IMD0033 - Probabilidade

Lesson 22 - Comparing Frequency Distributions

Ivanovitch Silva

November, 2018

Agenda

- Grouped bar plots
- Comparing histograms
- Kernel density estimates
- Strip plots
- Box plots
- Outliers

Atualizar o repositório

```
git clone https://github.com/ivanovitchm/imd0033_2018_2.git
```

Ou ....

```
git pull
```

PREVIOUSLY ON...

Data **source**

Getting **good** data

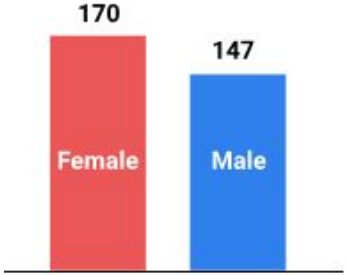
Id	Name	Salary	...	Gender
1	Mary Ann	$35 000	....	Female
2	Marc Downey	$55 000	....	Male
..	....	....	....	....
51	Juliet Ali	$45 000	....	Female
...	...	....	....	....
317	Jane Ace	$95 000	....	Female

Understand how the data is **structured** and **measured**

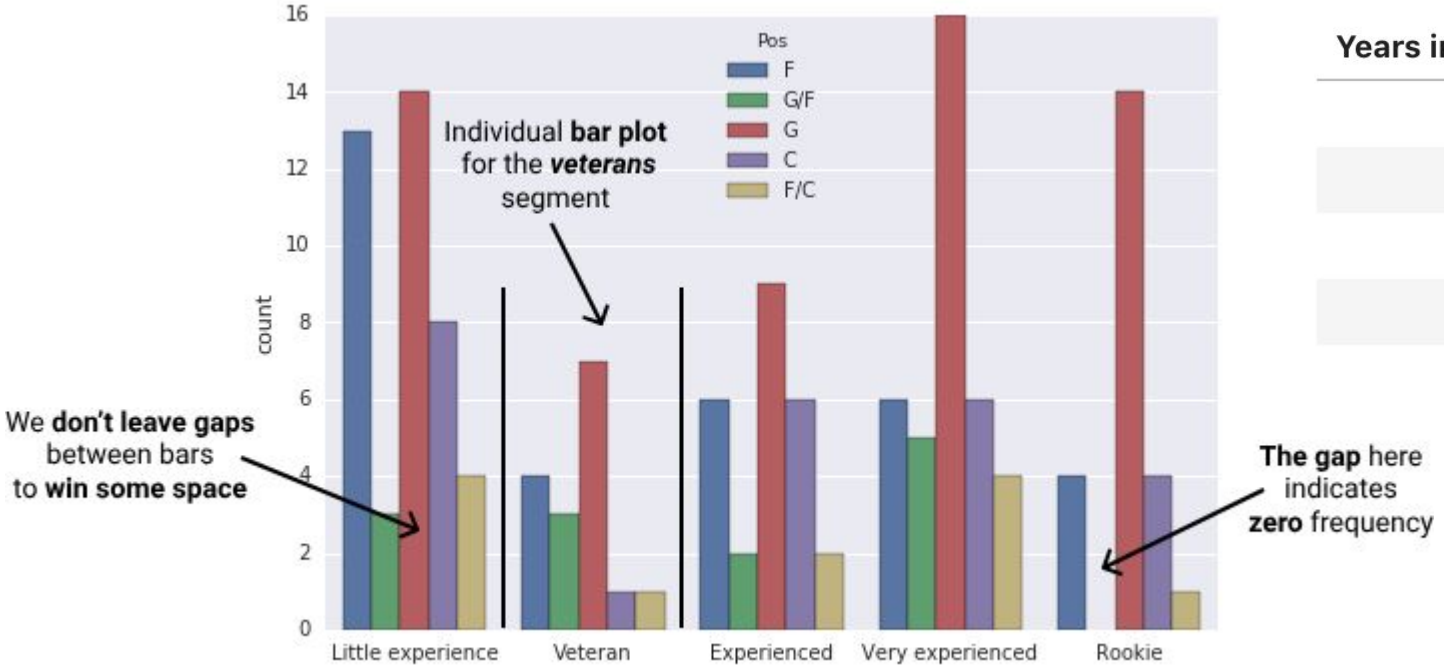
Organize the data in **comprehensible** forms to find patterns

