From Data To Discovery

We will start momentarily. In the meantime:

Class Starter (respond on pollev.com/tiasondjaja)

Suppose that a data frame called `results` stores the results of a 5k (3.1 mile) race as well as the gender and age group of the participating runners:

Gender	AgeGroup	RaceTime
Female	40-59	28
Male	20-39	35
Male	40-59	19
Other	20-39	25
⋮	⋮	⋮

Suppose that you would like to create a new column called `Pace`, which consists of the number of minutes it took each runner to run one mile during this race.
Which R code would accomplish this task?

A. `results$Pace <- 3.1 / RaceTime`

B. `Pace <- results$RaceTime / 3.1`

C. `results$Pace <- results$RaceTime / 3.1`

D. None of the above

Open the Lesson 04 Jupyter Notebook

- Our NYU Classes site > Lessons > 4. Exploring Data (part 2); Click link
- No Jupyter Notebook for Lesson 5


Update your name to **First name + Last Init.** **(+ pronouns,** optional)


Mute your microphone

Turn your camera on


Make sure you can raise your hand and give nonverbal responses


Open the chat window

Quick Concept Check

(respond on pollev.com/tiasondjaja)

Suppose that a data frame called `results` store the results of a 5k (3.1 mile) race as well as the gender and age group of the participating runners:

Gender	AgeGroup	RaceTime (minutes)
Female	40-59	28
Male	20-39	35
Male	40-59	19
Other	20-39	25
⋮	⋮	⋮

Suppose that you would like to find the average race time among runners of each gender.

Which of the following R commands would give you the answer?

A. `mean( filter(results, Gender == 'Female') )`

B.
```
groupedresults <- filter(results, Gender == 'Female' )
mean(groupedresults$RaceTime)
```

C.
```
groupedresults <- group_by(results)
summarize(groupedresults, aveRaceTime = mean(RaceTime))
```

D. None of the above works

Today's Plan

- Lesson02: Introduction to Jupyter Notebook and R
- Lesson03: Exploring Data
- Finish Lesson04: Exploring Data
- Start Lesson05: Visualizing data

Lesson 5: Visualizing Data

Goals and Key Ideas

1. Types of data visualizations
   - How to determine the right type of data visualization for a given type of variable.
2. Good and bad data visualizations
   - Bad data visualizations could mislead

Why visualize data?

Why visualize data?

- Help create a **visual summary** of data
- Help **identify patterns** in data
- Help **identify relationships** between variables

Why visualize data?

- Help create a **visual summary** of data
- Help **identify patterns** in data
- Help **identify relationships** between variables
- Help **communicate** or **describe** results of data analysis

Why visualize data?

- Help create a **visual summary** of data
- Help **identify patterns** in data
- Help **identify relationships** between variables
- Help **communicate** or **describe** results of data analysis

Bad data visualizations can **miscommunicate** or **misrepresent** information.

Breakout Activity: Types of Data Visualizations

Respond here: `https://pollev.com/tiasondjaja`

Types of Data Visualizations

Types of Data Visualizations

1. Bar Graphs

Types of Data Visualizations

1. Bar Graphs

   Two types:

   - (the most important one) Describe the **distribution** of a **categorical** variable
   - Describe the relationship between a **categorical** variable and a **numerical** variable

Types of Data Visualizations

1. Bar Graphs

   Two types:

   - (the most important one) Describe the **distribution** of a **categorical** variable
   - Describe the relationship between a **categorical** variable and a **numerical** variable

2. Histograms

Types of Data Visualizations

1. Bar Graphs

   Two types:

   - (the most important one) Describe the **distribution** of a **categorical** variable
   - Describe the relationship between a **categorical** variable and a **numerical** variable

2. Histograms

   Describe the **distribution** of a **numerical** variable.

Types of Data Visualizations

1. Bar Graphs

   Two types:

   - (the most important one) Describe the **distribution** of a **categorical** variable
   - Describe the relationship between a **categorical** variable and a **numerical** variable

2. Histograms

   Describe the **distribution** of a **numerical** variable.

3. Scatterplots

Types of Data Visualizations

1. Bar Graphs

   Two types:

   - (the most important one) Describe the **distribution** of a **categorical** variable
   - Describe the relationship between a **categorical** variable and a **numerical** variable

2. Histograms

   Describe the **distribution** of a **numerical** variable.

3. Scatterplots

   Describe the relationship between a **numerical** variable and another **numerical** variable.

Types of Data Visualizations

1. Bar Graphs

   Two types:

   - (the most important one) Describe the **distribution** of a **categorical** variable
   - Describe the relationship between a **categorical** variable and a **numerical** variable

2. Histograms

   Describe the **distribution** of a **numerical** variable.

3. Scatterplots

   Describe the relationship between a **numerical** variable and another **numerical** variable.

4. Time Series

Types of Data Visualizations

1. Bar Graphs

   Two types:

   - (the most important one) Describe the **distribution** of a **categorical** variable
   - Describe the relationship between a **categorical** variable and a **numerical** variable

2. Histograms

   Describe the **distribution** of a **numerical** variable.

3. Scatterplots

   Describe the relationship between a **numerical** variable and another **numerical** variable.

4. Time Series

   Describes how quantities change over time.

Types of Data Visualizations

1. Bar Graphs

   Two types:

   - (the most important one) Describe the **distribution** of a **categorical** variable
   - Describe the relationship between a **categorical** variable and a **numerical** variable

2. Histograms

   Describe the **distribution** of a **numerical** variable.

3. Scatterplots

   Describe the relationship between a **numerical** variable and another **numerical** variable.

4. Time Series

   Describes how quantities change over time.

5. Pie Charts

Types of Data Visualizations

1. Bar Graphs

   Two types:

   - (the most important one) Describe the **distribution** of a **categorical** variable
   - Describe the relationship between a **categorical** variable and a **numerical** variable

2. Histograms

   Describe the **distribution** of a **numerical** variable.

3. Scatterplots

   Describe the relationship between a **numerical** variable and another **numerical** variable.

4. Time Series

   Describes how quantities change over time.

5. Pie Charts

   Describe the proportion of observations that belong to each category.

Types of Data Visualizations

1. Bar Graphs

   Two types:

   - (the most important one) Describe the **distribution** of a **categorical** variable
   - Describe the relationship between a **categorical** variable and a **numerical** variable

2. Histograms

   Describe the **distribution** of a **numerical** variable.

3. Scatterplots

   Describe the relationship between a **numerical** variable and another **numerical** variable.

4. Time Series

   Describes how quantities change over time.

5. Pie Charts

   Describe the proportion of observations that belong to each category.

6. (and others!)

Examples of (good and bad) data visualizations

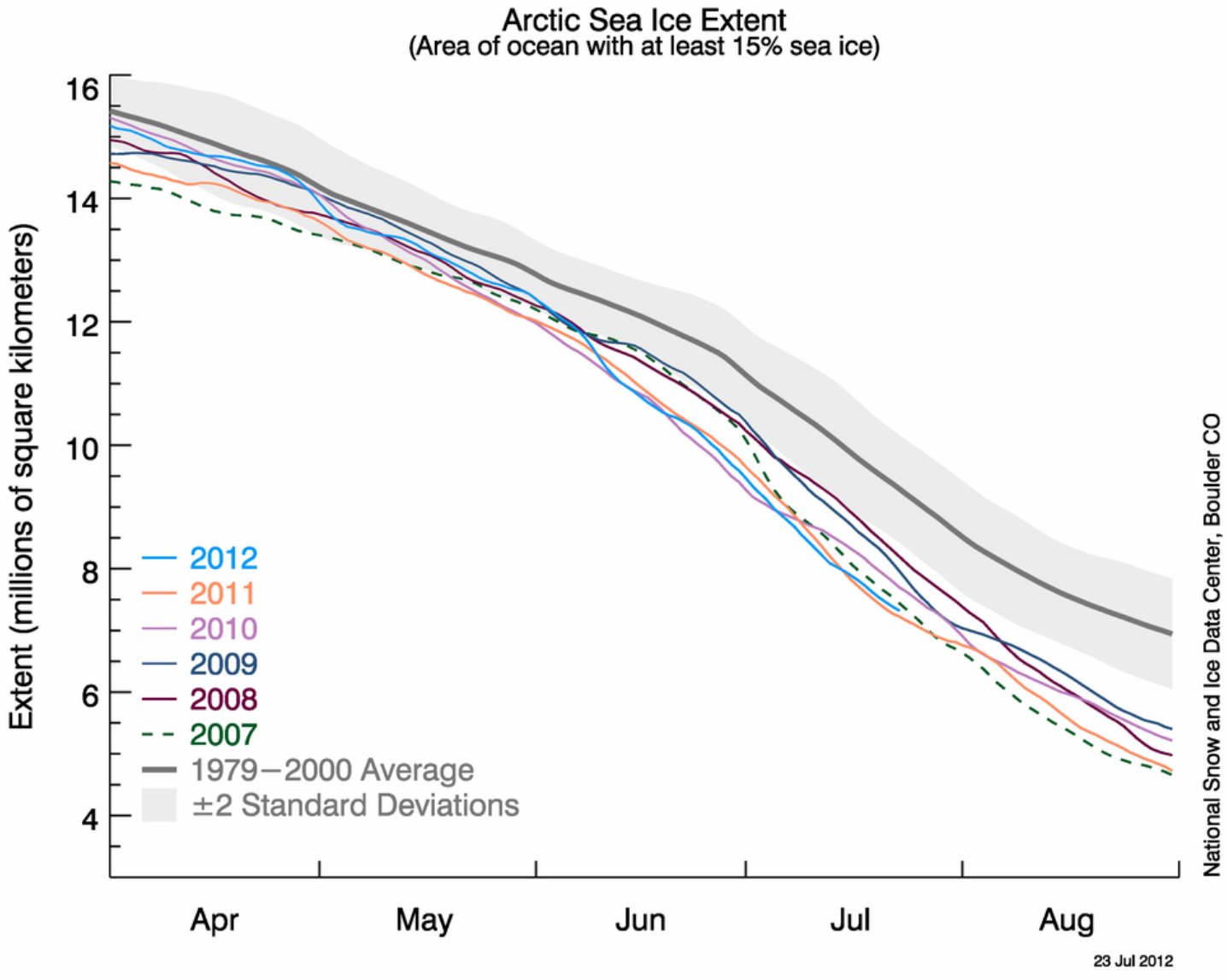
Handout

https://drive.google.com/file/d/1KLjLuK_5IzW2FyYR7Z10derewmsXL-LU/view?usp=sharing

Question & Task: Which of these examples are good data visualizations? Which are bad? Discuss!

