmaps

These visualizations are often used to show an event's severity and location. Such heatmaps generally work well if there should be an immediate visual impact on the viewer as humans can quickly identify areas of severity thanks to *heat bubbles*.

When to avoid using geospatial heatmaps

As already touched on in a previous section, these visualizations often don't work well when wanting to show the exact extent of an event. The *heat bubbles* show a gradient of colors — This makes it hard for users to pinpoint the actual value belonging to the color on the legend. Heatmaps often give the perception of continuous data, but in this case, the data (acres burned) is discrete to each county. The smooth transitions between colors could suggest a continuity that doesn't exist in the data. So users might lean towards thinking that the bubbles show the area of burning forest.

Chapter 3

Visual Perception

Making extremes stand out

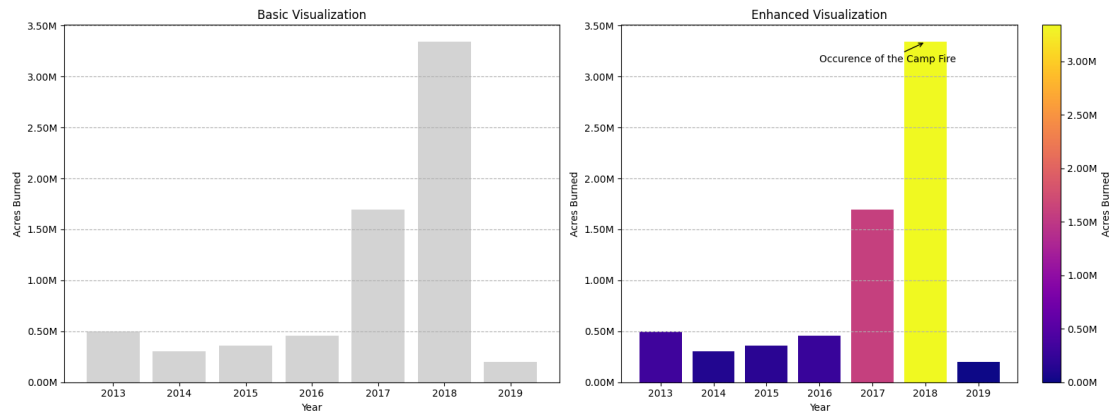


Figure 3.1: Color enhancement comparison

On the left, we have the **Basic Visualization** without colors and annotations, using a uniform light gray color for all bars.

On the right, we have the **Enhanced Visualization** with color gradients representing the intensity of acres burned, the year with the maximum acres burned highlighted in red, and an annotation providing context about the maximum value.

Of course one could say the first "Basic Visualization" is more than enough but the latter version even allows to capture the year of the most damage in a more dramatic way without having to examine axes and context closer.

3D and its perceptual problems

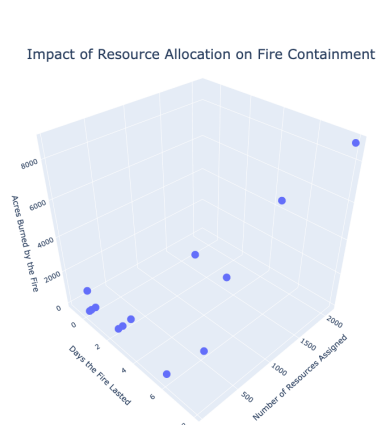


Figure 3.2: 3D Plot

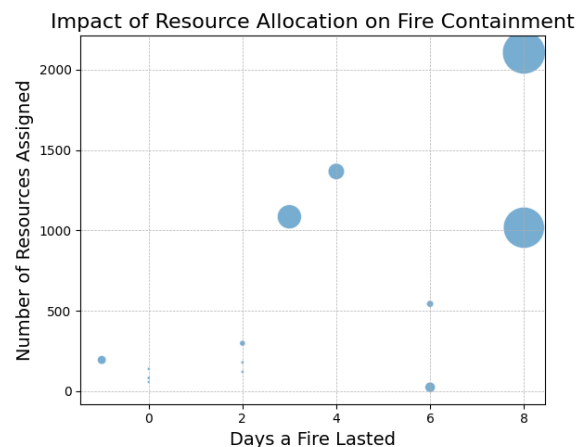


Figure 3.3: Simplified Plot

2D plots are often preferred for clarity and simplicity. They can be easier to read, especially when printed or viewed on a flat medium like paper or a non-interactive display. 3D plots, on the other hand, can provide additional perspective and are excellent for displaying complex datasets with three distinct variables. However, without interactivity, 3D plots can become confusing and may not effectively communicate the data due to issues with perspective, such as occlusion (where parts of the graph obscure other parts) or distortion (where the angle of view can make certain dimensions appear different than they are). Here, the 2D plot shows the exact same data as the 3D plot but we made use of another property, the **scatter size**.

