Exploratory Data Analysis

Part B - Visualizing the data : The drill

Pavlos Protopapas

Visualization

Visualization is incredibly important, both for EDA and for communicating our results to others.

> “The greatest value of a picture is when it forces us to notice what we never expected to see.”
>
> John Tukey

(American mathematical statistician, best known for the development of the Fast Fourier Transform algorithm and box plot.)

What's the need to visualize?

Anscombe’s Quartet

Set A

10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Set B

10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.11
7	7.26
5	4.74

Set C

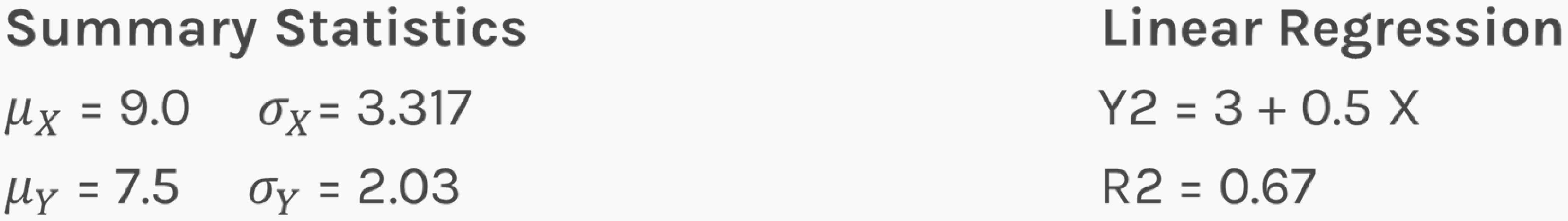
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
12	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Set D

8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

Anscombe’s Quartet

Anscombe's Quartet is a set of four datasets, where each produces the same summary statistics (mean, standard deviation, and correlation), which could lead one to believe the datasets are quite similar. However, after visualizing (plotting) the data, it becomes clear that the datasets are markedly different.



Summary Statistics

μ_X = 9.0 σ_X = 3.317

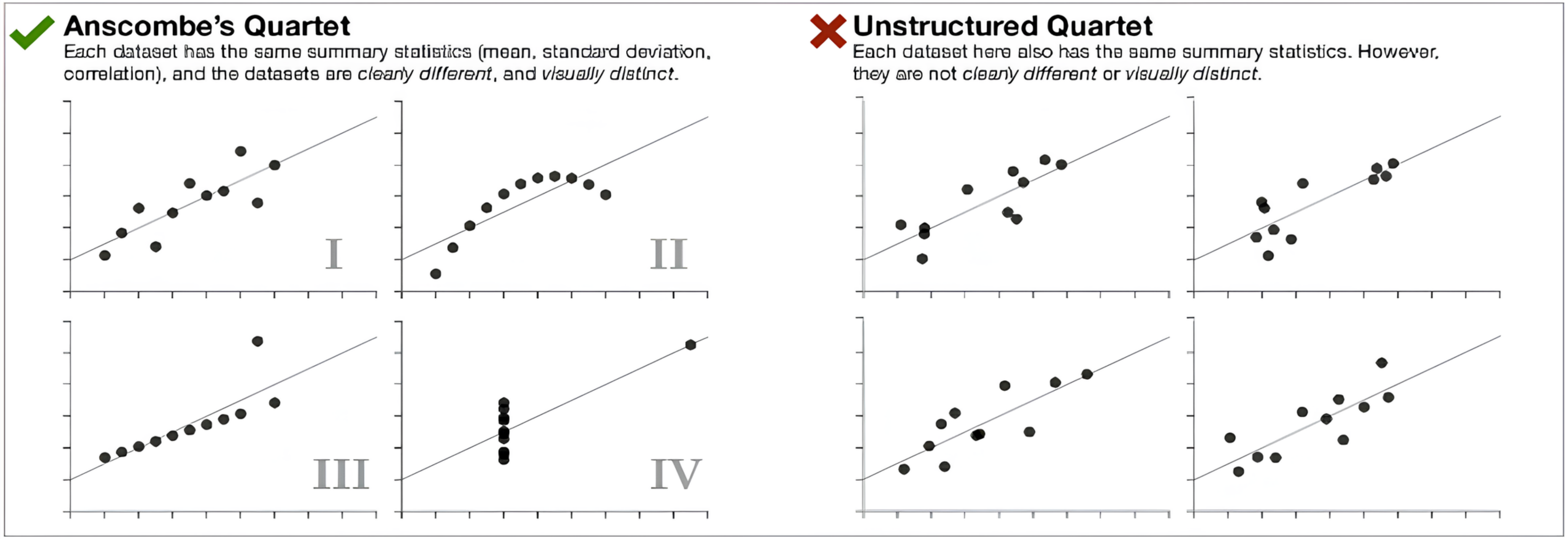
μ_Y = 7.5 σ_Y = 2.03

Linear Regression

Y2 = 3 + 0.5 X

R2 = 0.67

Make sure the statistics are not fooling you!