Breakout Activity

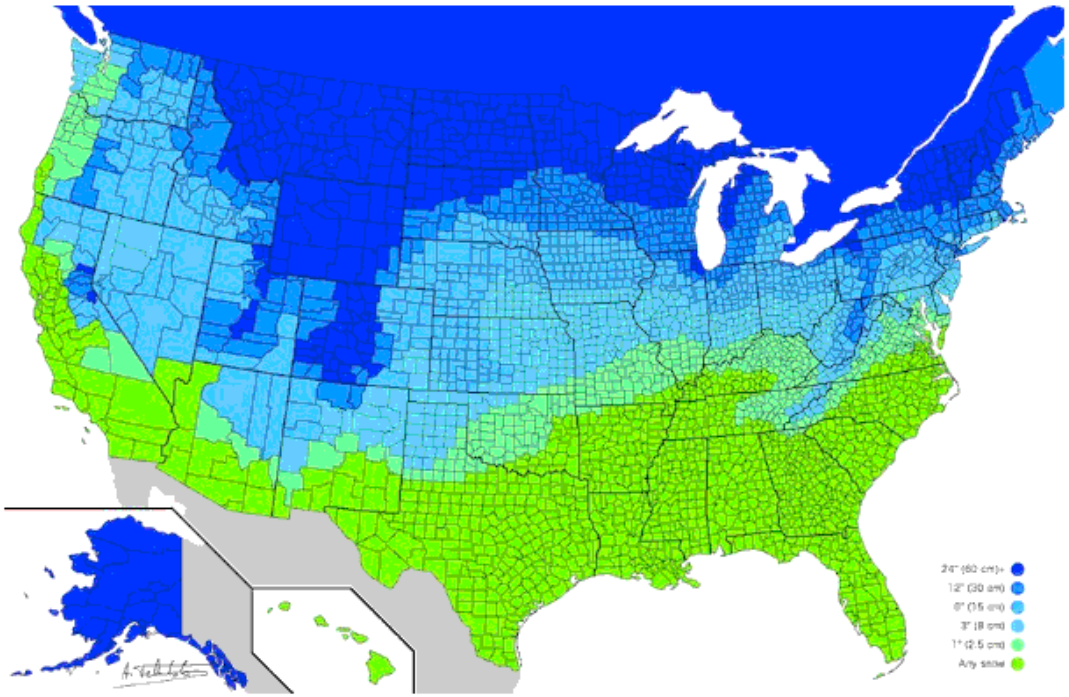
https://docs.google.com/document/d/10nDovYYmeyFXPAgySrBQQgYpbnoWosXDrlDy7Q8sMCs/edit?usp=sharing


Arctic Sea Ice Extent
(Area of ocean with at least 15% sea ice)
Extent (millions of square kilometers)
16
14
12
10
8
6
4
2012
2011
2010
2009
2008
2007
1979–2000 Average
±2 Standard Deviations
Apr
May
Jun
Jul
Aug
National Snow and Ice Data Center, Boulder CO
23 Jul 2012

How Much Snow Before America Cancels School?

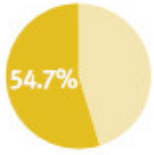
A map shows how many inches it takes before various regions call it off.

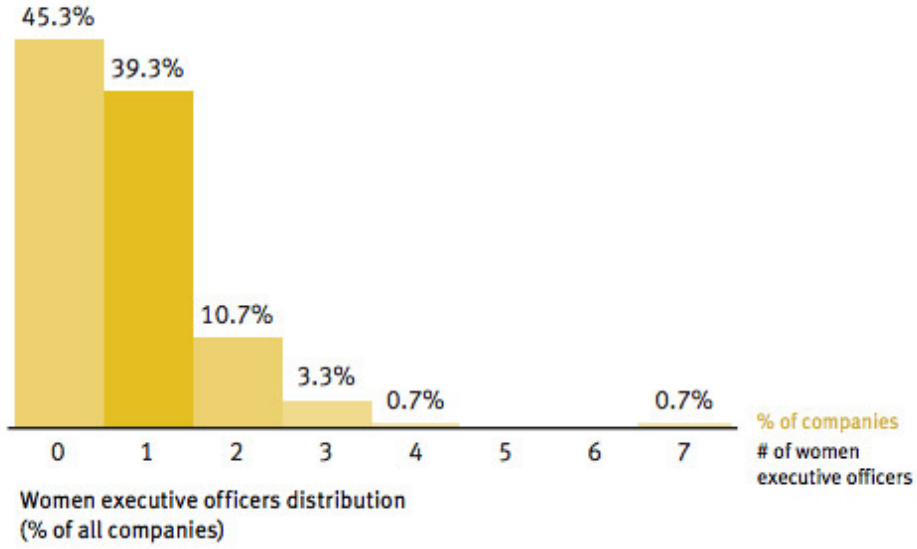
by ERIC RANDALL • 2/3/2014, 11:10 a.m.



WOMEN EXECUTIVE OFFICERS DISTRIBUTION — 2013 PROXY SEASON

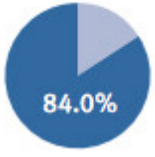
SV 150
2013
% of companies with at least 1 woman executive officer

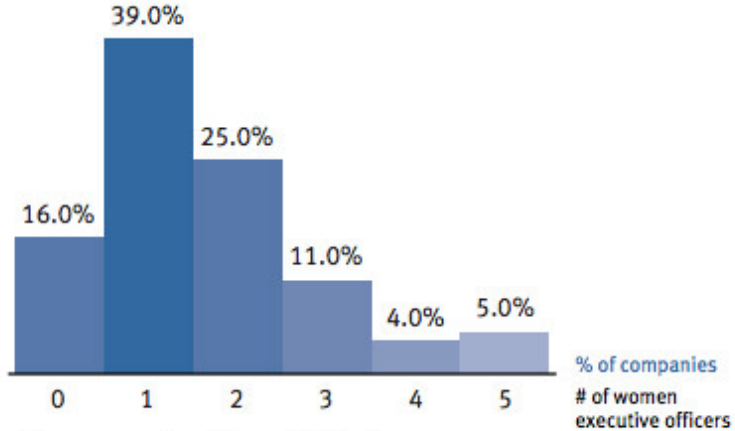




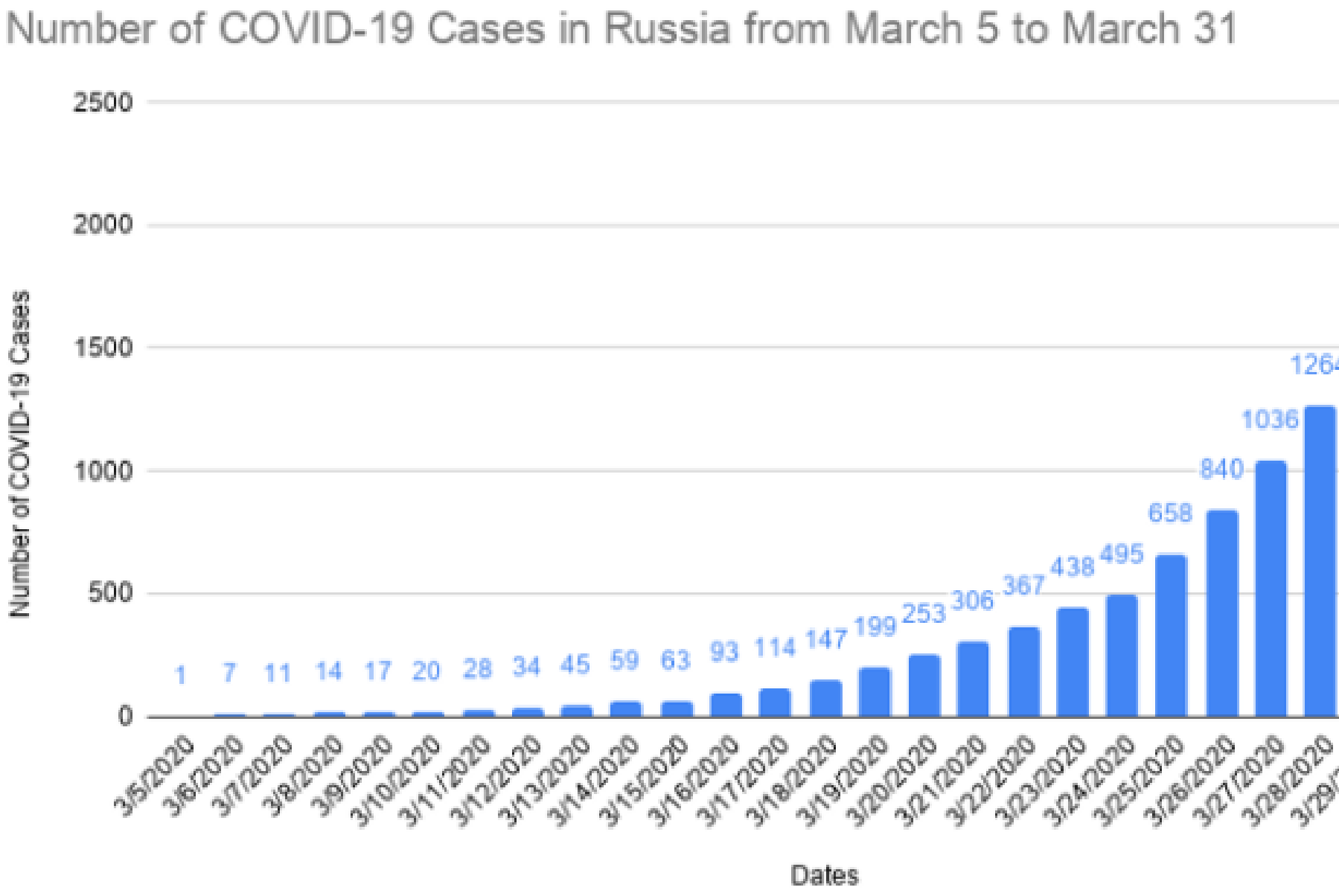
Women executive officers distribution
(% of all companies)

