Markov ACADEMY

Data Analysis & Visualization

Eman Raslan

WHAT IS DATA VISUALIZATION?

Why do We Visualize Data?

Which **MONTH** had the **HIGHEST** sales?

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
$44,703	$20,301	$58,872	$36,522	$44,261	$52,982	$45,264	$63,121	$87,867	$77,777	$118,448	$83,829

3...2...1...


NOOOOOO! NOOOOOOOO!
I WASN'T READYYYYYY!!!!
quickmeme.com

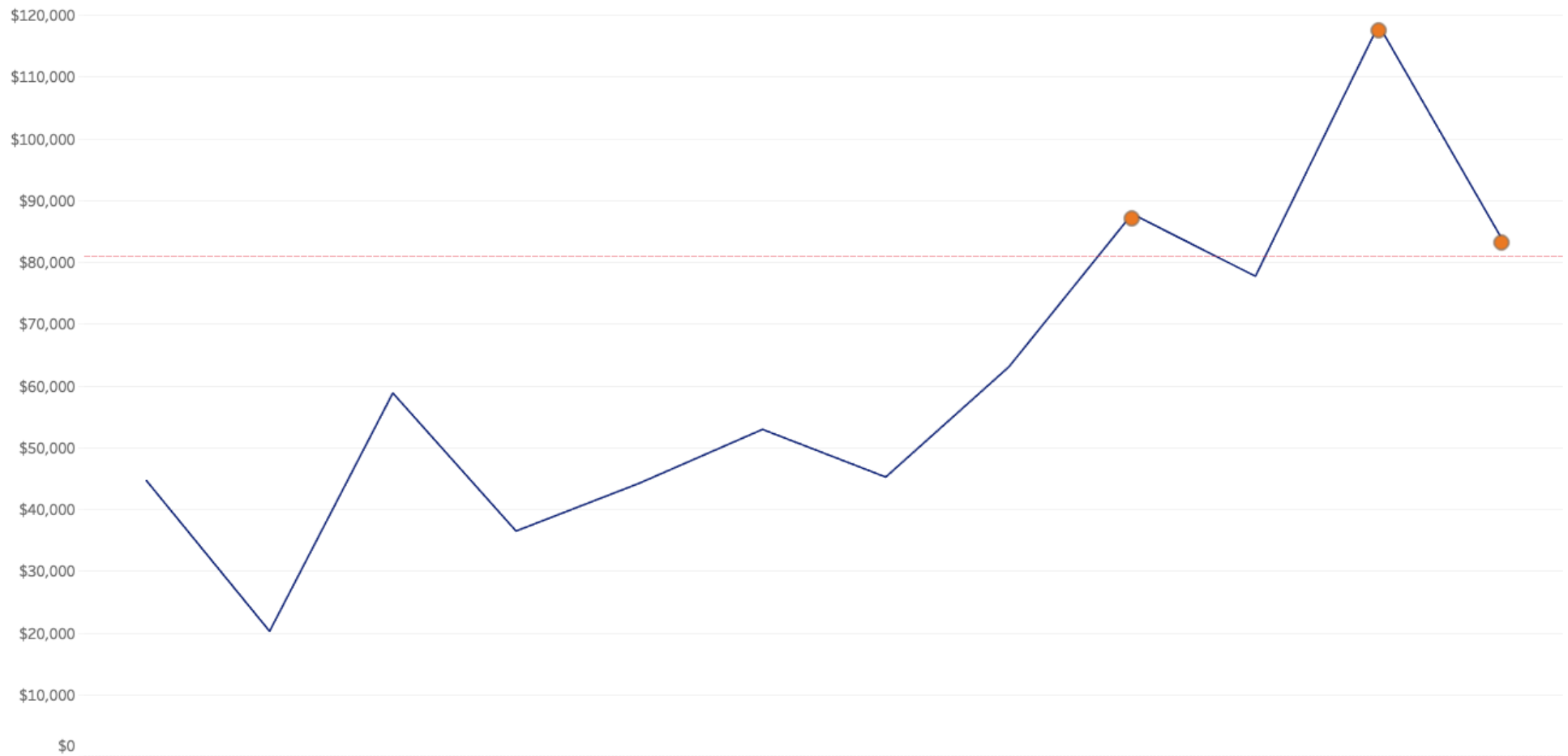
Can you name the **TOP 3** performing months?

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
$44,703	$20,301	$58,872	$36,522	$44,261	$52,982	$45,264	$63,121	$87,867	$77,777	$118,448	$83,829

3...2...1...


THAT MOMENT WHEN YOUR TEACHER CALLLS ON YOU
AND YOUR NOT PAYING ATTENTION
imgflip.com

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
$44,703	$20,301	$58,872	$36,522	$44,261	$52,982	$45,264	$63,121	$87,867	$77,777	$118,448	$83,829



Data Visualization in Data Science

The First Mile …

Data Exploration, Exploratory Data Analysis

Making many interactive graphics quickly for:

- ‘Getting a sense’ of the data.
- Looking for patterns & outliers.
- Validating data quality.

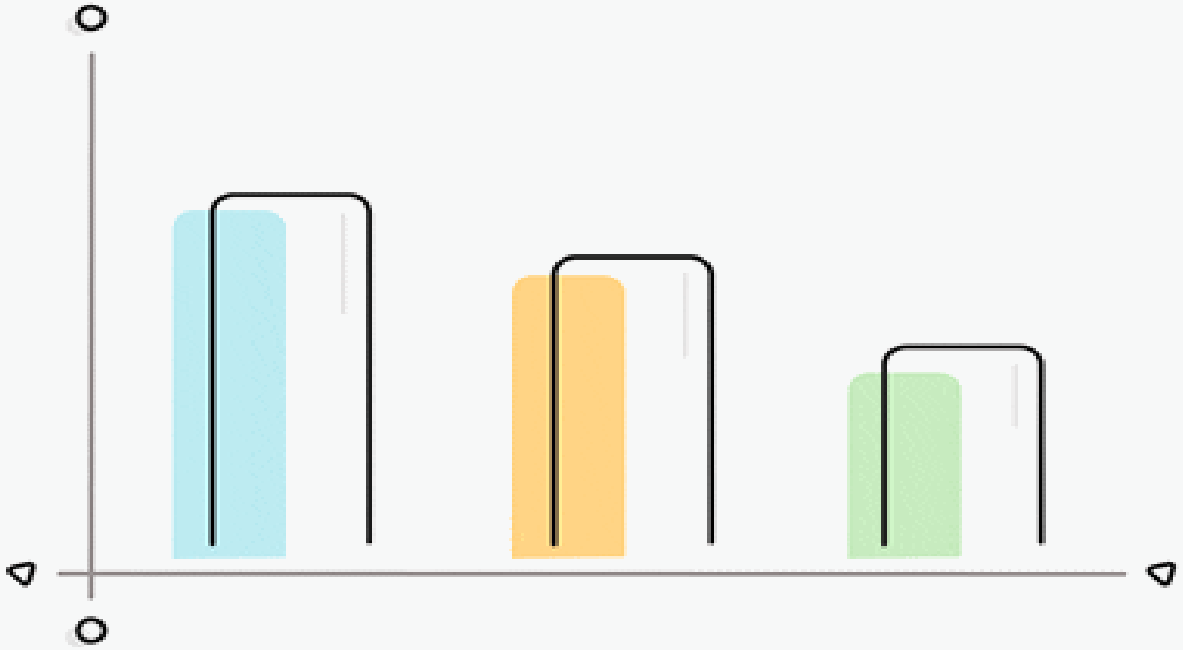
… and the Last Mile.

Communication, Explanation, Operationalization

Making fewer, more polished interactive graphics for:

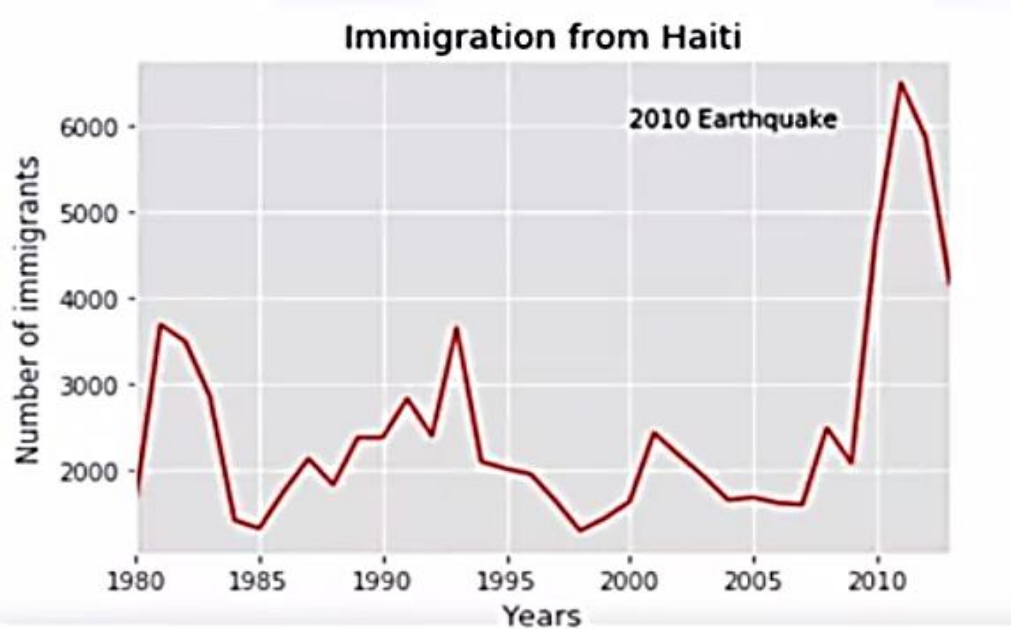
- Publication in articles or on the web.
- Non-data- scientists to interact with outputs.
- Making models actionable.

