What is statistics?

INTRODUCTION TO STATISTICS

George Boorman
Curriculum Manager, DataCamp

datacamp

What is statistics?

- **The field of statistics** - the practice and study of collecting and analyzing data

- Two main branches of statistics:
  - **Descriptive/summary statistics** - describing or summarizing our data
  - **Inferential statistics** - collect a sample of data, and apply the results to the population that the sample represents

Statistics is everywhere!

- Sports statistics

- Personal finances

[1] Image credits: https://unsplash.com/@jesusance; https://unsplash.com/@andretaissin; https://unsplash.com/@unarchive

What can statistics do?

- Allows us to answer practical questions:
  - What is the average salary in the USA?
  - How many customer inquiries is a company likely to receive per week?

- It has applications across society:
  - Developing safer products such as cars or airplanes
  - Help governments understand the needs of a population

- Validates scientific breakthroughs, such as Covid-19 vaccines

[1] source: https://www.bmj.com/content/373/bmj.n1088

Limitations of statistics

- Statistics requires specific, measurable questions:
  - Is rock music more popular than jazz?
  - On average, do women live longer than men?
- We can't use statistics to find out ***why*** relationships exist

[1] Image credit: https://unsplash.com/@mohammadmetri

Types of data: numeric

- Continuous data:
  - Stock prices

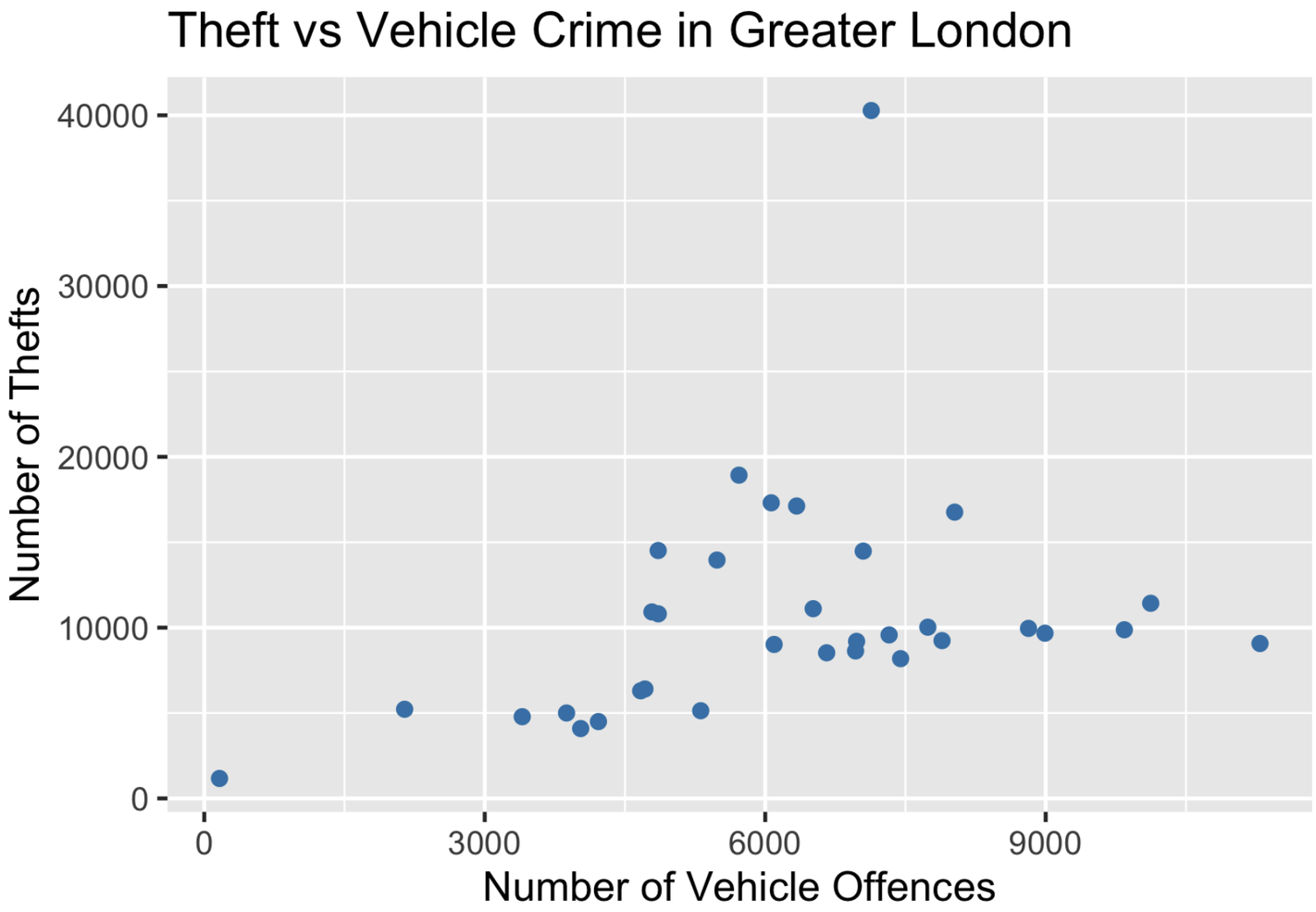
Stock	Opening Price ($)	Close Price {$}
Amazon.com, Inc	2328.14	2329.00
Apple Inc	156.77	157.04
Netflix Inc	188.32	188.75

- Interval/count data:
  - How many cups of coffee do people drink per day?

Name	Cups of coffee per day
Jessica	4
Andrew	2
Penny	3

[1] Image credits: Stocks https://unsplash.com/@behy_studio

Visualizing numeric data



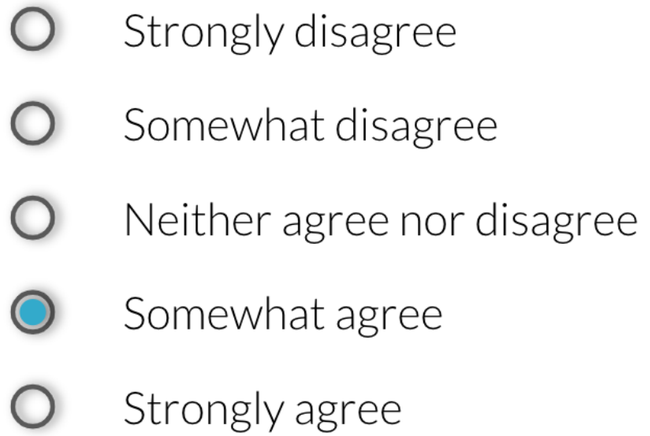
[1] Data source: Metropolitan Police Service, United Kingdom