S&P 100
2013
% of companies with at least 1 woman executive officer





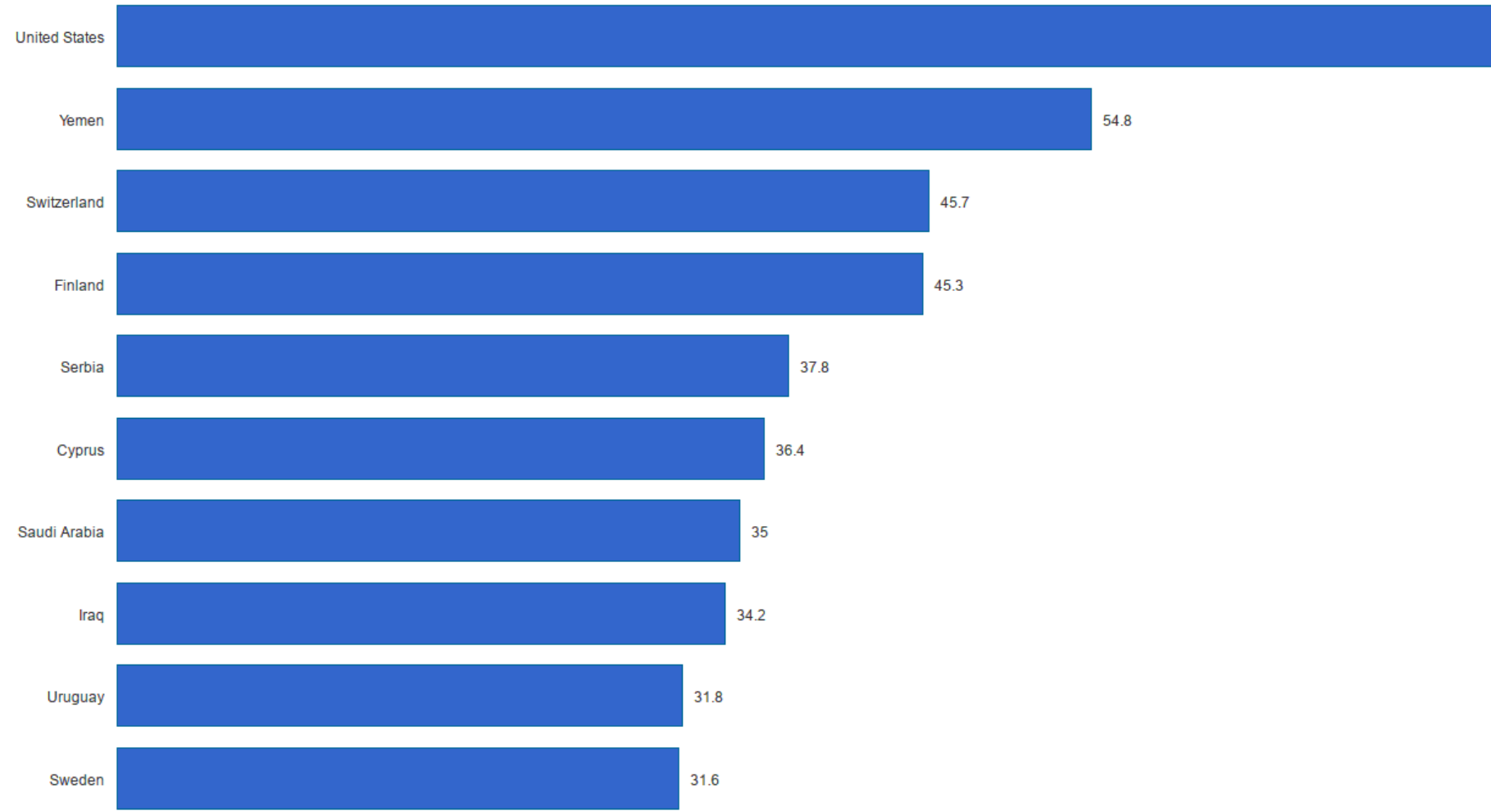
Women executive officers distribution
(% of all companies)

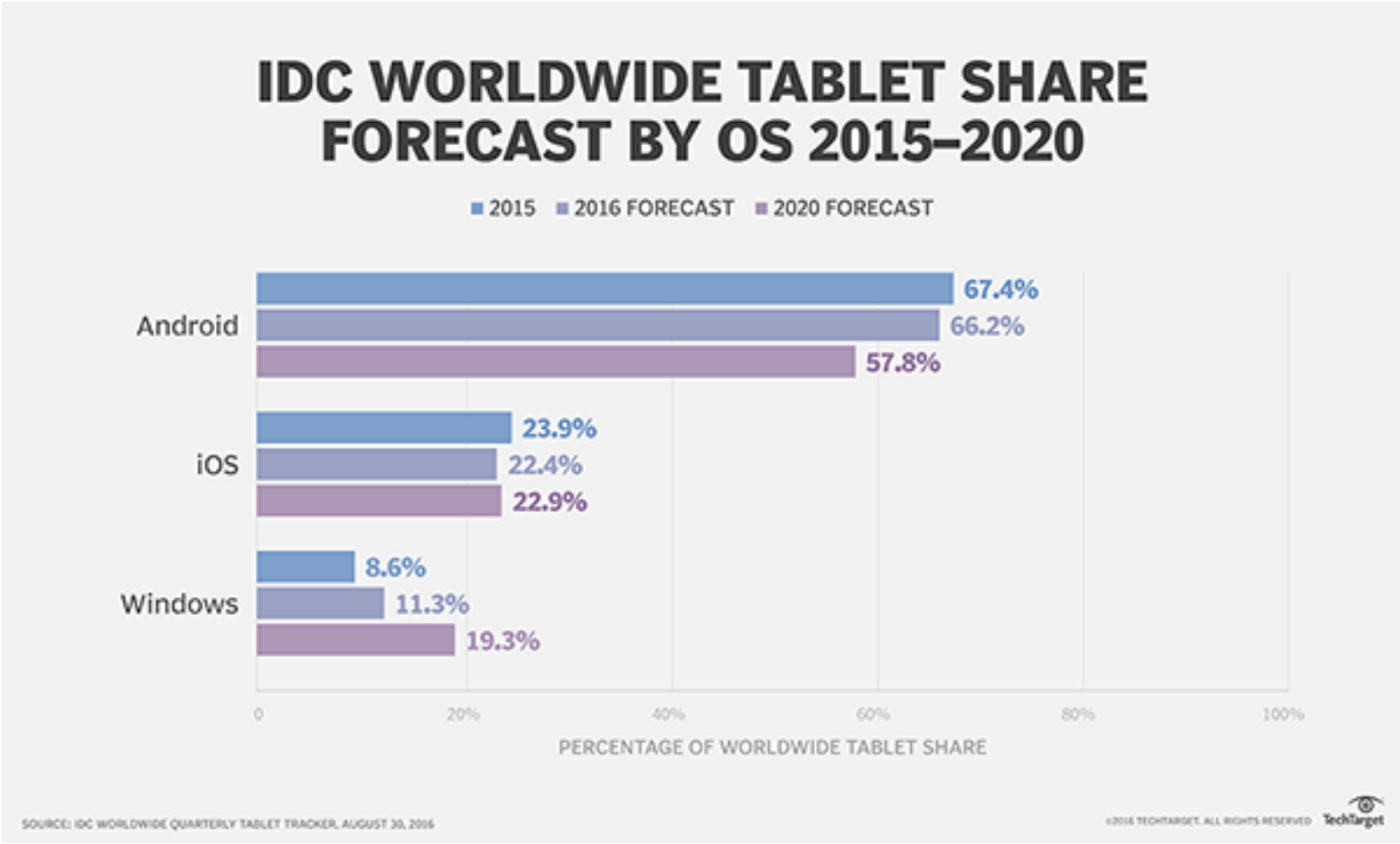

Number of COVID-19 Cases in Russia from March 5 to March 31
Number of COVID-19 Cases
2500
2000
1500
1000
500
0
1
7
11
14
17
20
28
34
45
59
63
93
114
147
199
253
306
367
438
495
658
840
1036
1264
3/5/2020
3/6/2020
3/7/2020
3/8/2020
3/9/2020
3/10/2020
3/11/2020
3/12/2020
3/13/2020
3/14/2020
3/15/2020
3/16/2020
3/17/2020
3/18/2020
3/19/2020
3/20/2020
3/21/2020
3/22/2020
3/23/2020
3/24/2020
3/25/2020
3/26/2020
3/27/2020
3/28/2020
Dates

