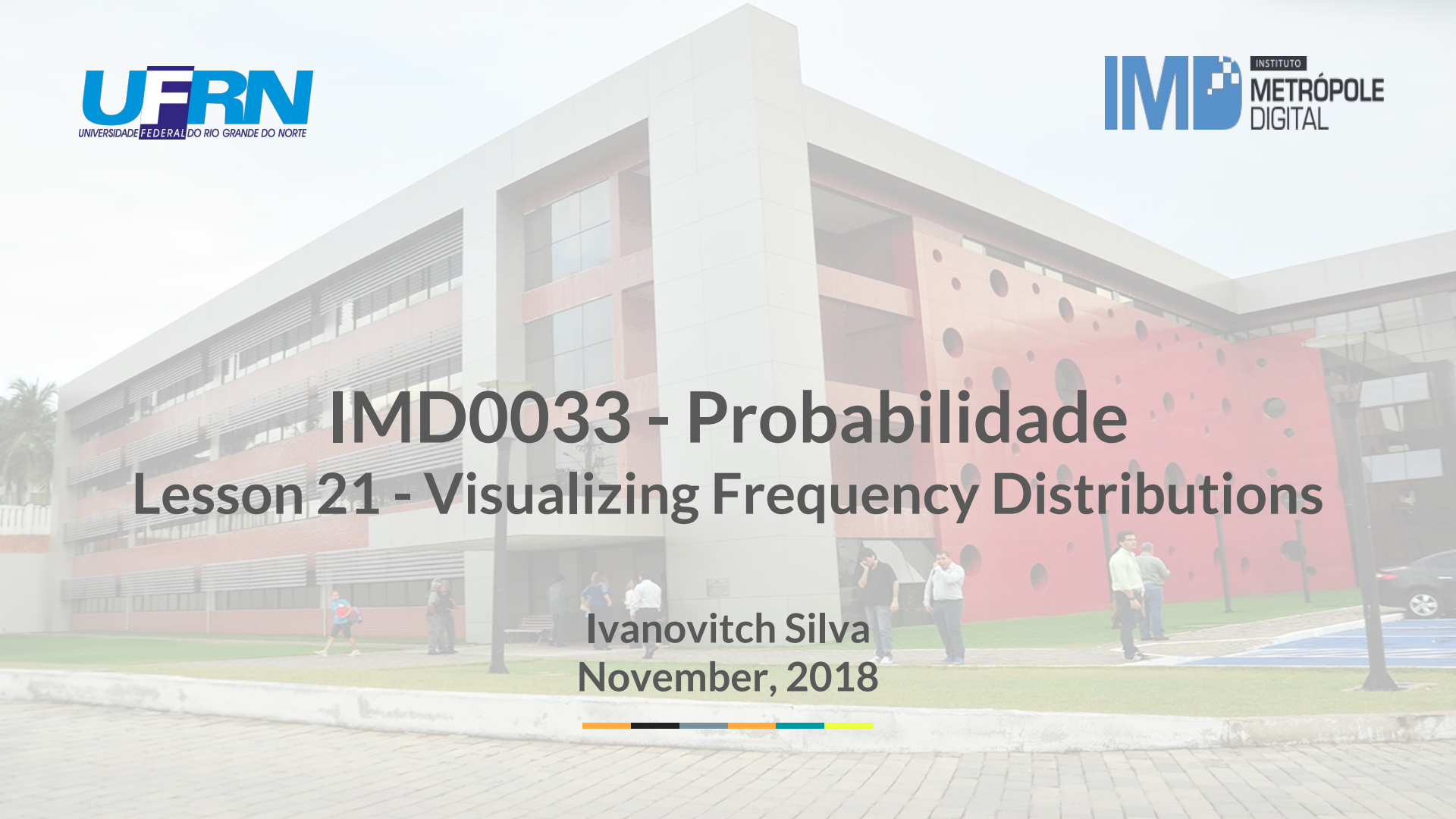


IMD0033 - Probabilidade

Lesson 21 - Visualizing Frequency Distributions

Ivanovitch Silva

November, 2018

Agenda

- Visualizing distributions
- Bar, Pie, Histograms plots
- Skewed distributions
- Symmetrical distributions

Atualizar o repositório

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

Ou ....

```
git pull
```

PREVIOUSLY ON...