Types of data: categorical

- Nominal data:
  - Eye color

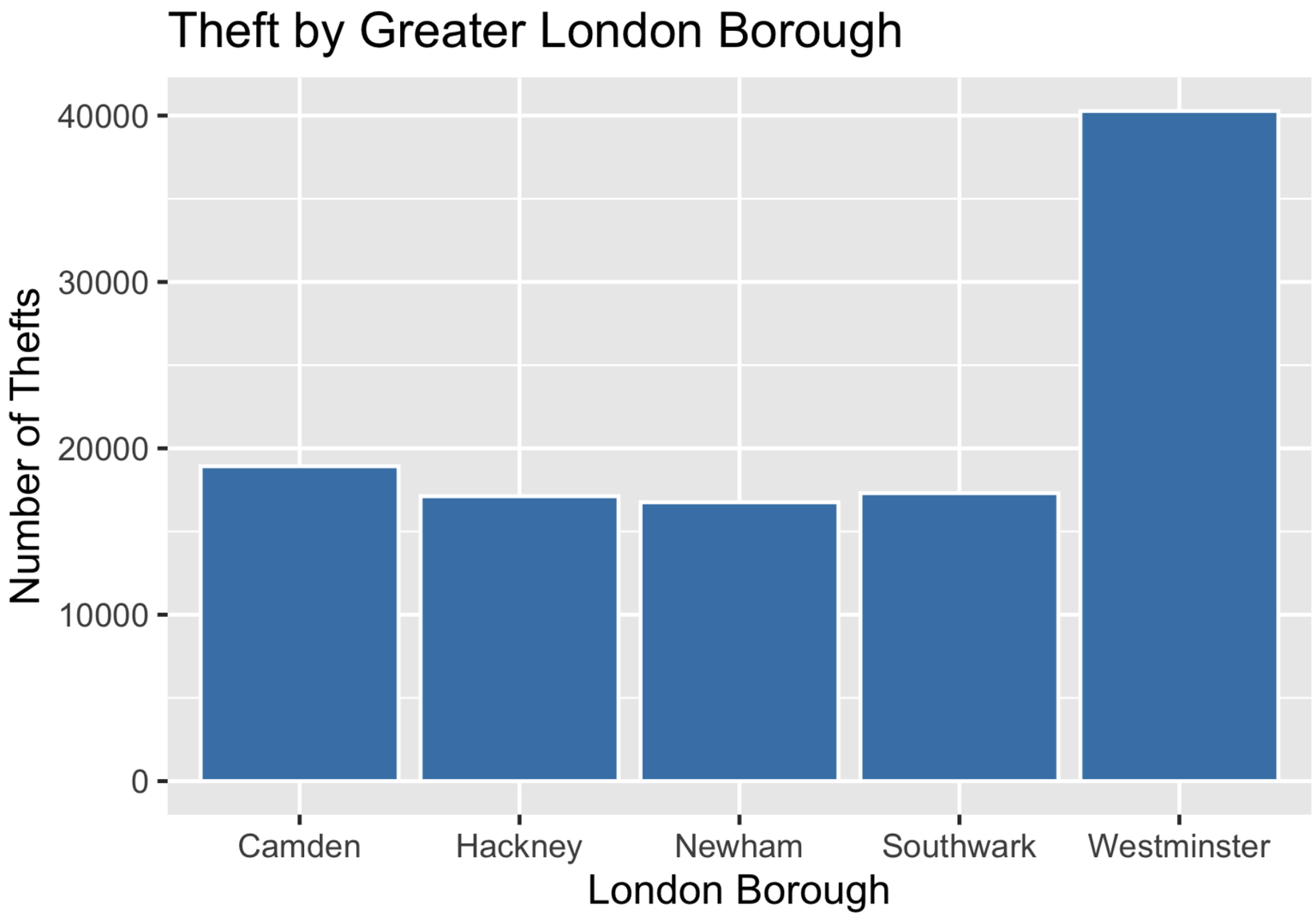
Name	Eye color
Jessica	Brown
Adam	Green
Sarah	Blue

- Ordinal data:
  - How strongly do you agree that basketball is the best sport?



[1] Image credit: https://unsplash.com/@mango_quan

Visualizing categorical data



Descriptive / Summary statistics

- *Describe* or *summarize* data

Borough	Number of Thefts	Percentage of Total
Westminster	40,278	**36.48%**
Camden	18,928	17.15%
Southwark	17,309	15.68%
Hackney	17,121	15.51%
Newham	16,762	15.18%

Inferential statistics

- Use a sample to *draw conclusions* about a population
- How many people purchase clothing following social media advertising?

[1] source: https://unsplash.com/@pickawood

Let's practice!

INTRODUCTION TO STATISTICS

Measures of center

INTRODUCTION TO STATISTICS

George Boorman
Curriculum Manager, DataCamp

datacamp

Why are measures of center useful?

- Average number of sales orders per month

- Typical cost of a house

- Most common hair color

[1] Image credits: https://unsplash.com/@arnosenoner; https://unsplash.com/@brenoassis;

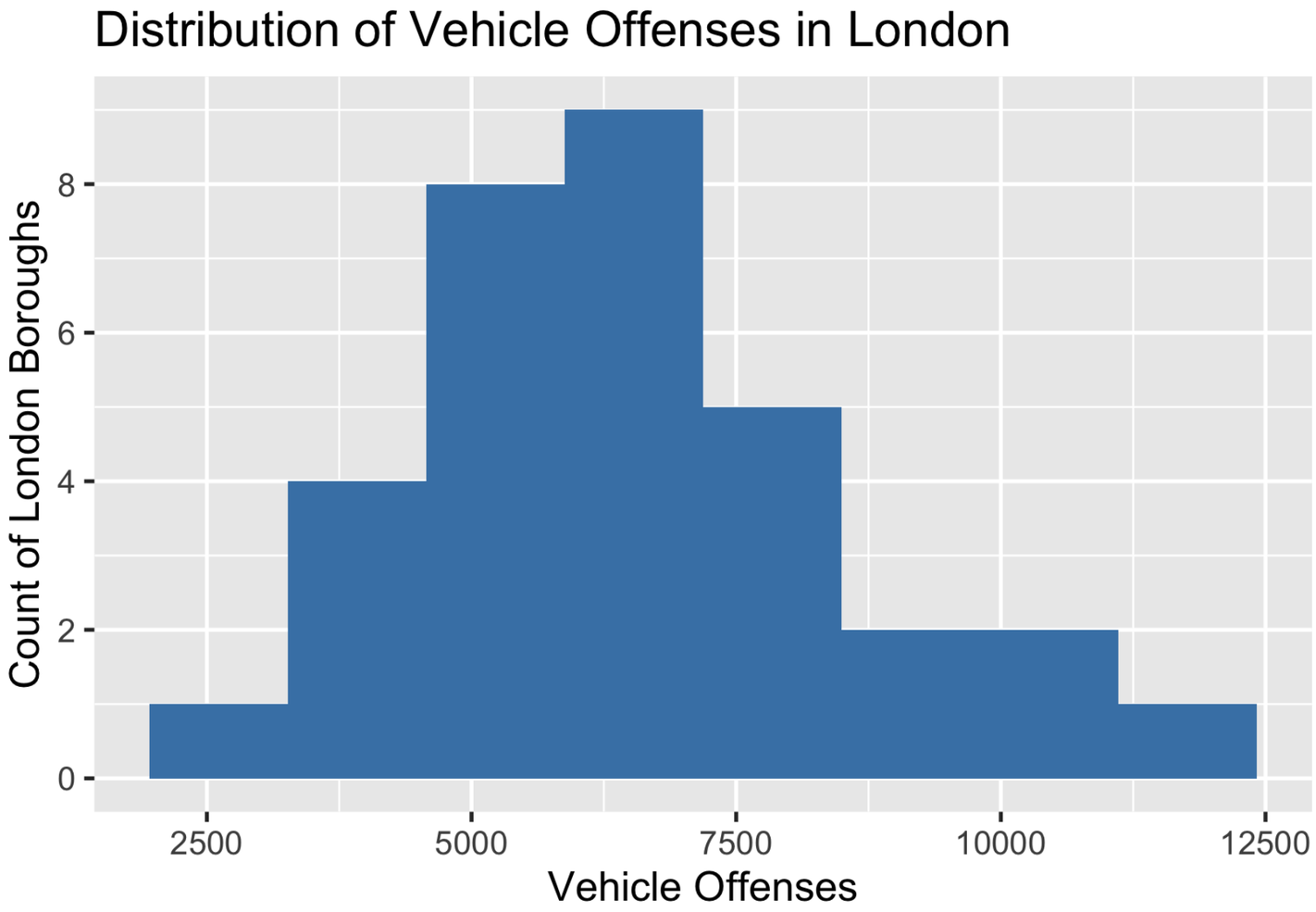
Crime data

Borough	Burglary	Robbery	Theft	Vehicle Crime	Total
Barking and Dagenham	2318	1265	6300	4668	14551
Barnet	5067	1369	9875	9841	26152
Bexley	1583	444	4500	4216	10743
Brent	3893	1650	10026	7739	23308
Bromley	3053	844	8635	6966	19498
...	...	...	...	...	...