✓ **Anscombe's Quartet**
Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *cleanly different*, and *visually distinct*.

✗ **Unstructured Quartet**
Each dataset here also has the same summary statistics. However, they are not *cleanly different* or *visually distinct*.

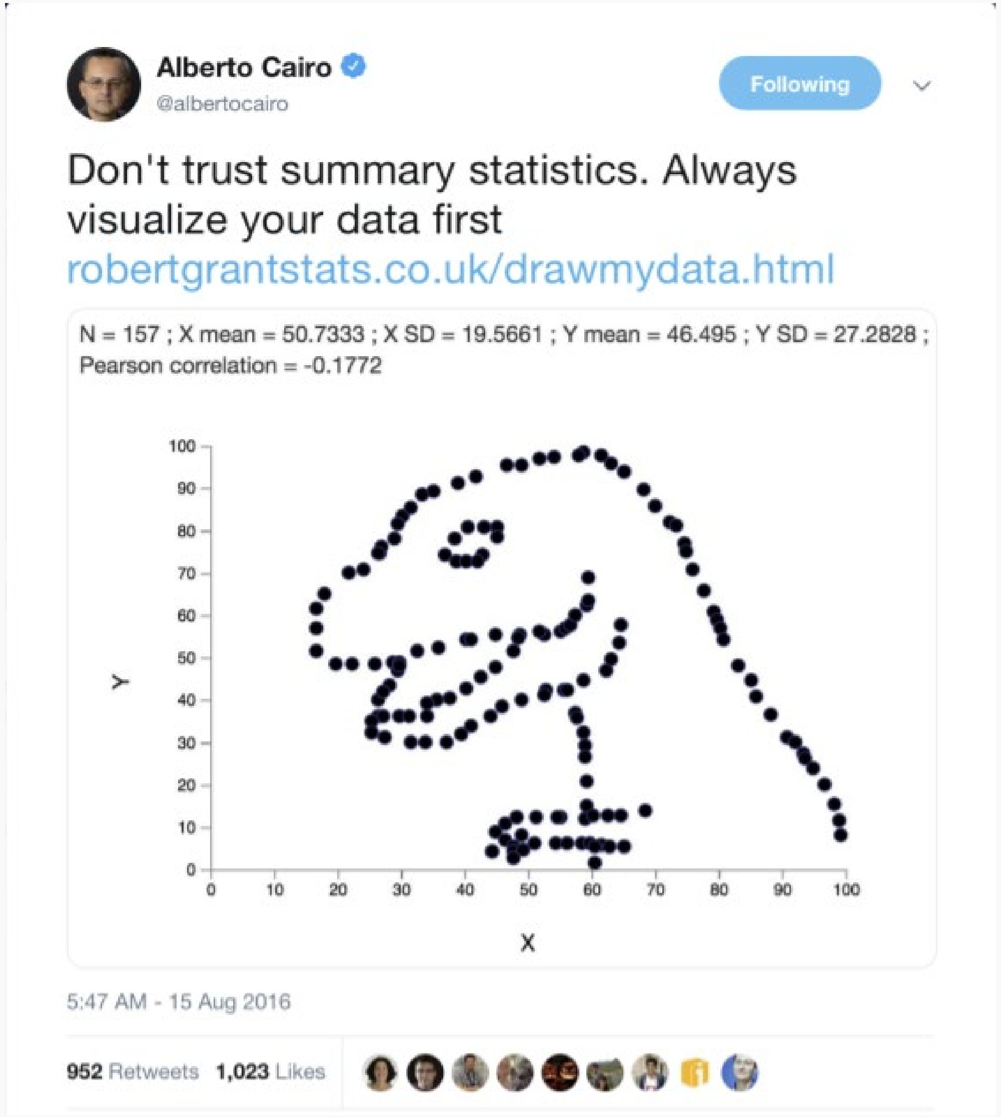
Fig 1. Anscombe's Quartet (left), and a "Unstructured Quartet" on the right, where the datasets have the same summary statistics as those in Anscombe's Quartet, but lack underlying structure or visual distinction.

Same stats do not imply same graphs

Same graphs do not imply same stats

Credits: Same Stats, Different Graphs. (Autodesk Research)

Same stats, different graphs!



