Data Visualization

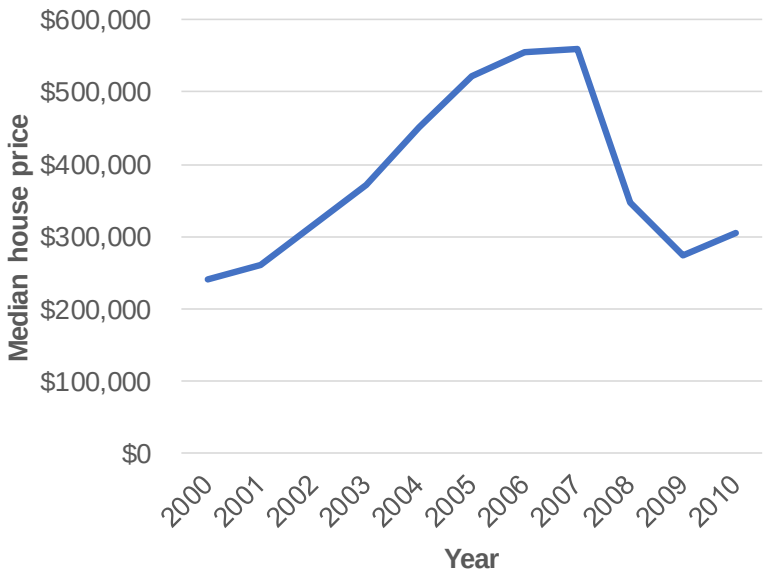
CS 418. Introduction to Data Science

Data Visualization
Introduction

- **Data visualization** is the display of data in a format, such as a **table** or **chart**, that conveys particular information to the viewer.
- **Data visualization** gives us a **better intuitive sense** of the data.
- **Data visualization** can **bring to light hidden patterns** in data.
- ***Example***:

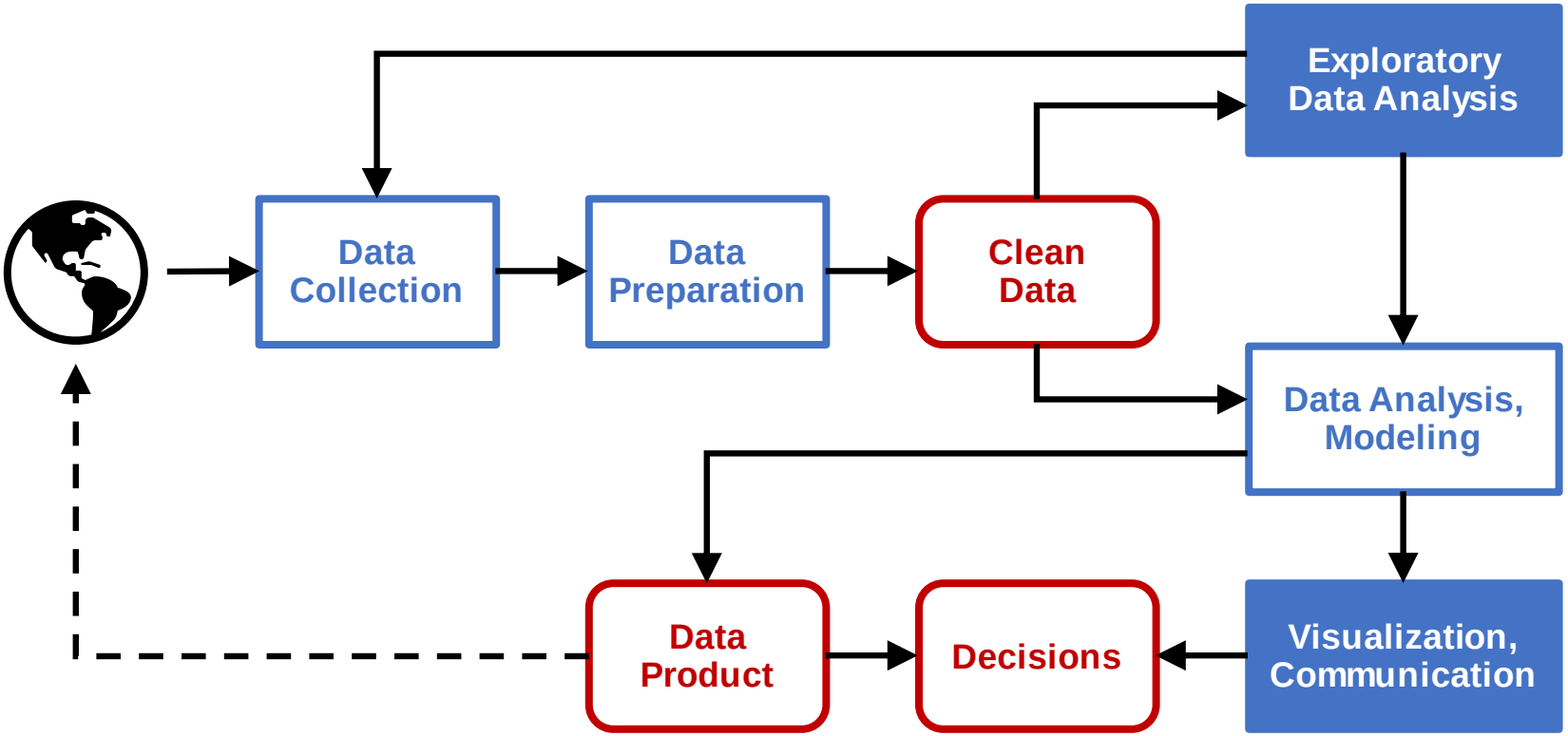
The following data represents California median house prices from 2000-2010: 2000 $241,000; 2001 $262,000; 2002 $316,000; 2003 $372,000; 2004 $451,000; 2005 $523,000; 2006 $556,000; 2007 $560,000; 2008 $348,000; 2009 $275,000; 2010 $305,000