Crime data

Borough	Burglary	Robbery	Theft	Vehicle Crime	Total
Barking and Dagenham	2318	1265	6300	4668	14551
Barnet	**5067**	1369	9875	9841	26152
Bexley	1583	444	4500	4216	10743
Brent	3893	1650	10026	7739	23308
Bromley	3053	844	8635	6966	19498
...	...	...	...	...	...

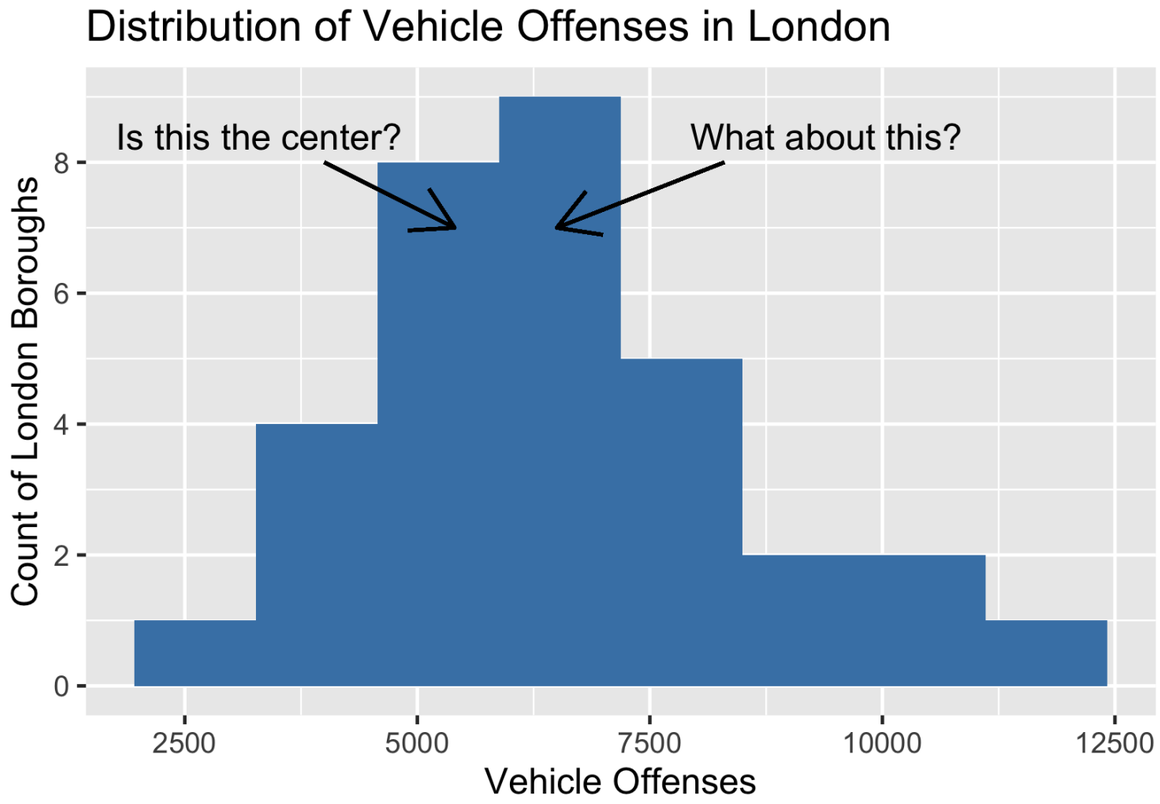
Histograms


Distribution of Vehicle Offenses in London
Count of London Boroughs
0
2
4
6
8
2500
5000
7500
10000
12500
Vehicle Offenses

What is the typical amount of vehicle crime in London?

Where is the center of the data?

- Mean
- Median
- Mode



Measures of center: mean

$$mean = \frac{Sum\ of\ all\ values}{Count\ of\ all\ values}$$

- Mean burglaries per London borough:

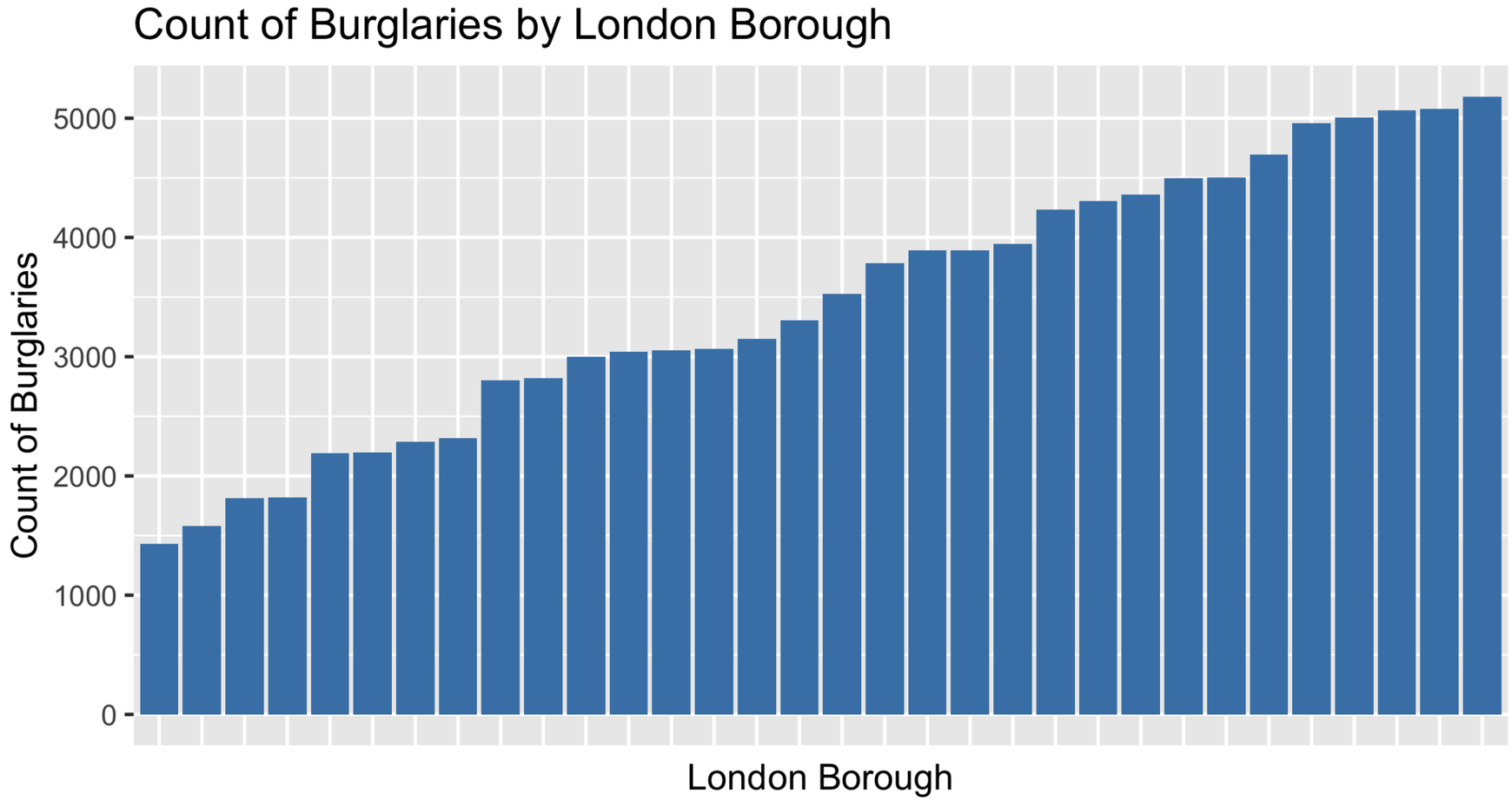
$$Mean(Burglary) = \frac{(2,318\ + 5,067\ + 1,583\ + 3,893\ + ...)}{32}\ =\ 3,462.8125$$