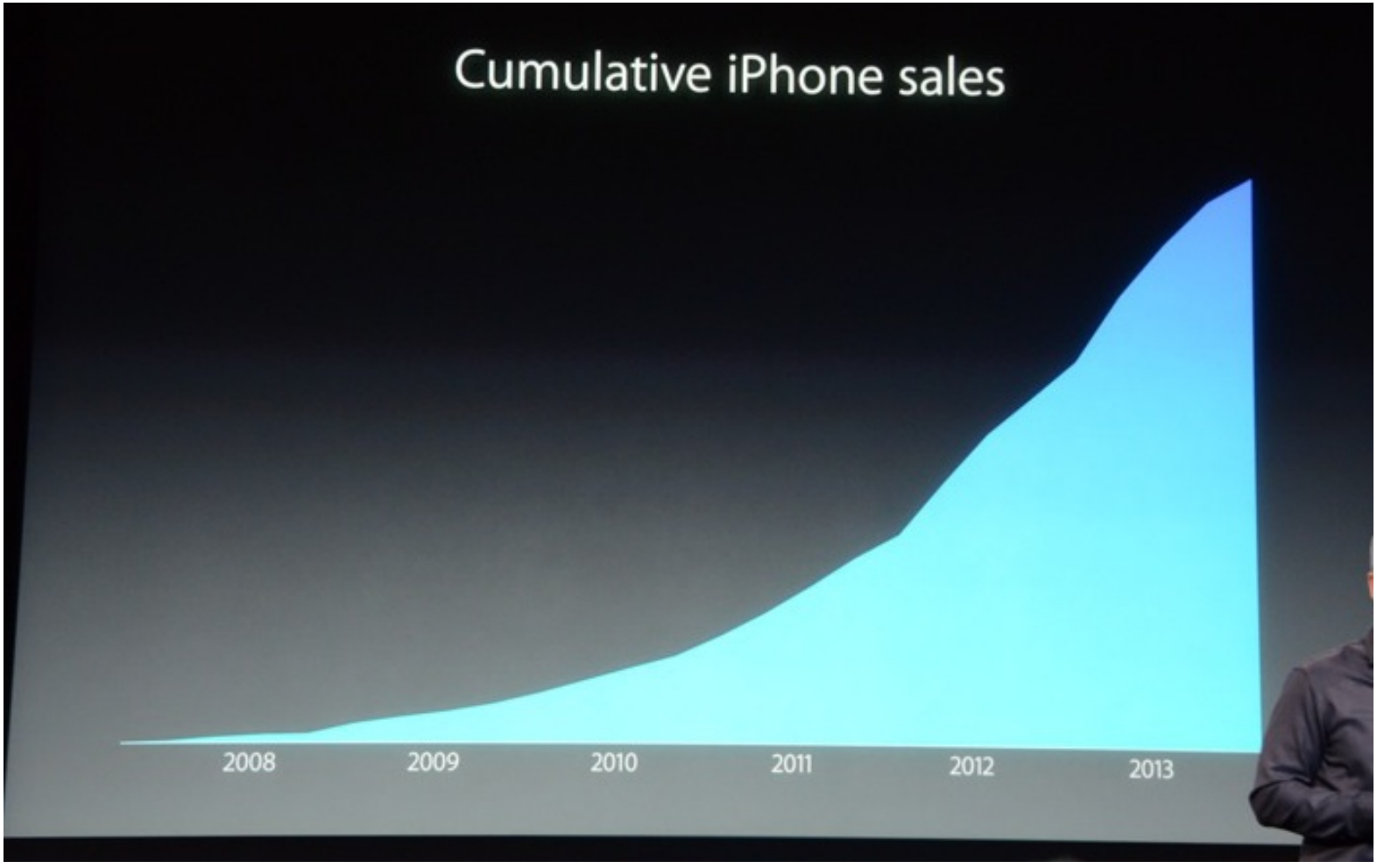
Gun ownership by country

Click to switch view

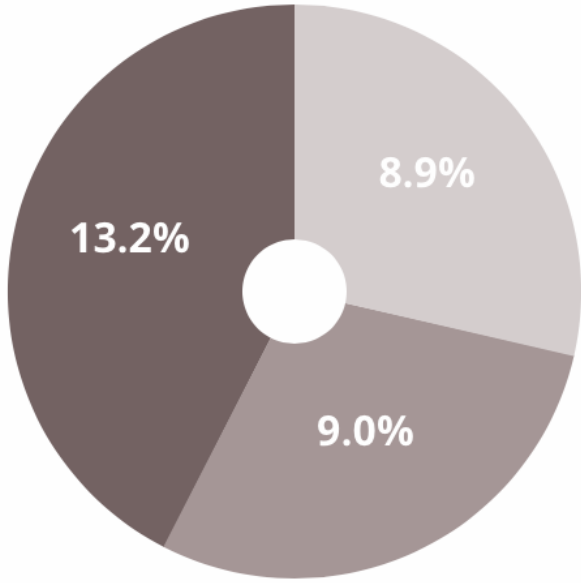
Average firearms per 100 people | Homicide by firearm rate per 100,000 pop