How to choose the right chart?!



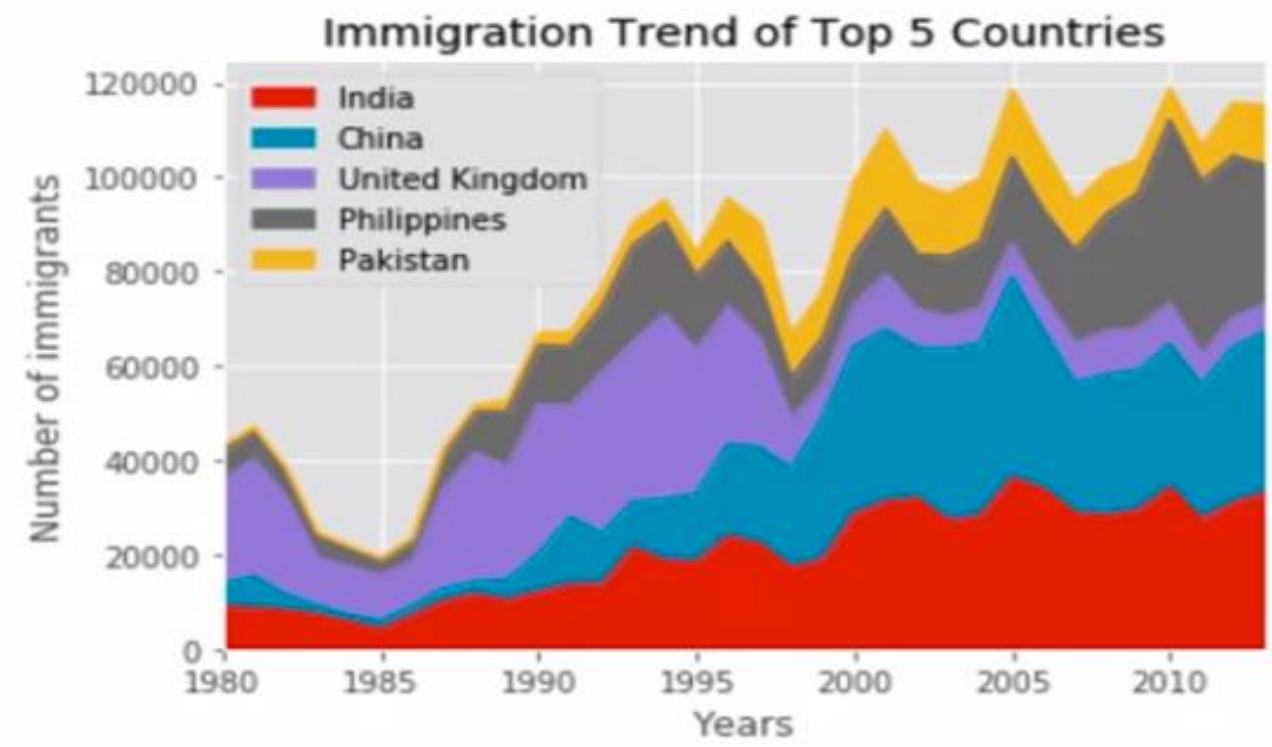
Line plots

- A line plot is a type of plot which displays information as a series of data points called ‘markers’ connected by straight line segments.
- The best use case for line plot is when you have a **continuous** dataset and you are interested in visualizing the data over a period of time.



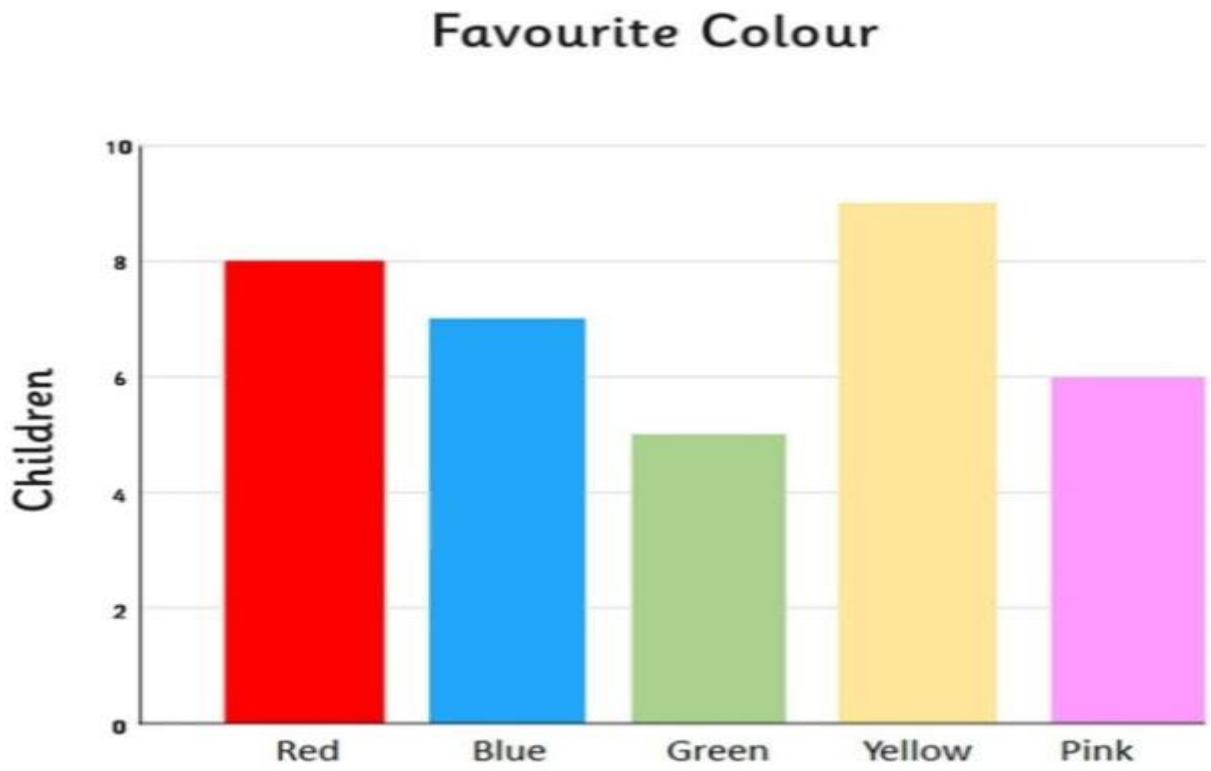
Area Plots

- An extension of line plots.
- Commonly used to represent and Compare cumulated totals using numbers or percentage over time.



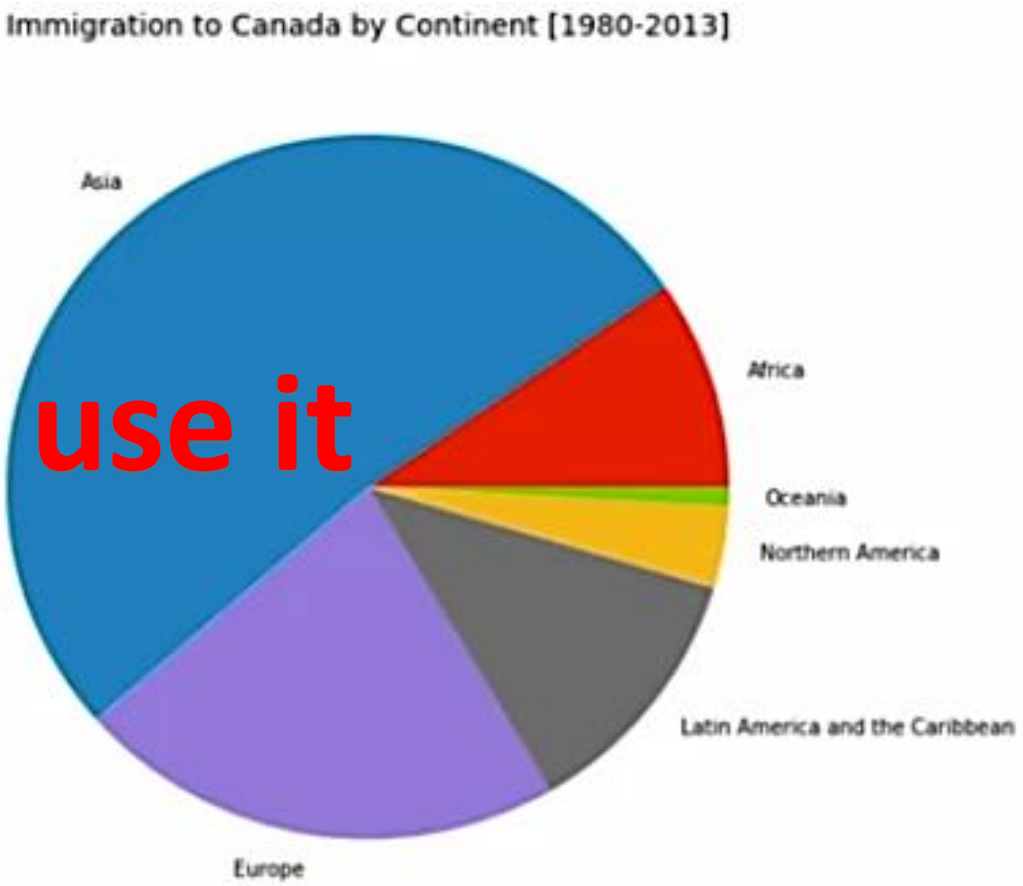
Bar Charts

- Unlike a histogram, a bar chart is commonly used to compare the values of a variable at a given point in time.