Counteracting Overlapping Data

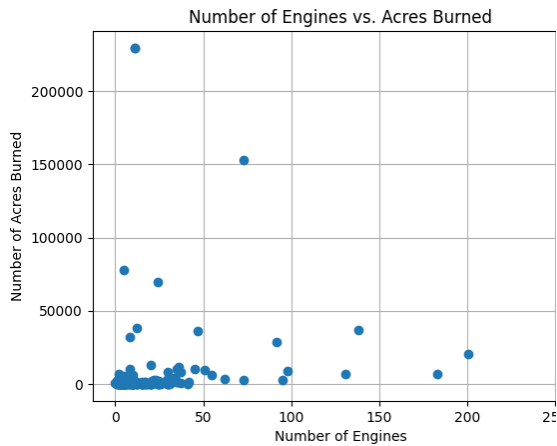


Figure 3.4: Default Scatter Plot

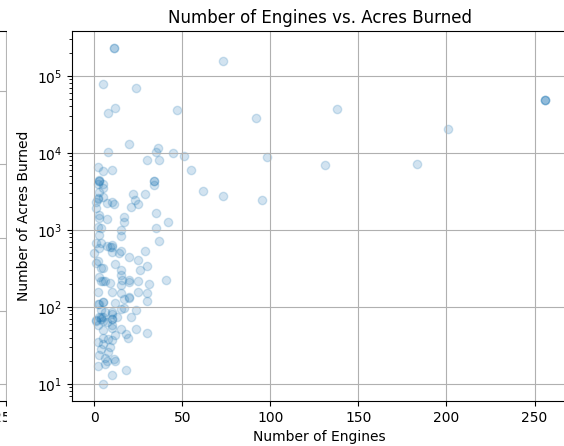


Figure 3.5: Enhanced Scatter Plot

The data points are clustered at the lower end of the engines axis, indicating that most observations have a relatively low number of engines involved. Overlapping data points, particularly where the concentration of points is high, can obscure individual data values and may indicate a need for more sophisticated plotting techniques. Let's see how we can modify the **scatter plot** to see what actually lies behind the dense cluster of points.

Applying a lower alpha value to the scatter points, as shown in the right plot, helps to mitigate the issue of overlapping data points by making the points semi-transparent, which allows for the visualization of density in regions with high data point concentration. Furthermore, transforming the y-axis to a logarithmic scale spreads out the points along the vertical axis, providing a clearer view of the data across a wider range of values. However, this approach can make it more difficult to interpret the actual values, especially for those unfamiliar with logarithmic scales, as the scale compresses the data range and changes the perception of the distances between points. The reduced emphasis on individual, more transparent points can make it challenging to notice outliers or standalone points which could be significant. This trade-off in clarity for density versus emphasis on individual data points requires careful consideration when visualizing data to ensure accurate interpretation.

Chapter 4

Design Principles vs. Data

In this chapter, we embark on a journey through the various stages of data visualization, focusing on geospatial data and its visualization. The goal of this demonstration is to create a **choropleth map** of California’s counties, showing the most frequently affected counties when it comes to wildfires.

Data Preparation

The goal of the first step of creating such a visualization is to decide what kind of data we need to display our desired graphic.

- Calculate the fire frequency for each county: The choropleth map is going to color in the county’s areas according to how many fires the data set has recorded.
- Next to the counties and their frequencies of wildfires we will be required to aggregate their locations on the map. The data set we have at hand provides us with latitude and longitude.
- Perhaps not every county is represented in the data set – We therefore need to make sure to add these counties with zero incidents so the desired geomapping framework ([Mapbox](#)) can handle our needs.

Select and Tailor Important Features

After preparing the data on a general level the data needs to be cropped to its needs.

- Implement a disambiguation process by cross-referencing county names with state identifiers. In the case of counties like “Lake County”, which may exist in multiple states, it is crucial to include a state-level filter in the data processing pipeline. This ensures that only the “Lake County” located within the boundaries of California is included in the dataset used for the choropleth map.
- Cross-check the county names and their associated data against a reliable source, such as a government geographical database or API. This step can help to confirm that the data being visualized corresponds correctly to the intended locations within California.