Gender	Frequency
Male	147
Female	170

Visualize the patterns

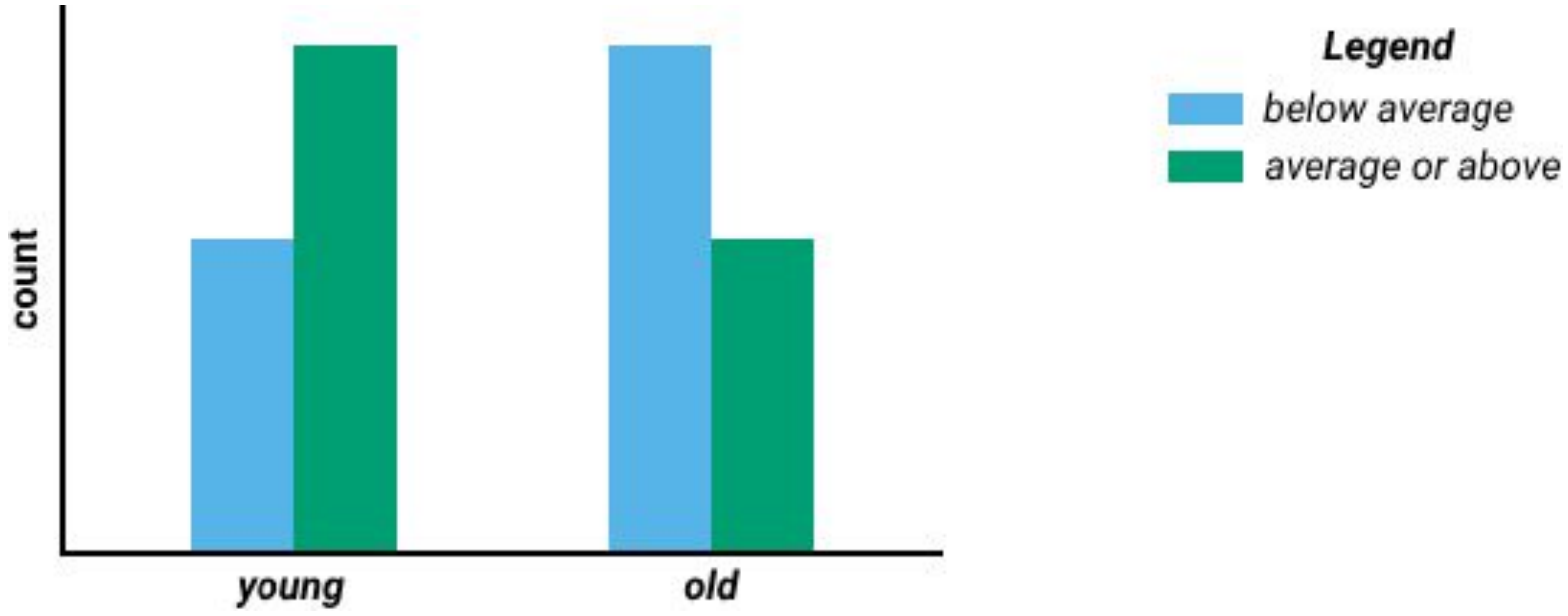


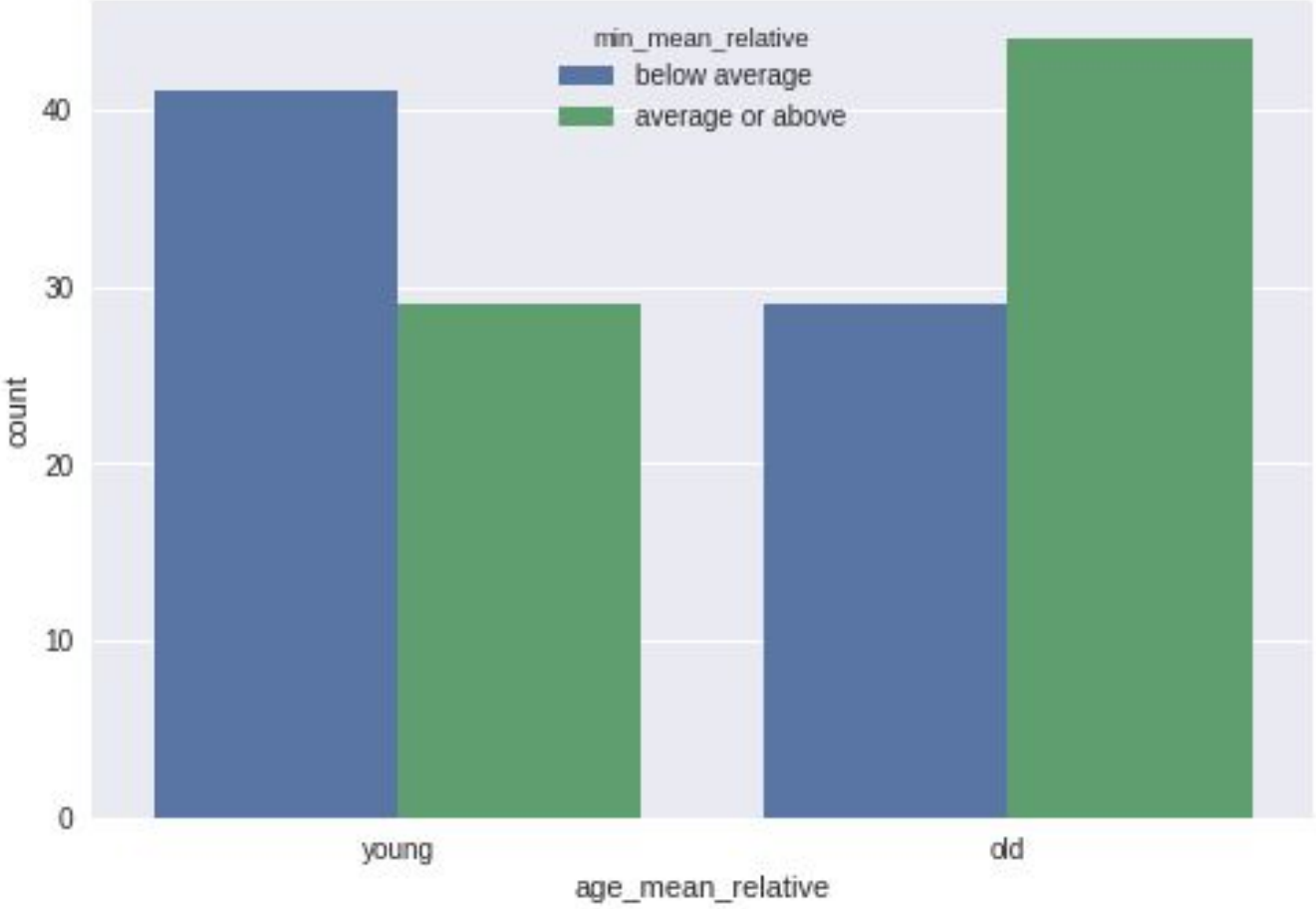
Comparing Frequency Distribution



Years in WNBA	Label
0	Rookie
1-3	Little experience
4-5	Experienced
5-10	Very experienced
>10	Veteran

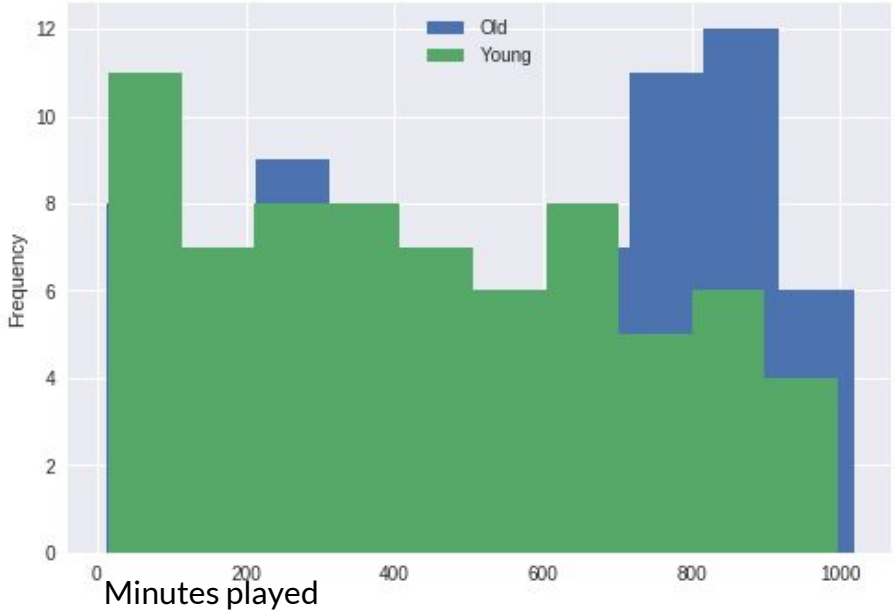
Challenge: Do older players play less?