Year	Median house price
2000	$241,000
2001	$262,000
2002	$316,000
2003	$372,000
2004	$451,000
2005	$523,000
2006	$556,000
2007	$560,000
2008	$348,000
2009	$275,000
2010	$305,000



Data Visualization
The Data Science Process

- **Data visualization** is as much a part of the **data exploration** step as the **data presentation** step.



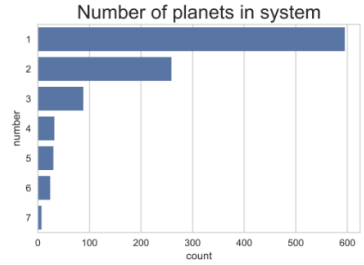
Adapted from: Cathy O'Neil and Rachel Schutt, *Doing Data Science* (2013)

Data Visualization
Exercise 6.1

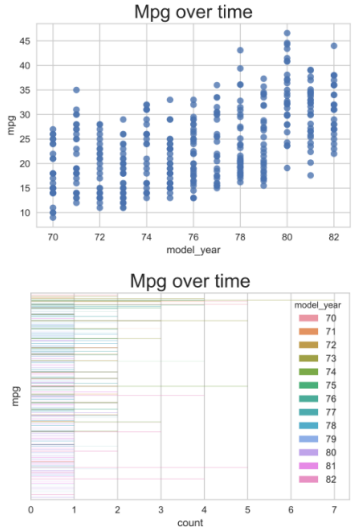
THINK PAIR SHARE

Which chart would you choose to visualize each of the following datasets?

