Exploratory Data Analysis
Graphical Summaries
and Data Cleaning

Colorado Data Science Team

Sign up on codata.colorado.edu

Curious about data science, or ready to apply your stats, math, or machine learning skills? The Colorado Data Science Team is an incredible opportunity to get hands-on experience with real problems and real data. All are welcome, including students without a computer science or statistics background!

Join us for our first meeting:
Tuesday, September 5, 5 p.m. ECCR 245

Join us for

Fun

Fame

Pizza

$$$

Students can also enroll in our optional 1-credit companion course CSCI 4802 and CSCI 5802

Administrivia

- Homework 1 is posted. Due at 5pm on Friday Sept 15th. Don't wait to start!!
- Please sign up for Moodle ASAP using the following enrollment keys
  - Chris' Section: csci3022_F17_001
  - Dan's Section: csci3022_F17_002

Old Faithful Data

216	108	110	235	108	105	258	268	261	113	288	283	134	270	242	118	248
260	132	260	149	262	284	138	157	244	112	250	254	134	302	240	132	249
115	125	113	275	288	110	126	261	143	282	112	256	245	145	249	141	132
267	286	272	109	268	200	137	272	173	282	216	117	261	252	105	282	130
105	288	96	255	207	184	272	216	118	245	231	266	202	242	230	121	112
290	110	287	274	105	272	199	230	126	278	120	110	290	104	293	223	100
274	259	105	288	109	264	250	282	124	282	270	240	119	304	121	274	233
216	246	158	244	296	237	271	130	240	112	289	110	258	280	225	112	294
126	270	243	112	282	107	291	221	294	265	102	278	139	276	109	265	255
118	276	226	115	270	136	279	168	260	110	263	113	296	122	224	272	289
260	119	278	121	306	108	144	276	214	240	270	245	108	238	120	230	210
275	142	300	116	277	275	200	250	260	270	145	240	250	255	226	122	266
245	110	265	131	288	246	238	254	210	262	135	280	248	112	276	107	262
231	116	270	112	230	205	254	144	288	120	249	105	269	240	247	245	256
235	273	251	133	267	113	111	257	237	140	296	174	275	230	125	262	128
261	214	270	249	229	235	267	120	257	111	255	119	135	285	247	129	265

Old Faithful Data

96	100	102	104	105	105	105	105	105	105	107	107	108	108	108	108	109
109	109	110	110	110	110	110	110	110	111	111	112	112	112	112	112	112
112	112	113	113	113	113	115	115	116	116	117	118	118	118	119	119	119
120	120	120	120	121	121	121	122	122	124	125	125	126	126	126	128	129
130	130	131	132	132	132	133	134	134	135	135	136	137	138	139	140	141
142	143	144	144	145	145	149	157	158	168	173	174	184	199	200	200	202
205	207	210	210	214	214	216	216	216	216	221	223	224	225	226	226	229
230	230	230	230	230	231	231	233	235	235	235	237	237	238	238	240	240
240	240	240	240	242	242	243	244	244	245	245	245	245	245	246	246	247
247	248	248	249	249	249	249	250	250	250	250	251	252	254	254	254	255
255	255	255	256	256	257	257	258	258	259	260	260	260	260	260	261	261
261	261	262	262	262	262	263	264	265	265	265	265	266	266	267	267	267
268	268	269	270	270	270	270	270	270	270	270	271	272	272	272	272	272
273	274	274	274	275	275	275	275	276	276	276	276	277	278	278	278	279
280	280	282	282	282	282	282	282	283	284	285	286	287	288	288	288	288
288	288	289	289	290	290	291	293	294	294	296	296	296	300	302	304	306

Old Faithful Data

96	100	102	104	105	105	105	105	105	105	107	107	108	108	108	108	109
109	109	110	110	110	110	110	110	110	111	111	112	112	112	112	112	112
112	112	113	113	113	113	115	115	116	116	117	118	118	118	119	119	119
120	120	120	120	121	121	121	122	122	124	125	125	126	126	126	128	129
130	130	131	132	132	132	133	134	134	135	135	136	137	138	139	140	141
142	143	144	144	145	145	149	157	158	168	173	174	184	199	200	200	202
205	207	210	210	214	214	216	216	216	216	221	223	224	225	226	226	229
230	230	230	230	230	231	231	233	235	235	235	237	237	238	238	240	240
240	240	240	240	242	242	243	244	244	245	245	245	245	245	246	246	247
247	248	248	249	249	249	249	250	250	250	250	251	252	254	254	254	255
255	255	255	256	256	257	257	258	258	259	260	260	260	260	260	261	261
261	261	262	262	262	262	263	264	265	265	265	265	266	266	267	267	267
268	268	269	270	270	270	270	270	270	270	270	271	272	272	272	272	272
273	274	274	274	275	275	275	275	276	276	276	276	277	278	278	278	279
280	280	282	282	282	282	282	282	283	284	285	286	287	288	288	288	288
288	288	289	289	290	290	291	293	294	294	296	296	296	300	302	304	306