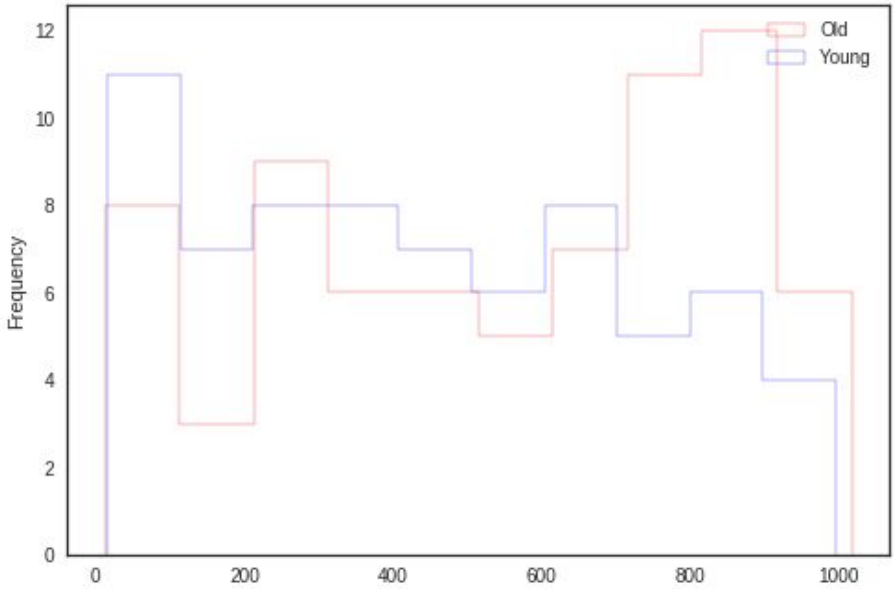



```
sns.countplot(x = 'age_mean_relative', hue = 'min_mean_relative', data = wnba)
```

Comparing Histograms

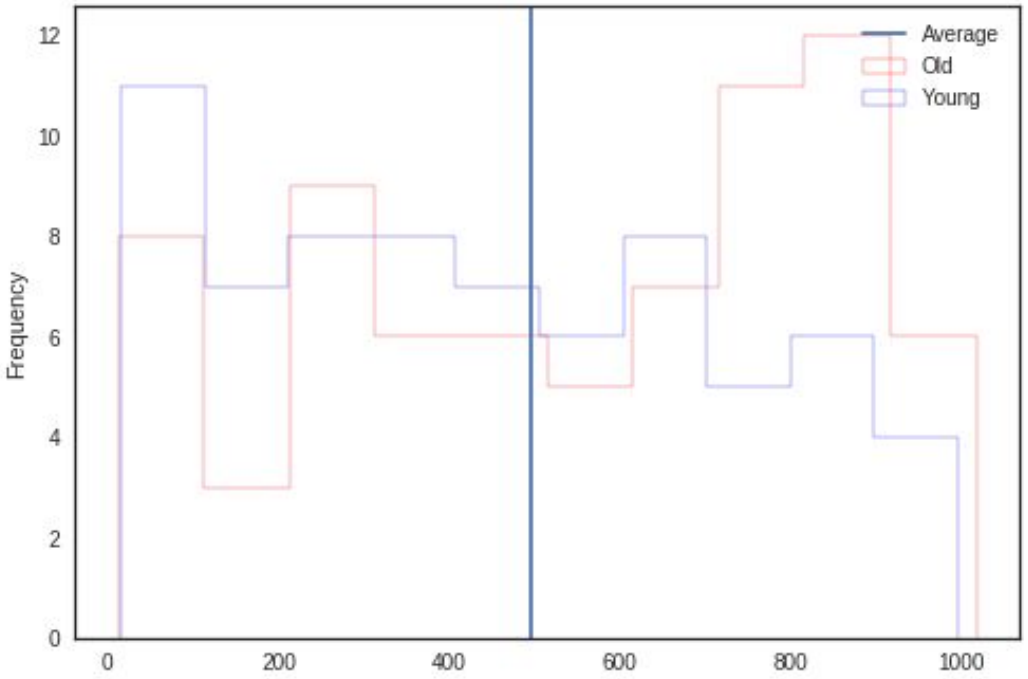



```
wnba[wnba.Age >= 27]['MIN'].plot.hist(label = 'Old', legend = True)
wnba[wnba.Age < 27]['MIN'].plot.hist(label = 'Young', legend = True)
```



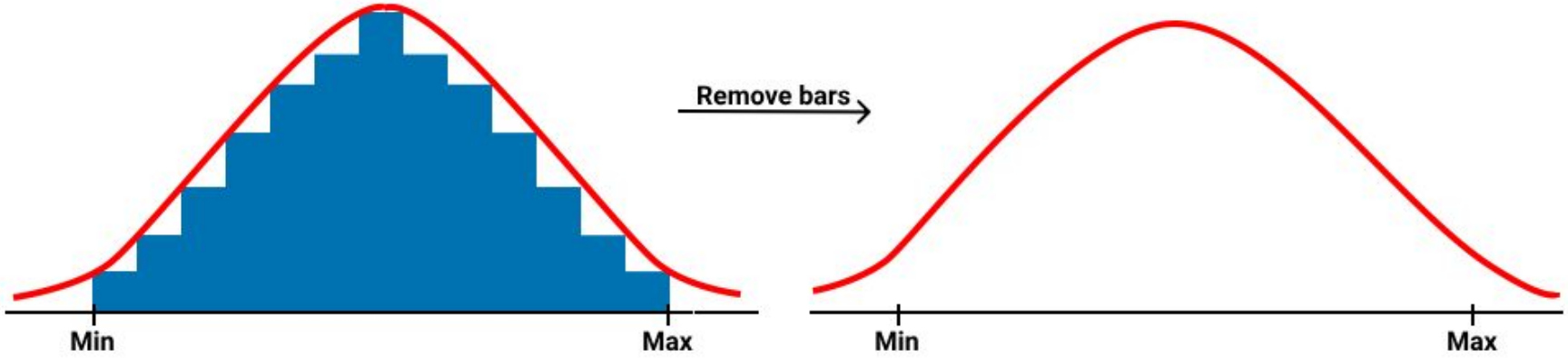

```
sns.set_style("white")
wnba[wnba.Age >= 27]['MIN'].plot.hist(histtype = 'step',
                                      label = 'Old',
                                      legend = True,color="red")
```