Measures of center: mean

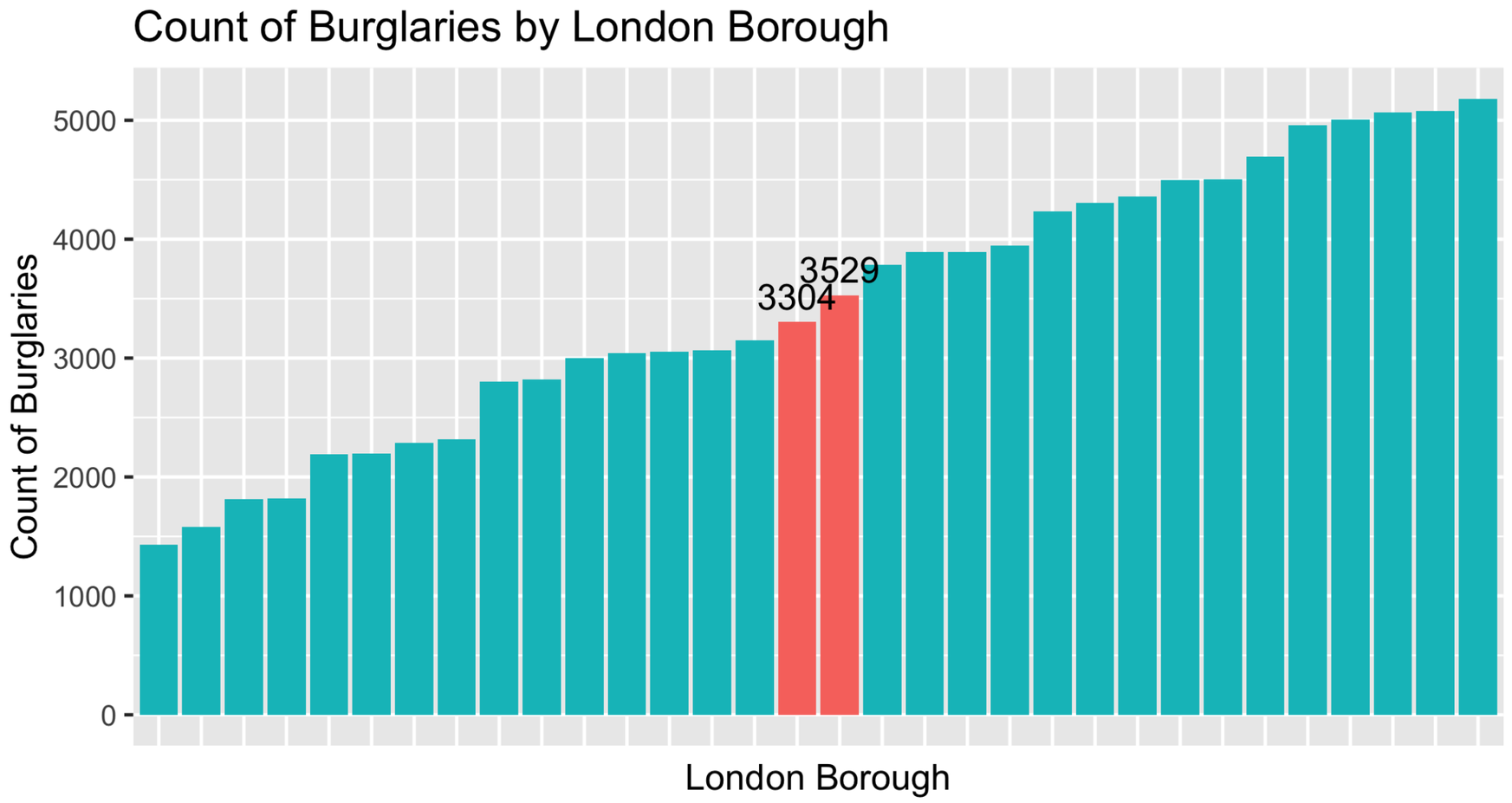
Type of Crime	Mean
Burglary	3358.06
Robbery	1450.85
Theft	10642.21
Vehicle Offenses	6227.27
Overall	47671.62