Data **source**

Getting **good** data

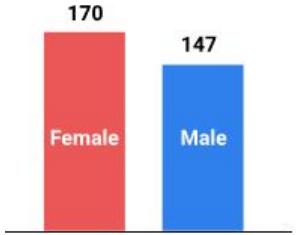
Id	Name	Salary	...	Gender
1	Mary Ann	$35 000	....	Female
2	Marc Downey	$55 000	....	Male
..	....	....	....	....
51	Juliet Ali	$45 000	....	Female
...	...	....	....	....
317	Jane Ace	$95 000	....	Female

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

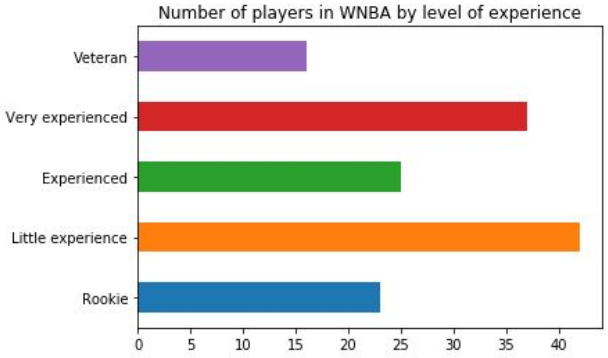
Organize the data in **comprehensible** forms to find patterns

