DSFA

Spring 2020

Lecture 5

Census & Charts

Announcements

- Homework 2 due Friday 2/7
- Pay for Vocareum
- If you’ve just joined…
  - Make sure you are on Piazza
  - Make sure you join Vocareum (come see me)
  - Course website at cornell-dsfa.org
- Today’s demo: tinyurl.com/dsfa2020-demos; lecture5/lec05.ipynb. Be sure to add `!pip install datascience` to first cell, and run the cell...

Sunday night I...

A. ... watched the game.
B. … watched the ads.
C. … watched Shakira/JLo.
D. What game? What are you talking about?
E. … was watching *real* football, not the American kind.

Worst commercial

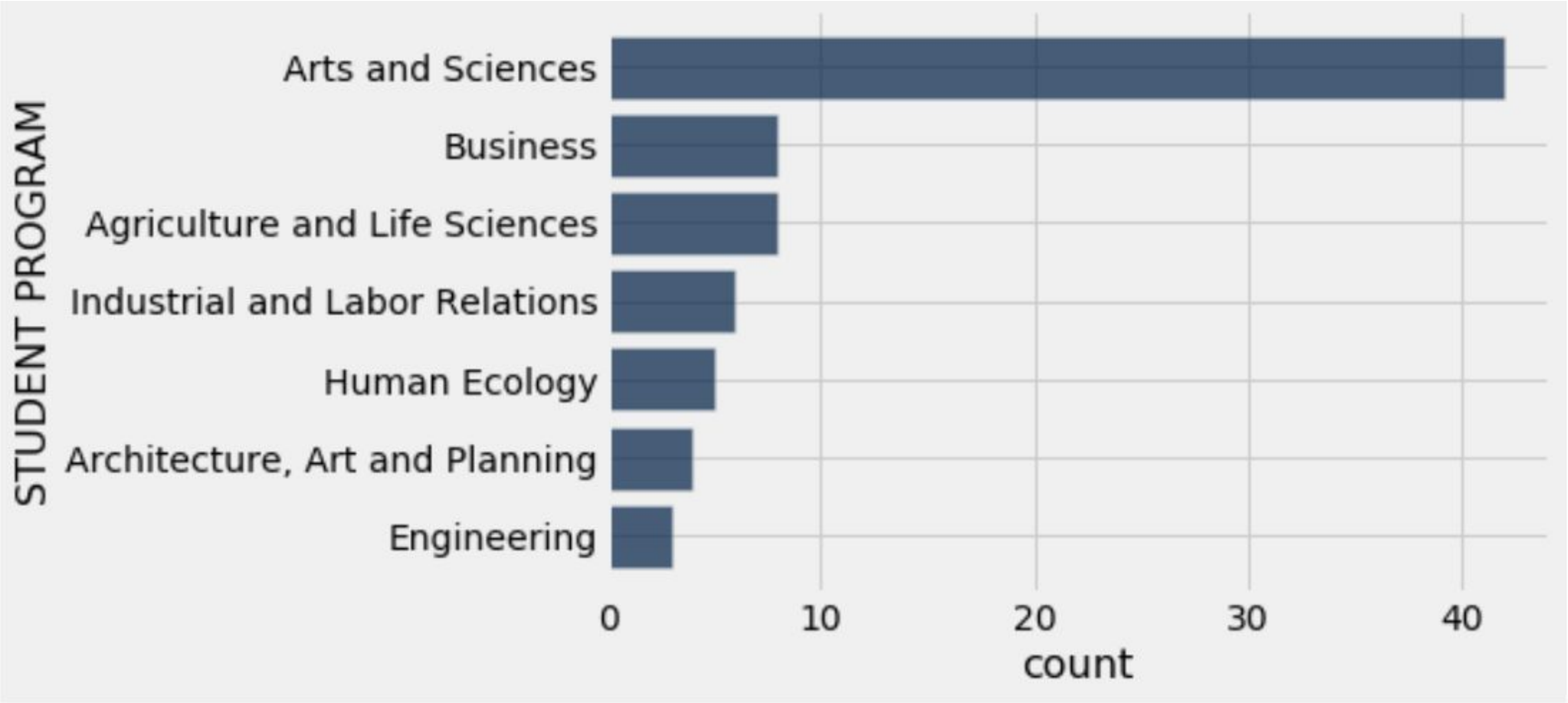
A. Avocado shopping network/Molly Ringwald
B. Rick and Morty/trapped in a Pringles commercial
C. TurboTax dancing
D. Poptart Pretzels
E. New York Life/Four Greek Words for Love