Comparing Histograms

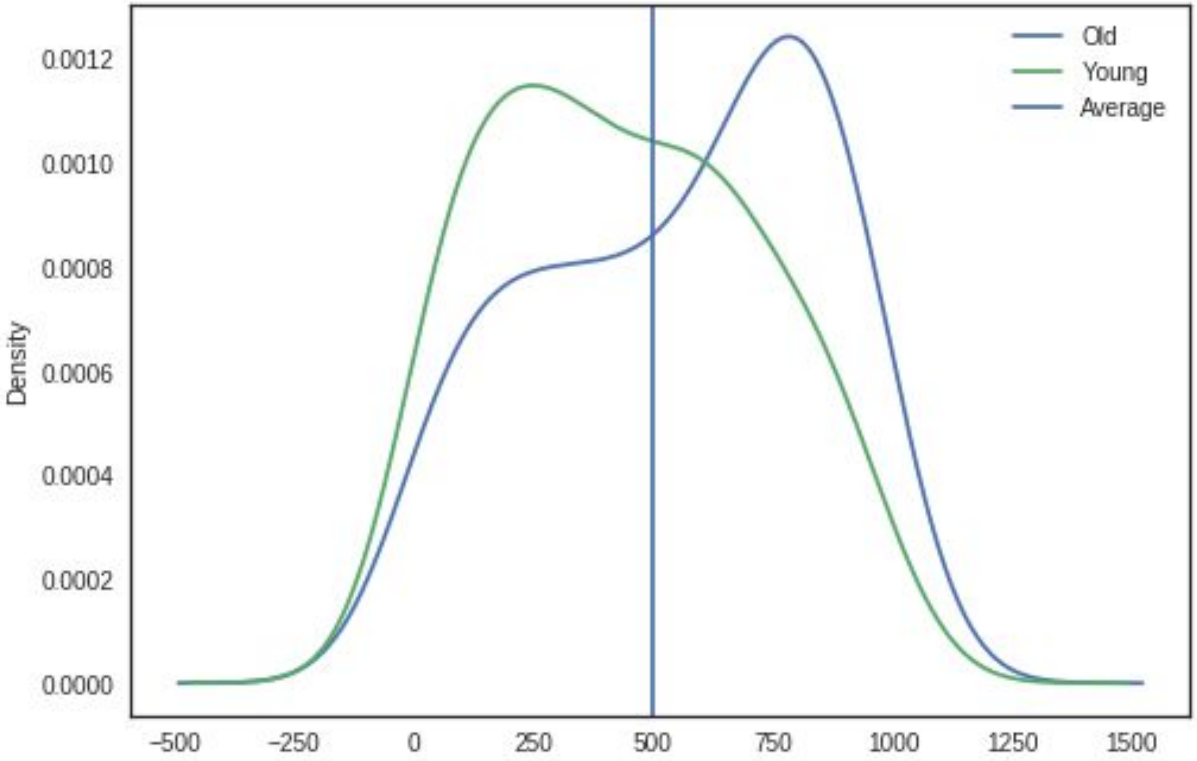



```
plt.axvline(497, label = 'Average')
```