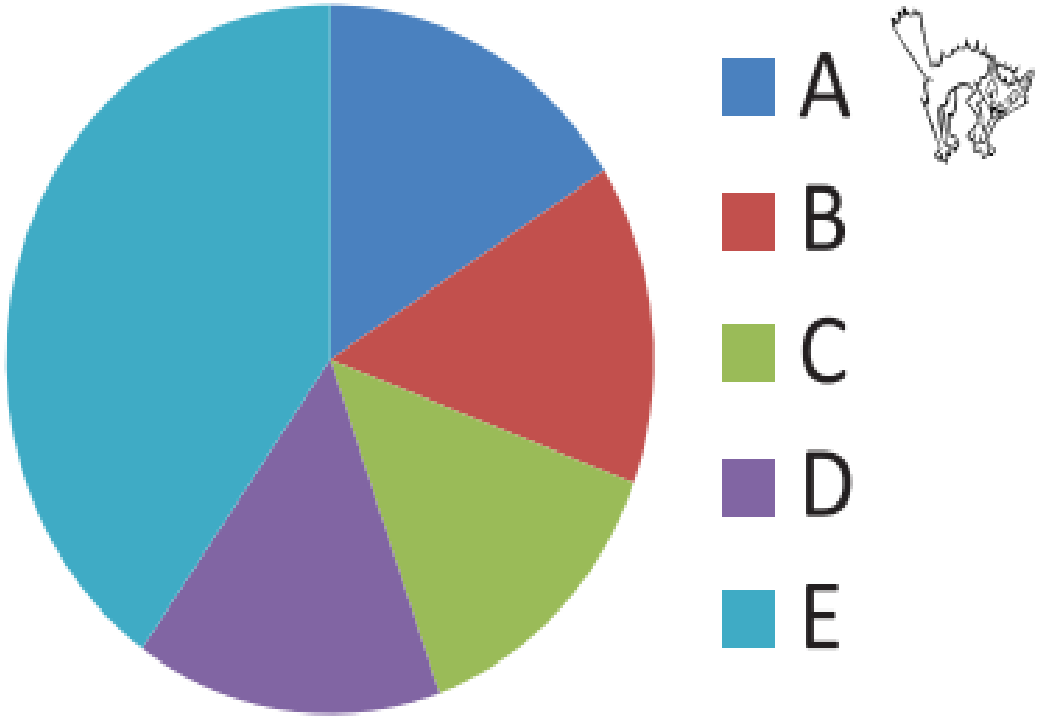
Pie Charts

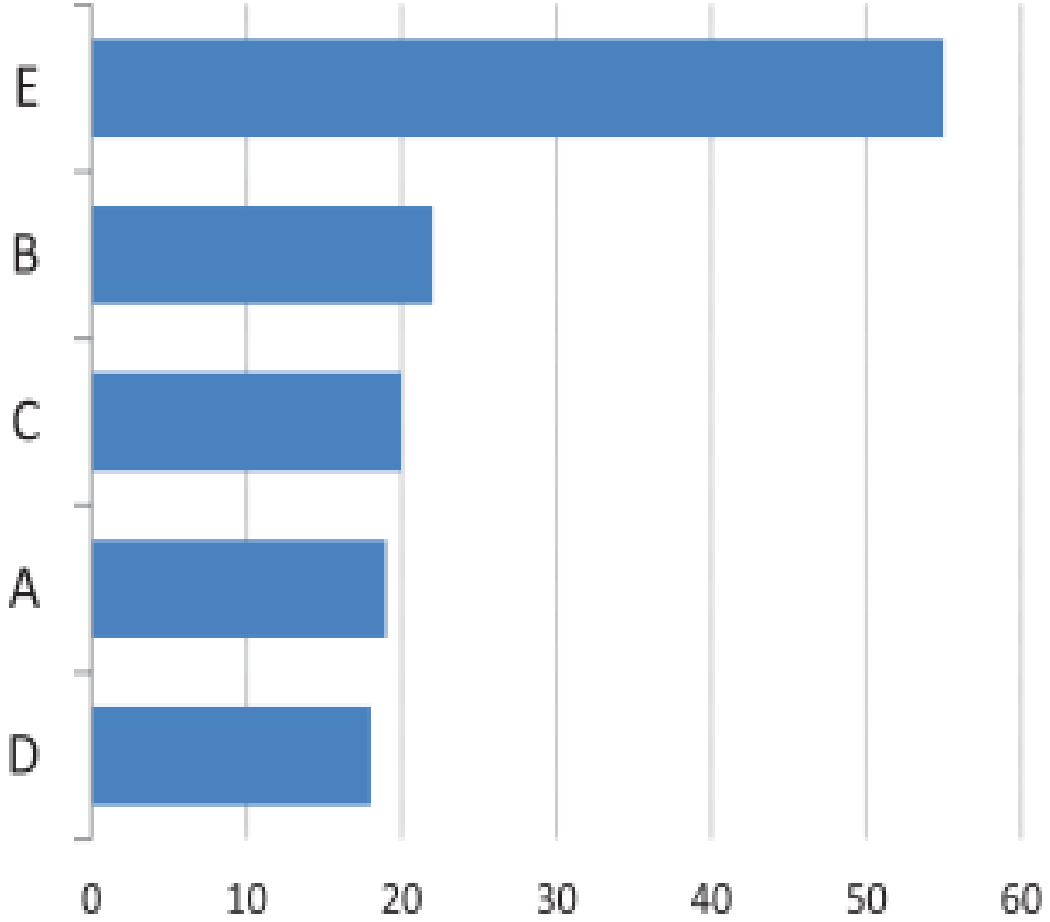
- Is a circular statistical graphic divided into slices to illustrate numerical proportion.

