02. Motivation for Data Visualization

Summary Statistics vs. Visualizations

Summary statistics like the mean and standard deviation can be great for attempting to quickly understand aspects of a dataset, but they can also be misleading if you make too many assumptions about how the data distribution looks.

Anscombe's Quartet Example

Consider we have the following four datasets of x, y pairs. You can download the data using the button below. A link to a Google Sheet with the data is also available here.

DOWNLOAD 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

X	Y	X	Y	X	Y	X	Y	
10	8.04	10	9.14	10	7.46	8	6.58	
8	6.95	8	8.14	8	6.77	8	5.76	
13	7.58	13	8.74	13	12.74	8	7.71	
9	8.81	9	8.77	9	7.11	8	8.84	
11	8.33	11	9.26	11	7.81	8	8.47	
14	9.96	14	8.1	14	8.84	8	7.04	
6	7.24	6	6.13	6	6.08	8	5.25	
4	4.26	4	3.1	4	5.39	19	12.5	
12	10.84	12	9.13	12	8.15	8	5.56	
7	4.82	7	7.26	7	6.42	8	7.91	
5	5.68	5	4.74	5	5.73	8	6.89	
Average	Average	Average	Average	Average	Average	Average	Average	
9	7.500909091	9	7.500909091	9	7.5	9	7.500909091	
			Make A copy of the data to calculate your statistics					
STD	STD	STD	STD	STD	STD	STD	STD	=STDEV.P(A2:A12)
3.16227766	1.937024215	3.16227766	1.937108691	3.16227766	1.935932944	3.16227766	3.16227766	

STDEV.P(**number1**, [number2], ...)

QUIZ QUESTION::

Use the data above to match an answer to each of the following questions. (Assume rounding to 2 digits)

ANSWER CHOICES:

They are the same. | They are the same. | They are the same. | They are different. | They are the same.

They are different. | They are different. | They are different.

Question	Answer
What is true for the means associated with any of the **X** columns?	Same
What is true for the means associated with any of the **Y** columns?	Same
What is true for the standard deviation associated with any of the **X** columns?	Same
What is true for the standard deviation associated with any of the **Y** columns?	Same

Next Concept

05. Quiz: Exploratory vs. Explanatory

QUIZ QUESTION::

Match each of the statements below to whether it is true for either **Exploratory** or **Explanatory** analyses.

ANSWER CHOICES:

Explanatory | Exploratory | Exploratory | Explanatory | Explanatory

Statement	Exploratory or Explanatory
These plots are used to tell stories.	**Explanatory**
This is the beginning of most data analyses processes.	**Exploratory**
This is the end of most data analyses processes.	**Explanatory**
When making these plots, you should pay attention to making the plot insightful to your audience.	**Explanatory**
Finding missing values in a dataset is a part of this analysis.	**Exploratory**

04. Quiz: Data Types (Quantitative vs. Categorical)

Data Types

QUIZ QUESTION::

For each variable below, identify each as either **quantitative** or **categorical**.

ANSWER CHOICES:

Categorical | Quantitative | Quantitative | Categorical | Quantitative

Variable	Data Type
Zip Code	**Categorical**
Age	**Quantitative**
Income	**Quantitative**
Marital Status (Single, Married, Divorced, etc.)	**Categorical**
Height	**Quantitative**

Data Types

QUIZ QUESTION::

For each variable below, identify each as either **quantitative** or **categorical**.

ANSWER CHOICES:

Quantitative | Categorical | Quantitative | Categorical | Quantitative

Variable	Data Types
Letter Grades (A+, A, A-, B+, B, B-, ...)	Categorical
Travel Distance to Work	Quantitative
Ratings on a Survey (Poor, Ok, Great)	Categorical
Temperature	Quantitative
Average Speed	Quantitative

Next Concept

Nominal vs. Ordinal

This quiz will assure you have a clear understanding of the differences between categorical nominal vs. categorical ordinal variables. All of the variables below are categorical. Your task is to select the **check** box next to each variable that is **nominal**; do not check the ordinal categorical variables.

- [ ] Letter Grades (A, B+, B, B-, etc.)
- [x] Types of Fruit (Apple, Banana, etc.)
- [ ] Ratings on a Survey (Poor, Ok, Great)
- [x] Types of Dog Breeds (German Shepherd, Collie, etc.)
- [x] Genres of Movies (Horror, Comedy, etc.)
- [x] Gender
- [x] Nationality
- [ ] Education (HS, Associates, Bachelors, Masters, PhD, etc.)