Following Design Principles

Choosing Appropriate Colors

Effective choropleth maps for visualizing Californian wildfires should highlight accurate and meaningful spatial patterns, ensuring clarity and readability both in grayscale print and colored digital displays. Moreover, these maps must be crafted with colorblind-friendly design principles, ensuring accurate interpretation by a diverse audience, including those with color vision deficiencies. This approach guarantees that the maps are not only informative but also inclusive, providing a comprehensive understanding of wildfire distribution and impact across California [\[1\]](#).

In this case the sequential color map “plasma” was used which is one of the out-of-the-box available colormaps from matplotlib. These default colormaps were made with all the common color design principles in mind, including colorblindness. The color map is also perceptually uniform,

meaning that the colors are evenly distributed across the color spectrum. This is important because it ensures that the colors are not biased towards a certain color, which could lead to misinterpretation of the data [2]. The chosen color map generally follows a linear sequential pattern which is what we want in our case as visual hierarchy is important when wanting to show frequency magnitudes of fires.

The Design Principles of Legends

These following few points are important to keep in mind when designing a legend for a choropleth map:

- The placement of the legend is crucial for the viewer to easily associate the labels with the corresponding data. Placing the legend below or parallel to the visualization ensures a clear connection between the two [3].
- The decision on whether to include a legend title is based on the need to provide additional context to the visualization. If the title adds value and clarity to the chart, it should be included; otherwise, it can be omitted [3].
- Directly labeling data representations instead of using a legend can enhance the understanding of the chart, as it reduces the cognitive load on the viewer. This principle aims to make the chart more intuitive and easier to interpret [4]. In the case of the choropleth map it is not possible to directly label the data in an aesthetically pleasing way as some counties are too small to fit the label in.

Showing Correct Scales

The Choropleth map we try to visualize in this chapter should show which counties or therefore regions are most critically affected by wildfires. Depending on this task we need to think about how we want to scale the data.

A common conception would be to assume that larger counties naturally have more wildfires, though that may not be true – A county may be large but if the overall forest coverage is low, the rate of fires would most likely also be low. To counteract these underlying causalities we could scale the data [5]: In this case it would mean to divide the number of fires by the area of forest and thus display the rate of fires per acre through our color palette.

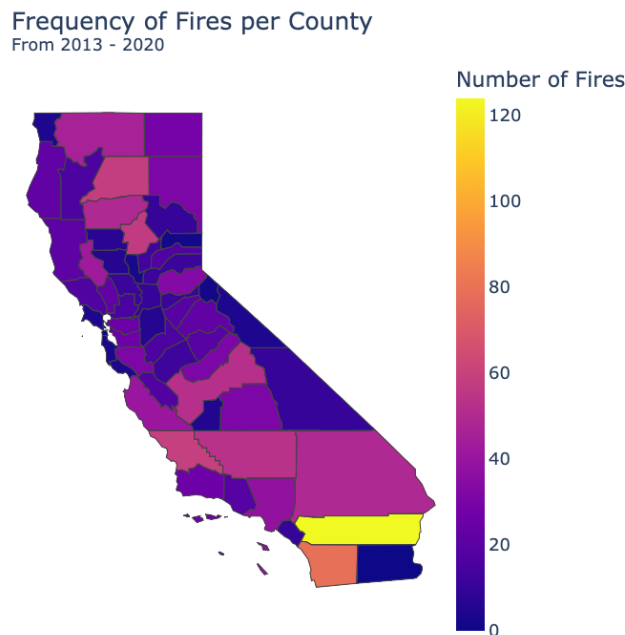


Figure 4.1: Choropleth

Chapter 5

Grammar of Graphics Tools

The Grammar of Graphics by Leland Wilkinson is a framework that provides a unique foundation for producing almost every quantitative graphic found in scientific journals, newspapers, statistical packages, and data visualization systems. It follows a layered approach to describe and construct visualizations or graphics, and it is a system of rules for generating valid statements in a language, making it possible for people to communicate accurately and concisely [6]. In this section we are going to take a look at the different layers of the Grammar of Graphics and construct visualizations using the framework at hand.

The Layers: An Overview

The Grammar of Graphics follows a layered approach to describe and construct visualizations or graphics. It provides a common language for thinking about the ways that design choices are made in visualization, describing everything from the data used to the visual channels displayed on the marks, and how data is converted into those channels.