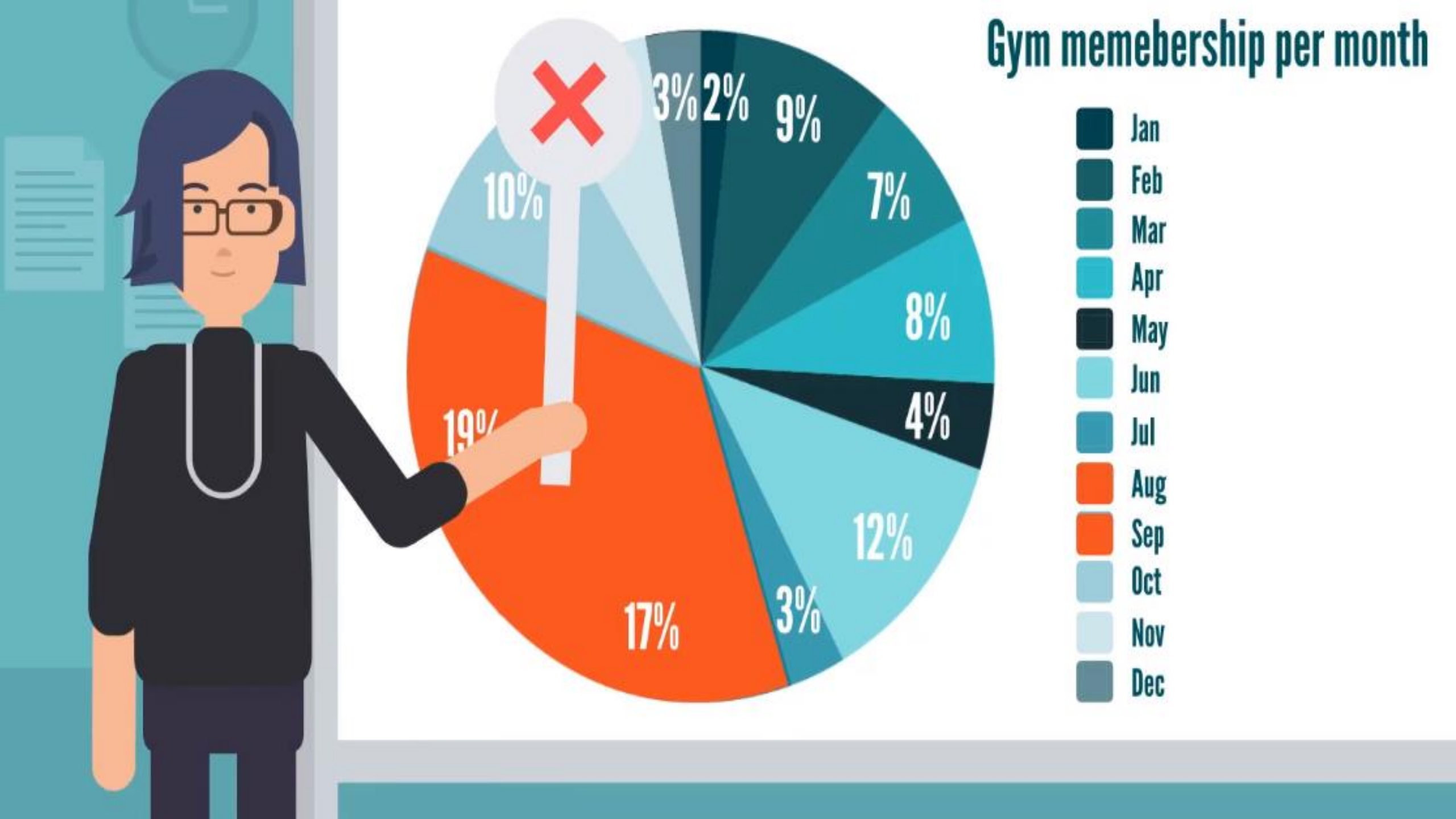


Don’t use it

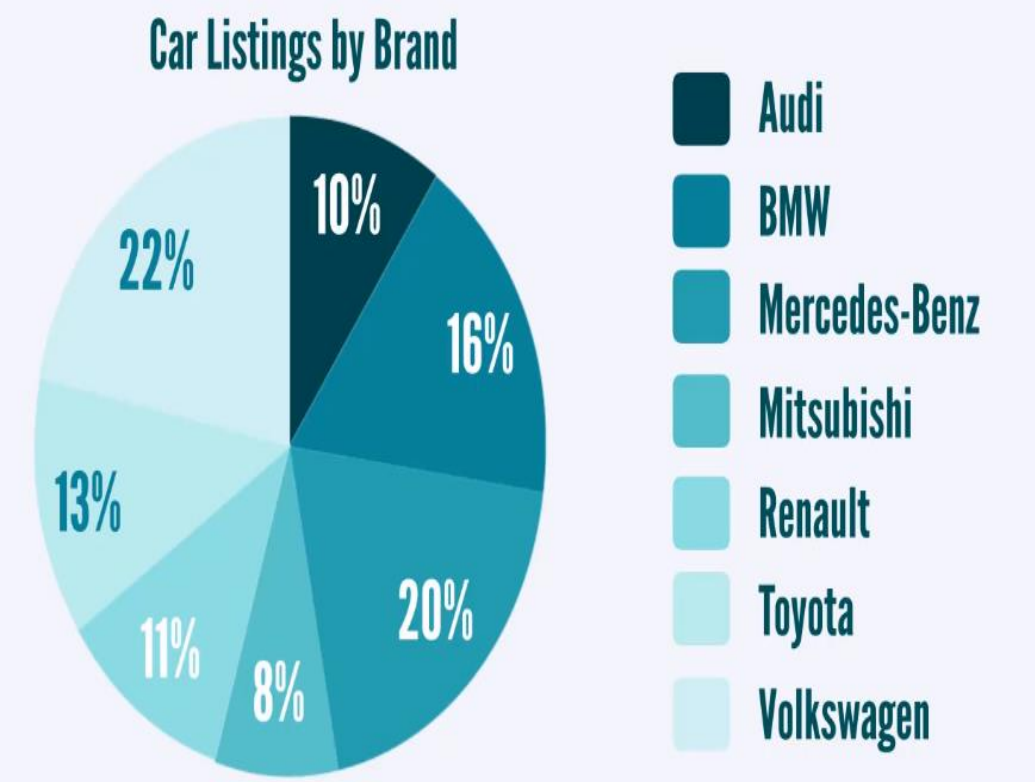
Can you order the slices from biggest to smallest?

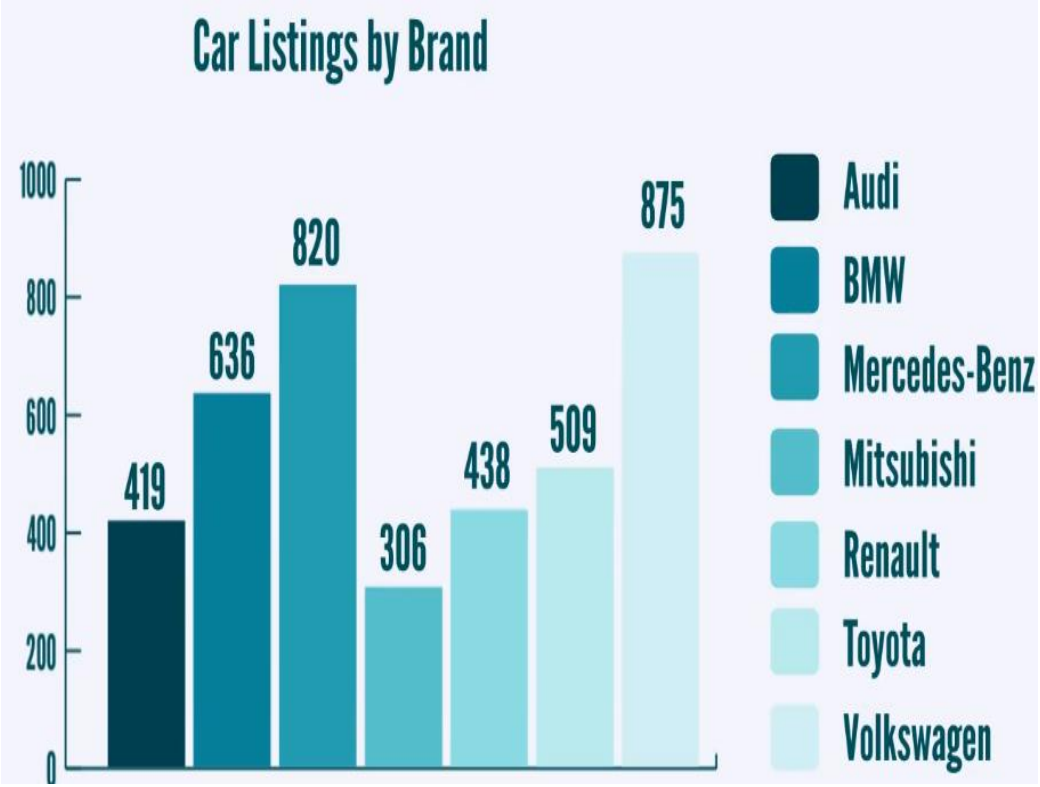




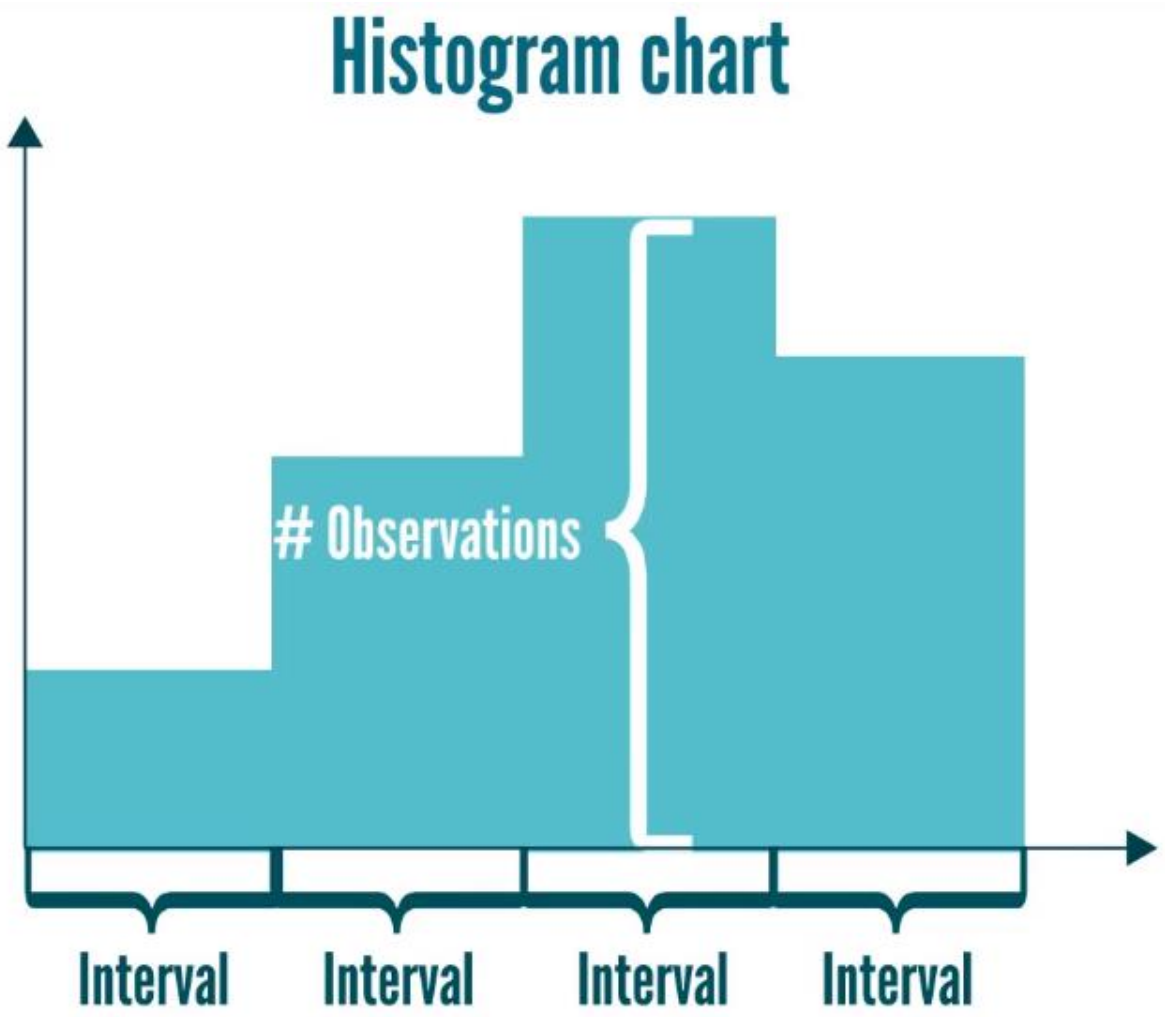

Gym memebership per month
3%
2%
9%
7%
8%
4%
12%
3%
17%
19%
10%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Which is better?