Don't trust summary statistics. Always visualize your data first robertgrantstats.co.uk/drawmydata.html

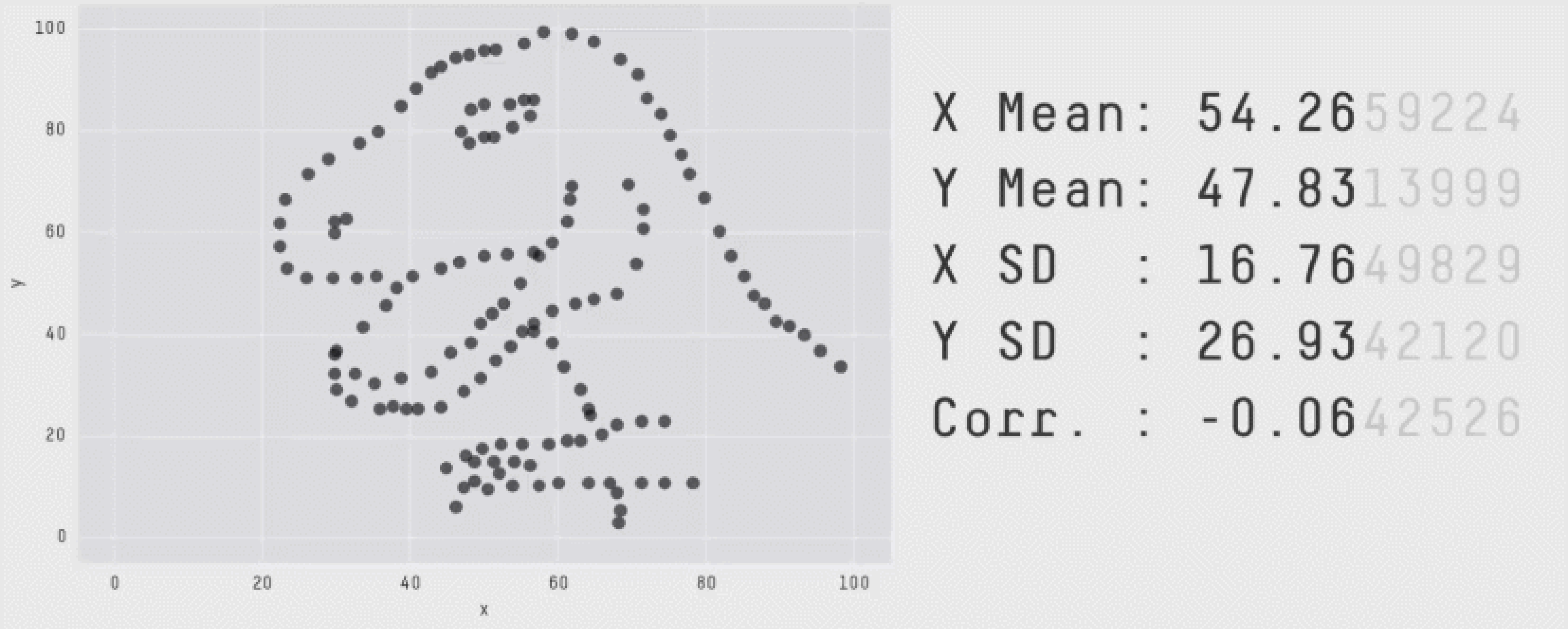
N = 157 ; X mean = 50.7333 ; X SD = 19.5661 ; Y mean = 46.495 ; Y SD = 27.2828 ; Pearson correlation = -0.1772

5:47 AM - 15 Aug 2016

952 Retweets 1,023 Likes

The Datasaurus Dozen

Datasets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data.



```
X Mean: 54.2659224
Y Mean: 47.8313999
X SD  : 16.7649829
Y SD  : 26.9342120
Corr. : -0.0642526
```

THE DATASAURUS DOZEN

Credits: Same Stats, Different Graphs. (Autodesk Research)

Visualization Goals

Analyze (Exploratory)