Best commercial

A. NFL 100/ “Take it to the house” running kid
B. Jason Momoa at home/Rocket Mortgage
C. Doritos Cool Ranch dance-off
D. Amazon “Before Alexa”/Ellen D.
E. Bill Murray/Groundhog Day/Jeep

What actor/actress has made the most money per movie made?

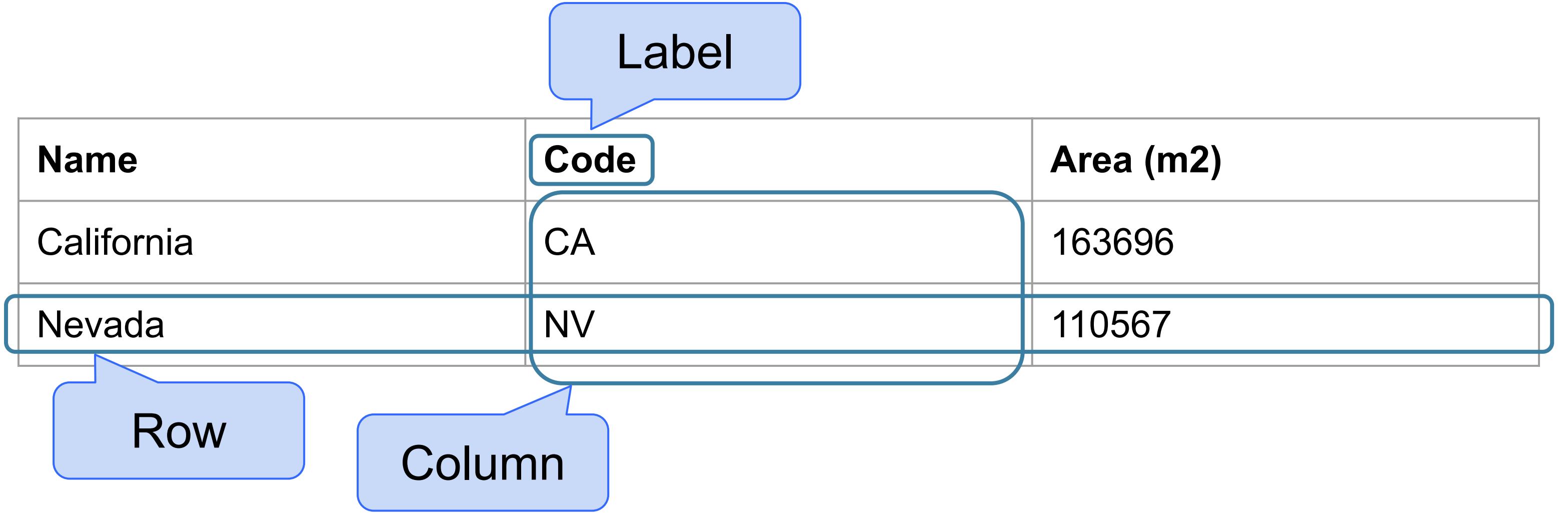
How can we make a chart like this?