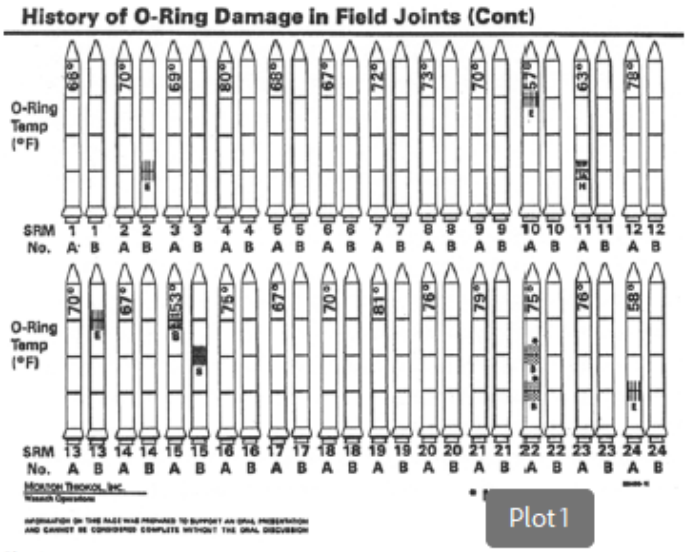
The plot above from Fox News claims to depict the change in the top tax rate bracket between the current level at the time, and after tax cuts were to expire. What is the lie factor for this chart? Some numbers to help: the small bar is 27 pixels tall and the large bar is 146 pixels tall.

- ○ 4.57
- ● 33.54
- ○ 1
- ○ .03

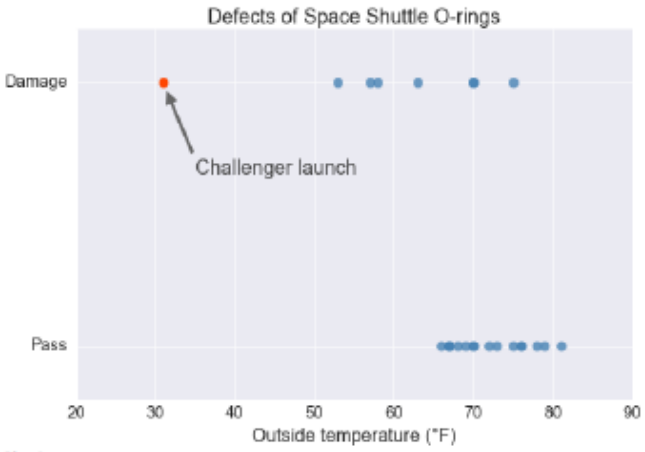
Which is Better?

Believe it or not, the next two plots are of the exact same data. Both of them depict information regarding flights of the USA's Space Shuttle program: whether or not a mechanical failure of O-Ring components occurred, as well as the temperature at the time of flight. A full background of the dataset can be found here.

Use these two plots to answer the quiz questions that follow.



Plot 1



Plot 2

Which visual best represents the underlying data?

- ○ Plot 1
- ● Plot 2

Use either of the two plots above to mark all the below that are true.

- ☐ Temperature appears to be associated with whether an O-ring is damaged or will pass.
- ☑ If the temperature is lower than 60 degrees F, no O-rings have ever passed.
- ☑ The **challenger** had the lowest temperature of any O-ring on record.
- ☑ There are 7 total damaged O-rings in the dataset.

What is the main data visualization violation for the first visual?

- ○ Data Integrity
- ○ High Data-Ink Ratio
- ● Chart Junk
- ○ Nothing, it is okay "as is".

Next Concept

The above pie chart violates a few rules of visual design, but which is the worst violation?

- [x] Chart Junk
- [ ] Design Integrity
- [x] ata-Ink Ratio
- [ ] This should be used for Exploratory analysis and not Explanatory analysis.

What all could be done to improve the above visual? Check all that apply.

- [x] Change the coloring to be less dramatic, while still relating to the different companies.
- [x] Remove 3D aspect.
- [x] Use a visual that uses length (bar chart) rather than area (pie chart) to demonstrate differences, as humans are better able detect differences in lengths.
- [ ] Remove the percentages, as they are redundant to the area of the pie chart slice.
- [x] Remove the legend, and put the names of the companies directly on the plot.

of:

17. Good Visual

QUIZ QUESTION::

Map each solution to each question/statement.

ANSWER CHOICES:

Position | Data-Ink Ratio | Color Hue | Chart Junk | Color

Question/Statement	Solution
What is the most appropriate visual encoding for adding a categorical variable to a scatterplot?	Color
Which visual encoding is most accurate for visual perception?	Position
The least accurate visual encoding for visual perception?	Color Hue
What are additional visuals that do not add to the message of the data?	Chart Junk
What do we want to have a high value of in our visuals?	Data-Ink Ratio

---

Color, shape, size, and other tools of data visualization are clutter that should never be used.

○ True

● False

Next Concept