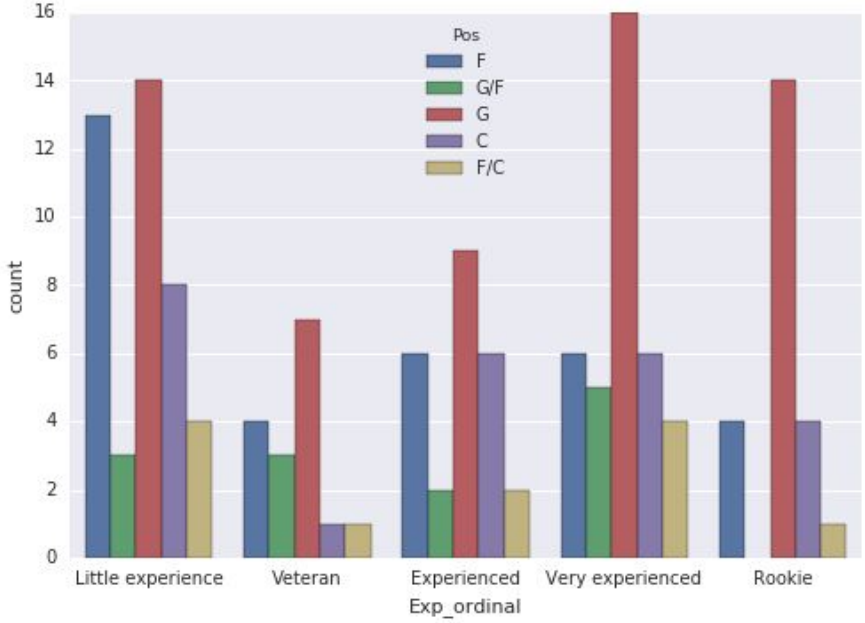
Kernel Density Estimate (KDE) Plots

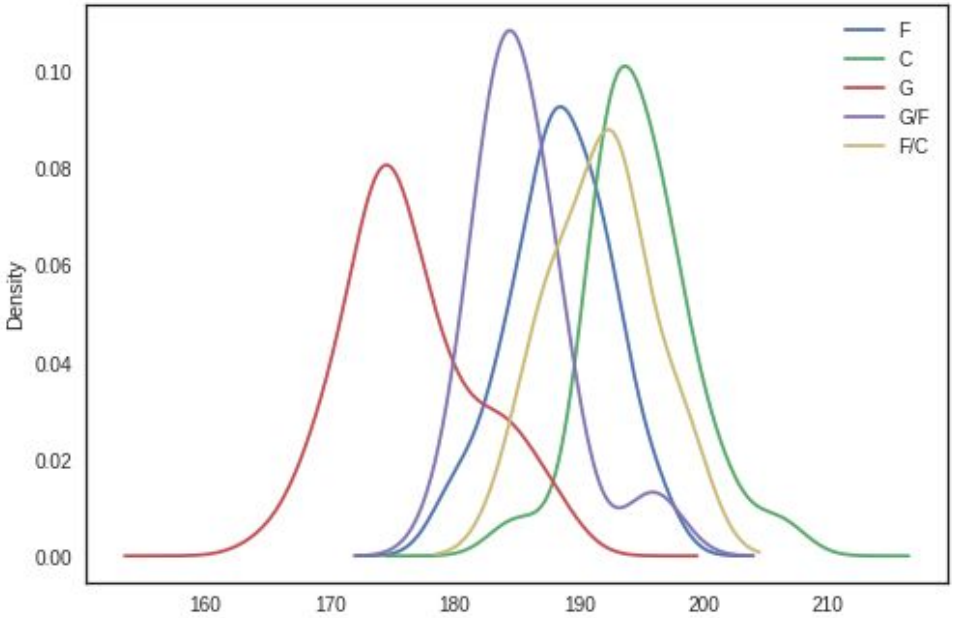


Kernel Density Estimate Plots

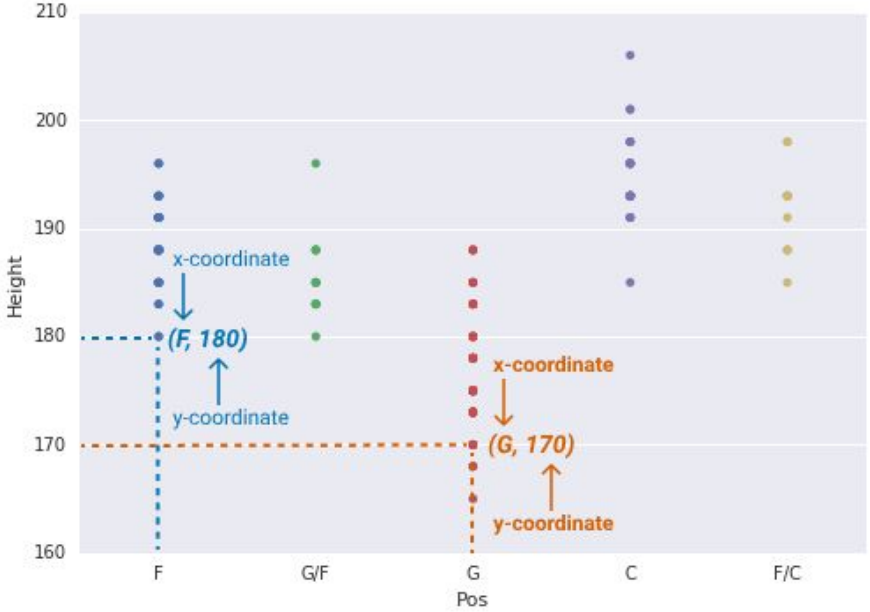


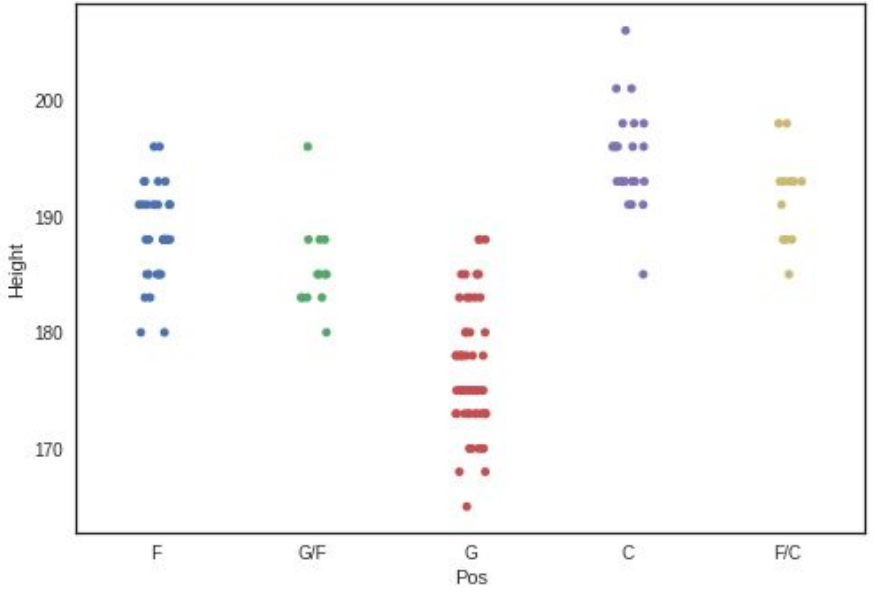
Drawbacks of Kernel Density Plots



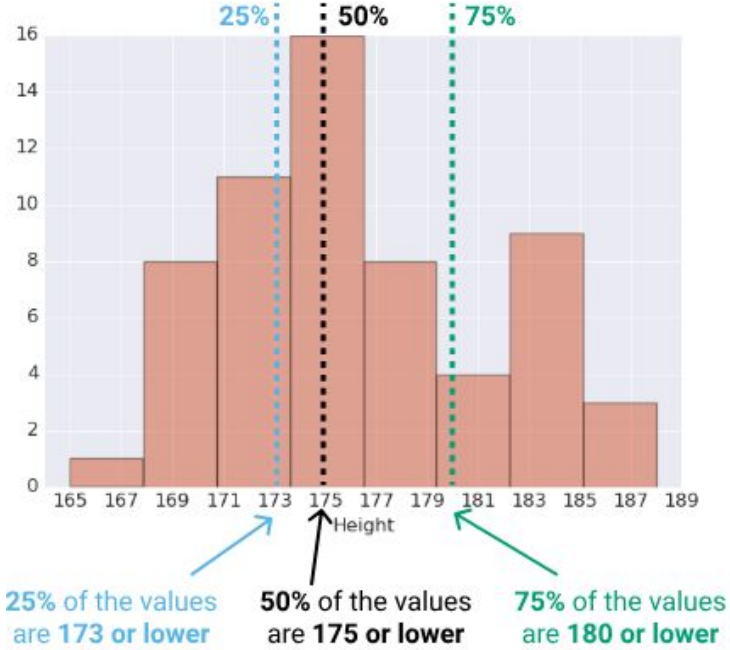


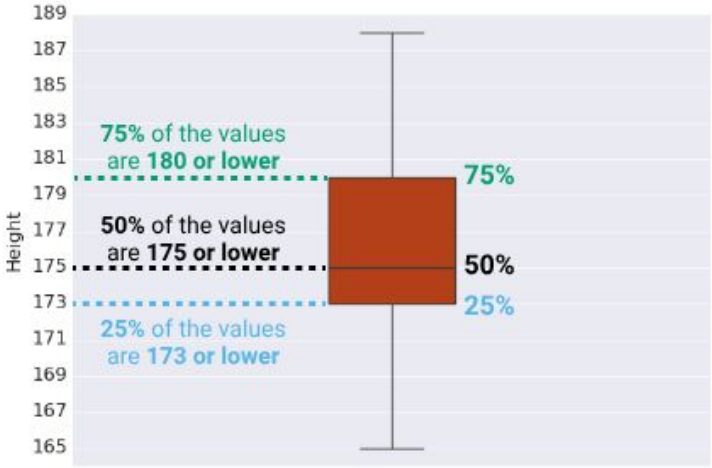
Strip Plots


210
200
190
180
170
160
Height
x-coordinate
(F, 180)
y-coordinate
x-coordinate
(G, 170)
y-coordinate
F
G/F
G
C
F/C
Pos

