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

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?	They are the same = 9

What is true for the means associated with any of the **Y** columns?

They are the same = 7.5

What is true for the standard deviation associated with any of the **X** columns?

They are the same

What is true for the standard deviation associated with any of the **Y** columns?

They are the same

Next Concept

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

Data Types

QUIZ QUESTION::

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

ANSWER CHOICES:

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**

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

Temperature

Quantitative

Average Speed. Quantitative

Next Concept

☰ 05. Text + Quiz: Data Types (Ordinal vs. Nominal)

Recap of Previous Video

The table below summarizes our data types. To expand on the information in the table, you can look through the text that follows.

Data Types		
Quantitative:	**Continuous**	**Discrete**
	Height, Age, Income	Pages in a Book, Trees in Yard, Dogs at a Coffee Shop
Categorical:	**Ordinal**	**Nominal**
	Letter Grade, Survey Rating	Gender, Marital Status, Breakfast Items

Below is a little more detail of the information shared in the above table.

Another Look

To break down our data types, there are two main blocks:

Quantitative and **Categorical**

Quantitative can be further divided into `Continuous` or `Discrete`.

Categorical data can be divided into `Ordinal` or `Nominal`.

You should have now mastered what types of data in the world around us falls into each of these four buckets: Discrete, Continuous, Nominal, and Ordinal. In the next sections, we will work through the numeric summaries that relate specifically to quantitative variables.

Quantitative vs. Categorical

Some of these can be a bit tricky - notice even though zip codes are a number, they aren't really a quantitative variable. If we add two zip codes together, we do not obtain any useful information from this new value. Therefore, this is a categorical variable.

Height, **Age**, the **Number of Pages in a Book** and **Annual Income** all take on values that we can add, subtract and perform other operations with to gain useful insight. Hence, these are `quantitative`.

Gender, **Letter Grade**, **Breakfast Type**, **Marital Status**, and **Zip Code** can be thought of as labels for a group of items or individuals. Hence, these are `categorical`.

Continuous vs. Discrete

To consider if we have continuous or discrete data, we should see if we can split our data into smaller and smaller units. Consider time - we could measure an event in years, months, days, hours, minutes, or seconds, and even at seconds we know there are smaller units we could measure time in. Therefore, we know this data type is continuous. **Height**, **age**, and **income** are all examples of `continuous data`. Alternatively, the **number of pages in a book**, **dogs I count outside a coffee shop**, or **trees in a yard** are `discrete data`. We would not want to split our dogs in half.

Ordinal vs. Nominal

In looking at categorical variables, we found **Gender**, **Marital Status**, **Zip Code** and your **Breakfast items** are `nominal variables` where there is no order ranking associated with this type of data. Whether you ate cereal, toast, eggs, or only coffee for breakfast; there is no rank ordering associated with your breakfast.

Alternatively, the **Letter Grade** or **Survey Ratings** have a rank ordering associated with it, as `ordinal data`. If you receive an A, this is higher than an A-. An A- is ranked higher than a B+, and so on... Ordinal variables frequently occur on rating scales from very poor to very good. In many cases we turn these ordinal variables into numbers, as we can more easily analyze them, but more on this later!

Final Words

In this section, we looked at the different data types we might work with in the world around us. When we work with data in the real world, it might not be very clean - sometimes there are typos or missing values. When this is the case, simply having some expertise regarding the data and knowing the data type can assist in our ability to 'clean' this data. Understanding data types can also assist in our ability to build visuals to best explain the data. But more on this very soon!

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.