Measures of center: median

- Median - middle value for a given variable



Measures of center: median



Measures of center: median

$$Median(Burglary) = \frac{(3{,}304 + 3{,}529)}{2}$$

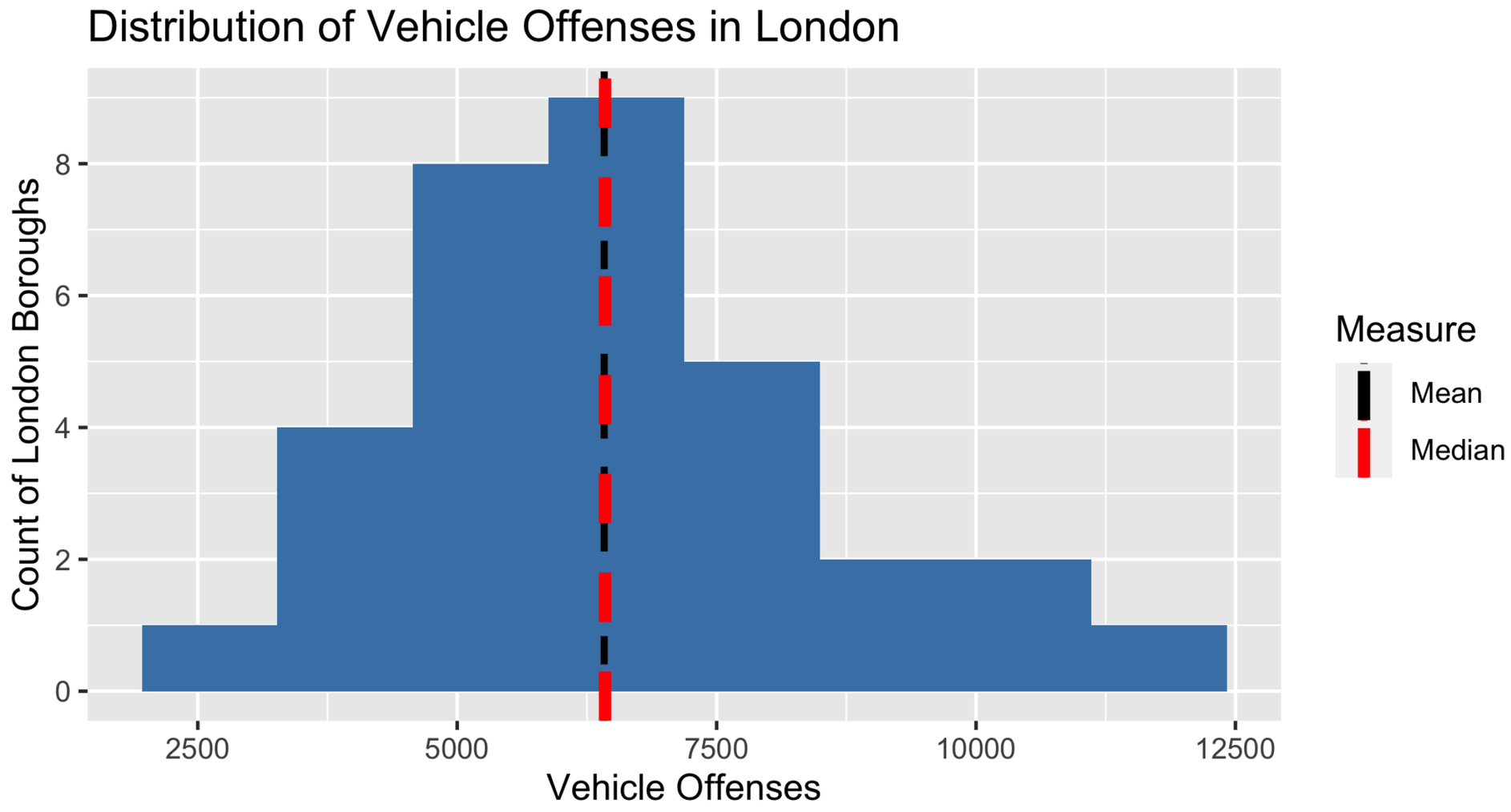
$$Median(Burglary) = 3,416.5$$

Measures of center: mode

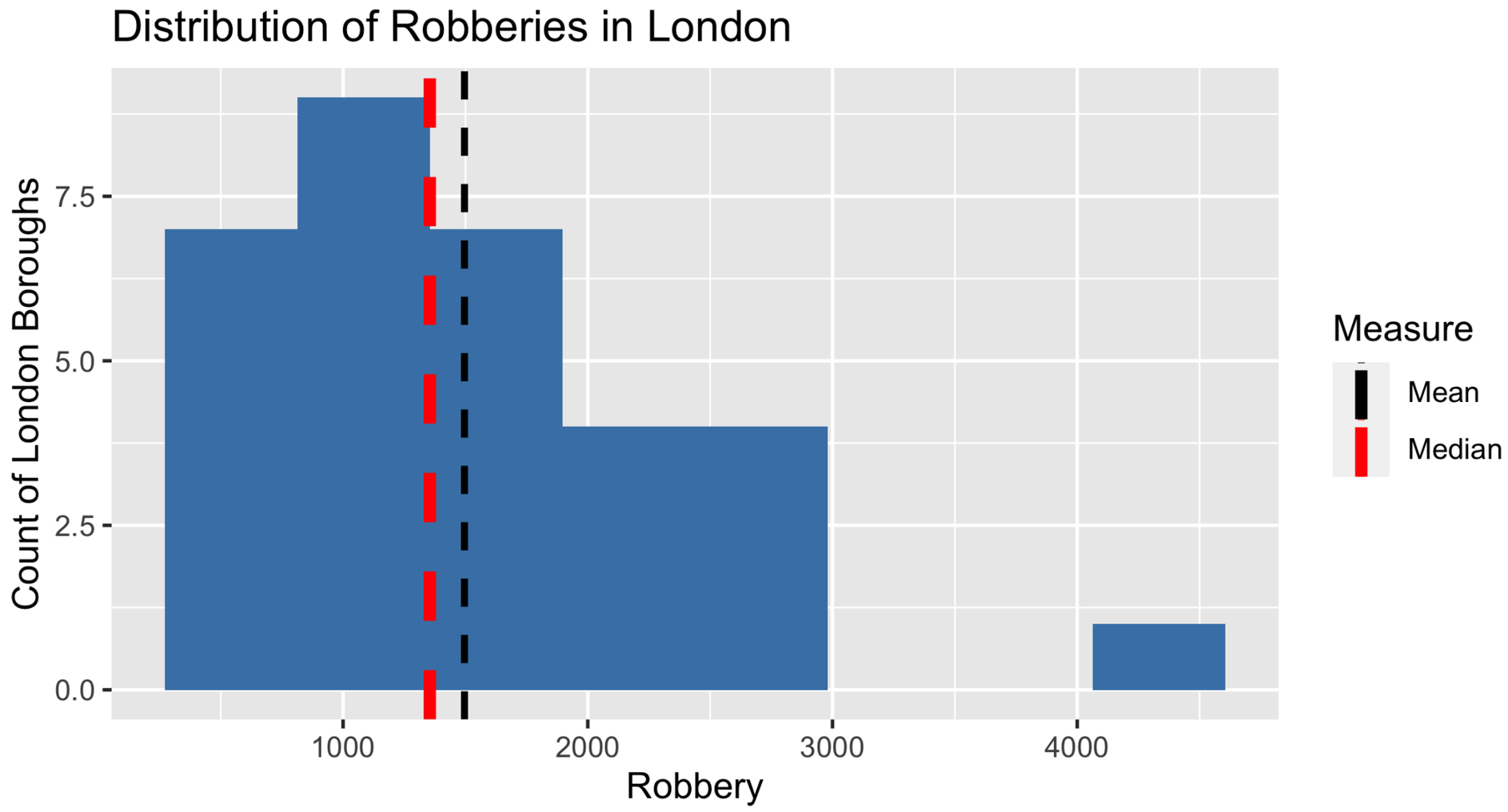
- Mode - *most frequent value*

Type of Crime	Count
Burglary	110,810
Robbery	47,877
Theft	350,025
Vehicle Offenses	205,337

Which measure to use?



Which measure to use?



Let's practice!