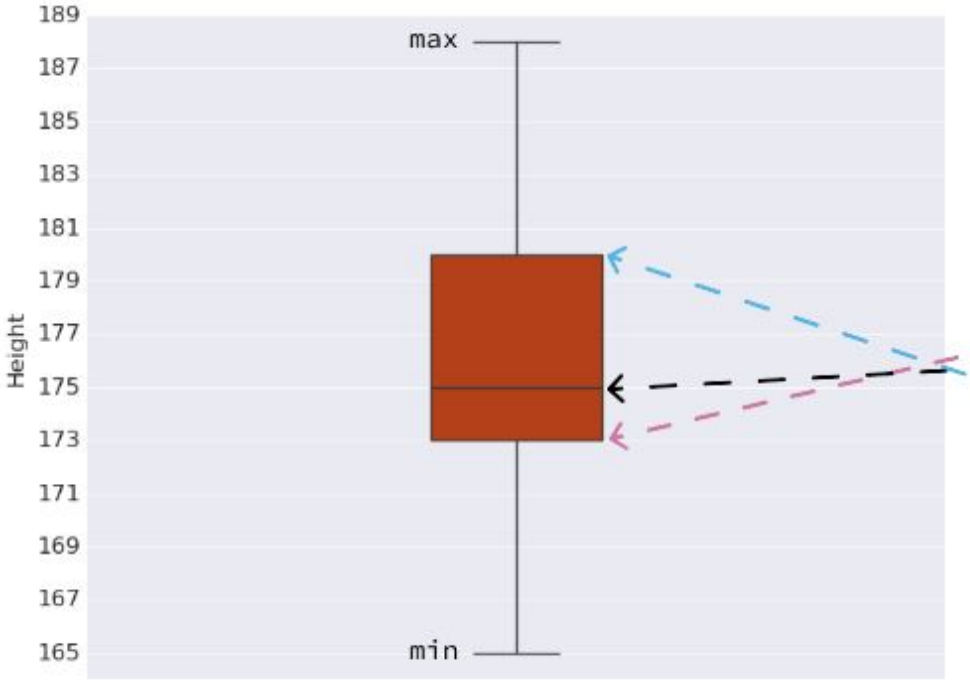

200
190
180
170
Height
F
G/F
G
C
F/C
Pos

Box Plots


25%
50%
75%
16
14
12
10
8
6
4
2
0
165 167 169 171 173 175 177 179 181 183 185 187 189
Height
25% of the values are 173 or lower
50% of the values are 175 or lower
75% of the values are 180 or lower

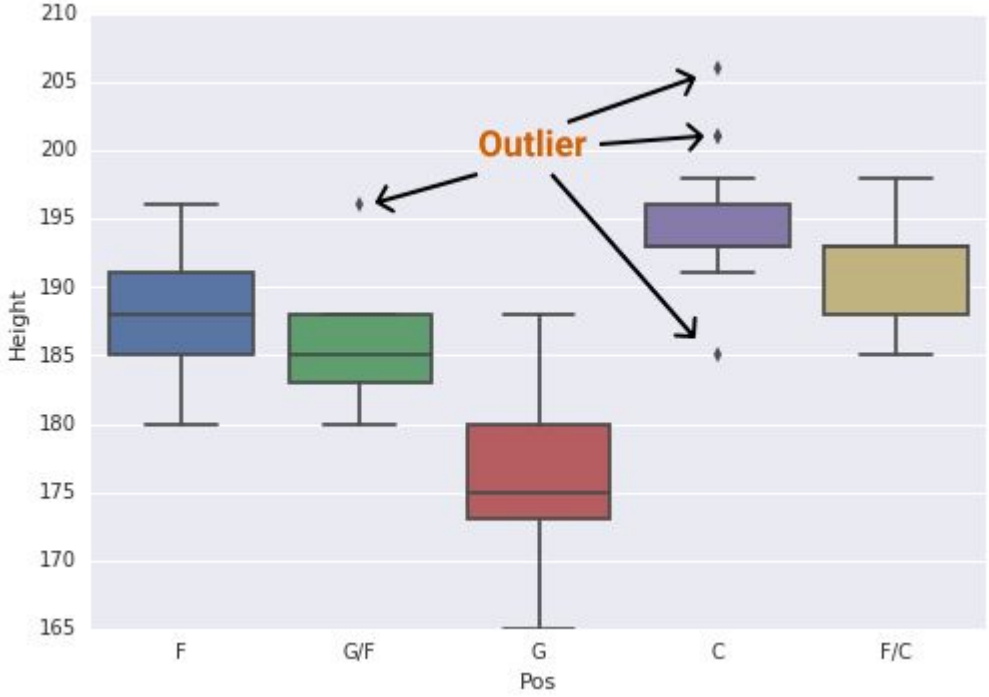

189
187
185
183
181
179
177
175
173
171
169
167
165
Height
75% of the values are 180 or lower
50% of the values are 175 or lower
25% of the values are 173 or lower
75%
50%
25%

