DM & Visualization - Exam 2021-02-15 - Solution

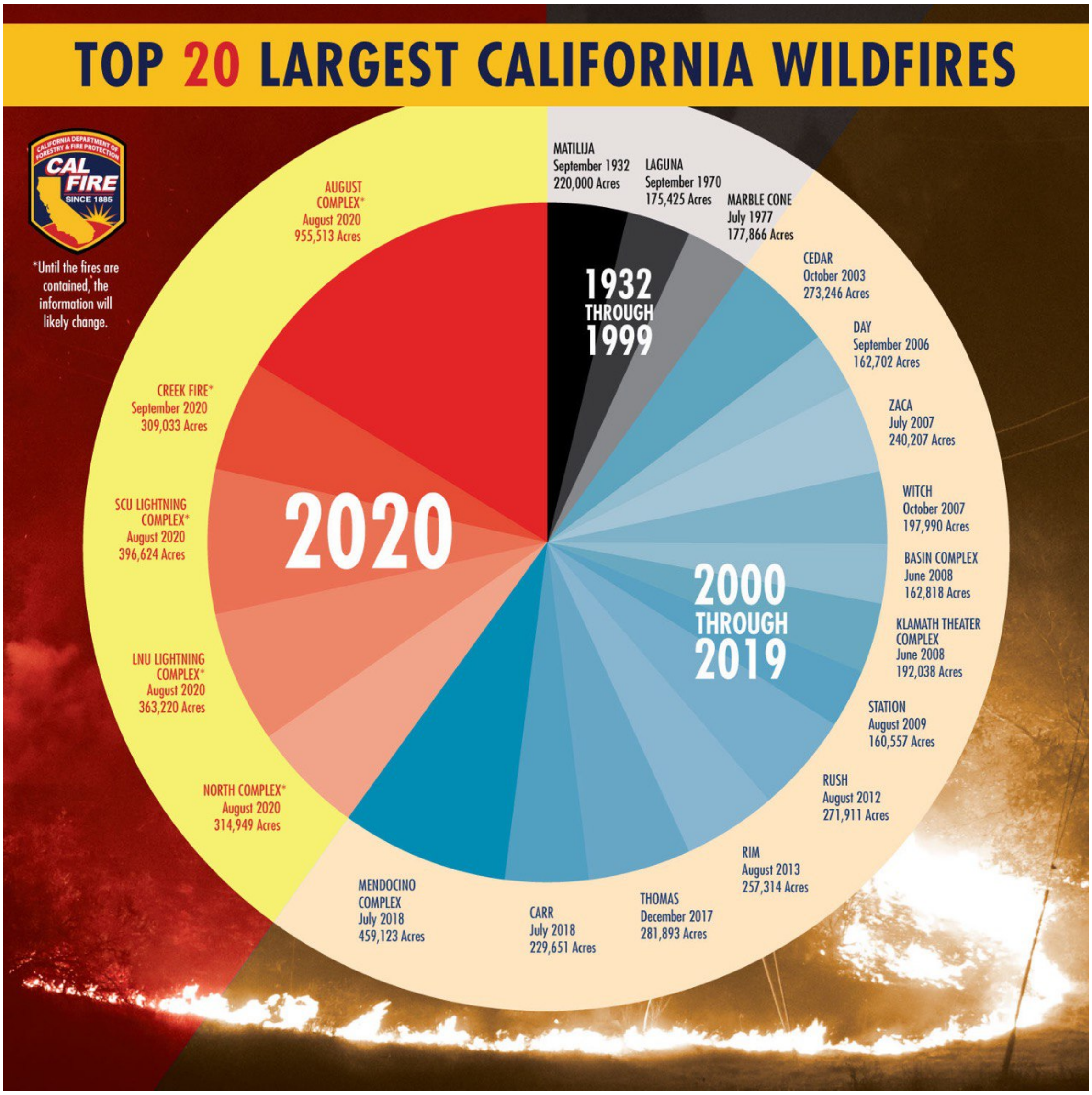


Figure 1: Top 20 largest California wildfires

Analysis

Analyze the above graph illustrating the top 20 largest California wildfires. The visualization was created by CAL FIRE, that is the California Department of Forestry and Fire Protection.

Question: Is there one (or more) question addressed by the visualization?

The question is quite clear: what is the size of wildfires that happened in 2020 in California and how does it compare with the size of wildfires from 2000-2019 and 1932-1999?