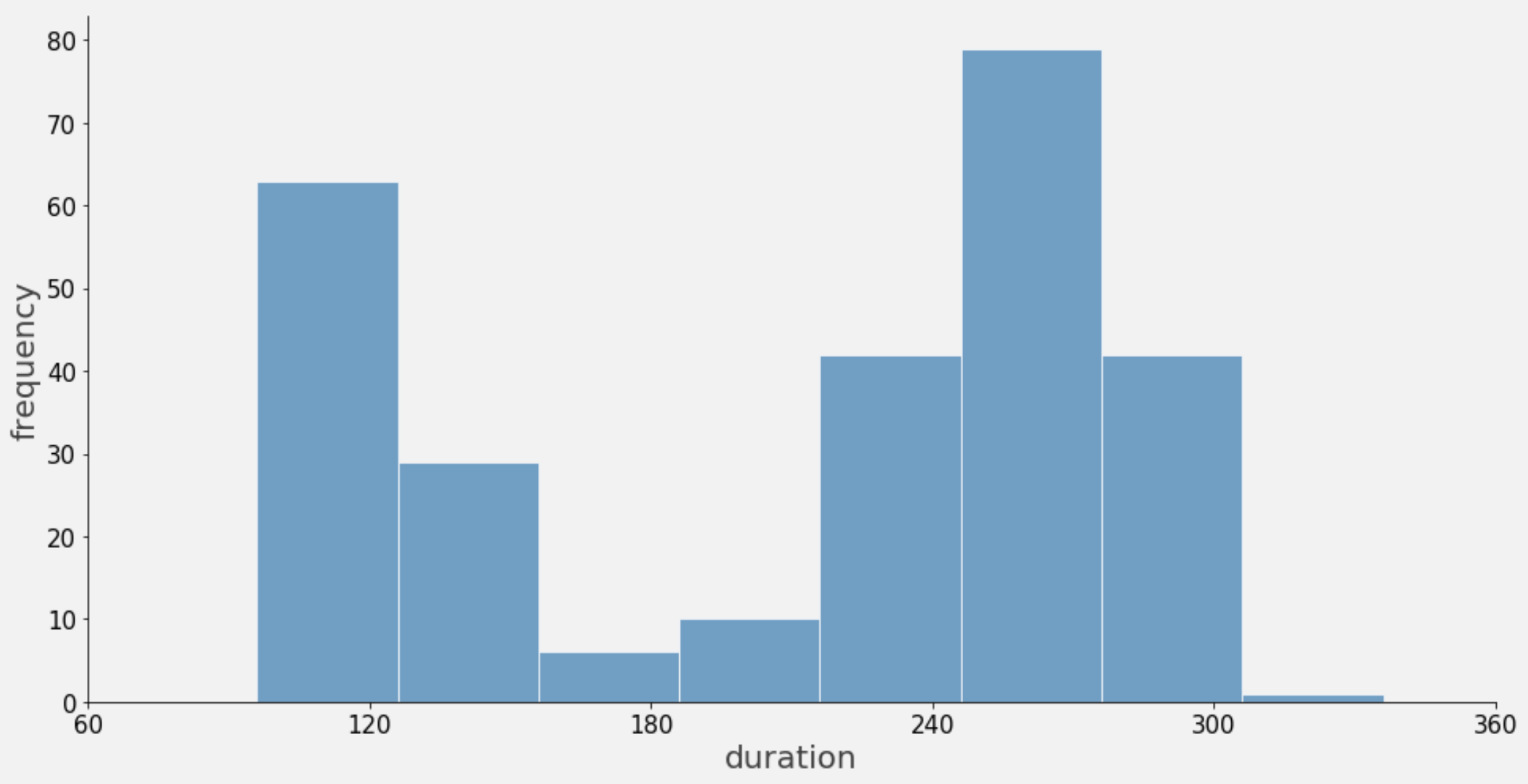
$\bar{x} = 209.3$

Old Faithful Data

96	100	102	104	105	105	105	105	105	105	107	107	108	108	108	108	109
109	109	110	110	110	110	110	110	110	111	111	112	112	112	112	112	112
112	112	113	113	113	113	115	115	116	116	117	118	118	118	119	119	119
120	120	120	120	121	121	121	122	122	124	125	125	126	126	126	128	129
130	130	131	132	132	132	133	134	134	135	135	136	137	138	139	140	141
142	143	144	144	145	145	149	157	158	168	173	174	184	199	200	200	202
205	207	210	210	214	214	216	216	216	216	221	223	224	225	226	226	229
230	230	230	230	230	231	231	233	235	235	235	237	237	238	238	240	240
240	240	240	240	242	242	243	244	244	245	245	245	245	245	246	246	247
247	248	248	249	249	249	249	250	250	250	250	251	252	254	254	254	255
255	255	255	256	256	257	257	258	258	259	260	260	260	260	260	261	261
261	261	262	262	262	262	263	264	265	265	265	265	266	266	267	267	267
268	268	269	270	270	270	270	270	270	270	270	271	272	272	272	272	272
273	274	274	274	275	275	275	275	276	276	276	276	277	278	278	278	279
280	280	282	282	282	282	282	282	283	284	285	286	287	288	288	288	288
288	288	289	289	290	290	291	293	294	294	296	296	296	300	302	304	306

$\bar{x} = 209.3 \quad Q_1 = 129.75 \quad Q_2 = 240 \quad Q_3 = 267.25$

Old Faithful Data

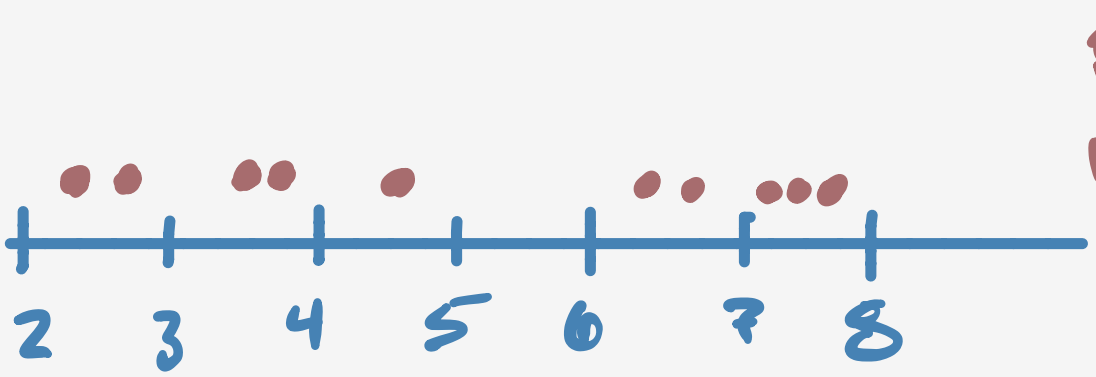


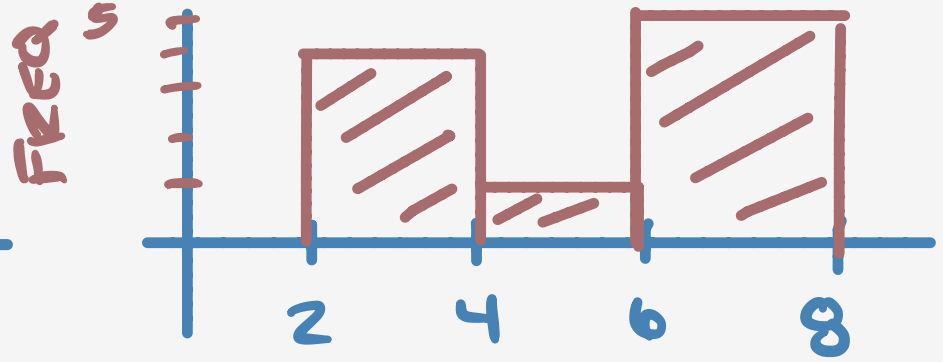
The Histogram

The **histogram** is a graphical representation of the distribution of numerical data