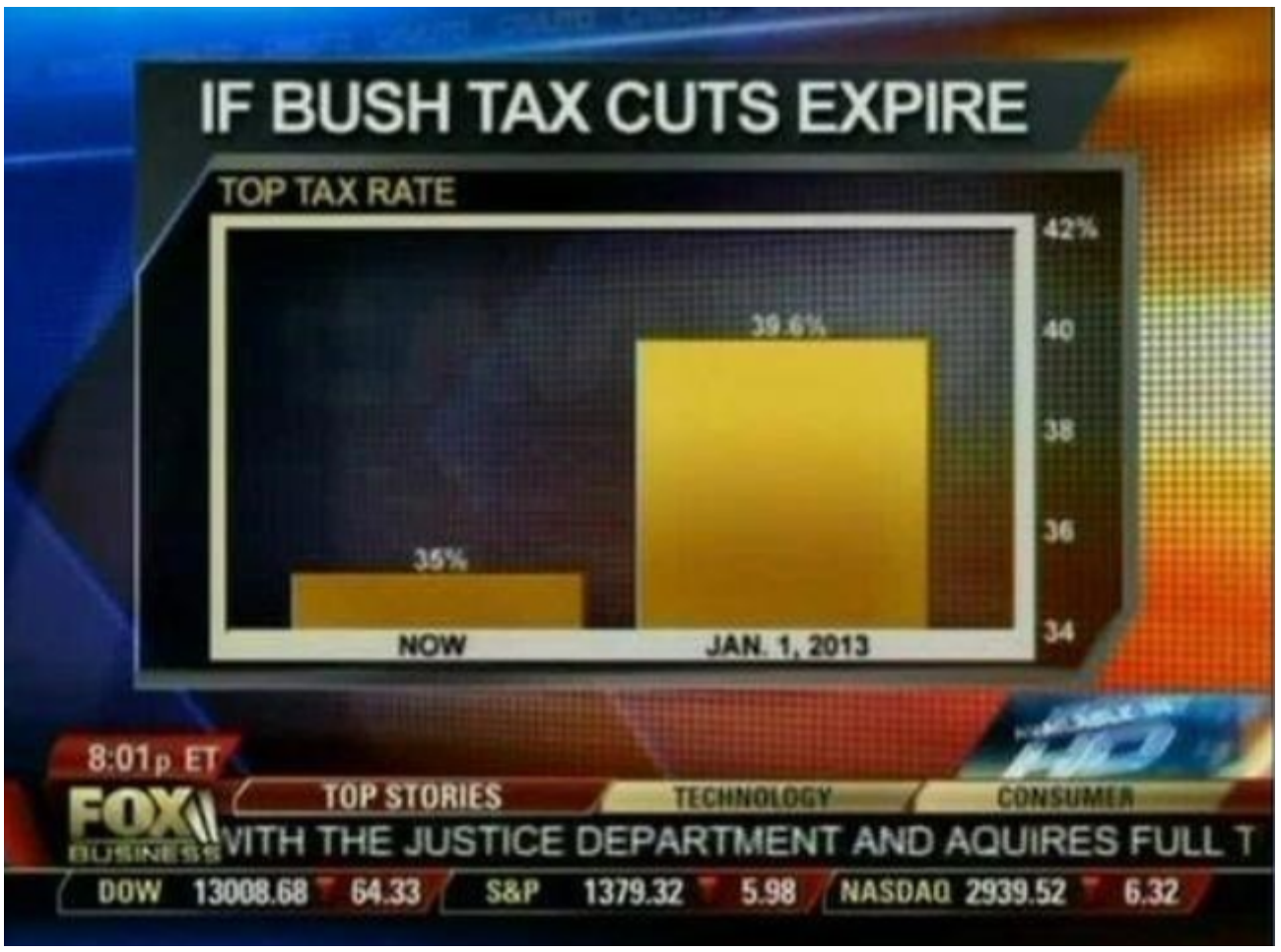
- [ ] Letter Grades (A, B+, B, B-, etc.)
- [ ] Types of Fruit (Apple, Banana, etc.)
- [ ] Ratings on a Survey (Poor, Ok, Great)
- [ ] Types of Dog Breeds (German Shepherd, Collie, etc.)
- [ ] Genres of Movies (Horror, Comedy, etc.)
- [ ] Gender
- [ ] Nationality
- [ ] Education (HS, Associates, Bachelors,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

Next Concept

12. Bad Visual Quizzes (Part I)

Practice, Practice, Practice

It is time to put your new skills in data visualization to practice. See if you can figure out the questions below!



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.