Data: Is the data quality appropriate?

Accuracy: the values reported are reasonable, the largest wildfire is about 0.9% the area of California.

Completeness: data are not complete, as only the top 20 wildfires are considered.

Consistency: the three different timeframes are of different lengths. The values of 2020 are estimated.

Currency: data were last updated in September 2020, but it was current when the visualization was created.

Credibility: the source is mentioned in the logo and it is trusted because it is a government agency.

Understandability: data are understandable in the USA, but in general it is better to use square kilometers.

Precision: precision is too detailed, apart the size of the first wildfire.

Visual Proportionality: Are the values encoded in a uniformly proportional way?

The slices of the piechart and the colors of the timeframes are proportional to the size of the wildfires. However, this visualization has serious perceptual problems because it is very difficult to compare areas and shades.

Visual Utility: All the elements in the graph convey useful information?

Several elements are useless: the image in the background, the CAL FIRE logo, the different shades for each timeframe, the donut around the piechart.

Visual Clarity: Are the data in the graph clearly identifiable and understandable (properly described)?

The usage of direct labeling is appropriate. The choice of piechart is wrong because the values are not part of a whole. The shades for each timeframe are difficult to interpret. It is not clear that only 20 wildfires are considered.

Design

Design the visualization based on the following data structure

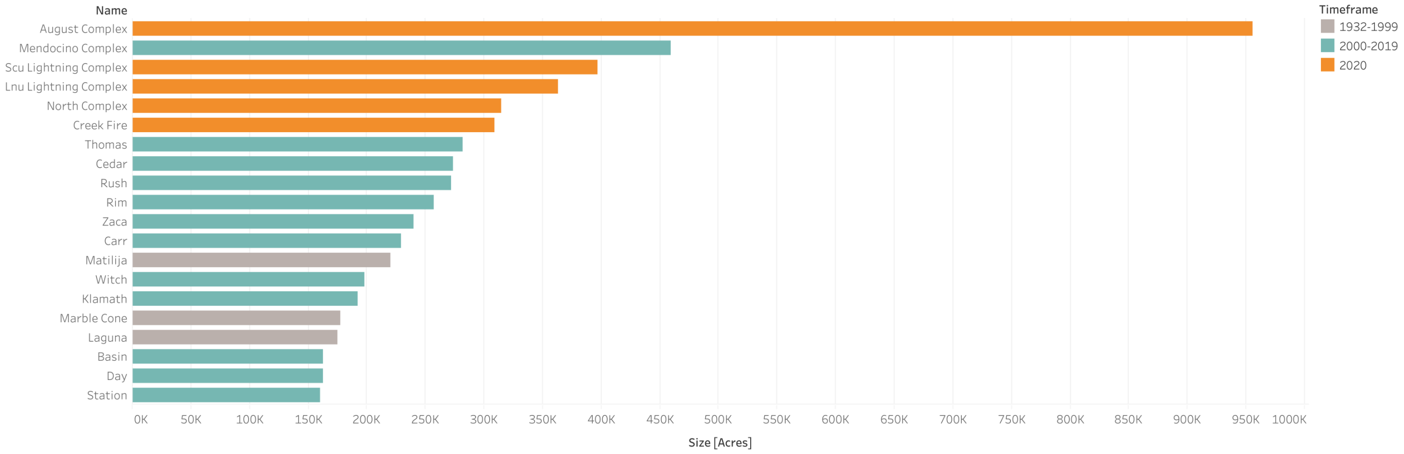
Field	Dim./Measure	Description
FIRE_NAME	Dimension	The name of the wildfire
FIRE_SIZE	Measure	The area affected by the wildfire
FIRE_DATE	Dimension	The month and year when wildfire happened
FIRE_TIMEFRAME	Dimension	The three timeframes: 1932-1999, 2000-2019, 2020

Design schema

Schema	Details
Columns:	SUM(FIRE_SIZE)
Rows:	FIRE_NAME
Graph type:	Bar
Color:	FIRE_TIMEFRAME
Size:	Default
Label:	Default