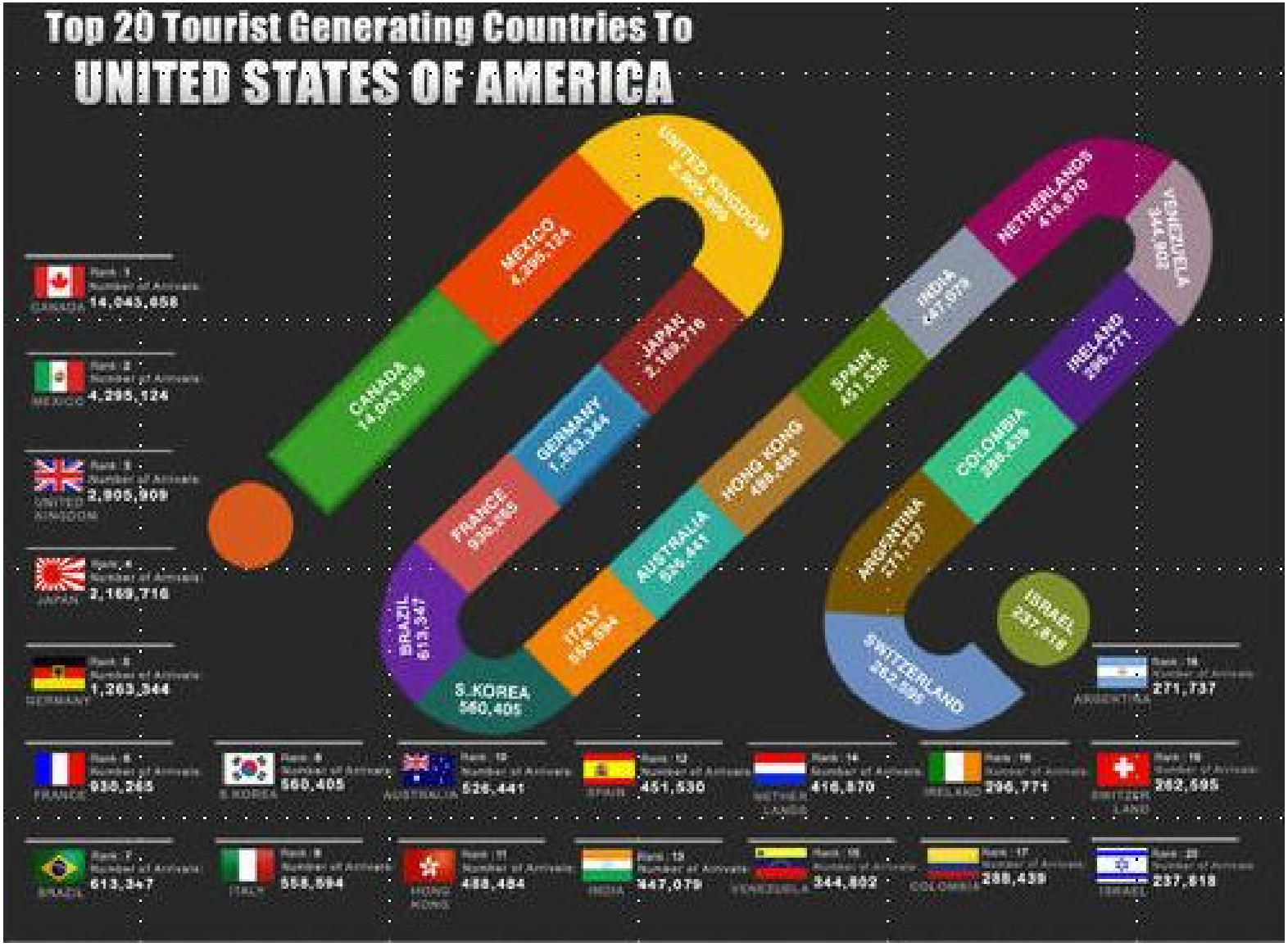


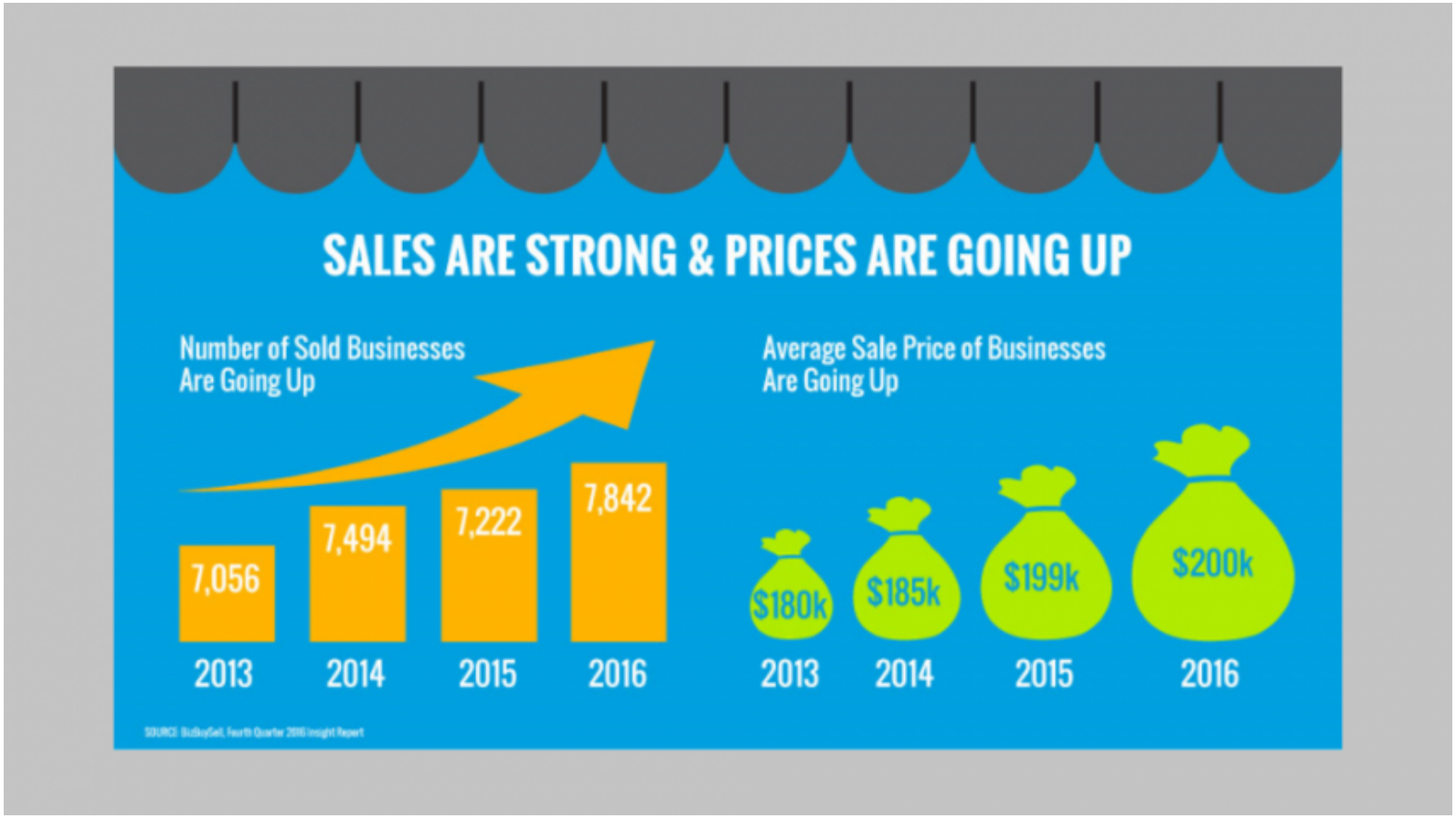

IDC WORLDWIDE TABLET SHARE FORECAST BY OS 2015–2020
2015
2016 FORECAST
2020 FORECAST
Android
67.4%
66.2%
57.8%
iOS
23.9%
22.4%
22.9%
Windows
8.6%
11.3%
19.3%
0
20%
40%
60%
80%
100%
PERCENTAGE OF WORLDWIDE TABLET SHARE
SOURCE: IDC WORLDWIDE QUARTERLY TABLET TRACKER, AUGUST 30, 2016
©2016 TECHTARGET. ALL RIGHTS RESERVED
TechTarget


Cumulative iPhone sales
2008
2009
2010
2011
2012
2013

PRETERM BIRTH BY RACE & ETHNICITY



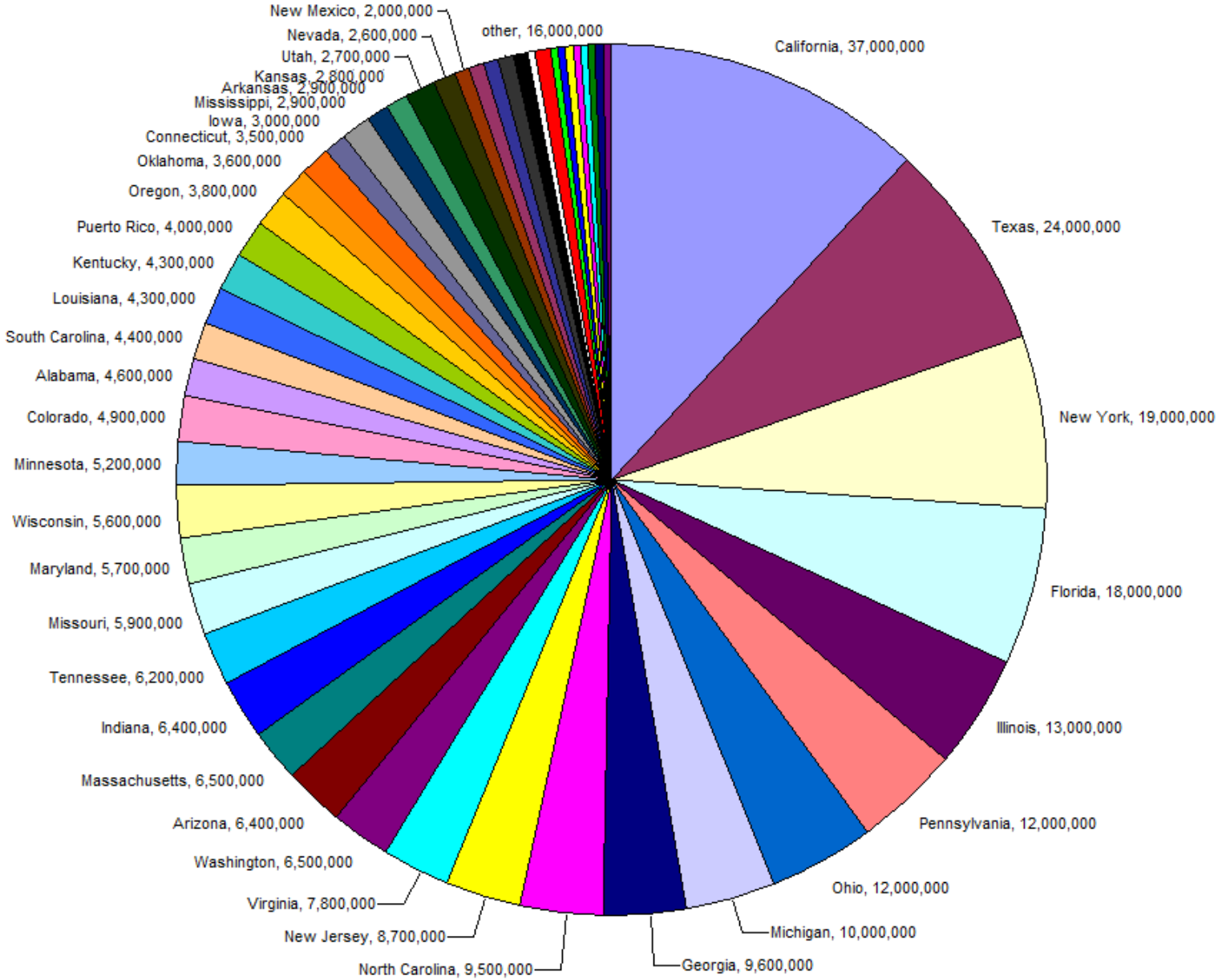

Top 20 Tourist Generating Countries To
UNITED STATES OF AMERICA
CANADA 14,043,658
MEXICO 4,295,124
UNITED KINGDOM 2,905,909
JAPAN 2,169,716
GERMANY 1,263,344
FRANCE 935,265
S.KOREA 560,405
BRAZIL 513,347
ITALY 558,594
HONG KONG 488,484
AUSTRALIA 526,441
INDIA 447,079
SPAIN 451,530
VENEZUELA 344,802
NETHERLANDS 416,870
COLOMBIA 288,439
IRELAND 296,771
ISRAEL 237,818
ARGENTINA 271,737
SWITZERLAND 282,595