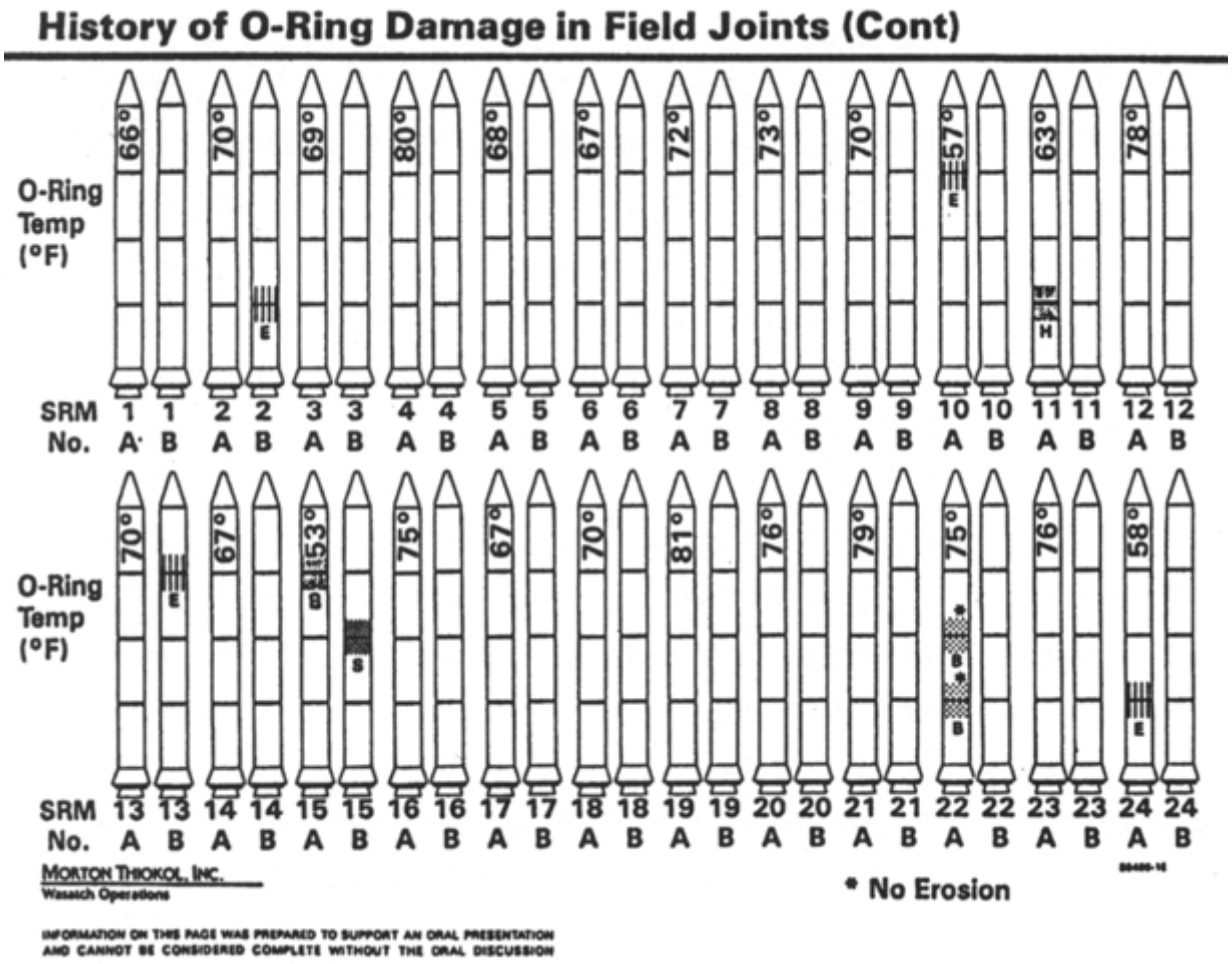
- 4.57
- 33.54
- 1
- .03

lie f = [(146-27)/27]/[(39.6-35)/35] =33.54

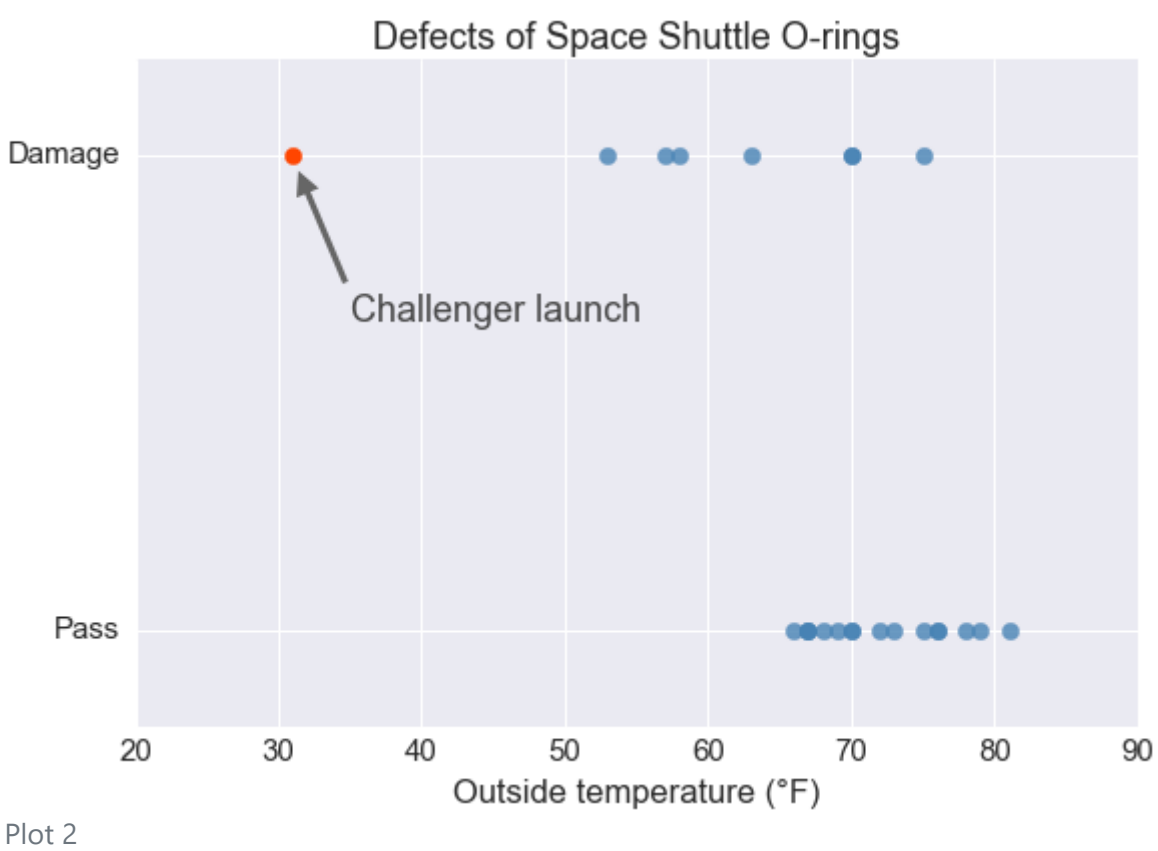
Which is Better?