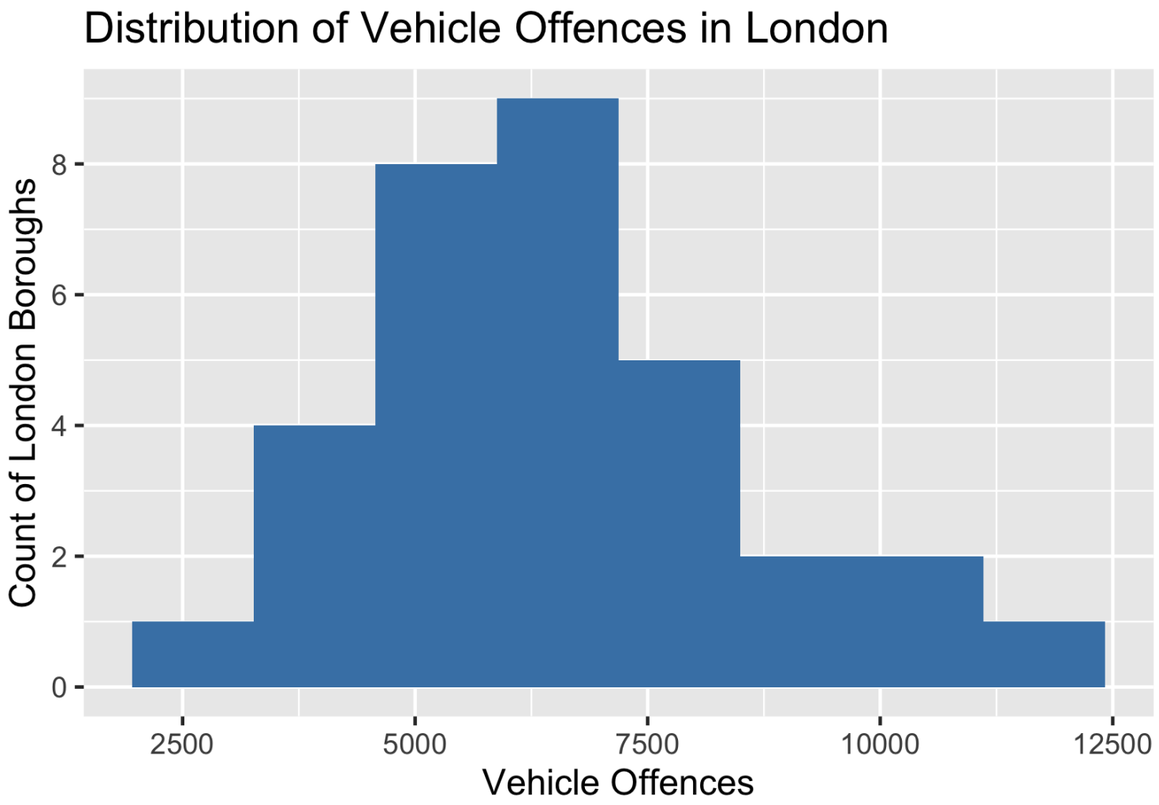
INTRODUCTION TO STATISTICS

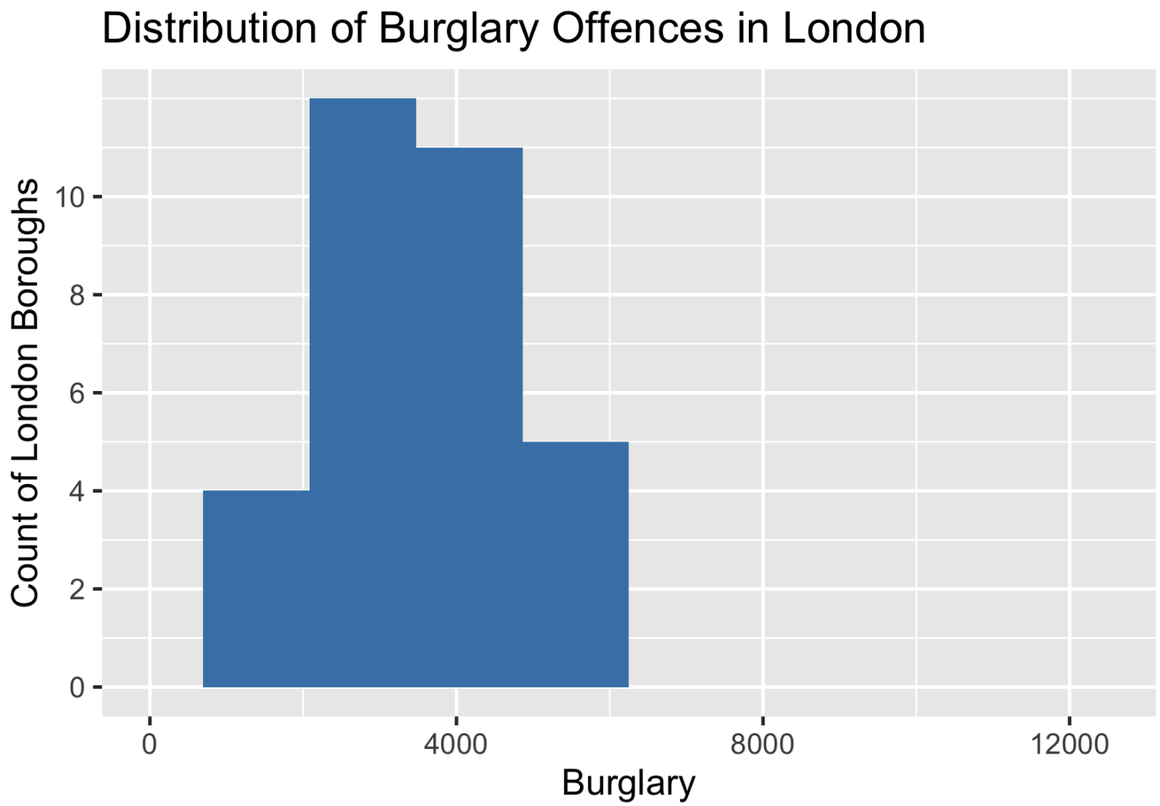
Measures of spread

INTRODUCTION TO STATISTICS

George Boorman
Curriculum Manager, DataCamp

What is spread?





Why is spread important?

- Spread measures the variety of our data

- T-shirts typically cost $30
  - Can cost between $10-200
  - How likely is it one will cost $30?
- If t-shirts were priced between $20-50
  - Does this change the likelihood of finding one for $30?