Tables Review

Table Structure

- A Table is a sequence of labeled columns
- Labels are strings
- Columns are arrays, all with the same length



Name	Code	Area (m2)
California	CA	163696
Nevada	NV	110567

Label

Row

Column

Table Methods

- Creating and extending tables:
  - `Table().with_columns` and `Table.read_table`
- Finding the size: `t.num_rows` and `t.num_columns`
- Referring to columns: labels, relabeling, and indices
  - `t.labels` and `t.relabeled`; column indices start at 0
- Accessing data in a column
  - `t.column` takes a label or index and returns an array
- Using array methods to work with data in columns
  - `a.item(row_index)` returns a value in an array
  - `a.sum()`, `a.min()`, `a.max()` or `sum(a)`, `min(a)`, `max(a)`
- Creating new tables containing some of the original columns:
  - `select`, `drop`

Manipulating Rows

- `t.sort(column)` sorts the rows in increasing order
- `t.take(row_numbers)` keeps the numbered rows
    - Each row has an index, starting at 0
- `t.where(column, are.condition)` keeps all rows for which a column's value satisfies a condition
- `t.where(column, value)` keeps all rows for which a column's value equals some particular value
- `t.with_row` makes a new table that has another row

(Demo)

Discussion Questions

The table `nba` has columns `NAME`, `POSITION`, `TEAM`, and `SALARY`.

a) Create an array containing the names of all point guards (`PG`) who make more than $15M/year

b) Create a table containing NAME, TEAM, and SALARY of all players whose name contains the letter ‘i’ and whose team contains the letter ‘o’ who make at most $1M/year.

c) What was the average salary?

Census Data

The Decennial Census

- Every ten years, the Census Bureau counts how many people there are in the U.S.
- In between censuses, the Bureau estimates how many people there are each year.
- Article 1, Section 2 of the Constitution:
    - “Representatives and direct Taxes shall be apportioned among the several States … according to their respective Numbers …”

Analyzing Census Data

Leads to the discovery of interesting features and trends in the population

(Demo)