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

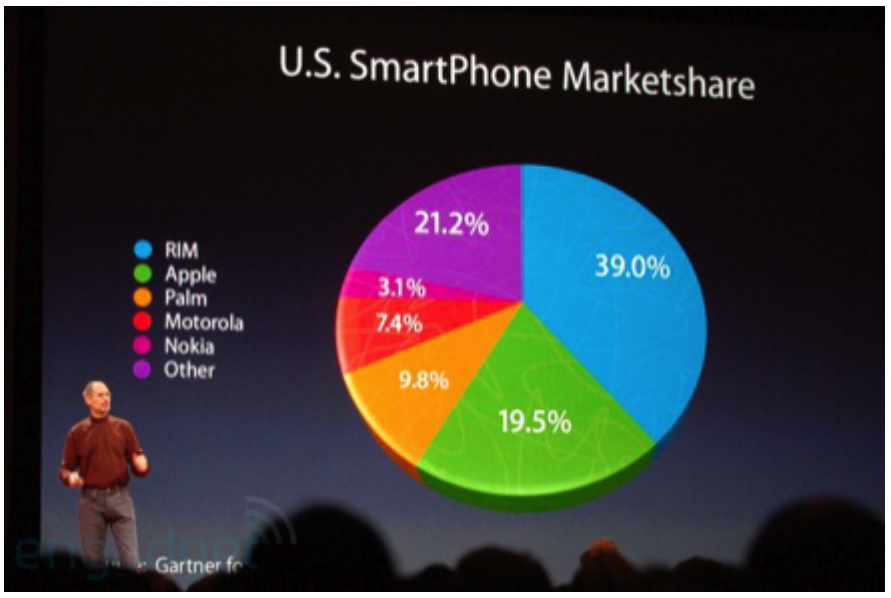
○ High Data-Ink Ratio

● Chart Junk

○ Nothing, it is okay "as is".