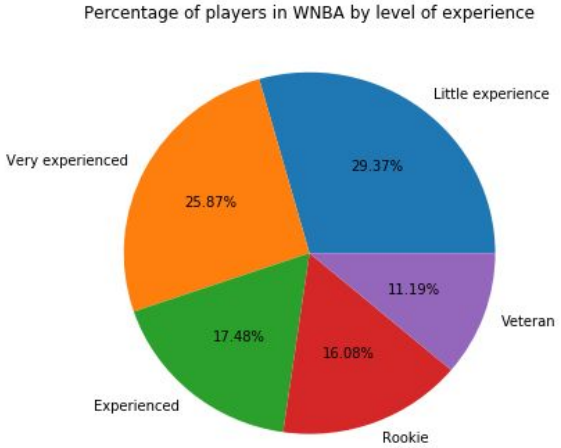
Gender	Frequency
Male	147
Female	170

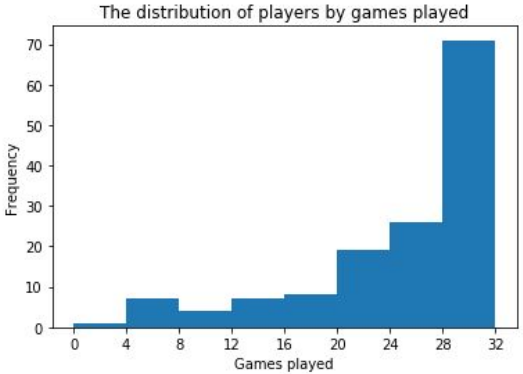
Visualize the patterns



Visualizing Distributions

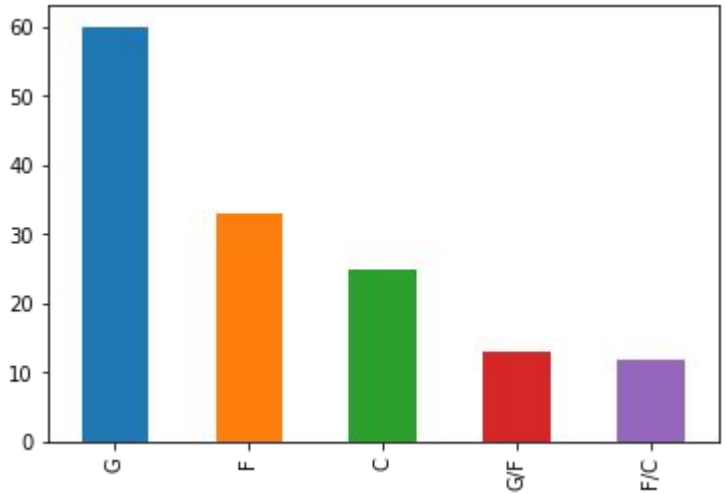






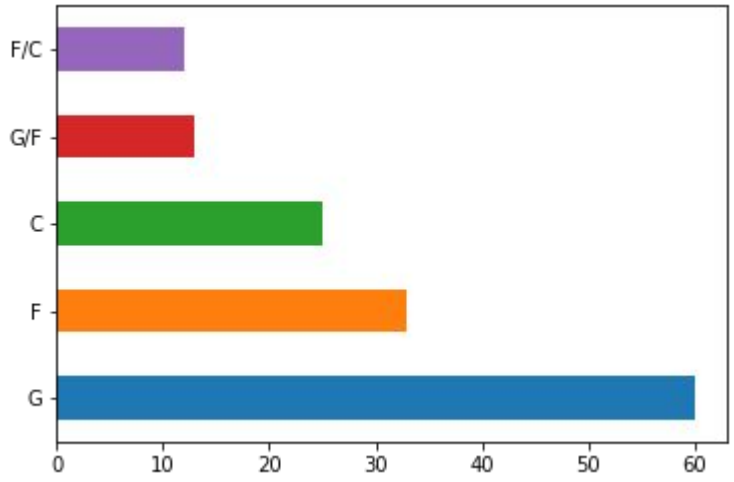
Graphs make easy to scan and compare frequencies, providing us with a single picture of the entire distribution of a variable (**nominal** or **ordinal scale**)

Bar Plots




```
wnba['Pos'].value_counts().plot.bar()
```

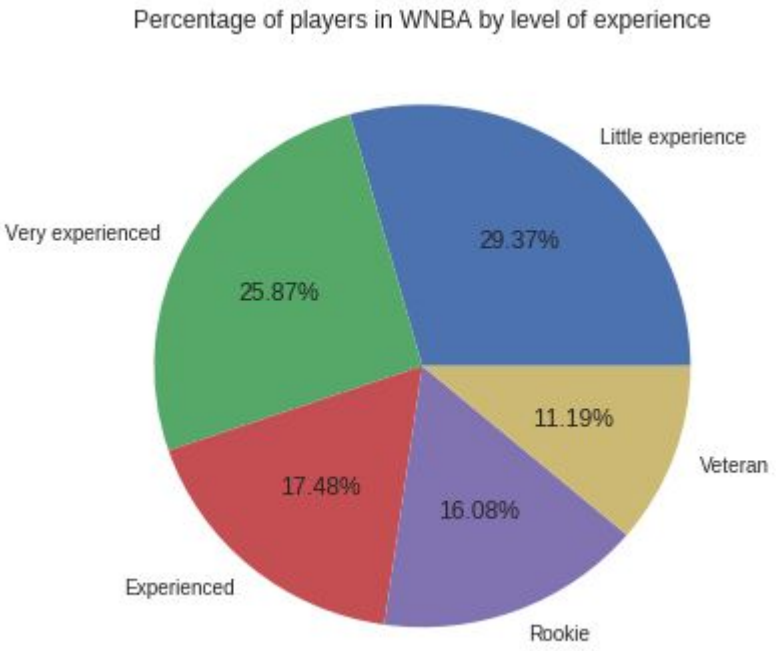
horizontal bar plots are ideal to use when the labels of the unique values are long




```
wnba['Pos'].value_counts().plot.barh()
```

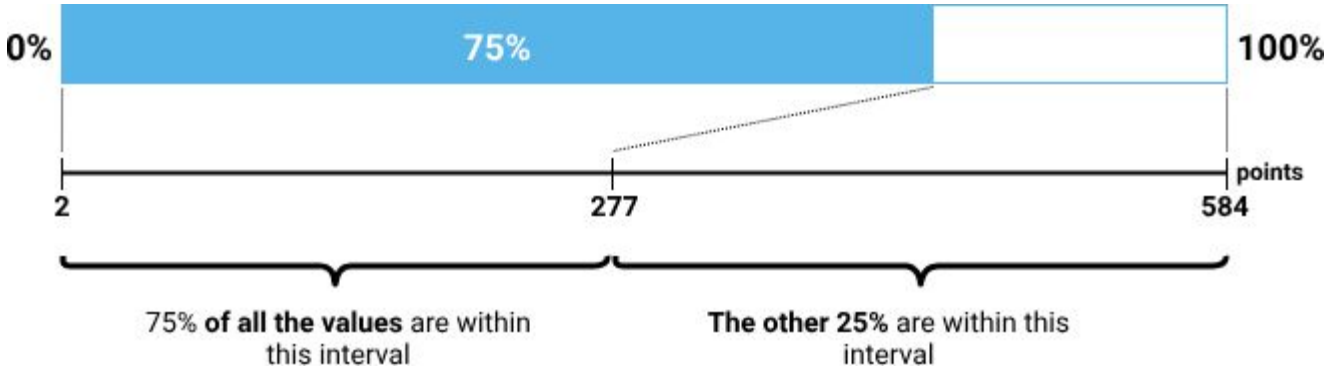
Pie Charts

Pie Charts




```
wnba['Exp_ordinal'].value_counts().\
plot.pie(figsize = (6,6),
         autopct = '%.2f%%',
         title = 'Percentage of players in \
         WNBA by level of experience')
plt.ylabel('')
```