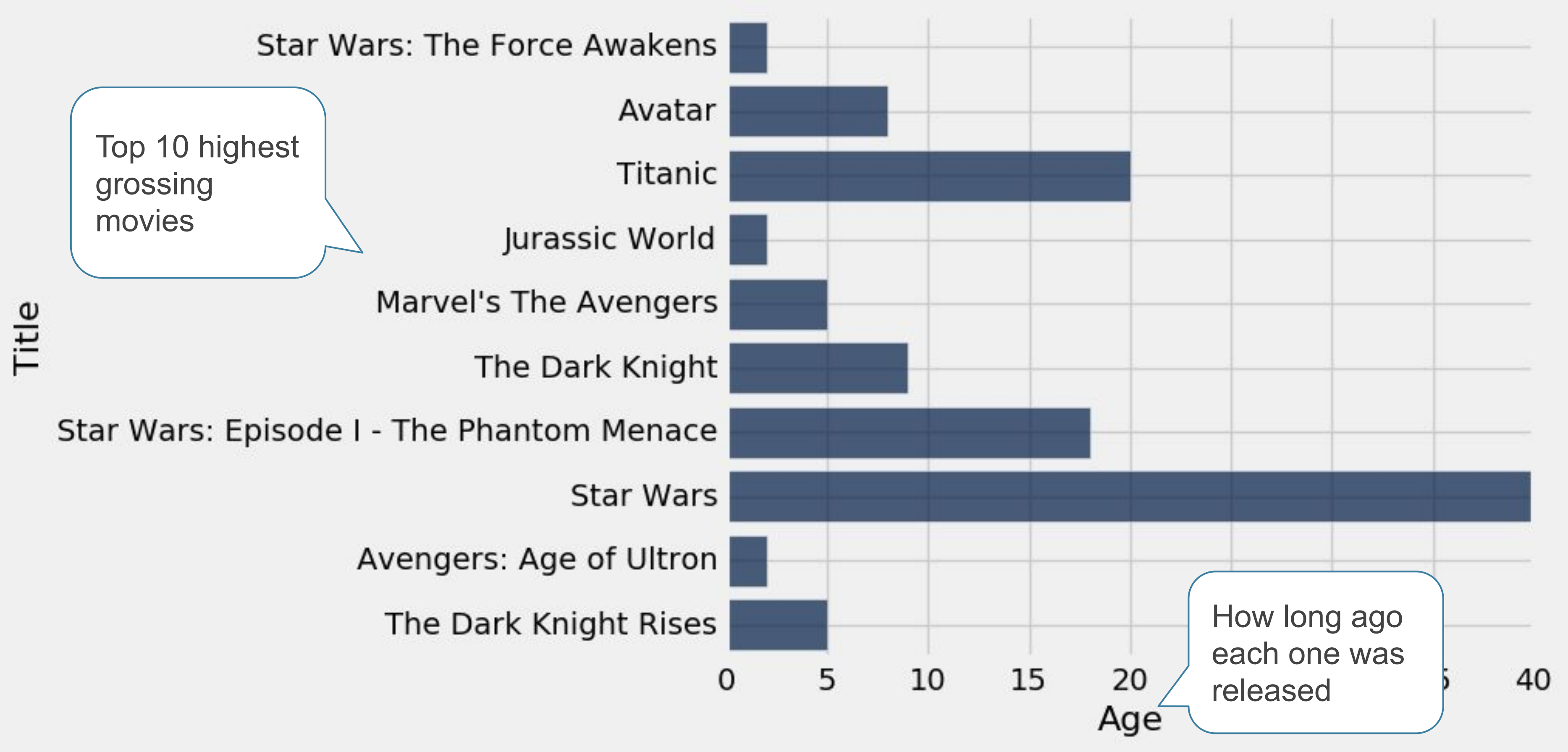
Census Table Description

- Values have column-dependent interpretations
  - The SEX column: 1 is *Male*, 2 is *Female*
  - The POPESTIMATE2010 column: *7/1/2010 estimate*
- In this table, some rows are sums of other rows
  - The SEX column: 0 is *Total* (of *Male* + *Female*)
  - The AGE column: 999 is *Total* of all ages
- Numeric codes are often used for storage efficiency
- Values in a column have the same type, but are not necessarily comparable (AGE 12 vs AGE 999)

http://www2.census.gov/programs-surveys/popest/datasets/2010-2015/national/asrh/nc-est2015-agesex-res.pdf

Data Visualization

How Do You Generate This Chart?



Top 10 highest grossing movies

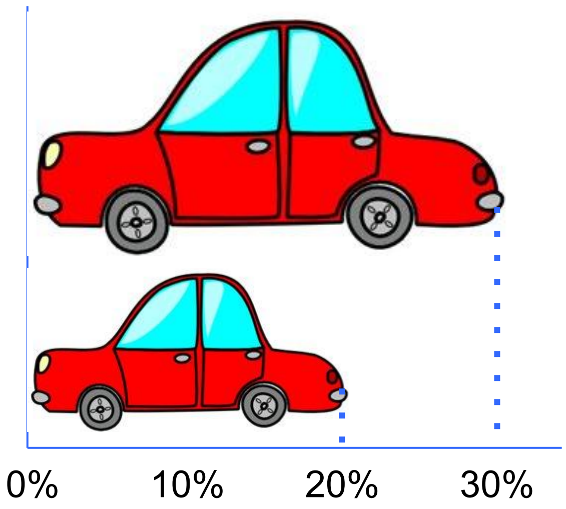
How long ago each one was released

Numerical Data

(Demo)

Area Principle

Areas should be proportional to the values they represent



In 2013,

30% of accidental deaths of males were due to automobile accidents

20% of accidental deaths of females were due to automobile accidents

Example from Tian Zheng

Types of Data

All values in a column should be both the same type **and** be comparable to each other in some way