Next Concept

13. Bad Visual Quizzes (Part II)



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

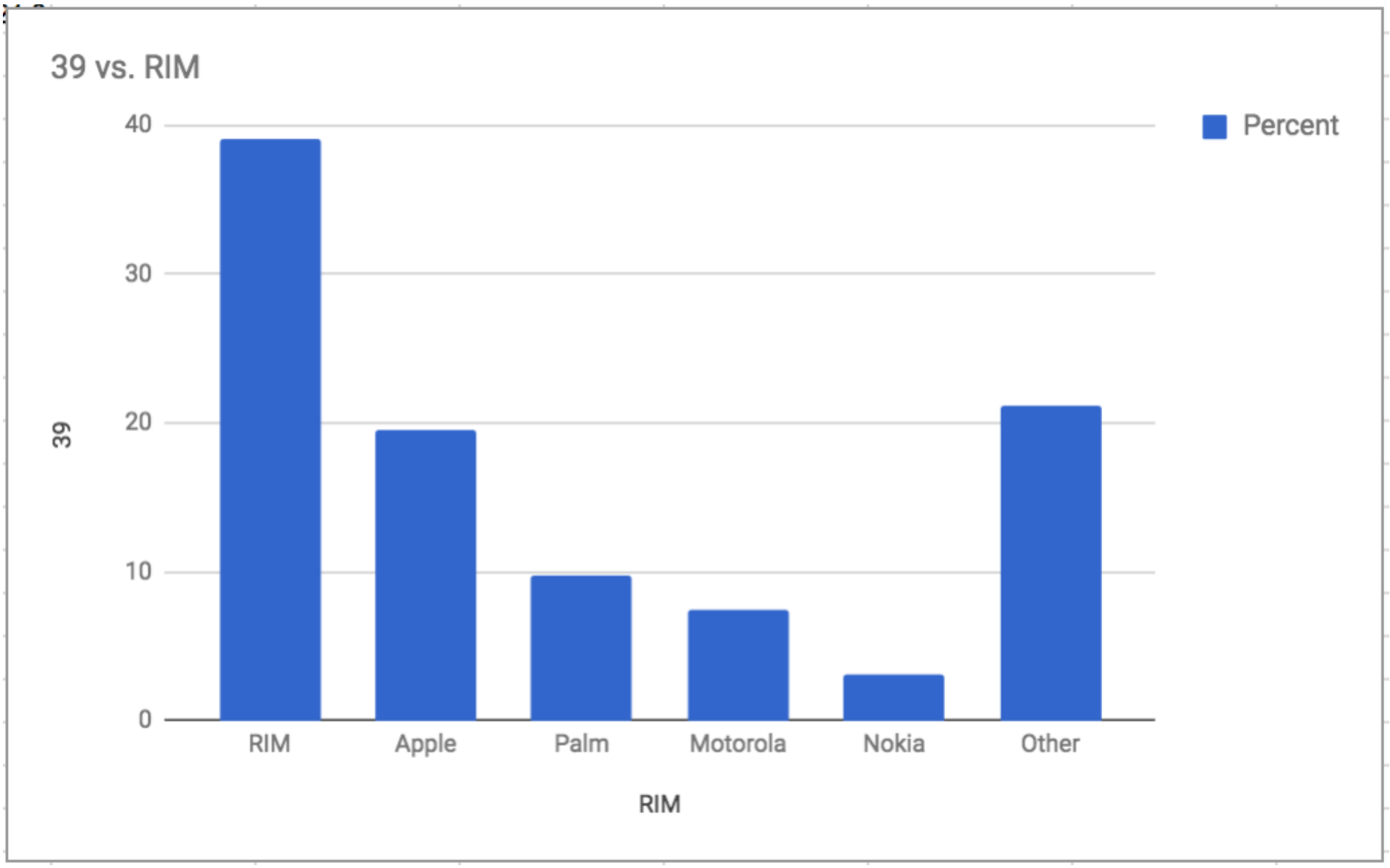
- ○ Chart Junk
- ● Design Integrity
- ○ Data-Ink Ratio
- ○ This should be used for Exploratory analysis and not Explanatory analysis.

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

- ☐ Change the coloring to be less dramatic, while still relating to the different companies.
- ● Remove 3D aspect.
- ● 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.
- ☐ Remove the legend, and put the names of the companies directly on the plot.

Updated Visual

The same data presented in the image above is recorded in the spreadsheet here. This data was used to create the following visual; the next question asks how this plot could be improved.