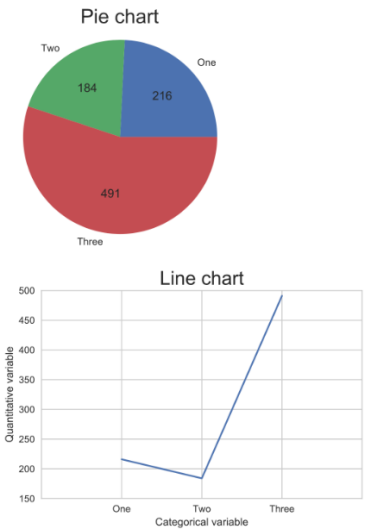
Dataset A



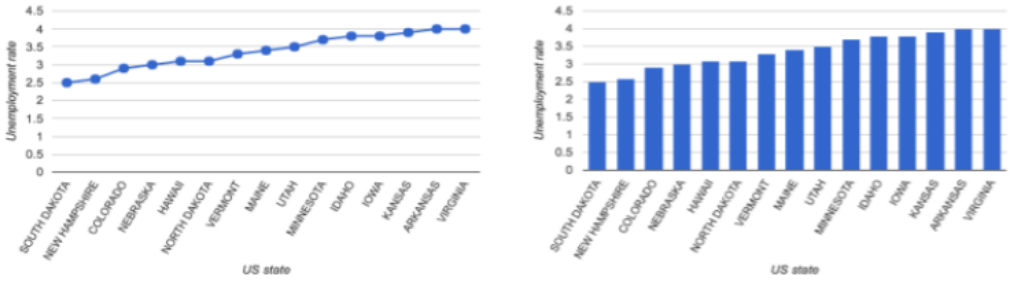
Dataset B



Dataset C



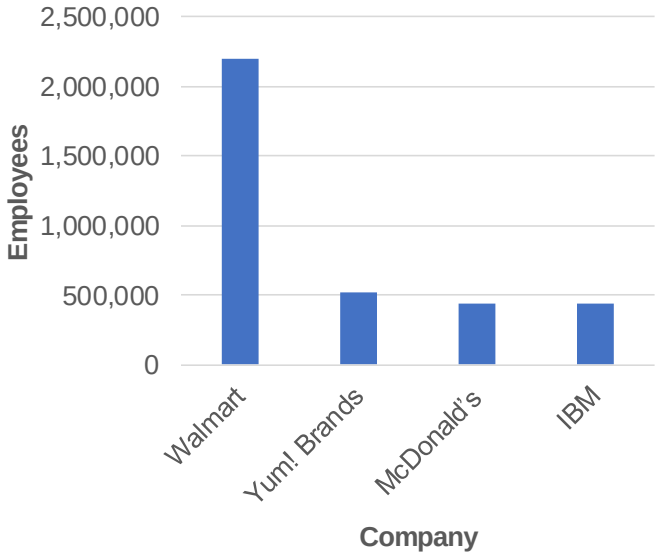
Dataset D

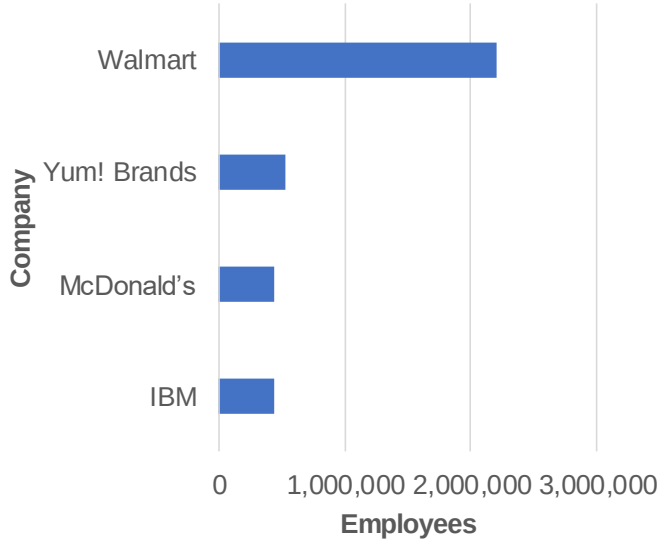


Data Visualization
Bar Charts

- **Bar charts** depict data values for a categorical variable with each category shown as a bar of appropriate length.
- We use **bar charts** to show **relative frequencies** for categories.
- **Bar charts** can be drawn **vertically** or **horizontally**.
- ***Example***:

Company	Employees
Walmart	2,200,000
Yum! Brands	523,000
McDonald's	440,000
IBM	434,246

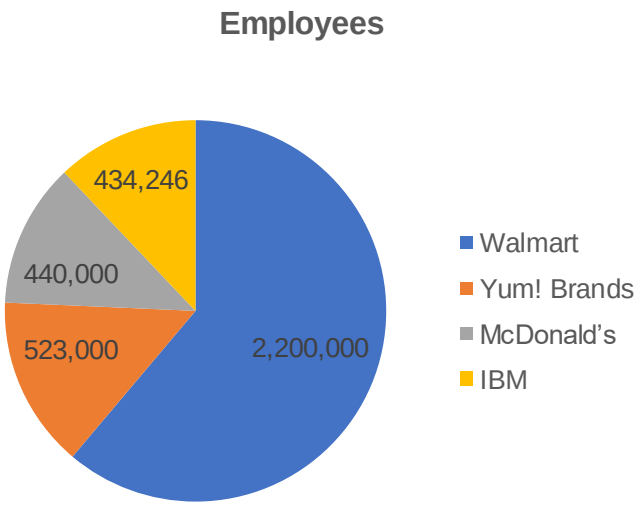




Data Visualization
Pie Charts

- **Pie charts** depict data values for a categorical variable using a circle with each category shown as a slice of appropriate size.
- We use **pie charts** to show **relative frequencies** for categories.
- ***Example***:

Company	Employees
Walmart	2,200,000
Yum! Brands	523,000
McDonald's	440,000
IBM	434,246

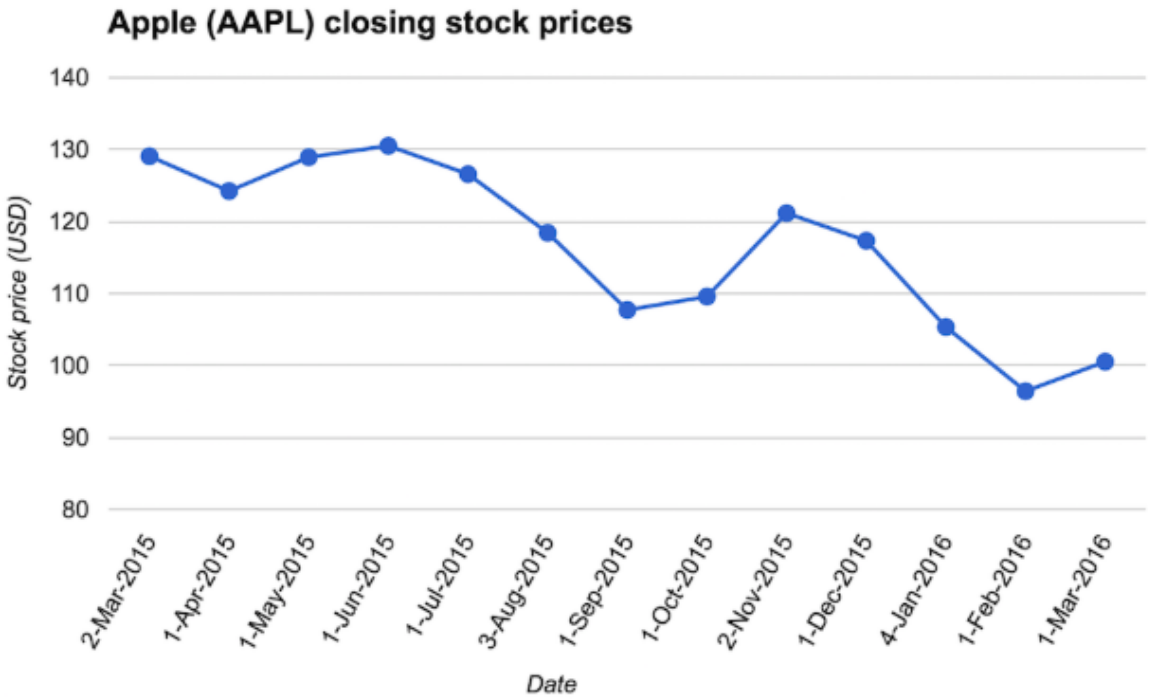


Data Visualization
Scatter Plots

- **Scatter plots** depict the relationship between two variables on a rectangular coordinate system, where each axis corresponds to one variable.
- ***Example*:**