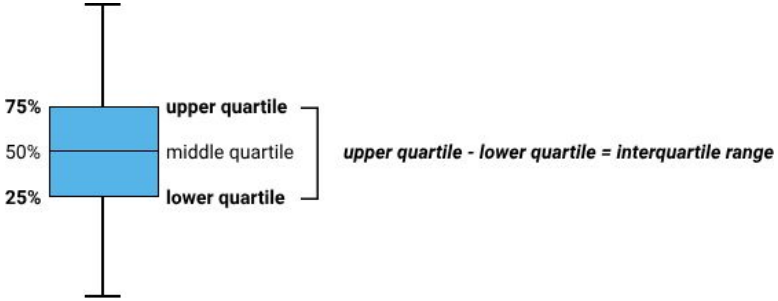
Box Plots



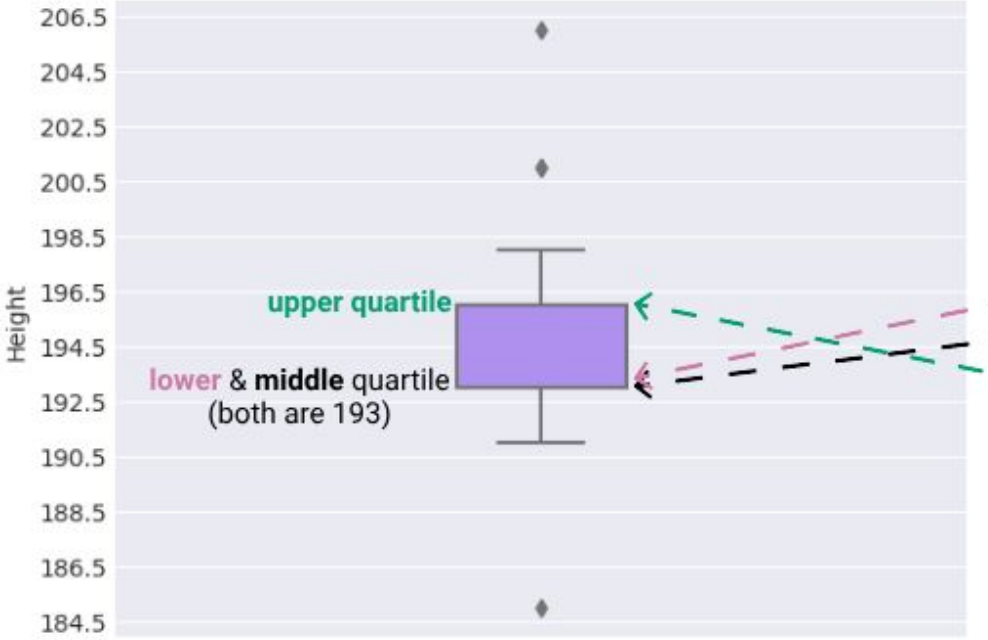

```
>> wnba[wnba['Pos'] == 'G']['Height'].describe()
count      60.000000
mean      176.450000
std         5.312043
min       165.000000
25%       173.000000
50%       175.000000
75%       180.000000
max       188.000000
Name: Height, dtype: float64
```

Outliers


Outlier
Height
210
205
200
195
190
185
180
175
170
165
F
G/F
G
C
F/C
Pos

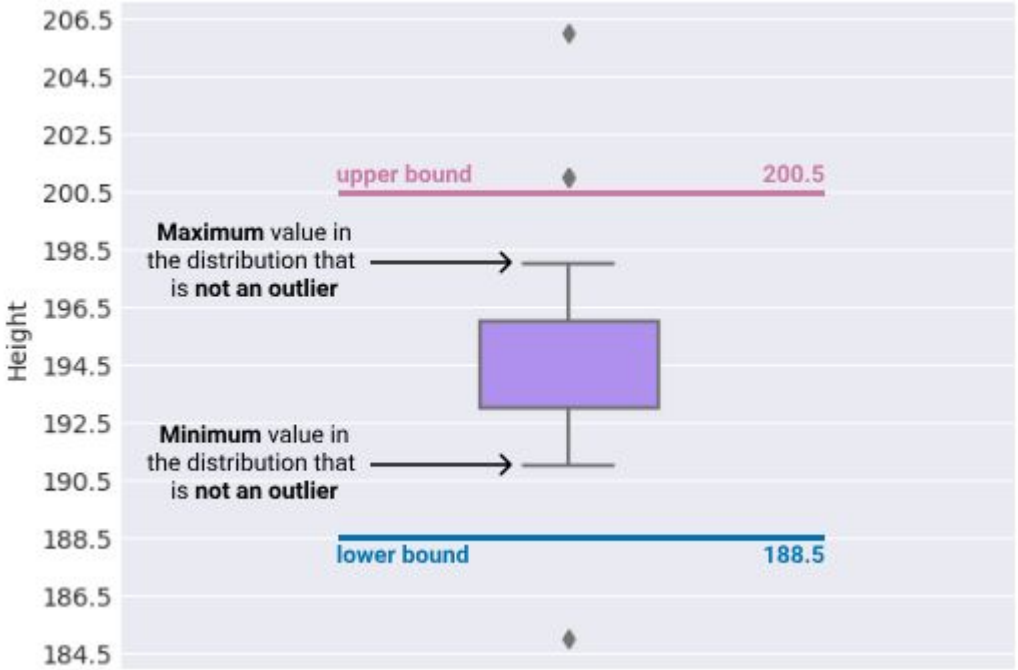

75%
upper quartile
50%
middle quartile
25%
lower quartile
upper quartile - lower quartile = interquartile range

Outliers




```
>> wnba[wnba['Pos'] == 'C']['Height'].describe()
count      25.000000
mean      194.920000
std         4.132392
min       185.000000
25%       193.000000
50%       193.000000
75%       196.000000
max       206.000000
Name: Height, dtype: float64
```

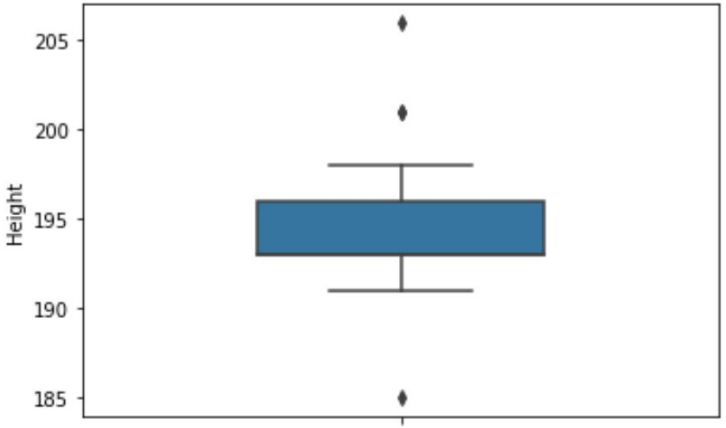
Outliers


206.5
204.5
202.5
200.5
198.5
196.5
194.5
192.5
190.5
188.5
186.5
184.5
Height
upper bound
200.5
Maximum value in the distribution that is not an outlier
Minimum value in the distribution that is not an outlier
lower bound
188.5