To construct a histogram:

- Lump or "Bin" the observed values of the VoI (bins are typically consecutive, non-overlapping, and equal in length)
- For a **Frequency Histogram**: count the number of values that fall into a bin and draw a rectangle over the bin with height equal to the count



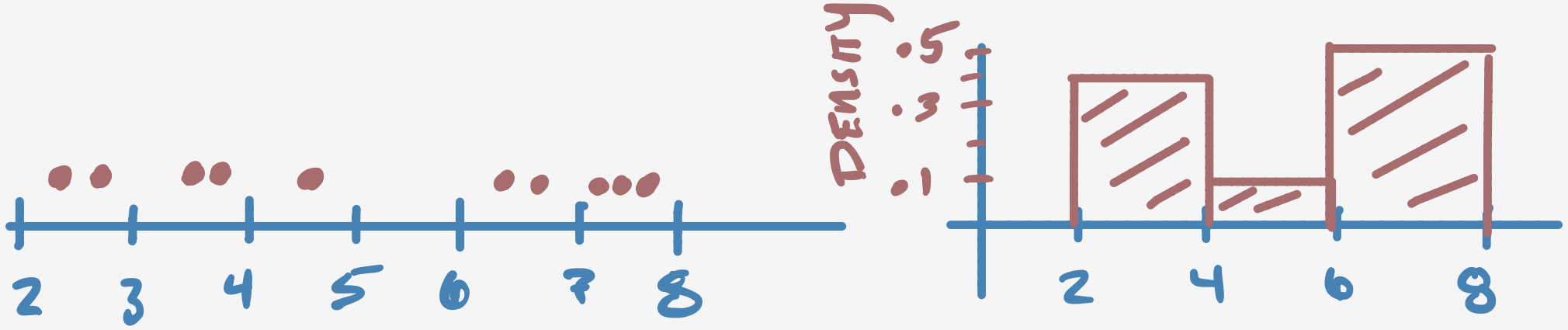


The Histogram

The **histogram** is a graphical representation of the distribution of numerical data

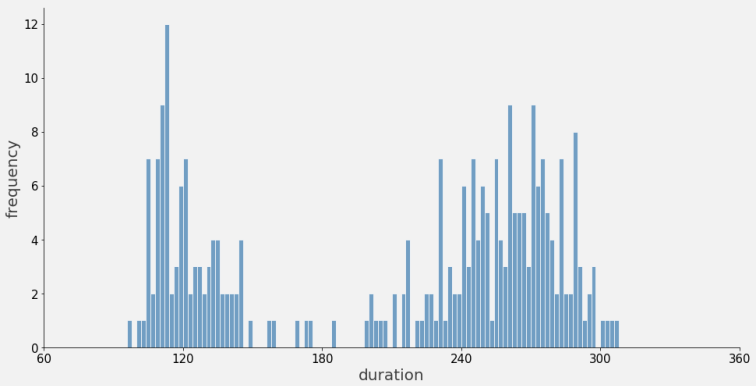
To construct a histogram:

- Lump or "Bin" the observed values of the VoI (bins are typically consecutive, non-overlapping, and equal in length)
- For a **Density Histogram**: count the number of values that fall into a bin and adjust the height such that the sum of the area of all bins is equal to 1

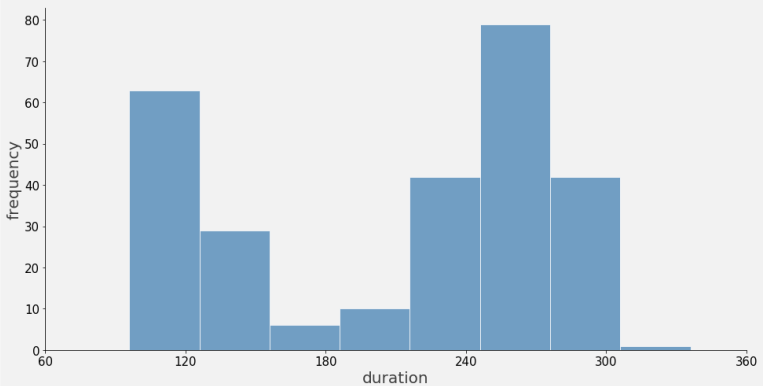


The Histogram

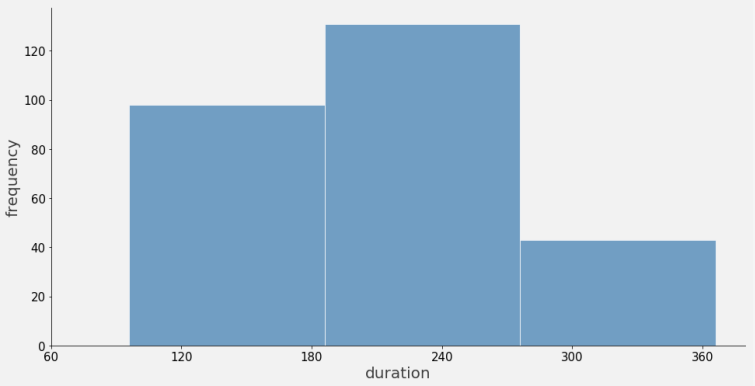
Note that choosing a different bin width can paint a very different picture