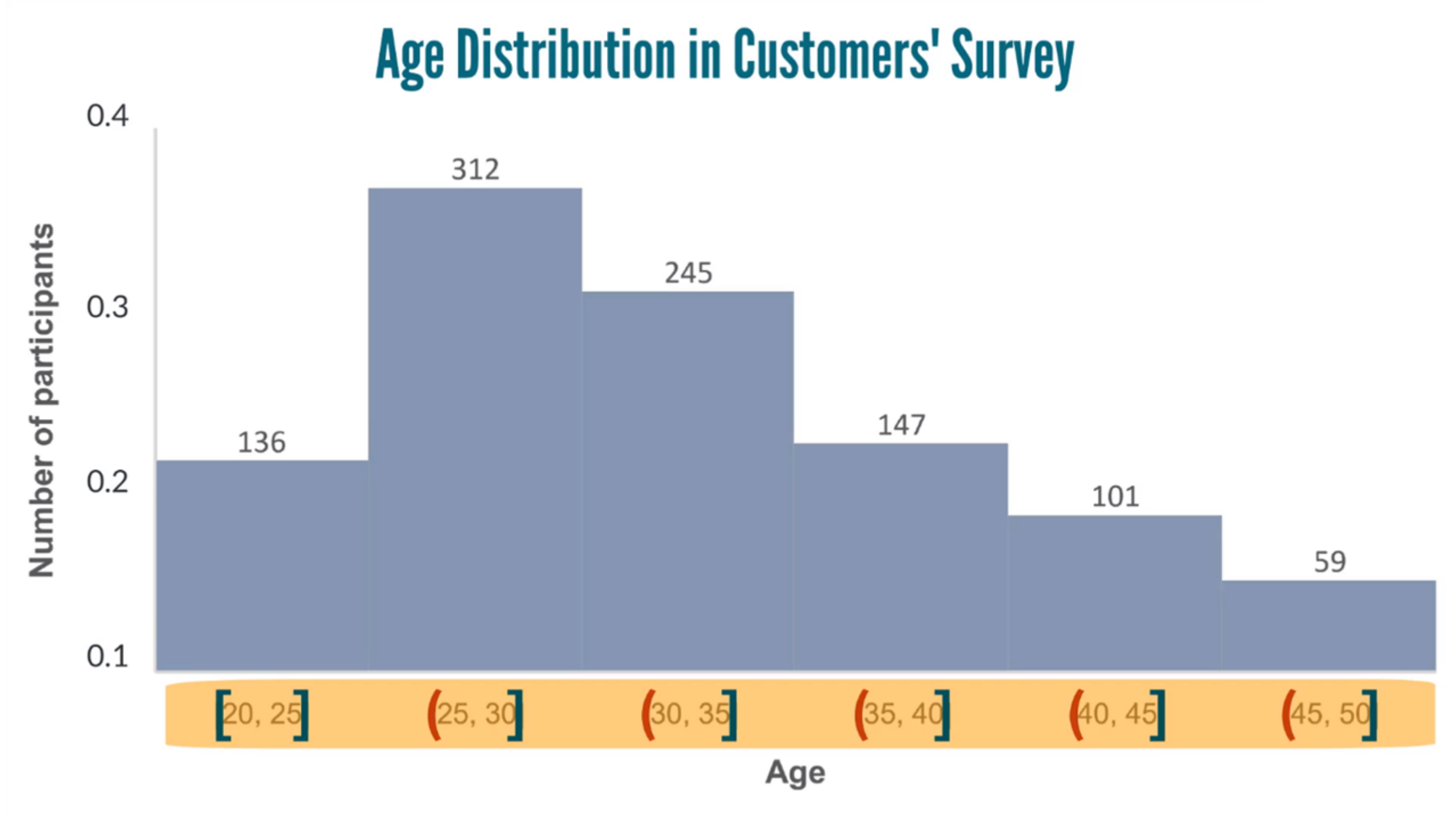




Histograms

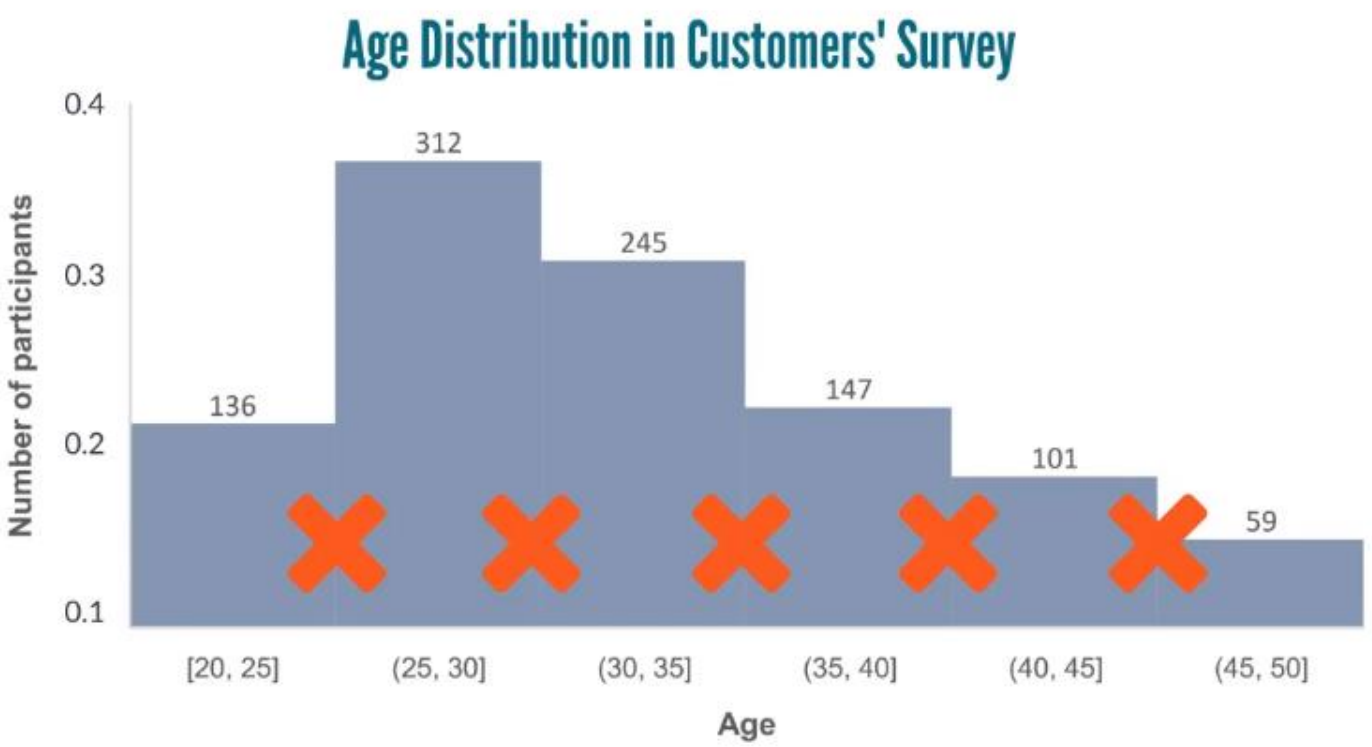


- Shows the distribution of a numeric variable
- The variable's range of values is split into intervals, represented by different bins
- The height of the bins shows the number of observations within an interval


Age Distribution in Customers' Survey
0.4
0.3
0.2
0.1
Number of participants
312
245
136
147
101
59
[20, 25]
(25, 30]
(30, 35]
(35, 40]
(40, 45]
(45, 50]
Age

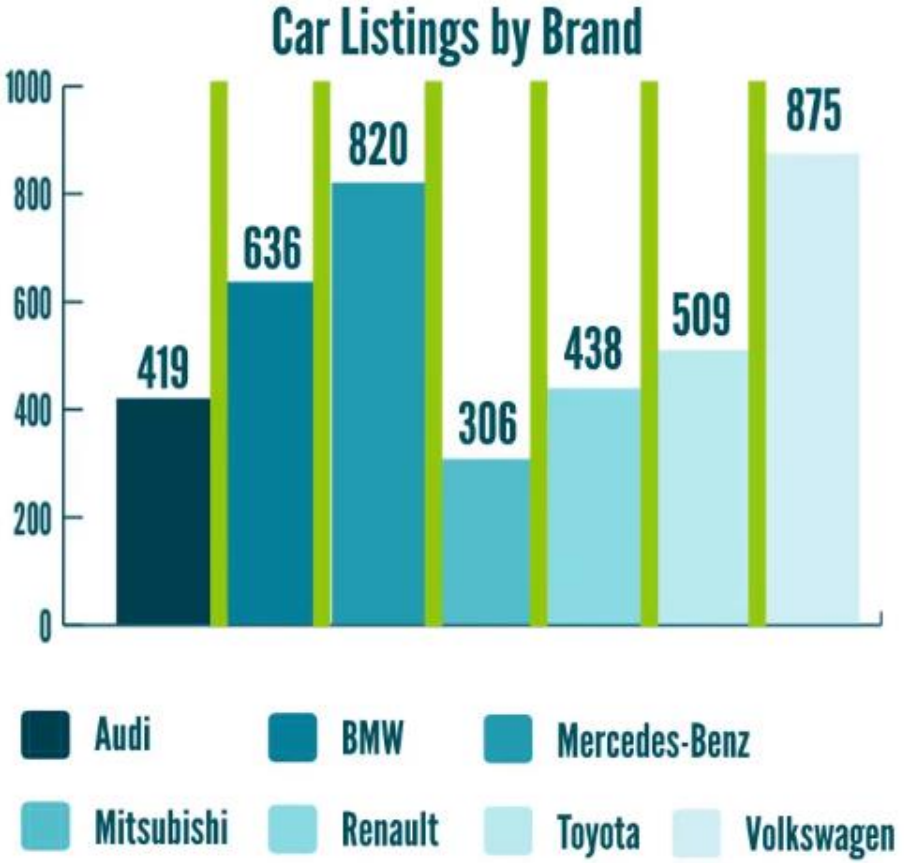
Histogram chart

- Represents bins (intervals)
- Bars are not separated



Bar chart

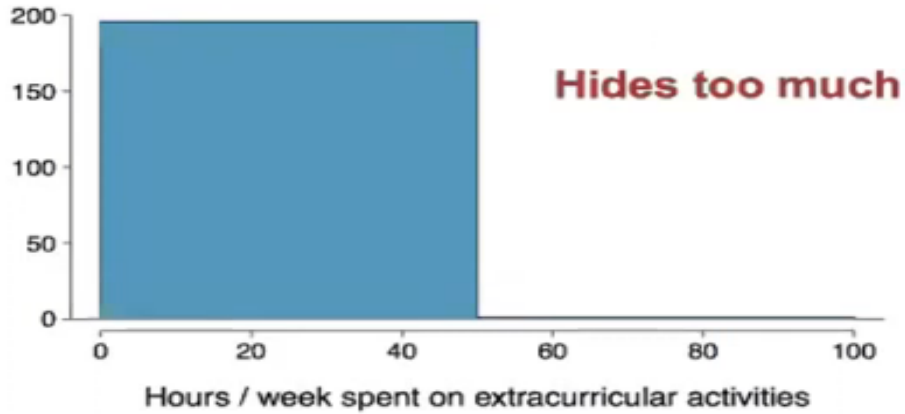
- Represents different categories
- Bars are separated