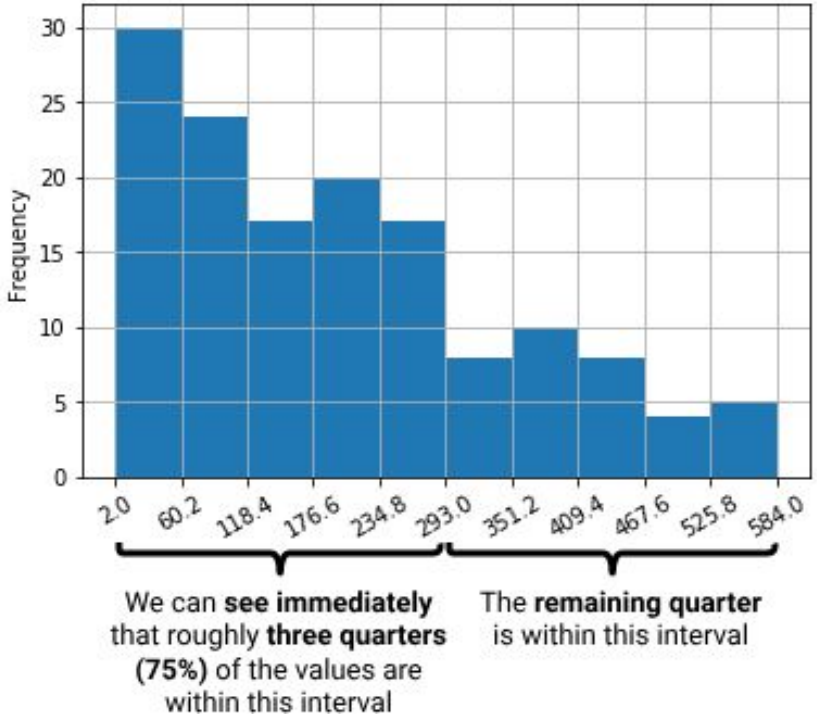
Histograms



We can see that 75% of the values are distributed within a relatively narrow interval (between 2 and 277), while the remaining 25% are distributed in an interval that's slightly larger.

```
>> wnba['PTS'].describe()
count    143.000000
mean     201.790210
std      153.381548
min        2.000000
25%       75.000000
50%      177.000000
75%      277.500000
max      584.000000
```