- **Numerical** — Each value is from a numerical scale
  - Numerical measurements are ordered
  - Differences are meaningful
- **Categorical** — Each value is from a fixed inventory
  - May or may not have an ordering
  - Categories are the same or different

“Numerical” Data

Just because the values are numbers, doesn’t mean the variable is numerical

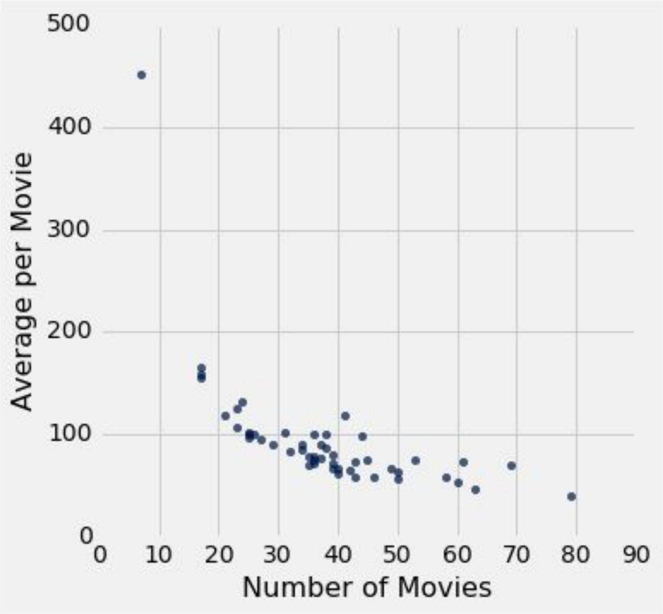
- Census example had numerical `SEX` code (0, 1, and 2)
- It doesn’t make sense to perform arithmetic on these “numbers”, e.g. 1 - 0 or (0+1+2)/3 are nonsense here
- The variable `SEX` is still categorical, even though numbers were used for the categories

Terminology

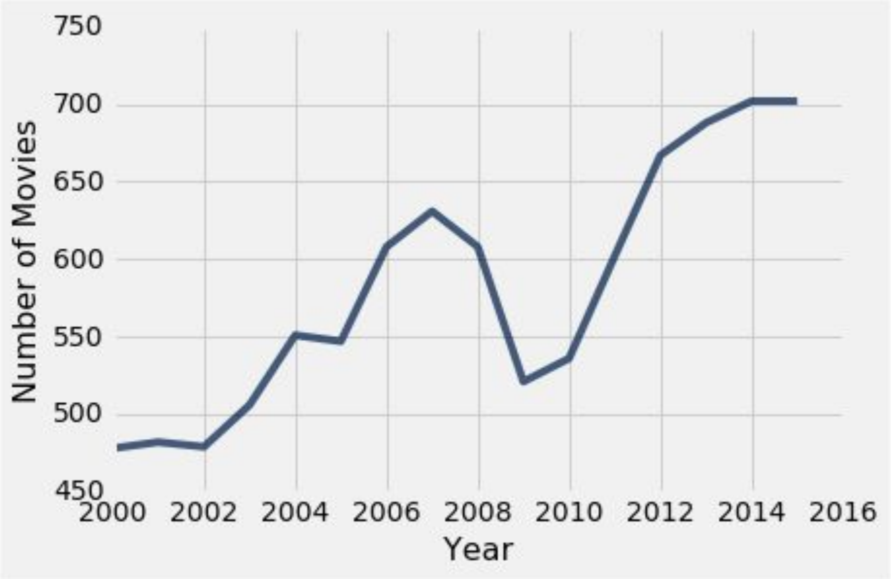
- **Individuals**: those whose features are recorded
- **Variables**: features; these vary across individuals
- Variables have different **values**
- Values can be **numerical**, or **categorical**, or of many other types

Plotting Two Numerical Variables

Scatter plot: `scatter`



Line graph: `plot`



Categorical Data

(Demo)