Sketch of the resulting graph

Bar chart by fire

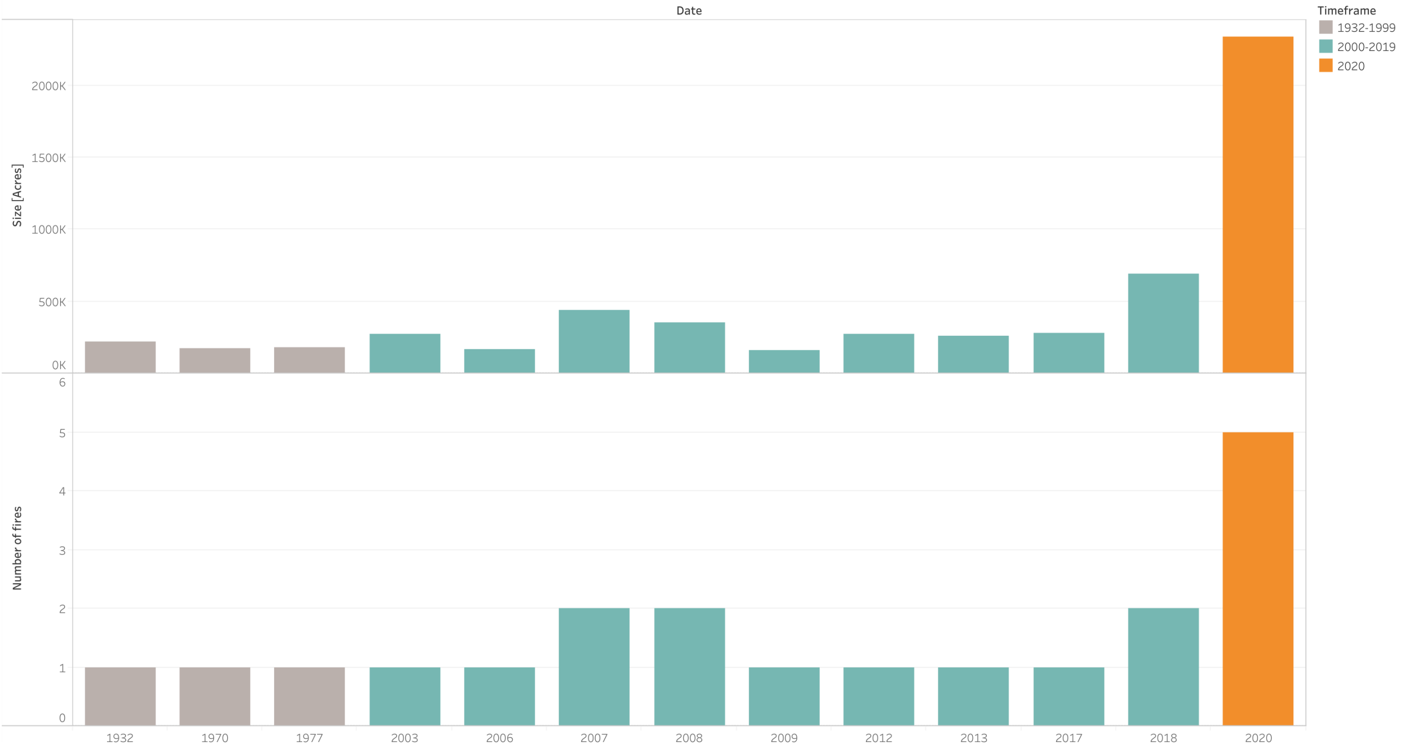


Design schema

Schema	Details
Columns:	YEAR(FIRE_DATE)
Rows:	SUM(FIRE_SIZE), CNT(FIRE_SIZE)
Graph type:	Bar
Color:	FIRE_TIMEFRAME
Size:	Default
Label:	Default