bin size 2



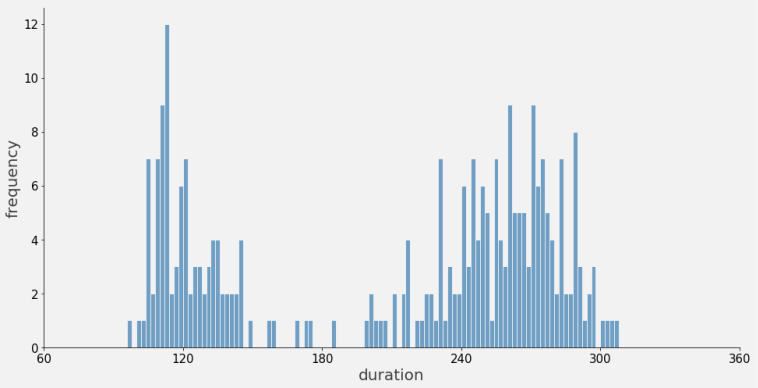
bin size 30



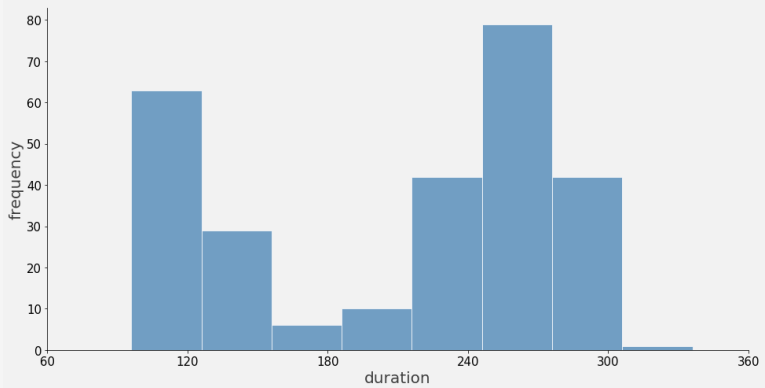
bin size 90

The Histogram

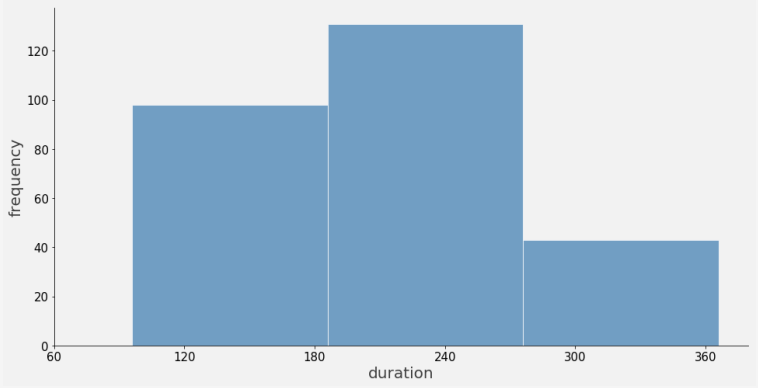
Note that choosing a different bin width can paint a very different picture