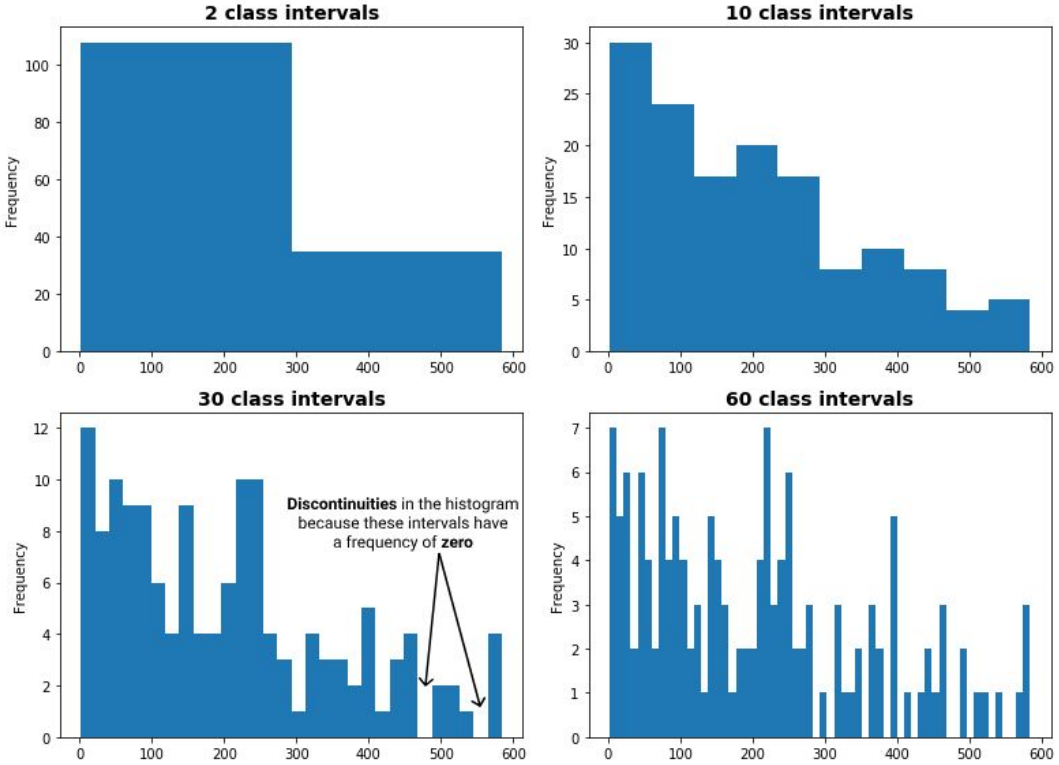
The Statistics Behind Histograms

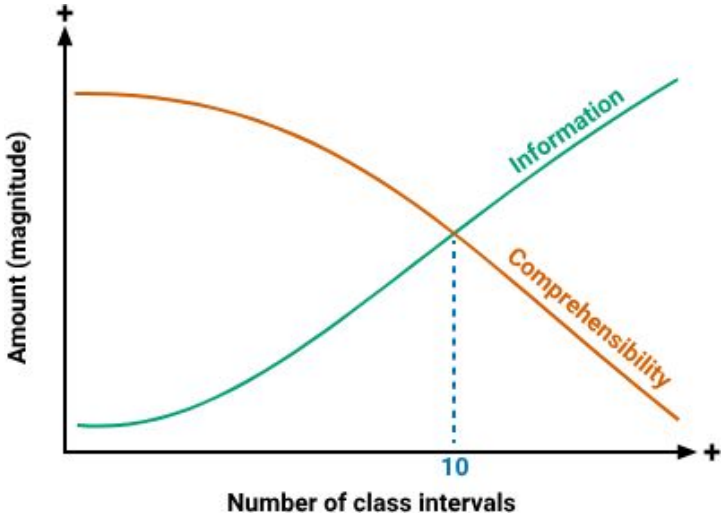



```
>> wnba['PTS'].describe()
count    143.000000
mean     201.790210
std      153.381548
min        2.000000
25%       75.000000
50%      177.000000
75%      277.500000
max      584.000000
Name: PTS, dtype: float64
```