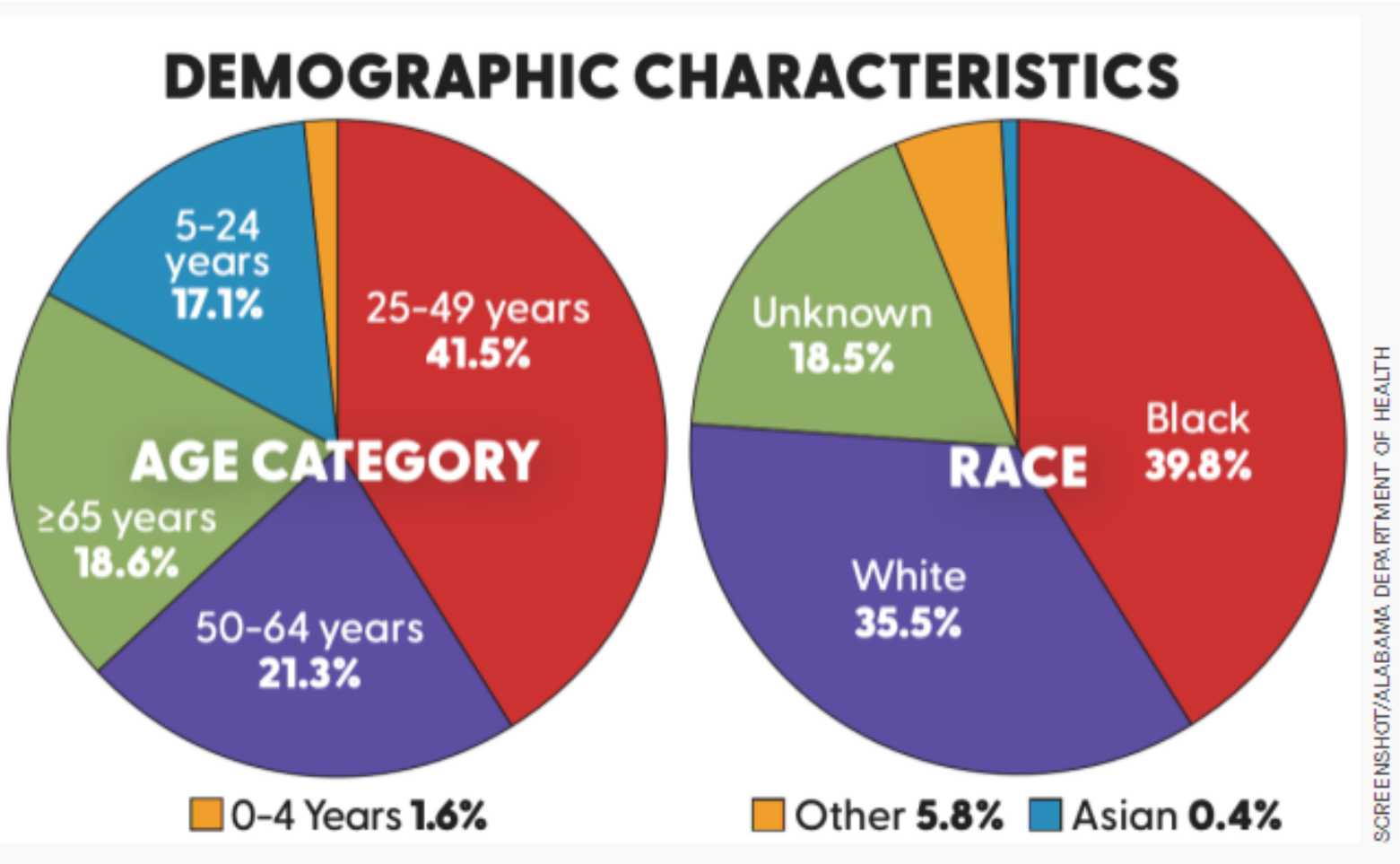

SALES ARE STRONG & PRICES ARE GOING UP
Number of Sold Businesses Are Going Up
7,056
7,494
7,222
7,842
2013
2014
2015
2016
Average Sale Price of Businesses Are Going Up
$180k
$185k
$199k
$200k
2013
2014
2015
2016

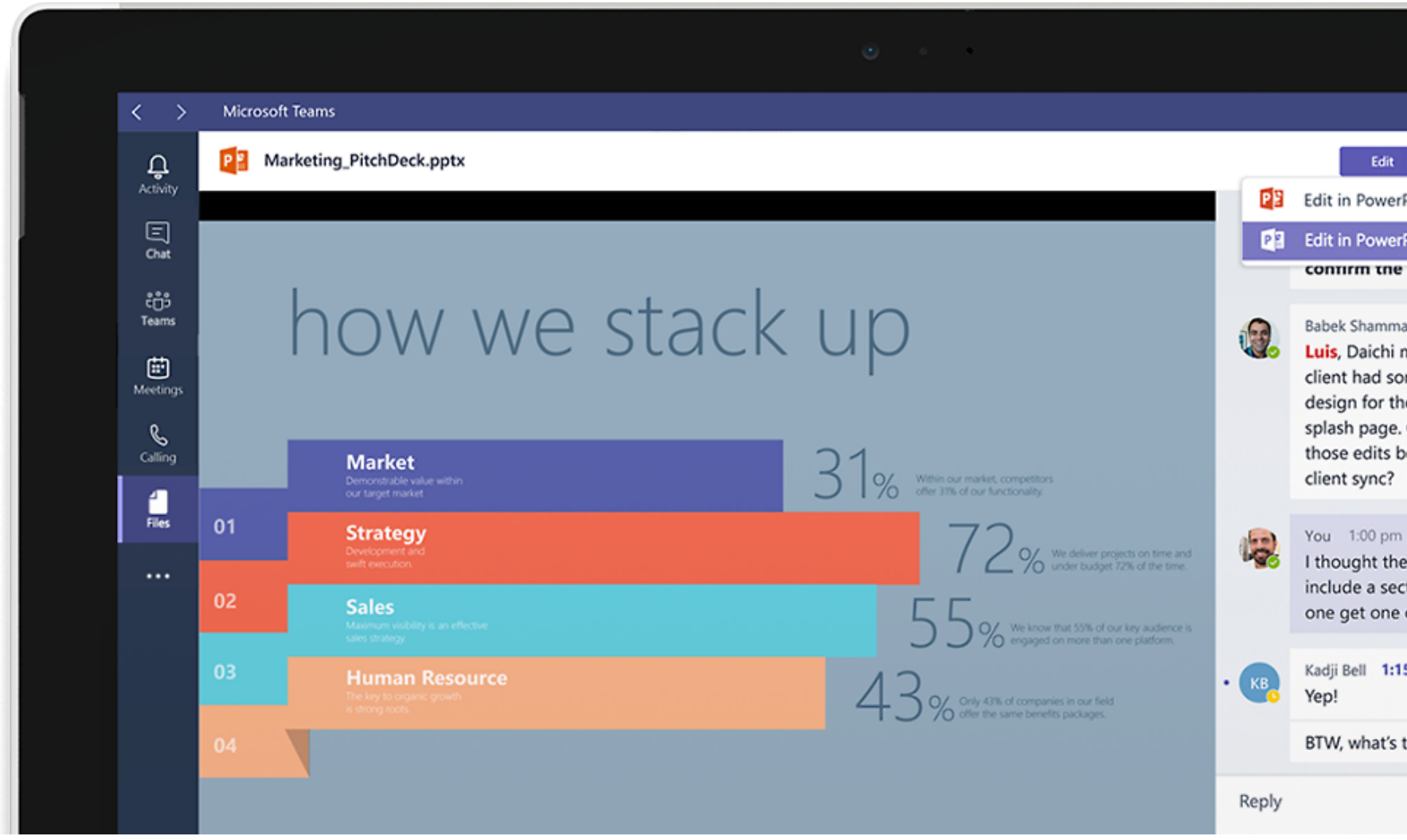

C5N #QuedateEnCasa
12:56 23°2
TESTEOS POR MILLÓN DE HABITANTES
258
330
7.000
14.100
15.700
22.300
BRASIL
ARGENTINA
EEUU
ITALIA
ALEMANIA
NORUEGA
ANTE LA PRESENCIA DE SÍNTOMAS LLAMAR AL 107 SAN LUIS