bin size 2



bin size 30

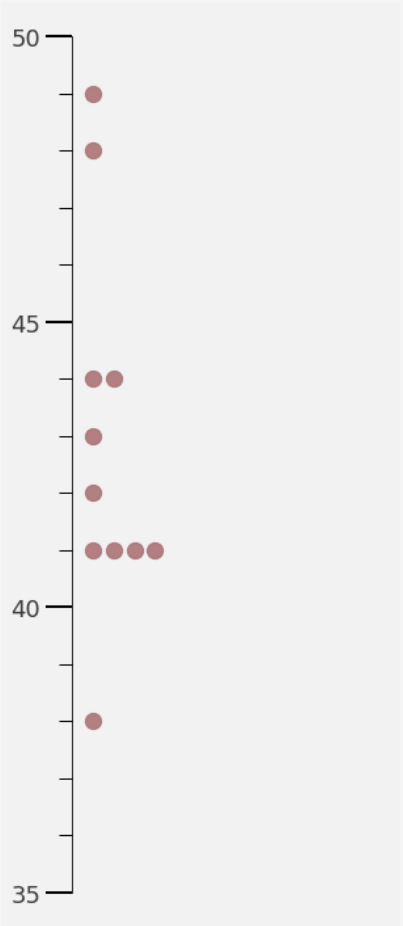


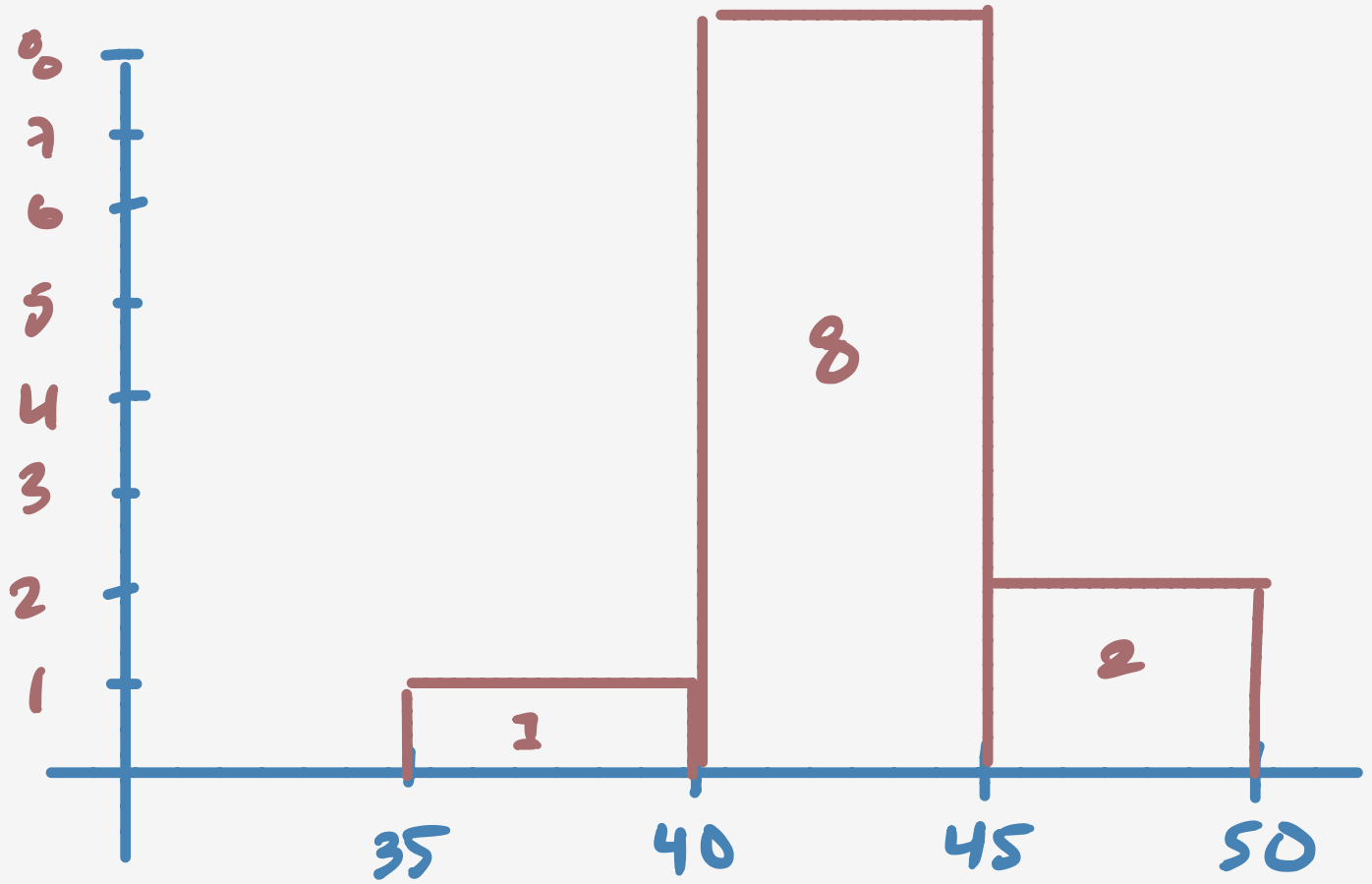
bin size 90

Common Choice: Freedman–Diaconis Rule, Bin Size $= 2\,\frac{IQR}{n^{1/3}} = 2\,\frac{Q_3 - Q_1}{n^{1/3}}$

The Histogram

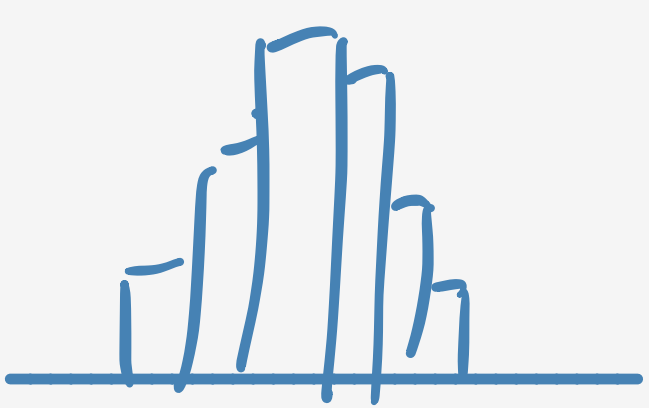
Example: Compute Frequency and Density histograms with Bin Width 5 of data on left