[1] Image credit: https://unsplash.com/@uyk

Range

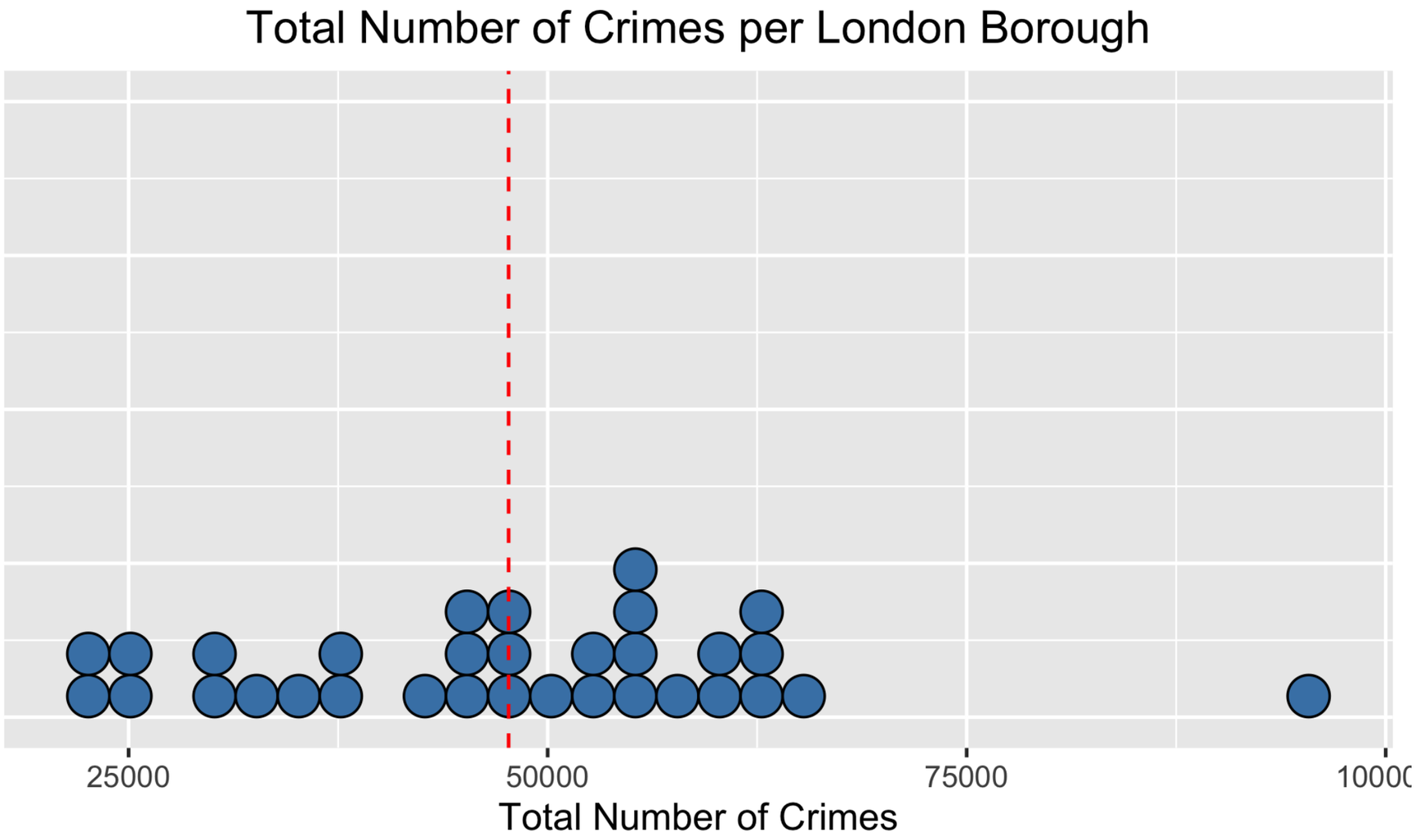
$$range = maximum - minimum$$

$$range(Burglaries) = 5,183 - 1,432$$

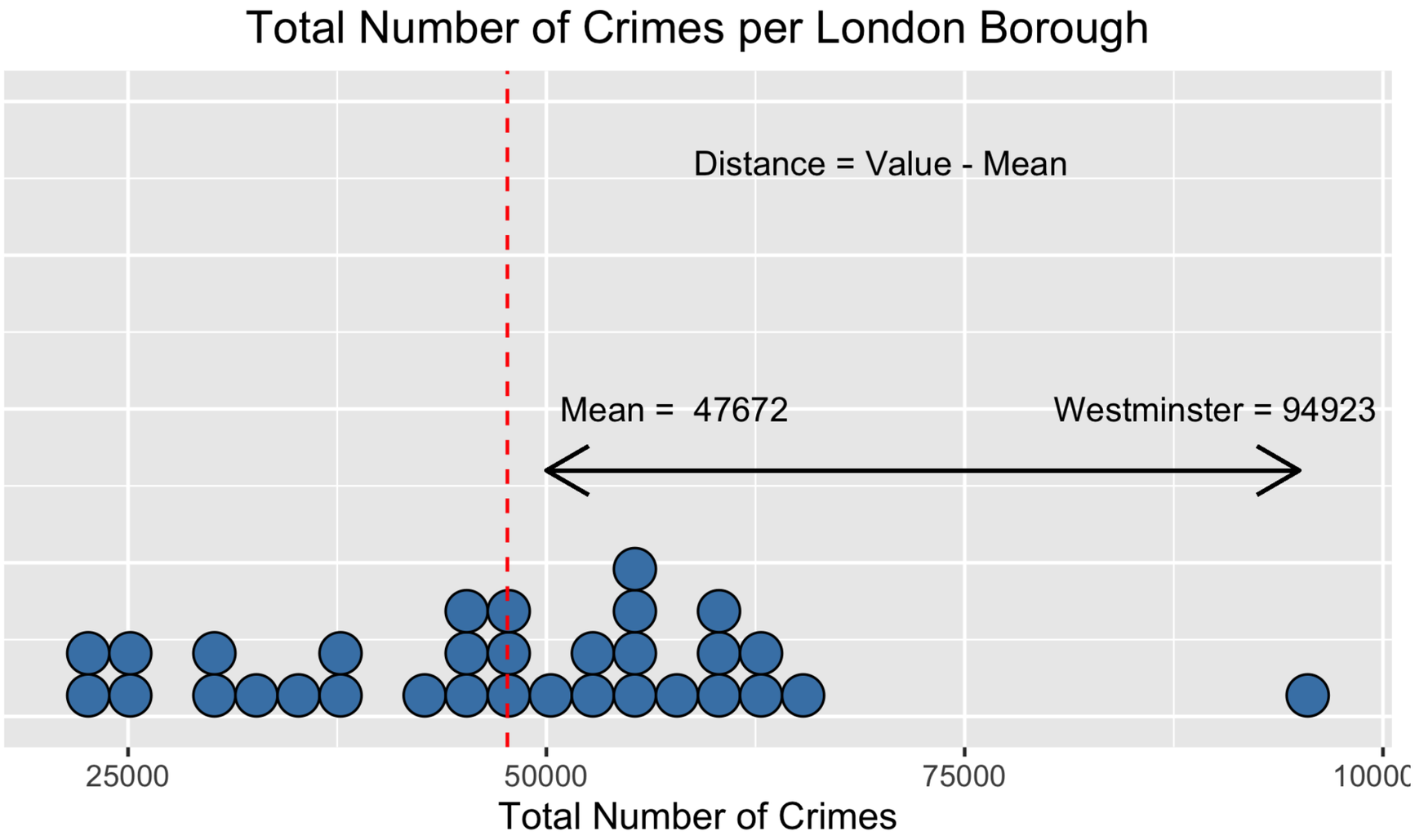
$$range(Burglaries) = 3,751$$

Borough	Burglary
Tower Hamlets	5,183
Hackney	5,079
Barnet	5,067
...	...
Sutton	1,815
Bexley	1,583
Kingston upon Thames	1,432

Variance


Total Number of Crimes per London Borough
25000
50000
75000
10000
Total Number of Crimes

Variance


Total Number of Crimes per London Borough
Distance = Value - Mean
Mean = 47672
Westminster = 94923
25000
50000
75000
10000
Total Number of Crimes

Variance

Borough	Total Crime	Mean	Distance
Barking and Dagenham	37,939	47,672	-9,733
Barnet	52,421	47,672	4,749
Bexley	29,285	47,672	-18,387
Brent	55,465	47,672	7,793
Bromley	42,982	47,672	-4,690
Camden	54,806	47,672	7,134
...	...	...	...
Total	**1,525,492**	**1,525,492**	**0**

Variance

Borough	Total Crime	Mean	Distance	Squared Distance
Barking and Dagenham	37,939	47,672	-9,733	94,731,289
Barnet	52,421	47,672	4,749	22,553,001
Bexley	29,285	47,672	-18,387	338,081,769
Brent	55,465	47,672	7,793	60,730,849
Bromley	42,982	47,672	-4,690	21,996,100
Camden	54,806	47,672	7,134	50,893,956
...	...	...	...	...
Total	**1,525,492**	**1,525,492**	**0**	**7,509,750,824**

Variance

$$variance(total\ crime) = \frac{7,509,750,824}{32}$$

$$variance(total\ crime) = 234,679,713$$

Standard deviation

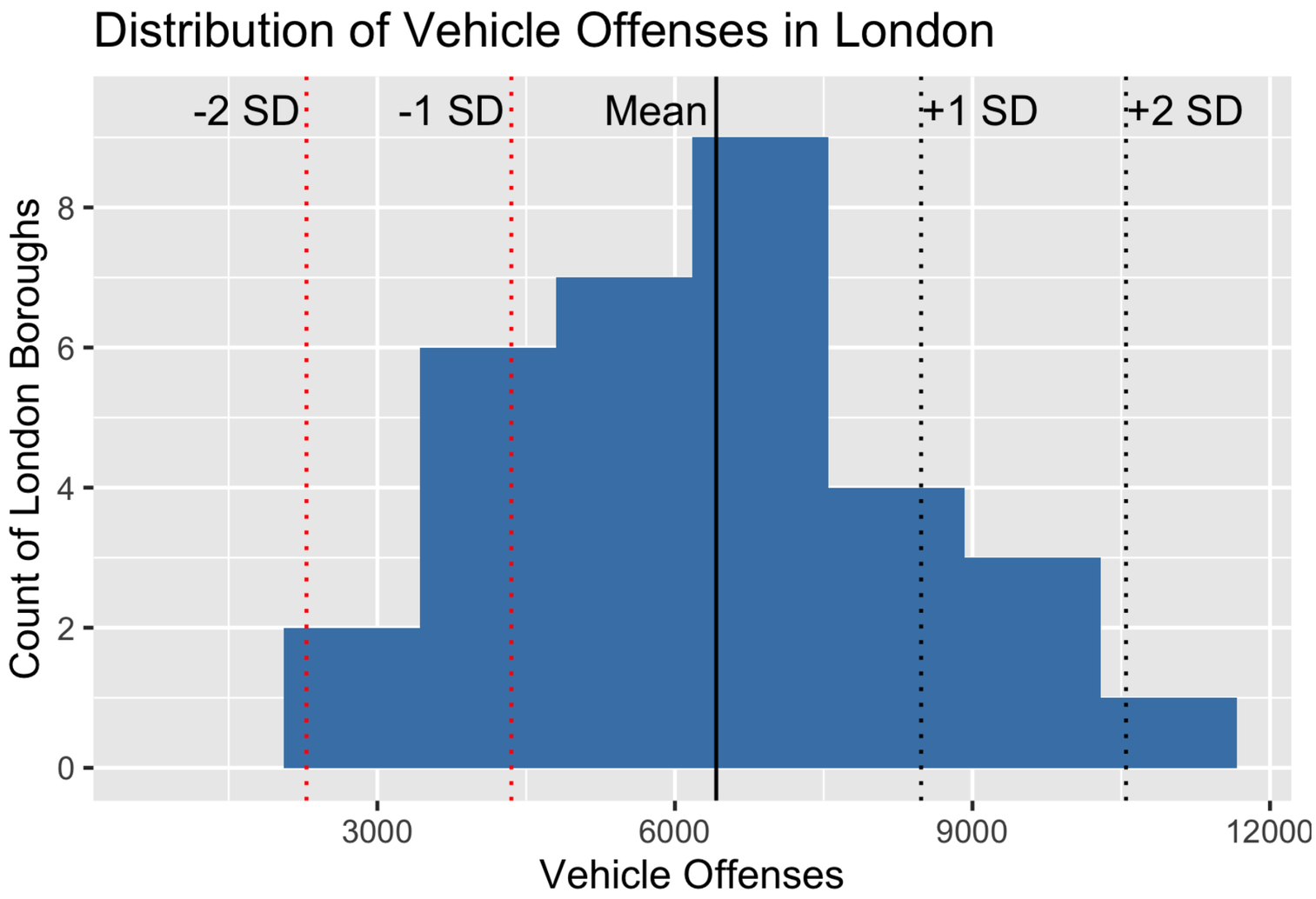
$$standard\ deviation(total\ crime) = \sqrt{(variance(total\ crime))}$$