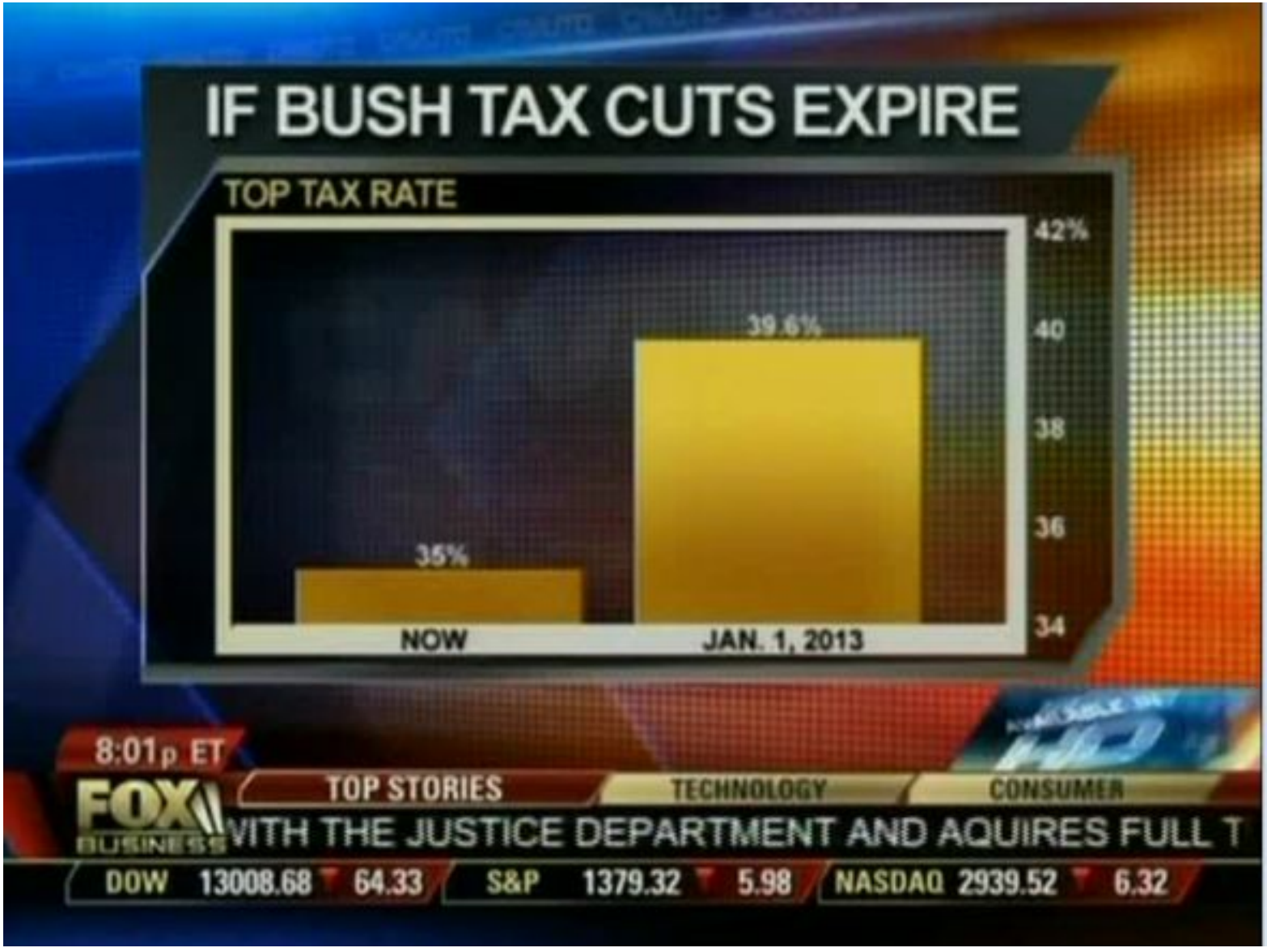
Cryptocurrencies Transaction Speeds Compared to Visa & Paypal

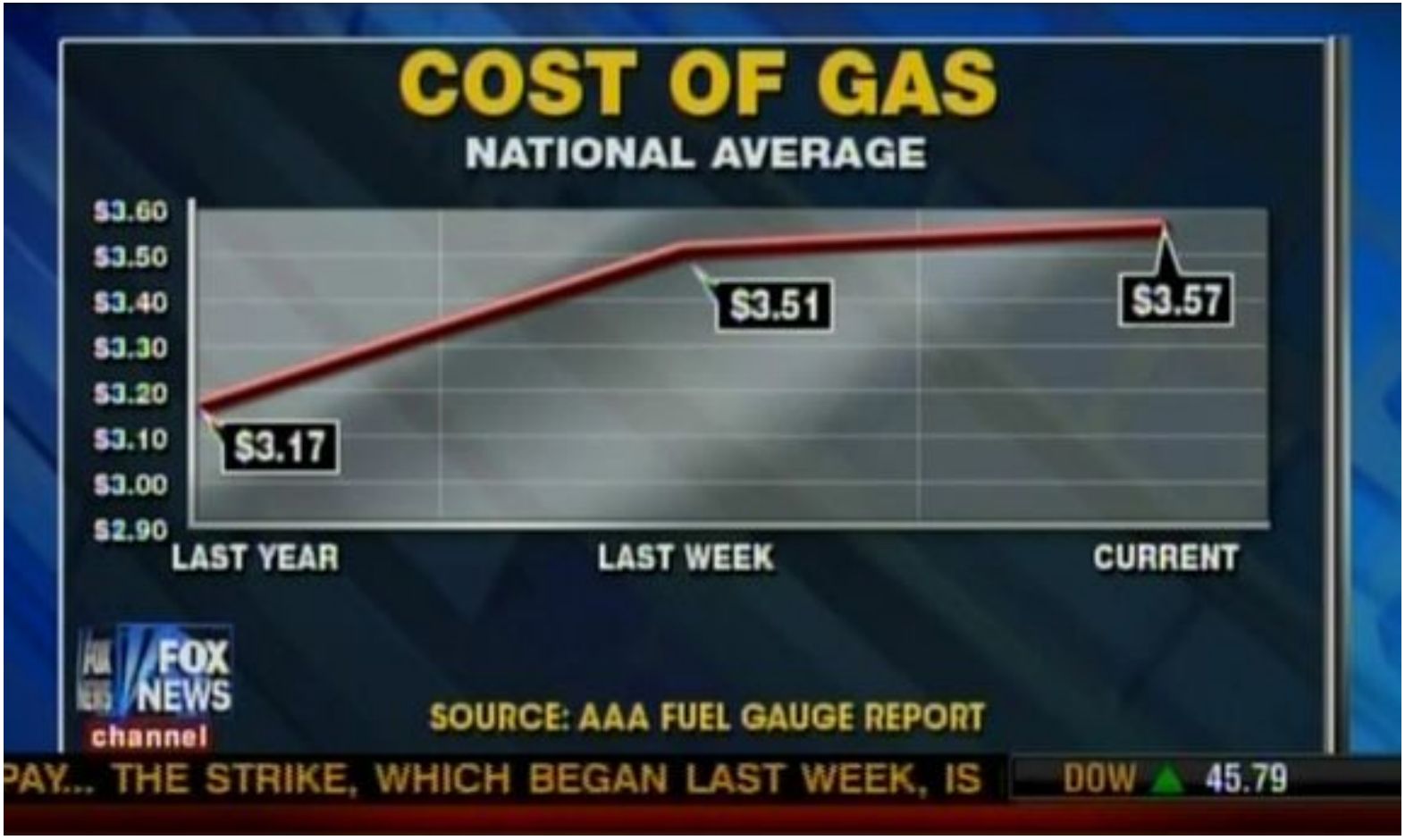
Article & Sources:
https://howmuch.net/articles/crypto-transaction-speeds-compared
https://howmuch.net/sources/crypto-transaction-speeds-compared