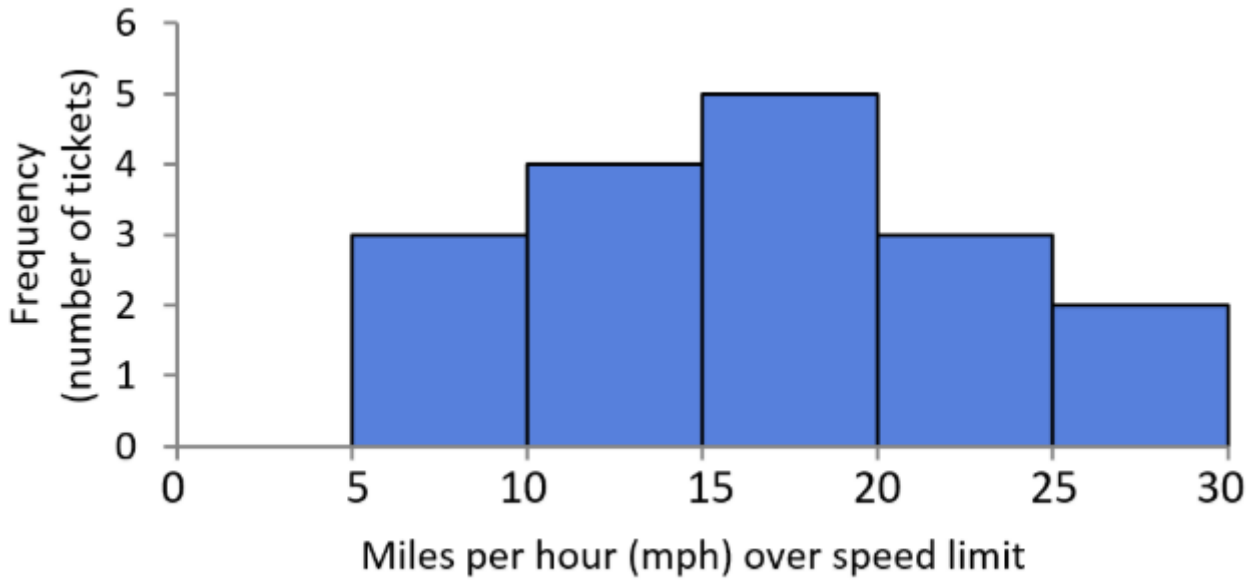
Data Visualization
Line Plots

- **Line plots** depict data trends by using straight lines to connect successive data points in a **scatter plot**.
- ***Example***:



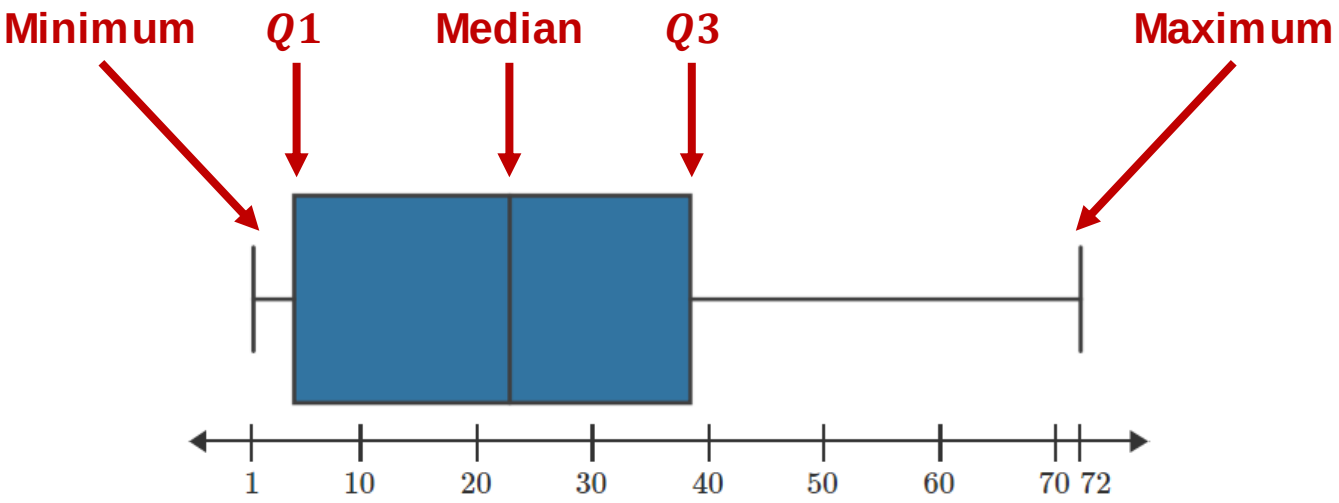
Data Visualization
Histograms

- **Histograms** depict the distribution of data values by splitting a continuous variable into consecutive, non-overlapping intervals or **bins**.
- The simplest type of **histogram** has **bins** of equal size.
- ***Example*:**



Data Visualization
Box Plots

- **Box plots** depict the distribution of data values using **boxes** and **whiskers**.
- **Box plots** may also show the presence of **outliers**.
- ***Example*:**



Data Visualization
References

- Daniel Chen. *Pandas for Everyone* (2018).
- Joel Grus. *Data Science from Scratch* (2015).