$$standard\ deviation(total\ crime) = \sqrt{(234,679,713)}$$

$$standard\ deviation(total\ crime) = 15,319.26$$

- Standard deviation close to zero = data clustered around the mean

Standard deviation in a histogram



Quartiles

- Quartiles:
  - splitting the data into four equal parts

Crime	0%	25%	50%	75%	100%
Burglary	1,432.00	2,681.75	3,416.50	4,392.00	5,183.00
Robbery	363.00	895.75	1,354.50	1,976.50	4,156.00
Theft	4,090.00	7,739.75	9,624.00	12,059.00	40,278.00
Vehicle Offenses	2,143.00	4,838.25	6,424.50	7,520.75	11,292.00

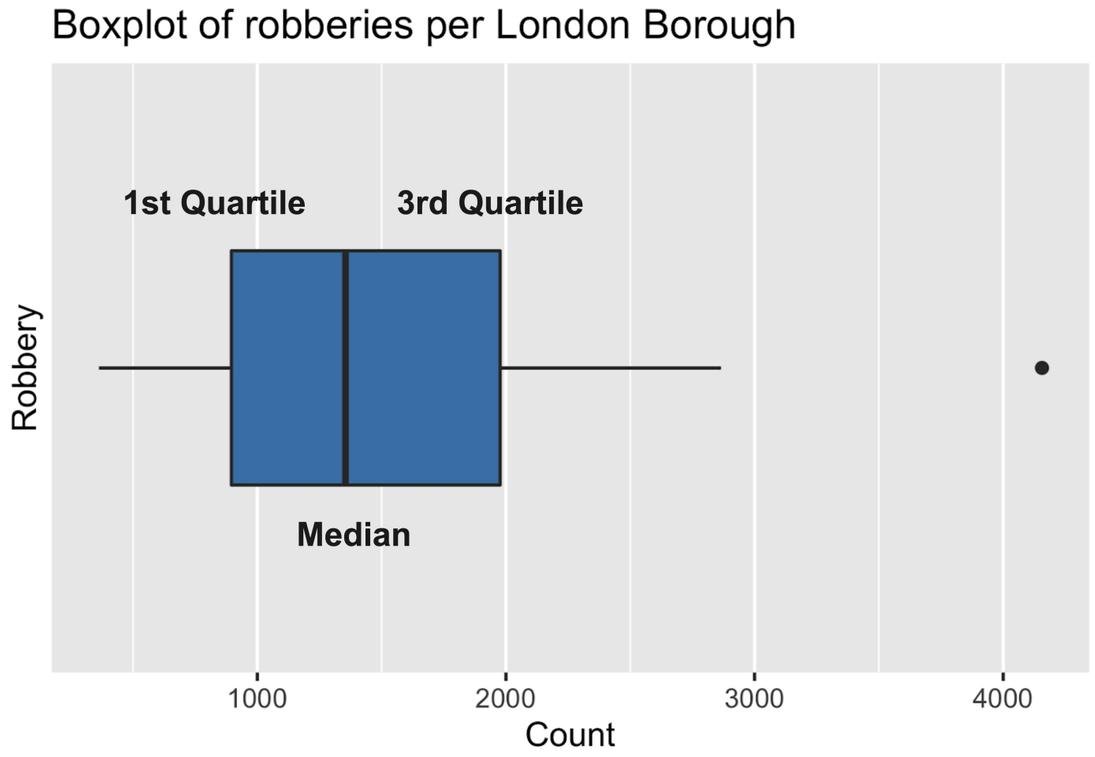
Quartiles

- Quartiles:
  - splitting the data into four equal parts

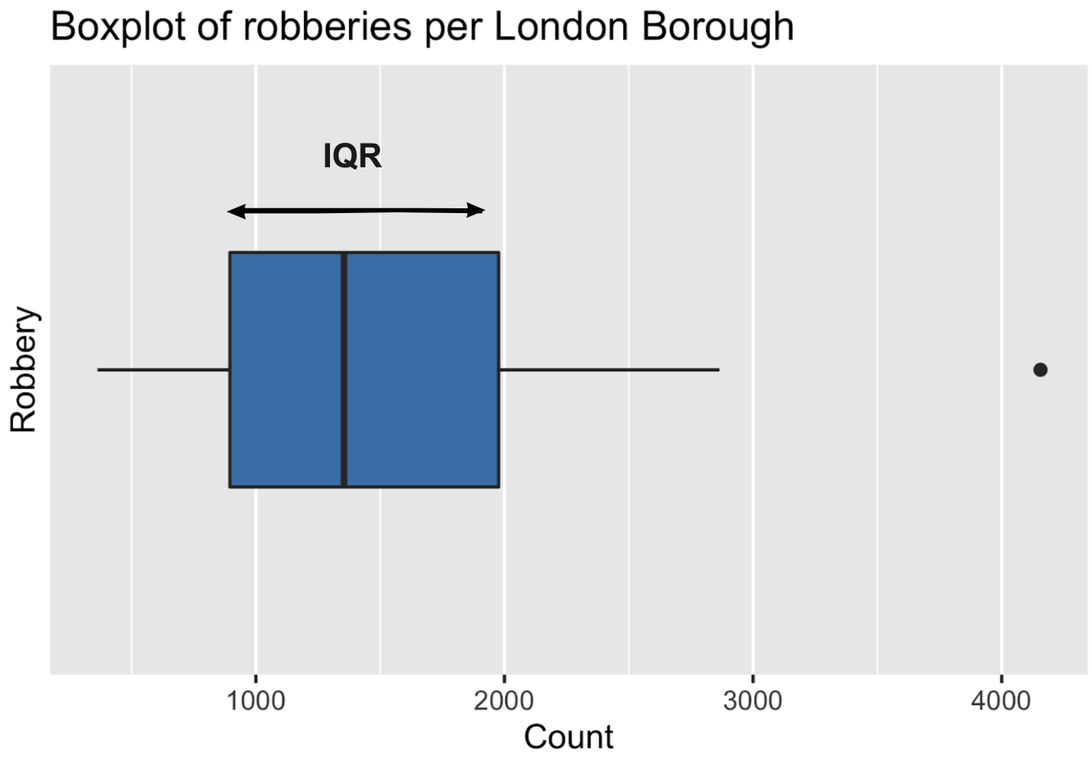
Crime	0%	25%	50%	75%	100%
Burglary	1,432.00	2,681.75	3,416.50	**4,392.00**	5,183.00
Robbery	363.00	895.75	1,354.50	1,976.50	4,156.00
Theft	4,090.00	7,739.75	9,624.00	12,059.00	40,278.00
Vehicle Offenses	2,143.00	4,838.25	6,424.50	7,520.75	11,292.00

- Second quartile (50%) = median

Box plots


Boxplot of robberies per London Borough
1st Quartile
3rd Quartile
Median
Robbery
1000
2000
3000
4000
Count

Interquartile range (IQR)



IQR = 3rd Quartile - 1st Quartile

IQR = 1976.50 - 895.75

IQR = 1080.75

- IQR is less affected by extreme values

Let's practice!

INTRODUCTION TO STATISTICS