Sketch of the resulting graph

Bar chart by year

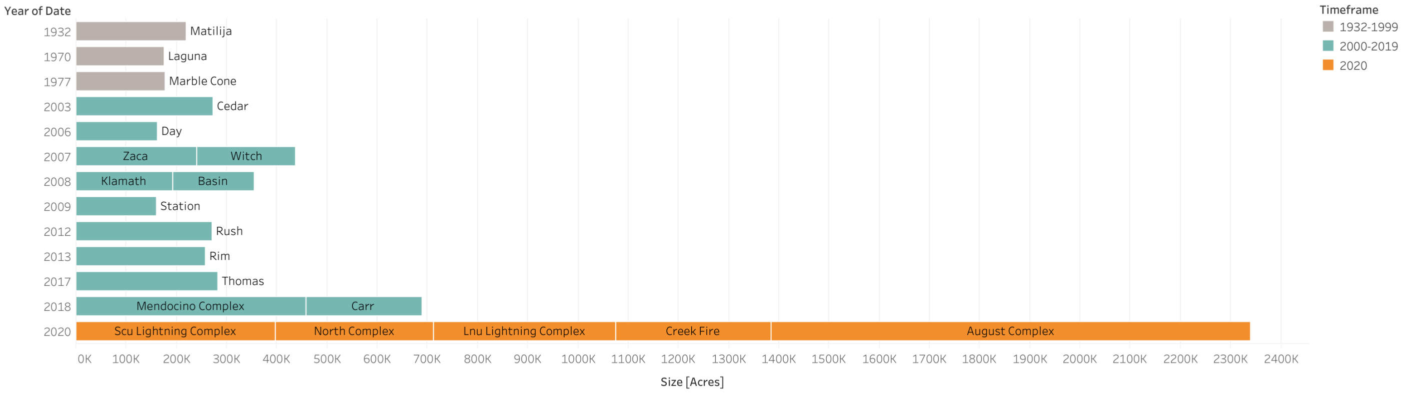


Design schema

Schema	Details
Columns:	SUM(FIRE_SIZE)
Rows:	YEAR(FIRE_DATE)
Graph type:	Bar
Color:	FIRE_TIMEFRAME
Size:	Default
Label:	FIRE_NAME

Sketch of the resulting graph

Stacked bar chart by year

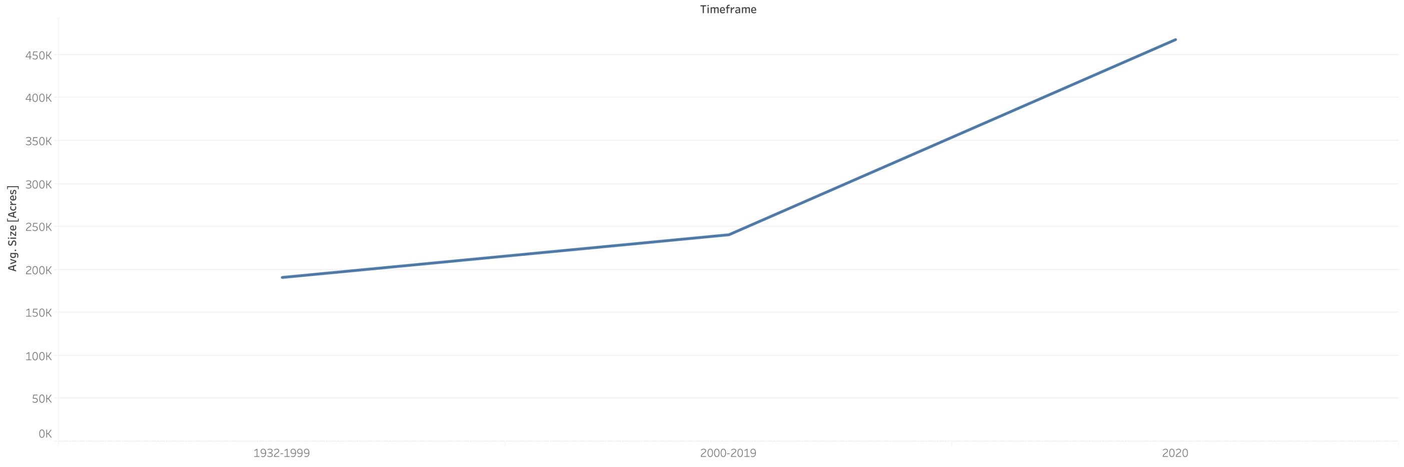


Design schema

Schema	Details
Columns:	FIRE_TIMEFRAME
Rows:	AVG(FIRE_SIZE)
Graph type:	Line
Color:	Default
Size:	Default
Label:	Default

Sketch of the resulting graph

Line chart per avg. size



Theory

Which one of the following visualizations is the most appropriate one for representing a measure as a statistical distribution, a dimension with a high cardinality and another dimension with a low cardinality? For example, think about a visualization representing incomes (measure), level of education (dimension with high cardinality), and gender (dimension with low cardinality).

- Gauges
- *Multiple box plots*
- Stacked bars
- Pie charts
- Heatmaps