Outliers

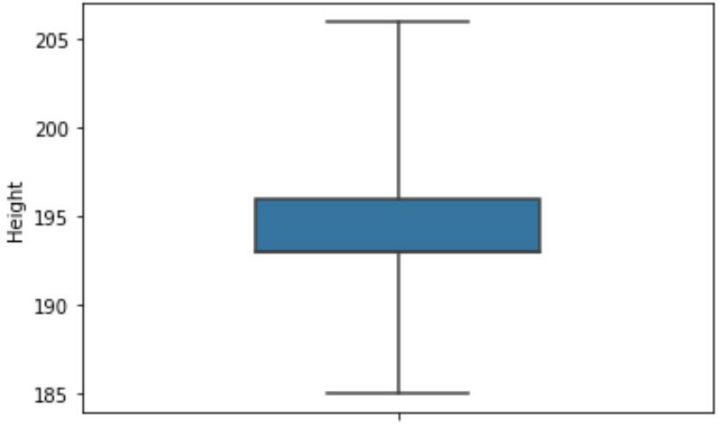
```
sns.boxplot(wnba[wnba['Pos'] == 'C']['Height'], whis = 1.5,
            orient = 'vertical', width = .45)
```

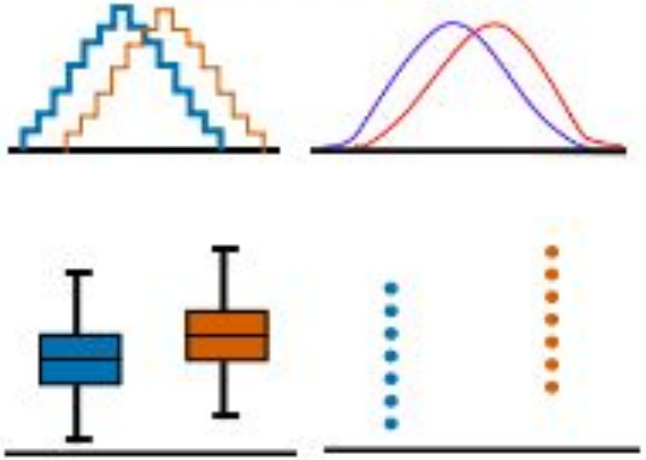
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a180c4518>
```

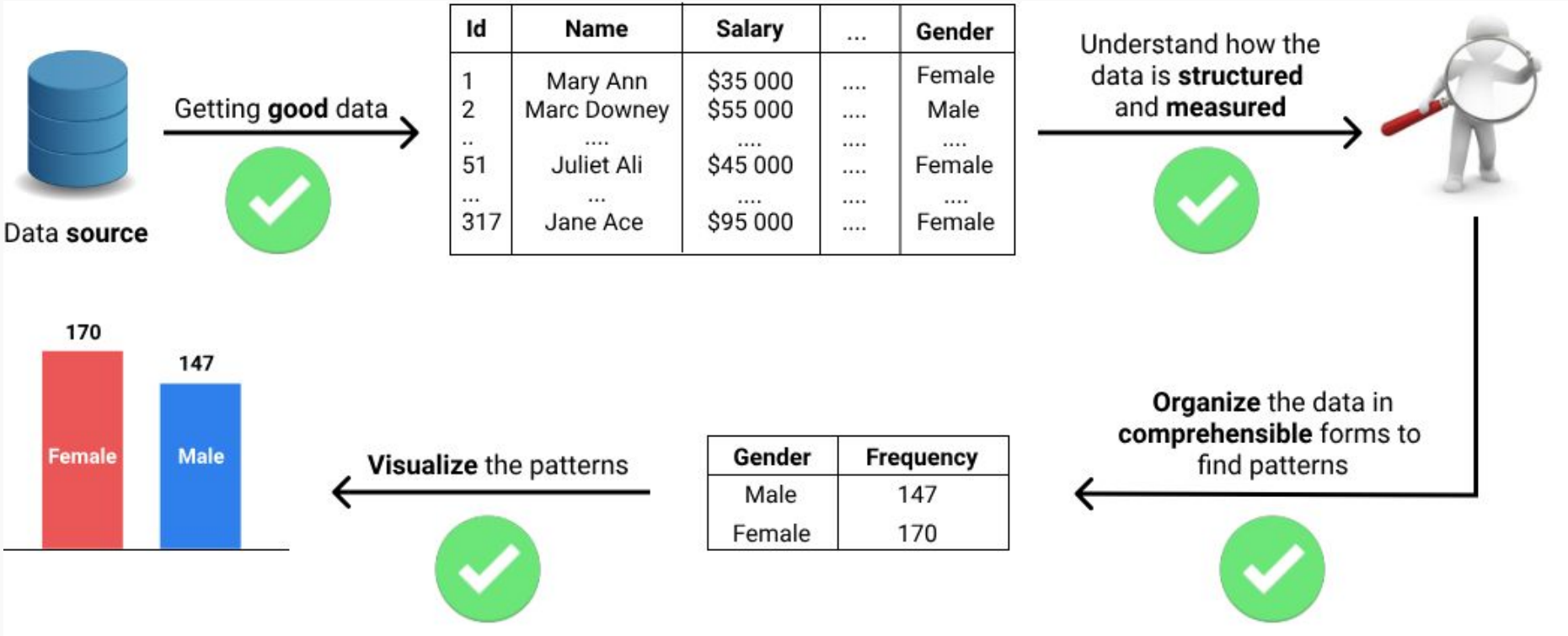



```
sns.boxplot(wnba[wnba['Pos'] == 'C']['Height'], whis = 4,
            orient = 'vertical', width = .45)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a18180208>
```



Scale of measurement	Graphs we can use to compare distributions
Nominal	
Ordinal	
Interval & Ratio	


Data source
Getting good data
Id
Name
Salary
...
Gender
1
Mary Ann
$35 000
....
Female
2
Marc Downey
$55 000
....
Male
..
....
....
....
....
51
Juliet Ali
$45 000
....
Female
...
...
....
....
....
317
Jane Ace
$95 000
....
Female
Understand how the data is structured and measured
Organize the data in comprehensible forms to find patterns
Gender
Frequency
Male
147
Female
170
Visualize the patterns
170
147
Female
Male