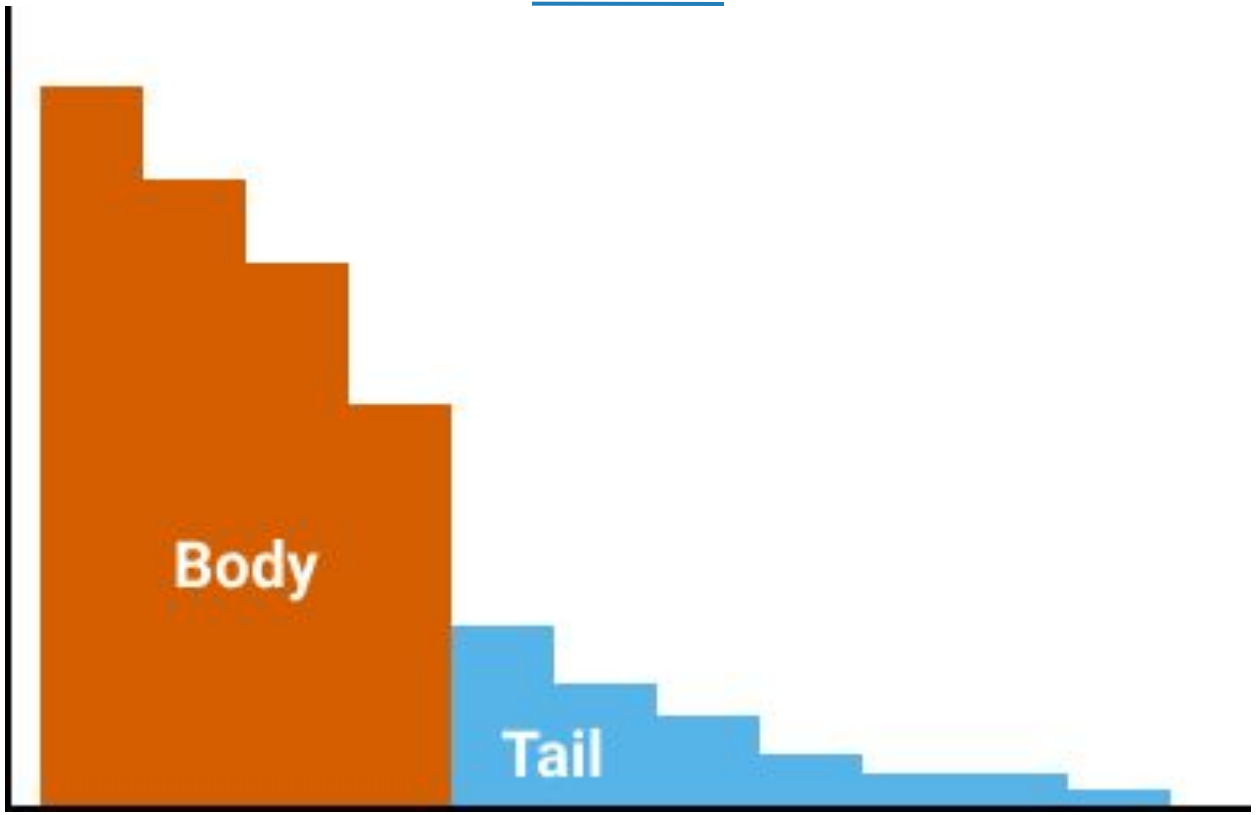
```
>> wnba['PTS'].plot.hist()
```

Binning for Histograms

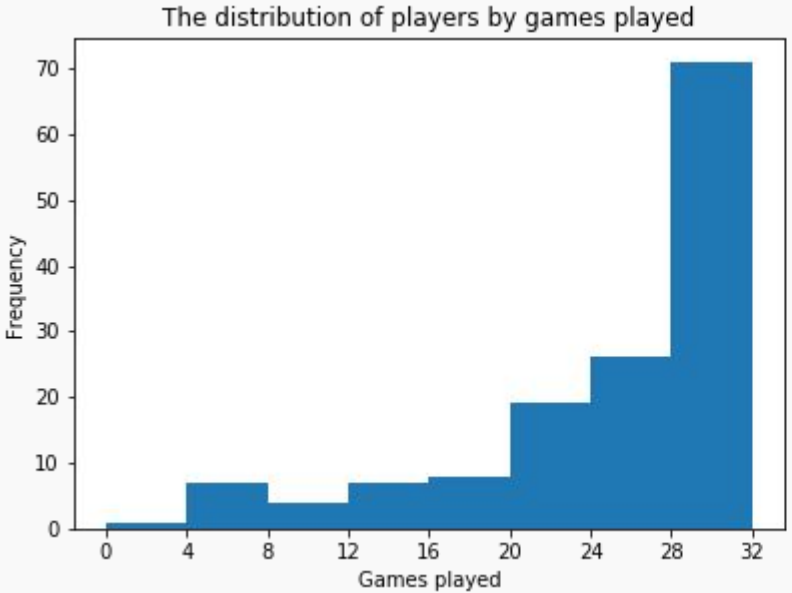


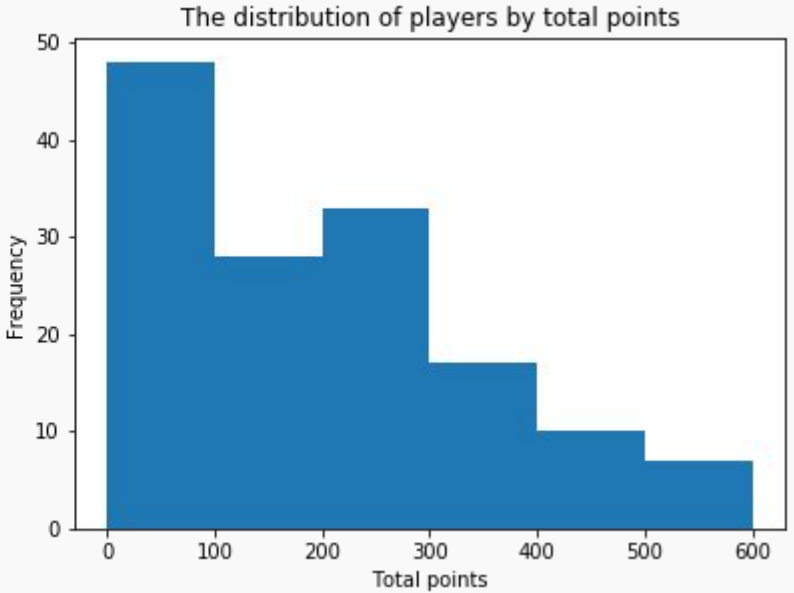