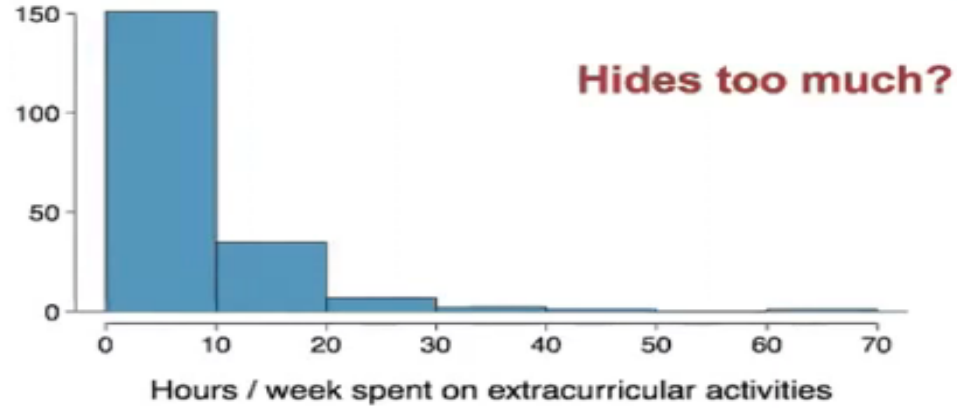
Bin Width

Which one(s) of these histograms are useful?

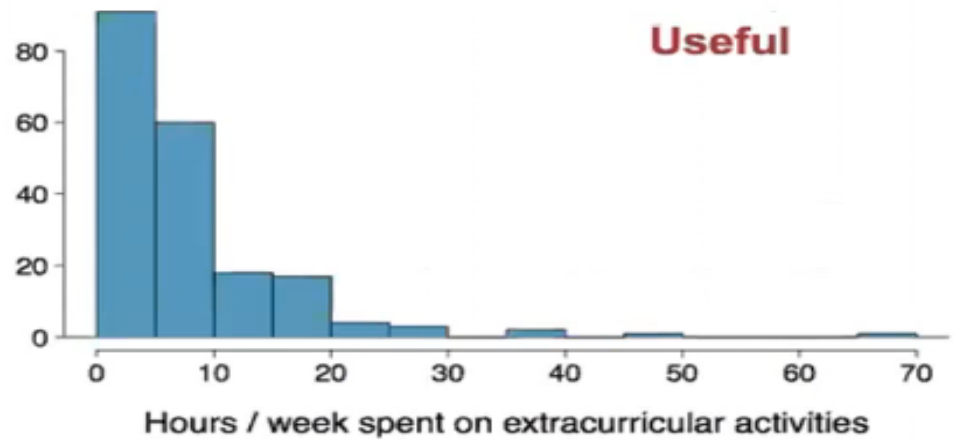
Which reveal too much about the data? Which hide too much?



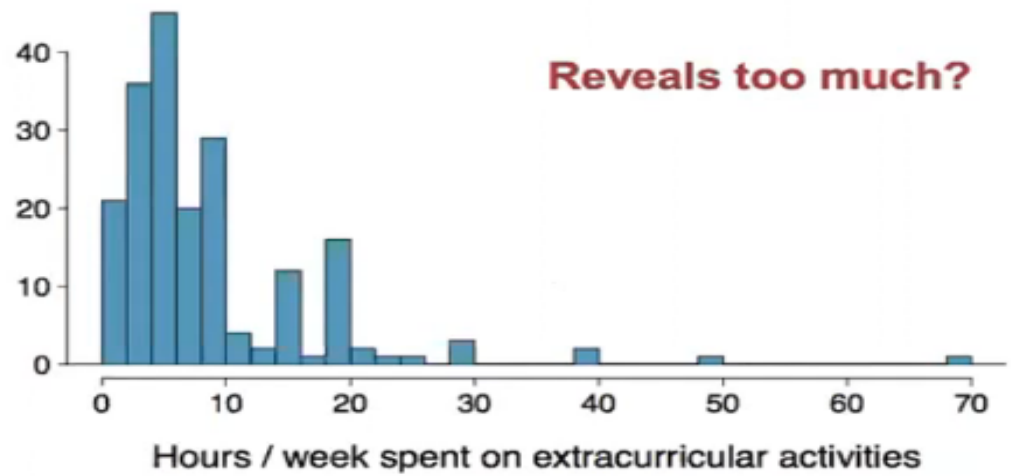
Hides too much



Hides too much?



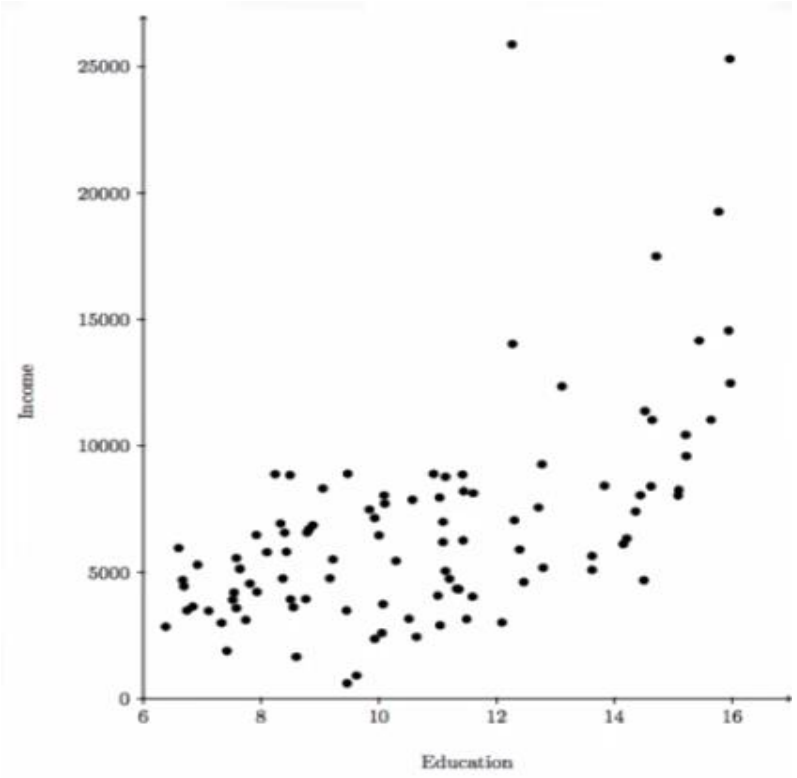
Useful



Reveals too much?

Scatter Plots

- Displays values pertaining to typically two variables against each other.
- Usually it is a dependent variable to be plotted against independent variable in order to determine if any correlation between the two variables exists.