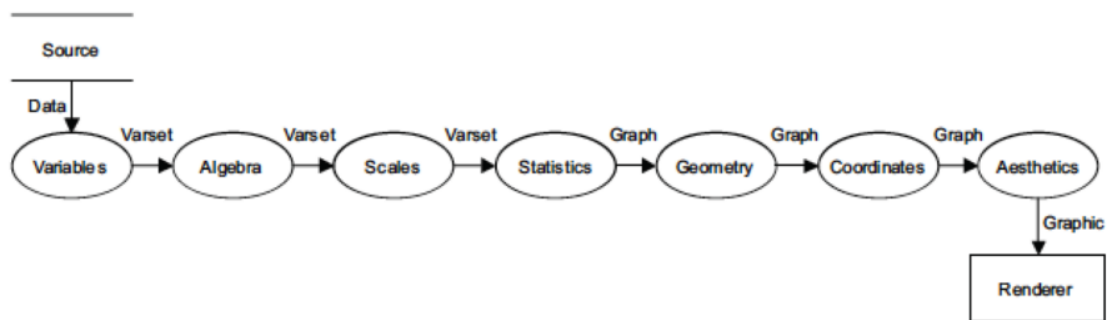


Figure 5.1: Grammar of Graphics Flow [7]

In the graphic above the layered approach is displayed. After the data was loaded, different transformations get applied to stick together the features that should be displayed. Each step yields a new “varset” up until the Statistics layer. This layer is responsible for calculating the statistical properties of the data. The succeeding layers are responsible for visual modifications which at the end form a graphic.

Data

In the first step, as described above, we aggregate the data we want to visualize.

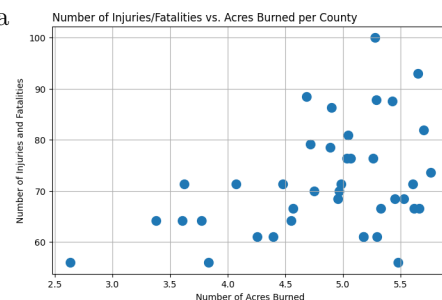


Figure 5.2: Grammar of Graphics:
Data

Aesthetics

In the Grammar of Graphics, the Aesthetics Layer refers to the mapping of one or more variables to one or more visual elements on the graph. This includes mapping variables to the x-axis, y-axis, and using color to differentiate different attributes. Aesthetics are essential in creating meaningful and effective visualizations [8]. In this example we map the `AcresBurned` variable to the x-axis and the sum of `Injuries` and `Fatalities` variables to the y-axis.

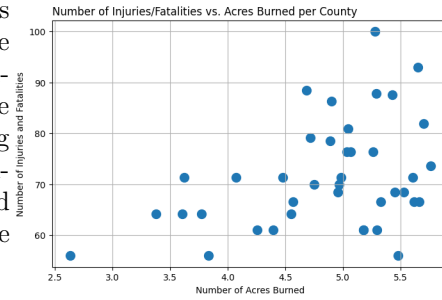


Figure 5.3: Grammar of Graphics: Aesthetics

Scale

Geometry

Statistics

Facets

Coordinates

Chapter 6

Evaluation

Bibliography

- [1] Jack Dougherty and Ilya Ilyankou. Design Choropleth Colors & Intervals. In *Hands-On Data Visualization*. Picturedigits Ltd., 2021.
- [2] L.D. Bergman, B.E. Rogowitz, and L.A. Treinish. A rule-based tool for assisting colormap selection. In *Proceedings Visualization '95*, pages 118–125,, Atlanta, GA, USA, 1995. IEEE Comput. Soc. Press.
- [3] Legends | Data Visualization Standards. <https://xdgov.github.io/components/legends>.
- [4] Carbon Design System. <https://www.carbondesignsystem.com>.
- [5] Jack Dougherty and Ilya Ilyankou. Normalize Choropleth Map Data. In *Hands-On Data Visualization*. Picturedigits Ltd., 2021.
- [6] *The Grammar of Graphics*. Statistics and Computing. Springer-Verlag, New York, 2005.
- [7] How To Make a Pie. In Leland Wilkinson, editor, *The Grammar of Graphics*, Statistics and Computing, pages 23–40. Springer, New York, NY, 2005.
- [8] Aesthetics. In Leland Wilkinson, editor, *The Grammar of Graphics*, Statistics and Computing, pages 255–318. Springer, New York, NY, 2005.