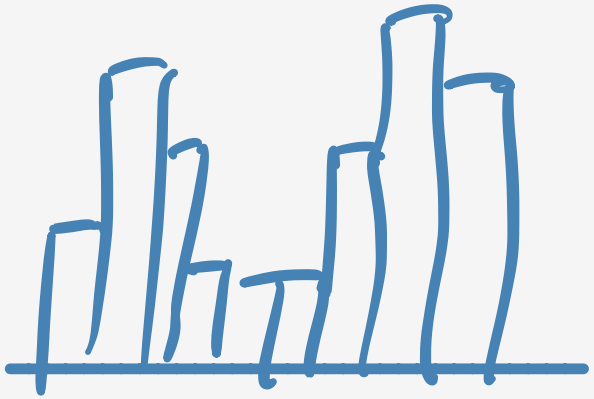


The Histogram

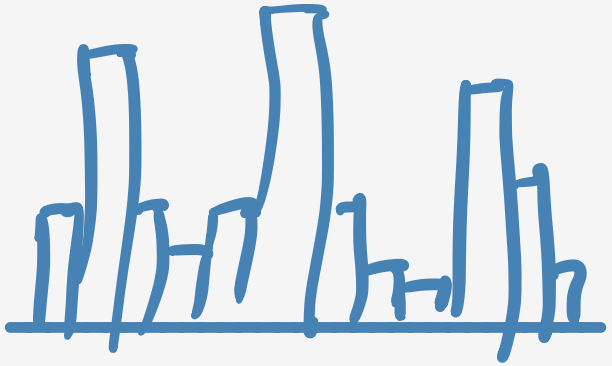
Histograms come in a variety of shapes



unimodal



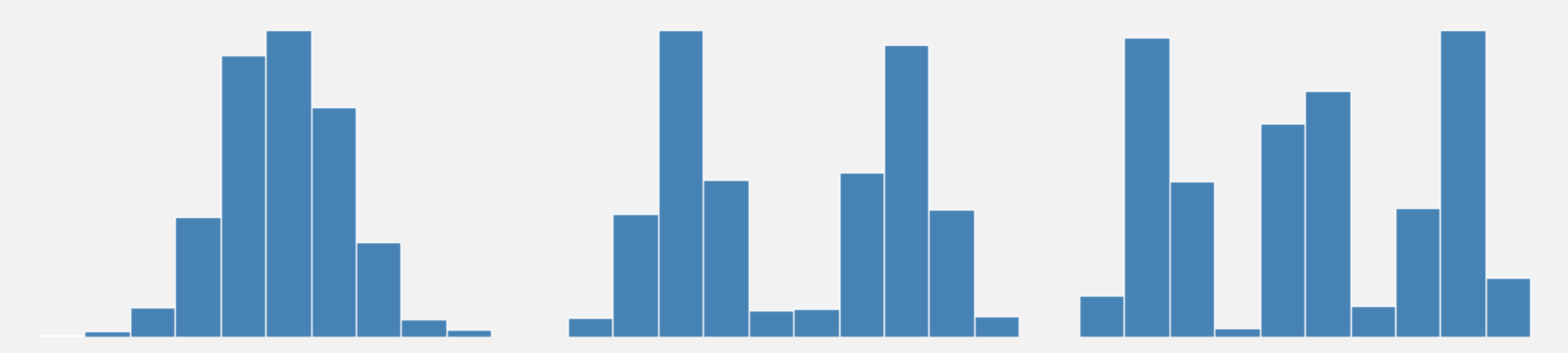
bimodal



multimodal

The Histogram

Histograms come in a variety of shapes



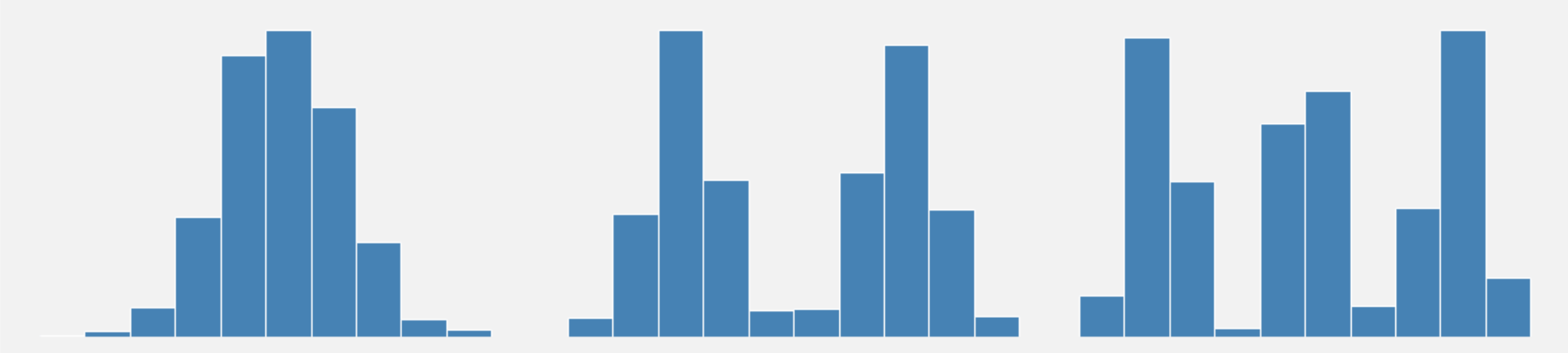
unimodal

bimodal

multimodal

The Histogram

Histograms come in a variety of shapes



unimodal

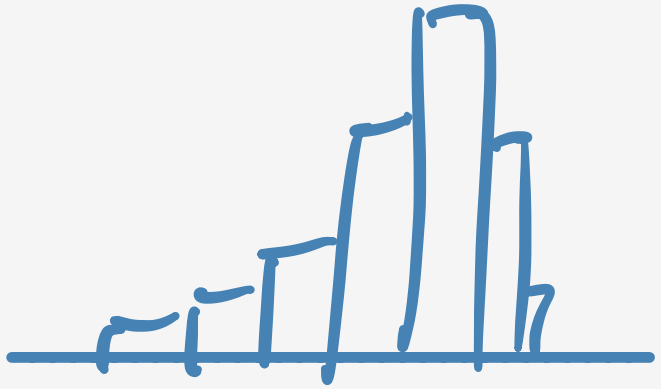
bimodal

multimodal

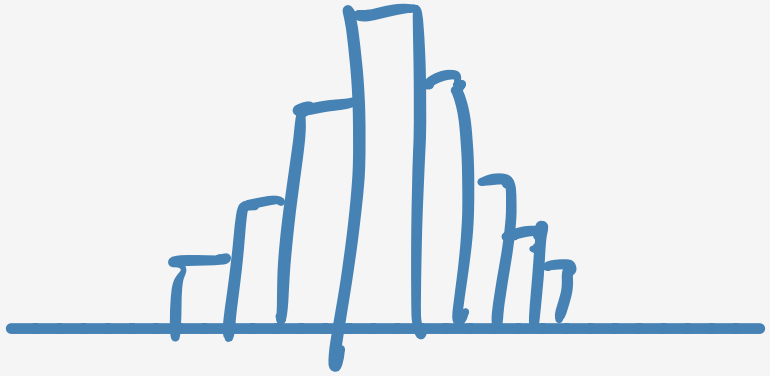
Question: What can you say about the data if histogram is bimodal?