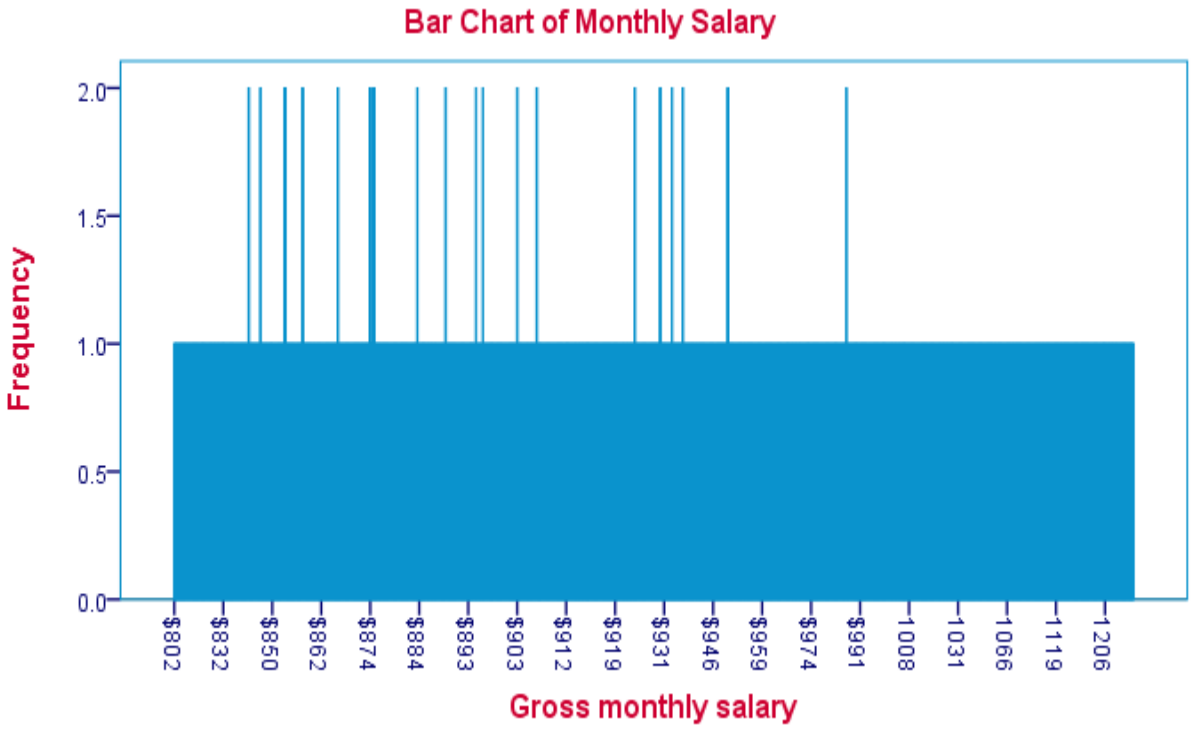


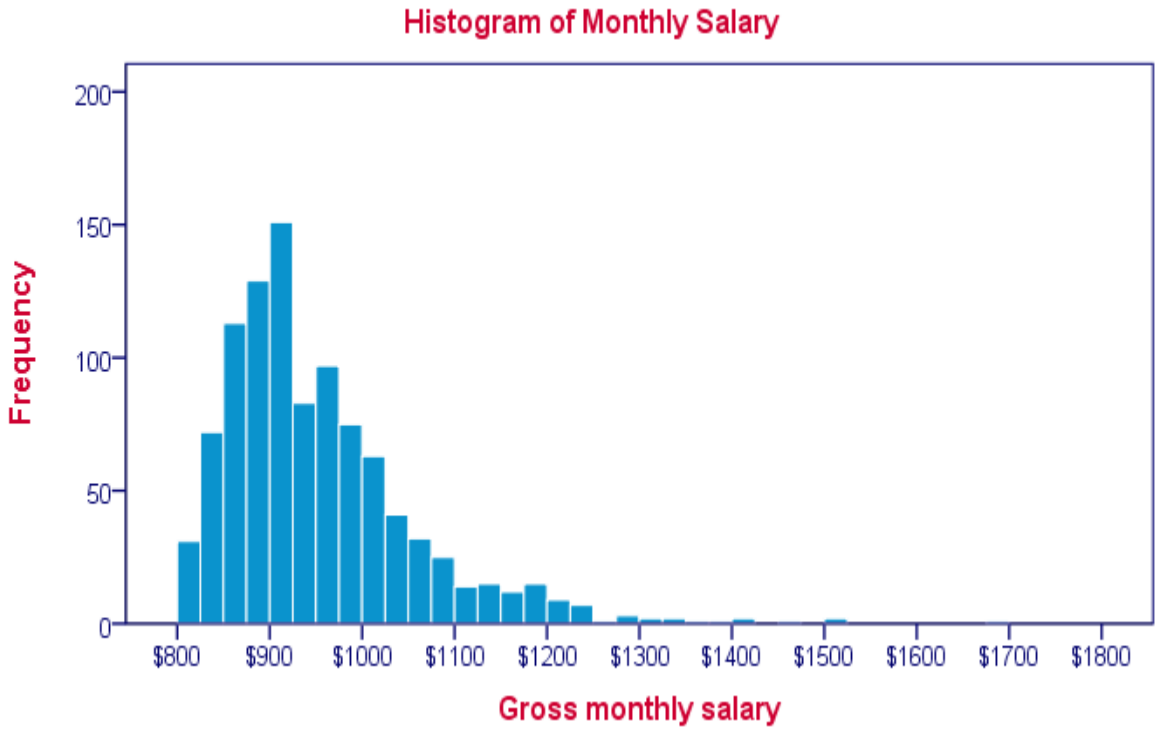
What is a good visualization?

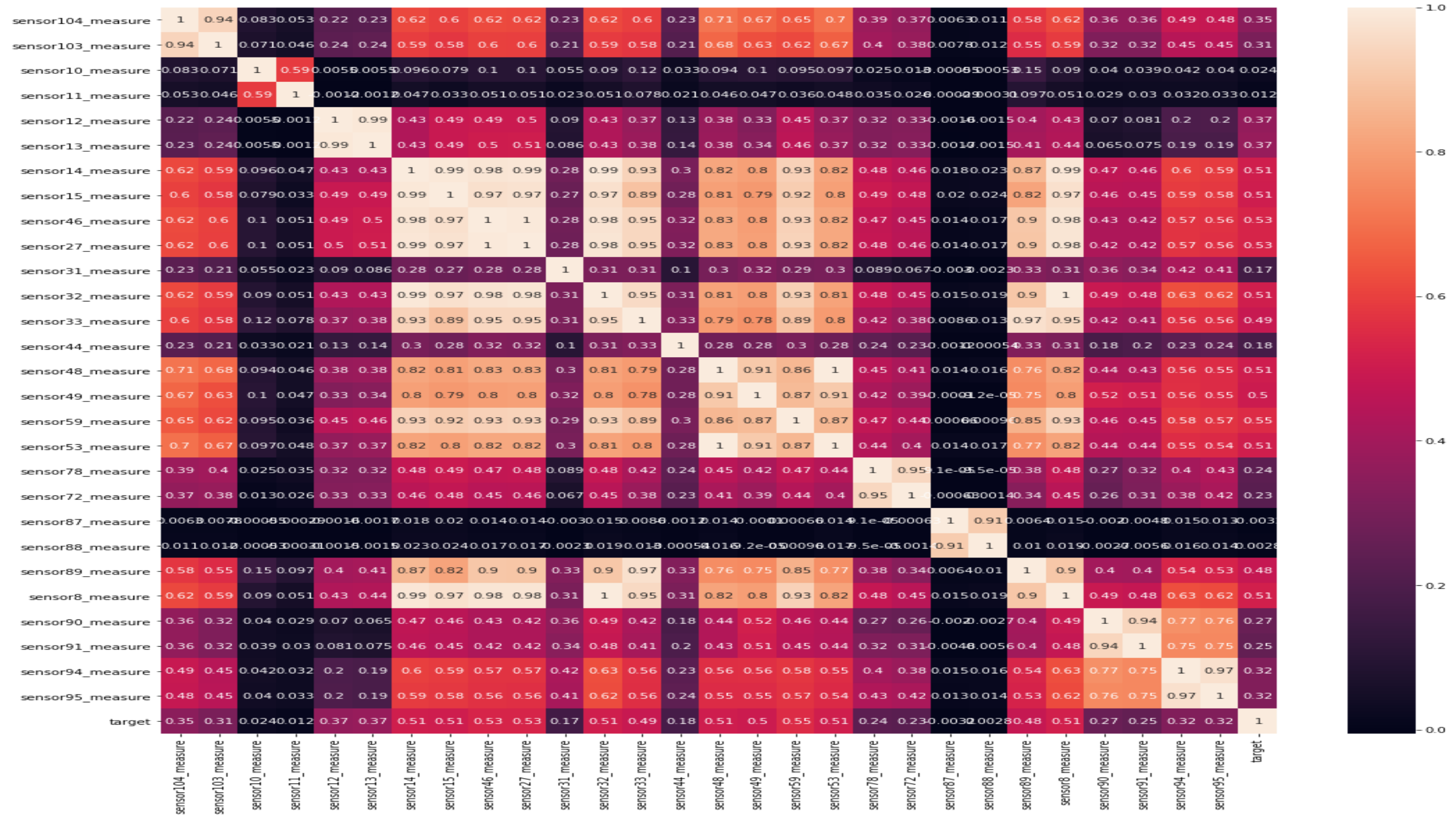
- A company wants to know how monthly salaries are distributed over 1,110 employees having operational, middle or higher management level jobs. The screenshot below shows what their raw data look like.

	first_name	gender	level	salary
1	Charlotte	female	Operational	$802.29
2	Eli	male	Operational	$804.73
3	Angel	male	Operational	$806.44
4	Juan	female	Operational	$807.66

What is a good visualization?