Skewed Distributions

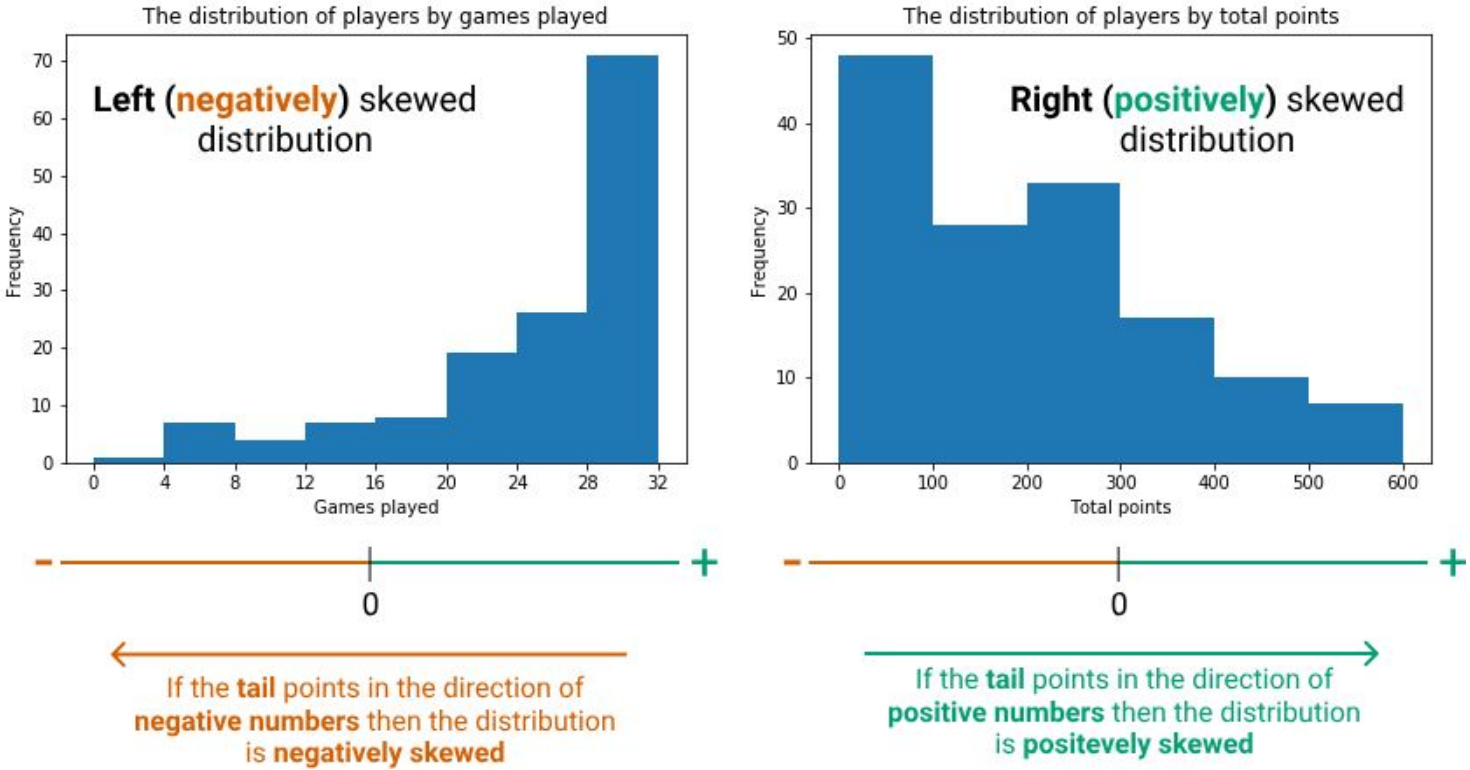


Skewed Distributions

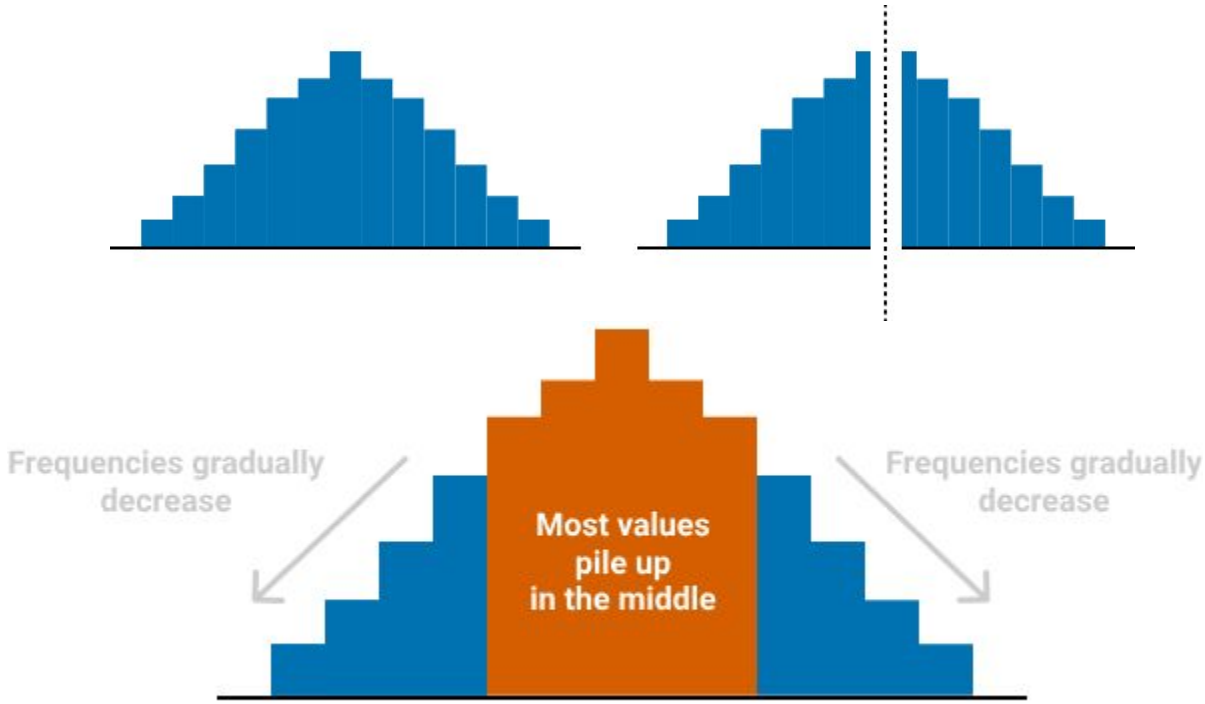




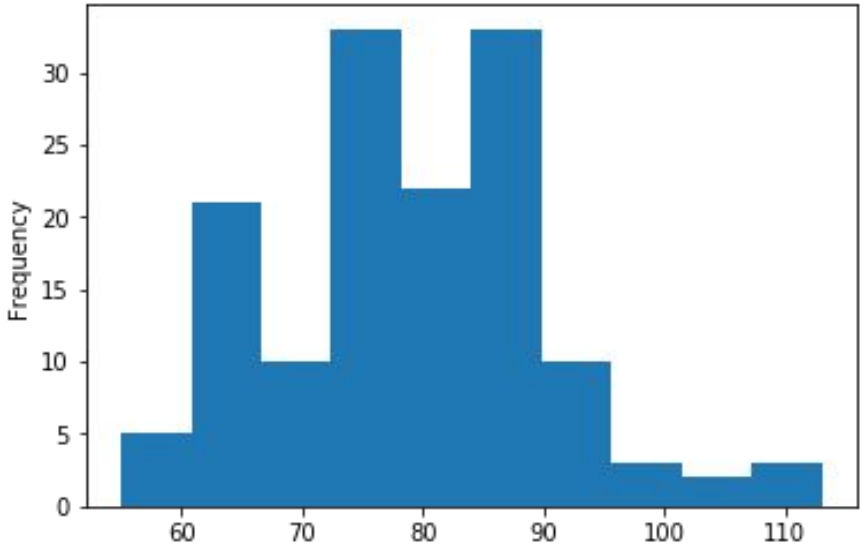
Skewed Distributions



Symmetrical Distributions



Symmetrical Distribution (uniform)



Scale of measurement	Graphs we can use to show the distribution
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