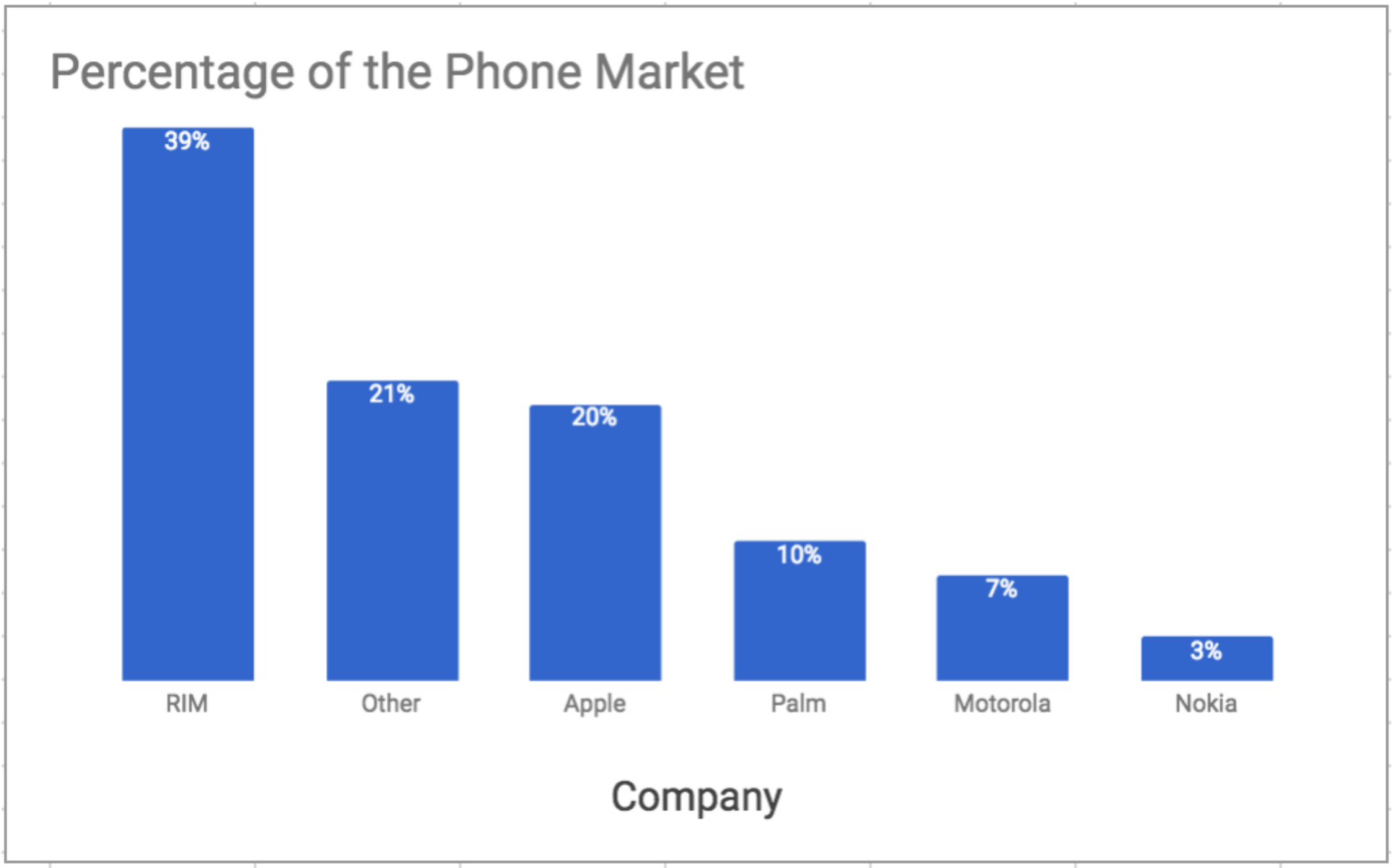


Using the above bar chart, select all that apply for improving this visual. Check all that apply.

- [ ] Rename axis labels and title.
- [ ] Order the bars in alphabetical order for an easier comparison.
- [ ] Order the bars according to height for an easier comparison.
- [ ] Remove legend, as percent isn't a useful legend label.
- [ ] Color the bars to highlight the different companies.
- [ ] Add to the legend the company names and color the bars as represented in the legend.

There are always personal preferences. The visual below is a **good visualization** of this data from a design, following the principles of:

1. reducing chart junk,
2. maintaining a high data-ink ratio,
3. maintaining data integrity, and
4. using length to show changes and differences rather than areas.



Next Concept

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