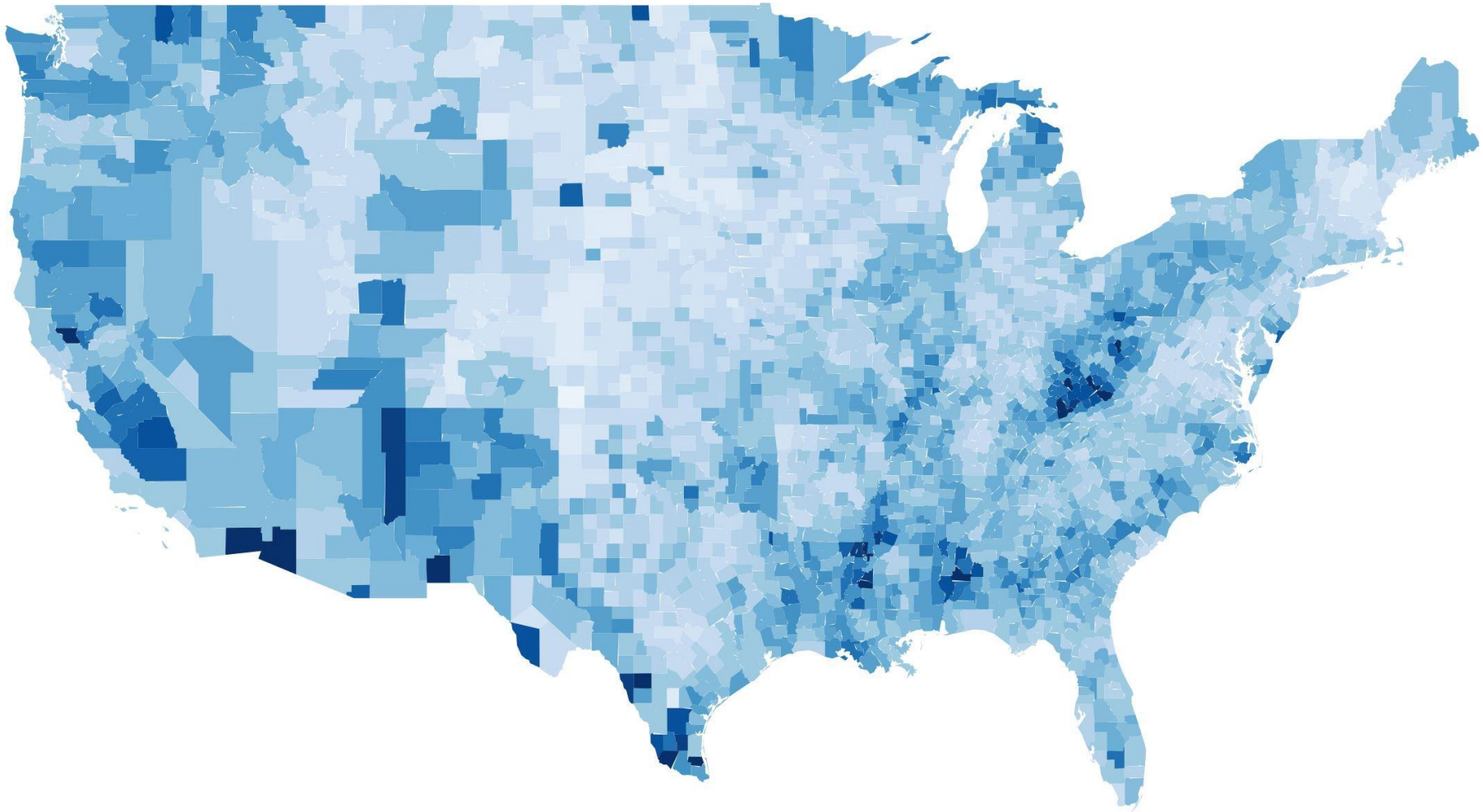



sensor104_measure
sensor103_measure
sensor10_measure
sensor11_measure
sensor12_measure
sensor13_measure
sensor14_measure
sensor15_measure
sensor46_measure
sensor27_measure
sensor31_measure
sensor32_measure
sensor33_measure
sensor44_measure
sensor48_measure
sensor49_measure
sensor59_measure
sensor53_measure
sensor78_measure
sensor72_measure
sensor87_measure
sensor88_measure
sensor89_measure
sensor8_measure
sensor90_measure
sensor91_measure
sensor94_measure
sensor95_measure
target
1.0
0.8
0.6
0.4
0.2
0.0

Word Cloud


INFECTED
INFECTION
DIAGNOSTIC
CONTAMINATION
BIOHAZARD
NATION
HEALTH
CARE
SAFETY
ATTENTION
HEALTH
PANDEMIC
HOSPITAL
CONTROL
REPORT
TERM
LOSE
OFFICIAL
SITUATION
TESTING
CHEST
REPORT
CONFIRMED
MASK
ILLNESS
PROTECTION
MEDICAL
QUARANTINE
PNEUMONIA
COVID-19
RISK
SPREADING
HUMAN
HAZARDOUS
DANGER
STOP
CORONAVIRUS
FEAR
VIRUS
SPREAD
GLOBAL
EPIDEMIC
RESPIRATORY
CLINIC
SICKNESS
CAUTION
WORLD
DEATH
PREVENTION
PATHOGEN
SCIENCE
INFECTIOUS
FEVER
SICK
CARE
PATHOLOGY
IMMUNITY
VACCINE
VIROLOGY
OUTBREAK
DIAGNOSIS
PEOPLE
HEALTHCARE
DANGEROUS
ALERT
PATIENT
BIOHAZARD
WARNING
SYMPTOM
CONTAGIOUS
LABORATORY
HAZARD
MEDICINE
DISTANCING
FEVER
DOCTOR
ANALYZING

Choropleth Maps



What would you like to show?

Comparison

- Among items
  - Two variables per item
    - Variable width chart
  - One variable per item
    - Many categories
      - Table or tables with embedded charts
    - Few categories
      - Bar chart horizontal
      - Bar chart vertical
- Over time
  - Many periods
    - Cyclical data
      - Circular area chart
    - Non-cyclical data
      - Line chart
  - Few periods
    - Single or few categories
      - Bar chart vertical
    - Many categories
      - Line chart

Relationship

- Two variables
  - Scatter plot
- Three variables
  - Scatter plot bubble size

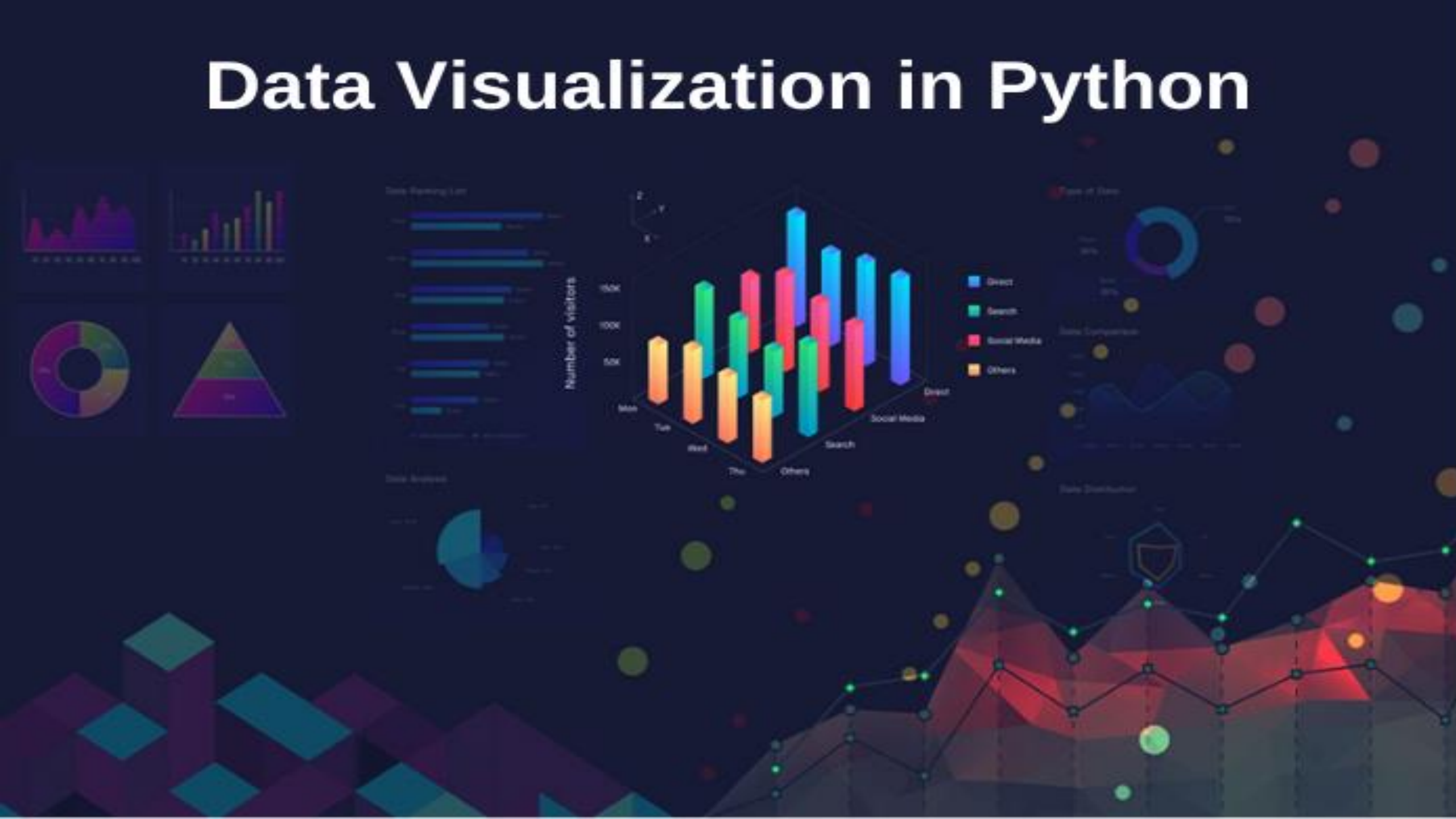
Distribution

- Single variable
  - Few data points
    - Bar histogram
  - Many data points
    - Line histogram
- Two variables
  - Scatter plot

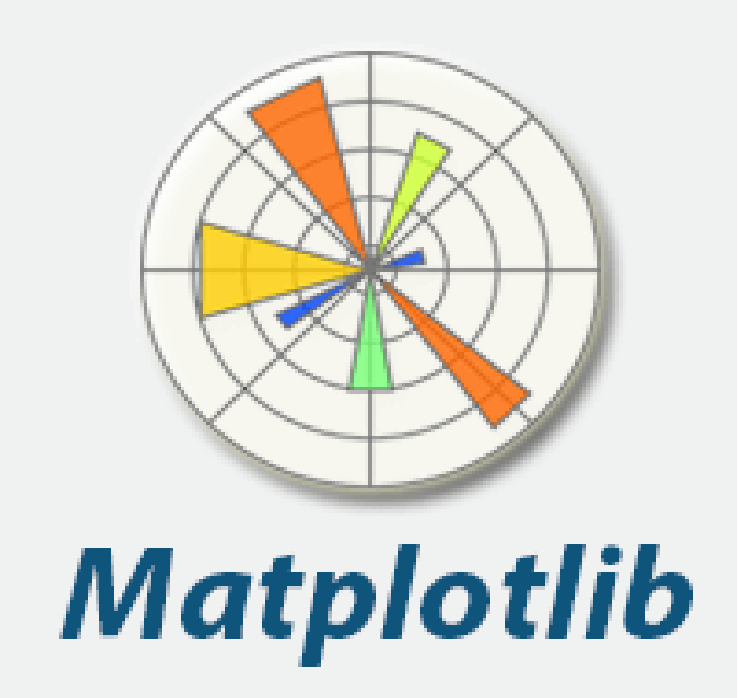
Composition

- Changing over time
  - Few periods
    - Only relative differences matter
      - Stacked 100% bar chart
    - Relative and absolute differences matter
      - Stacked bar chart
  - Many periods
    - Only relative differences matter
      - Stacked 100% area chart
    - Relative and absolute differences matter
      - Stacked area chart
- Static
  - Simple share of total
    - Pie chart
  - Accumulation or subtraction to total
    - Waterfall chart
  - Components of components
    - Stacked 100% bar chart w/subcomponents
  - Accumulation to total & absolute difference matters
    - Tree map

Data Visualization in Python



Data Visualization Library



THANK YOU