The Histogram

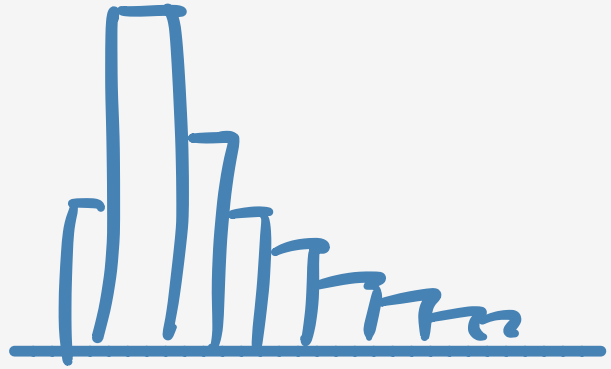
Histograms come in a variety of shapes



negative skew



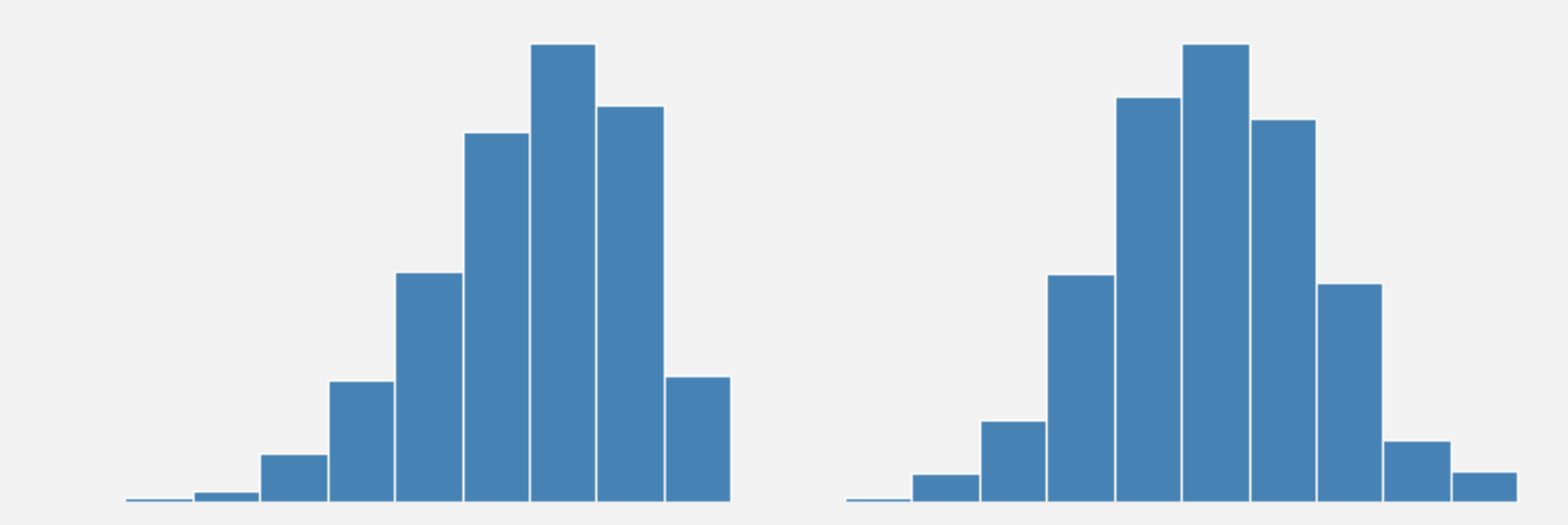
symmetric



positive skew

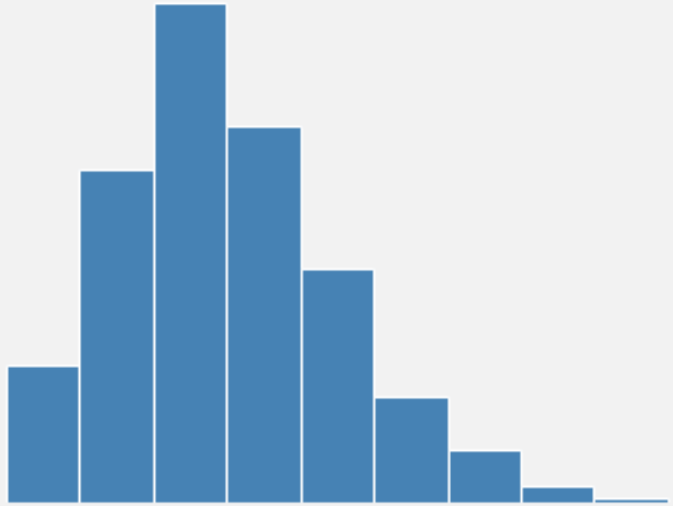
The Histogram

Histograms come in a variety of shapes



negative skew

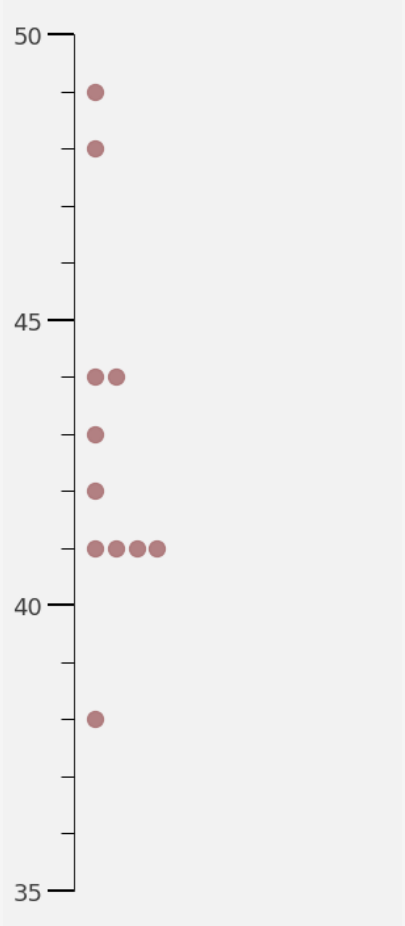
symmetric



positive skew

Quartile Refresher

Example: Compute the Quartiles of the data on the left



38, 41, 41, 41, 41, 42, 43, 44, 44, 48, 49

$n = 11$

$Q_2 = 42$

38, 41, 41, 41, 41, 42

42, 43, 44, 44, 48, 49

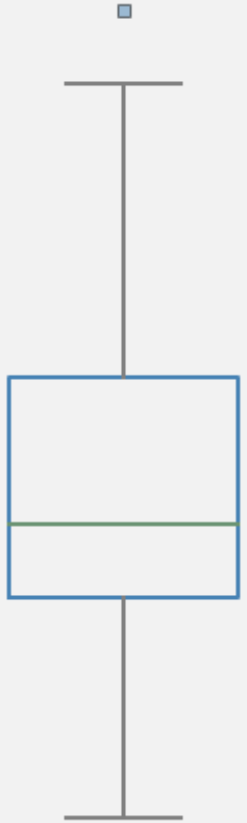
$Q_1 = \frac{41+41}{2} = 41$

$Q_3 = \frac{44+44}{2} = 44$

$IQR = Q_3 - Q_1 = 44 - 41 = 3$

The Box-and-Whisker Plot

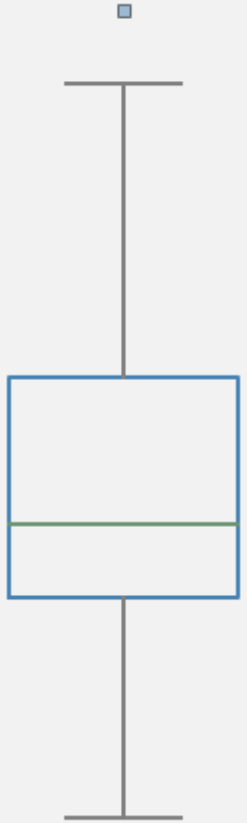
Box-and-Whisker plots are convenient ways of visualizing data through Quartiles



- The **Box** extends from Q_1 **to** Q_3
- The **Median Line** goes through median $\widetilde{x}$
- The **Whiskers** extend to farthest point within $1.5 \times IQR$ **of quartile**
- The **Fliers** or outliers are any points outside of whiskers
- The width of the box is **unimportant**