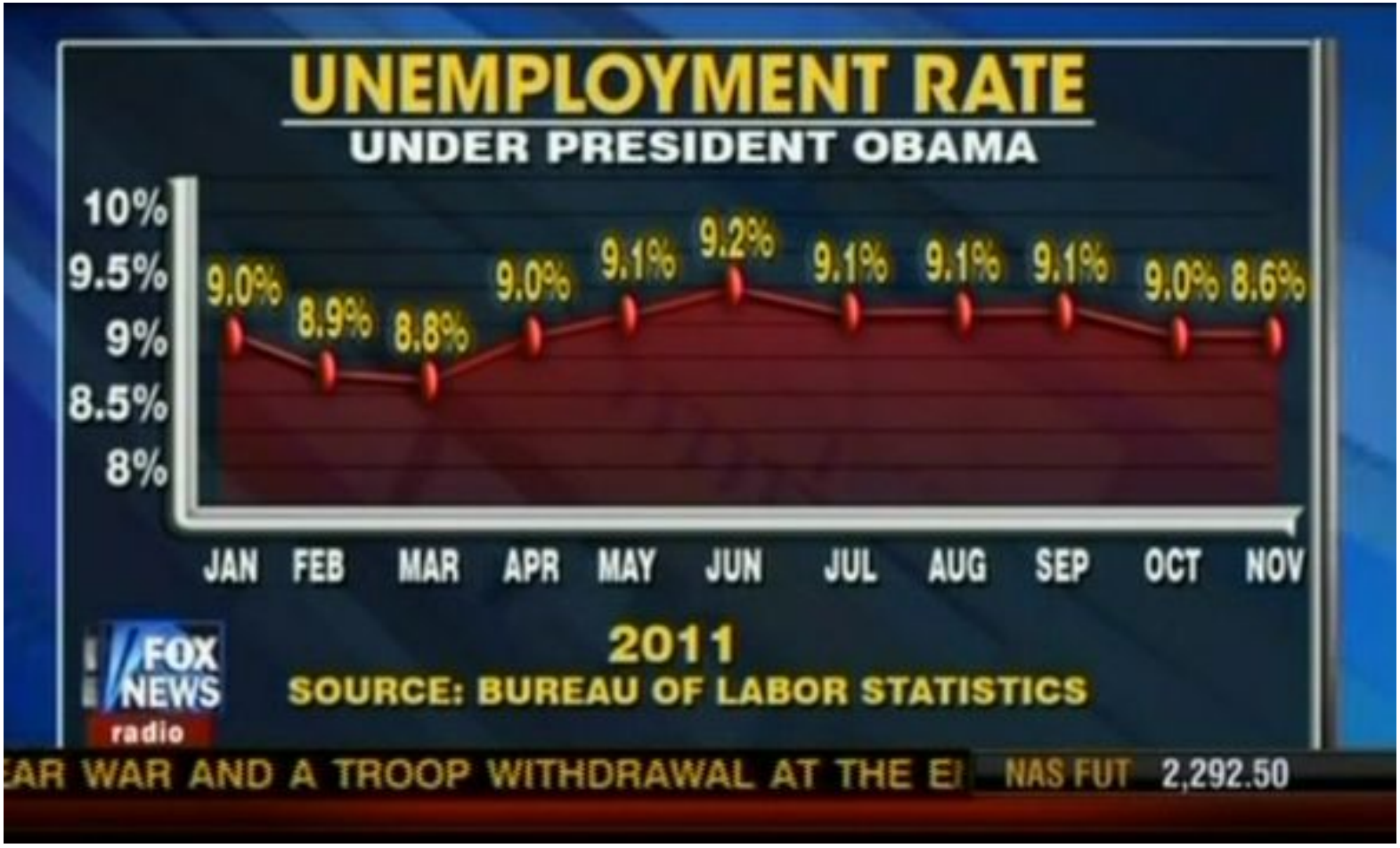

California, 37,000,000
Texas, 24,000,000
New York, 19,000,000
Florida, 18,000,000
Illinois, 13,000,000
Pennsylvania, 12,000,000
Ohio, 12,000,000
Michigan, 10,000,000
Georgia, 9,600,000
North Carolina, 9,500,000
New Jersey, 8,700,000
Virginia, 7,800,000
Washington, 6,500,000
Arizona, 6,400,000
Massachusetts, 6,500,000
Indiana, 6,400,000
Tennessee, 6,200,000
Missouri, 5,900,000
Maryland, 5,700,000
Wisconsin, 5,600,000
Minnesota, 5,200,000
Colorado, 4,900,000
Alabama, 4,600,000
South Carolina, 4,400,000
Louisiana, 4,300,000
Kentucky, 4,300,000
Puerto Rico, 4,000,000
Oregon, 3,800,000
Oklahoma, 3,600,000
Connecticut, 3,500,000
Iowa, 3,000,000
Mississippi, 2,900,000
Arkansas, 2,900,000
Kansas, 2,800,000
Utah, 2,700,000
Nevada, 2,600,000
New Mexico, 2,000,000
other, 16,000,000


DEMOGRAPHIC CHARACTERISTICS
5-24 years 17.1%
25-49 years 41.5%
AGE CATEGORY
≥65 years 18.6%
50-64 years 21.3%
0-4 Years 1.6%
Unknown 18.5%
Black 39.8%
RACE
White 35.5%
Other 5.8%
Asian 0.4%
SCREENSHOT/ALABAMA DEPARTMENT OF HEALTH


Microsoft Teams
Marketing_PitchDeck.pptx
Edit
Activity
Chat
Teams
Meetings
Calling
Files
how we stack up
01
Market
Demonstrable value within our target market
31%
Within our market, competitors offer 31% of our functionality.
02
Strategy
Development and swift execution.
72%
We deliver projects on time and under budget 72% of the time.
03
Sales
Maximum visibility is an effective sales strategy.
55%
We know that 55% of our key audience is engaged on more than one platform.
04
Human Resource
The key to organic growth is strong roots.
43%
Only 43% of companies in our field offer the same benefits packages.
Edit in PowerP
Edit in PowerP
Babek Shamma
Luis, Daichi n
client had so
design for th
splash page.
those edits b
client sync?
You 1:00 pm
I thought the
include a sec
one get one
KB
Kadji Bell 1:15
Yep!
BTW, what's t
Reply


IF BUSH TAX CUTS EXPIRE
TOP TAX RATE
42%
40
38
36
34
39.6%
35%
NOW
JAN. 1, 2013
8:01p ET
FOX BUSINESS
TOP STORIES
TECHNOLOGY
CONSUMER
VITH THE JUSTICE DEPARTMENT AND AQUIRES FULL T
DOW 13008.68 64.33
S&P 1379.32 5.98
NASDAQ 2939.52 6.32

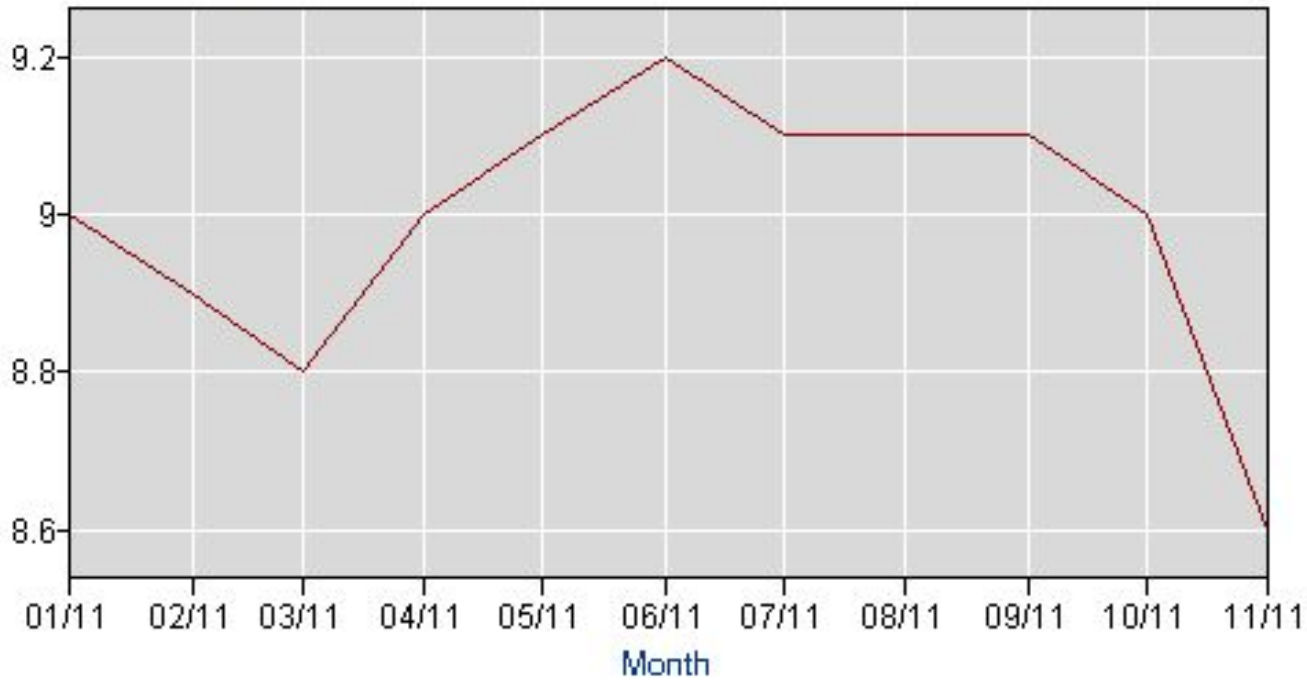

COST OF GAS
NATIONAL AVERAGE
$3.60
$3.50
$3.40
$3.30
$3.20
$3.10
$3.00
$2.90
$3.17
$3.51
$3.57
LAST YEAR
LAST WEEK
CURRENT
FOX NEWS channel
SOURCE: AAA FUEL GAUGE REPORT
PAY... THE STRIKE, WHICH BEGAN LAST WEEK, IS
DOW ▲ 45.79


UNEMPLOYMENT RATE
UNDER PRESIDENT OBAMA
10%
9.5%
9%
8.5%
8%
9.0%
8.9%
8.8%
9.0%
9.1%
9.2%
9.1%
9.1%
9.1%
9.0%
8.6%
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
2011
SOURCE: BUREAU OF LABOR STATISTICS
FOX NEWS radio
AR WAR AND A TROOP WITHDRAWAL AT THE E
NAS FUT 2,292.50

Data extracted on: December 12, 2011 (9:50:59 AM)

Labor Force Statistics from the Current Population Survey

Series Id:	LNS14000000
Seasonally Adjusted	
Series title:	(Seas) Unemployment Rate
Labor force status:	Unemployment rate
Type of data:	Percent or rate
Age:	16 years and over



Partial list of image sources:

- https://qz.com/1872980/how-bad-covid-19-data-visualizations-mislead-the-public/
- https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9

Some principles for good data visualizations

- Use the appropriate type of graphs/charts (consistent with type of data)

Some principles for good data visualizations

- Use the appropriate type of graphs/charts (consistent with type of data)
- Clearly label and explain the axes

Some principles for good data visualizations

- Use the appropriate type of graphs/charts (consistent with type of data)
- Clearly label and explain the axes
- Axes should generally start from 0 and consistent in scale

Some principles for good data visualizations

- Use the appropriate type of graphs/charts (consistent with type of data)
- Clearly label and explain the axes
- Axes should generally start from 0 and consistent in scale
- Avoid pie charts or other shapes; use bar graphs

Some principles for good data visualizations

- Use the appropriate type of graphs/charts (consistent with type of data)
- Clearly label and explain the axes
- Axes should generally start from 0 and consistent in scale
- Avoid pie charts or other shapes; use bar graphs
- In bar charts and histograms, the area of each bar should be proportional to the quantity represented.

Some principles for good data visualizations

- Use the appropriate type of graphs/charts (consistent with type of data)
- Clearly label and explain the axes
- Axes should generally start from 0 and consistent in scale
- Avoid pie charts or other shapes; use bar graphs
- In bar charts and histograms, the area of each bar should be proportional to the quantity represented.
- When visualizing proportions or percentages, clearly state what the population is.