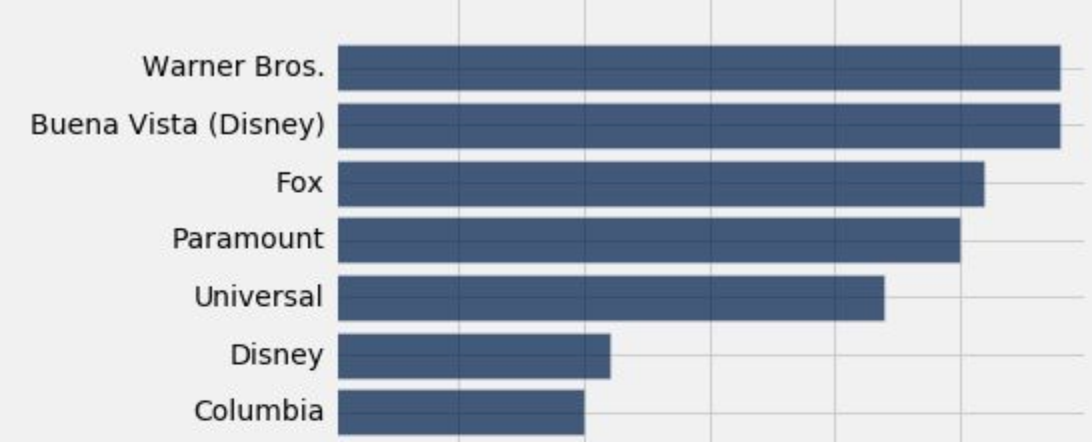
Bar Charts

- Can visualize categorical data via bar charts.
  - E.g. gross of top grossing movies.
- The `group` method counts the number of rows for each value in a column
  - E.g. count of how many top movies were released by each studio

(Demo)

Categorical Distributions

bar chart: `barh`



Displays a categorical distribution

Discussion Question

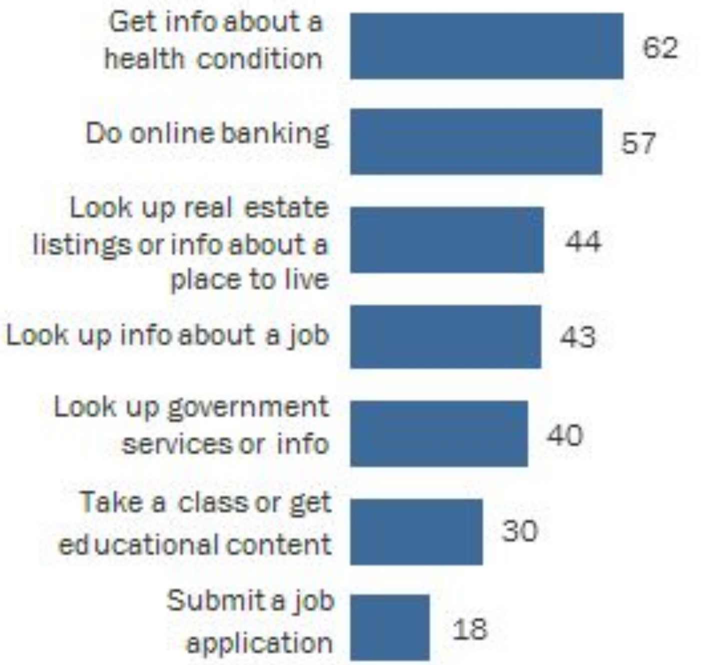
Which of the following questions can be answered by this chart?

Among survey responders...

- What proportion did **not** use their phone for online banking?
- What proportion either used their phone for online banking or to look up real estate listings?
- Did everyone use their phone for at least one of these activities?
- Did anyone use their phone for both online banking and real estate?

More than Half of Smartphone Owners Have Used Their Phone to get Health Information, do Online Banking

% of smartphone owners who have used their phone to do the following in the last year



Pew research center, 2014