The Box-and-Whisker Plot

Box-and-Whisker plots are convenient ways of visualizing data through Quartiles



Box-and-Whisker plots are good because they

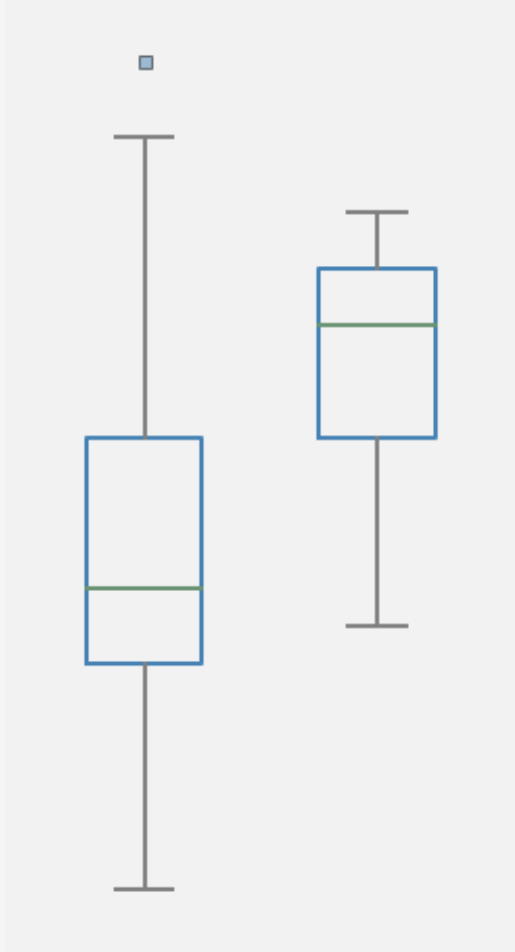
- Depict the center of the data
- Depict the range and IQR
- Depict symmetry / skewness
- Show likely outliers

When might a Box-and-Whisker plot be **misleading**?

Box-and-Whisker plots are particularly good at …

The Box-and-Whisker Plot

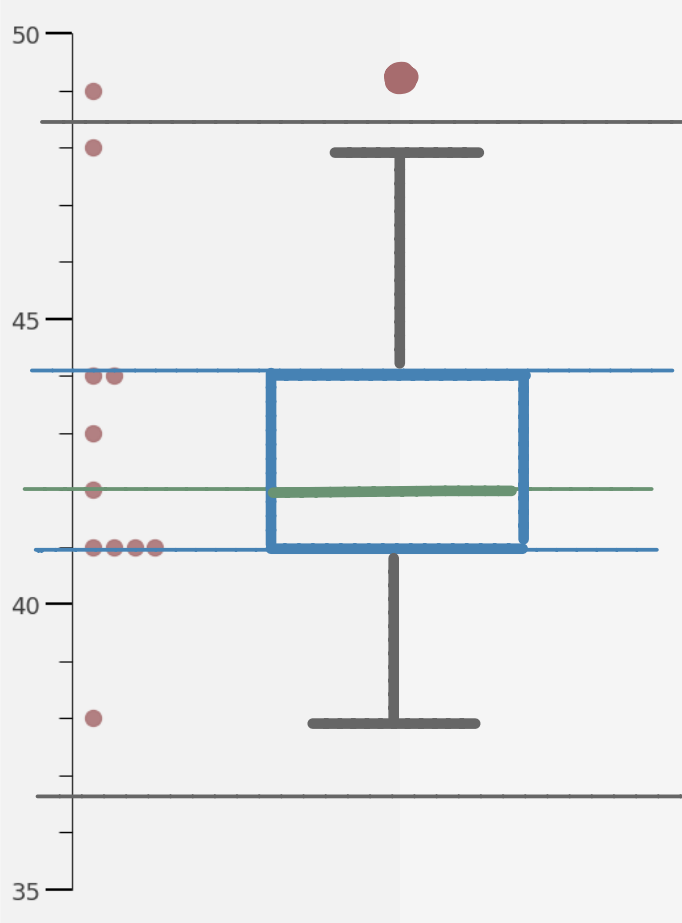
Box-and-Whisker plots are convenient ways of visualizing data through Quartiles



Box-and-Whisker plots are particularly good at **comparing data**

The Box-and-Whisker Plot

Example: Draw the Box-and-Whisker plot for the data on the left



Q_3

Q_1

$Q_3 + 1.5 \times IQR = 48.5$

$\tilde{x} = Q_2 = 42$

$Q_1 - 1.5 \times IQR = 36.5$

$Q_1 = 41$

$Q_2 = 42$

$Q_3 = 44$

$IQR = 3$

$1.5 \times IQR = 4.5$

Cleaning and Wrangling Data

Example: Dirty Titanic Data. What looks *wrong* to you?

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0.0	3.0	Braund, Mr. Owen Harris	male	22yrs	1	0	A/5 21171	£7.5s	NaN	S
1	2	1.0	1.0	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38yrs	1	0	PC 17599	£71.5s	C85	C
2	3	1.0	3.0	Heikkinen, Miss. Laina	female	26yrs	0	0	STON/O2. 3101282	£7.18s	NaN	S
3	4	1.0	1.0	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35yrs	1	0	113803	£53.2s	C123	S
4	5	0.0	3.0	Allen, Mr. William Henry	male	35yrs	0	0	373450	£8.1s	NaN	S

Cleaning and Wrangling Data

Example: Dirty Titanic Data. What looks *wrong* to you?

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0.0	3.0	Braund, Mr. Owen Harris	male	22yrs	1	0	A/5 21171	£7.5s	NaN	S
1	2	1.0	1.0	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38yrs	1	0	PC 17599	£71.5s	C85	C
2	3	1.0	3.0	Heikkinen, Miss. Laina	female	26yrs	0	0	STON/O2. 3101282	£7.18s	NaN	S
3	4	1.0	1.0	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35yrs	1	0	113803	£53.2s	C123	S
4	5	0.0	3.0	Allen, Mr. William Henry	male	35yrs	0	0	373450	£8.1s	NaN	S

Today's In-Class Notebook:

- Remove rows and columns with too many missing values
- Derive new columns from values of other columns using apply( ) and custom functions
- Replace string values with associated numerical values