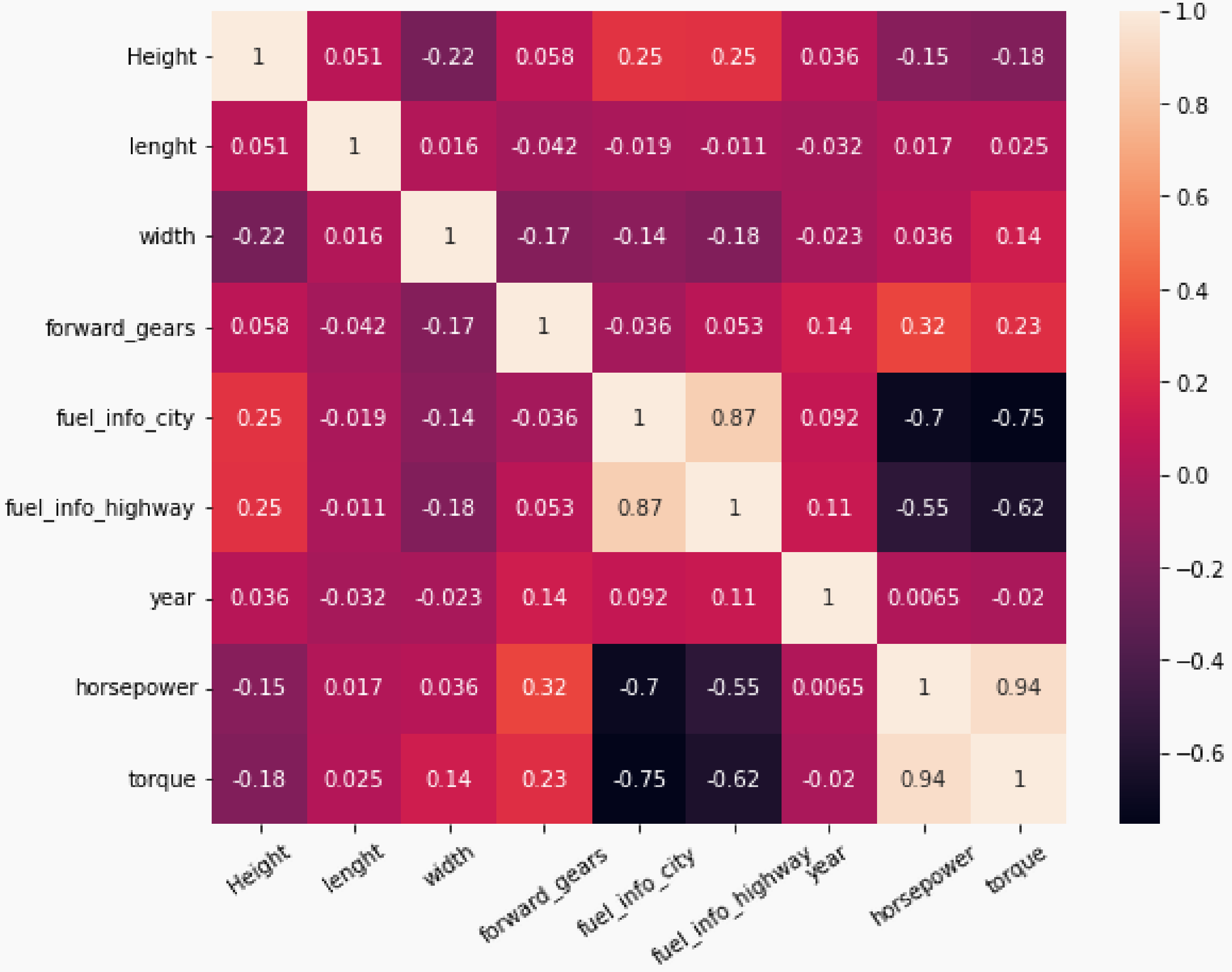
1. Explore the data.
2. Assess the situation.
3. Determine how to proceed.
4. Decide what to do.

Analyze (Exploratory)

Exploring data

The figure illustrates the correlation plot of numerical variables using a heat map.

The correlation plot is used to drop variables that are highly correlated..



Correlation between features

	Height	lenght	width	forward_gears	fuel_info_city	fuel_info_highway	year	horsepower	torque
Height	1	0.051	-0.22	0.058	0.25	0.25	0.036	-0.15	-0.18
lenght	0.051	1	0.016	-0.042	-0.019	-0.011	-0.032	0.017	0.025
width	-0.22	0.016	1	-0.17	-0.14	-0.18	-0.023	0.036	0.14
forward_gears	0.058	-0.042	-0.17	1	-0.036	0.053	0.14	0.32	0.23
fuel_info_city	0.25	-0.019	-0.14	-0.036	1	0.87	0.092	-0.7	-0.75
fuel_info_highway	0.25	-0.011	-0.18	0.053	0.87	1	0.11	-0.55	-0.62
year	0.036	-0.032	-0.023	0.14	0.092	0.11	1	0.0065	-0.02
horsepower	-0.15	0.017	0.036	0.32	-0.7	-0.55	0.0065	1	0.94
torque	-0.18	0.025	0.14	0.23	-0.75	-0.62	-0.02	0.94	1

Visualization Goals

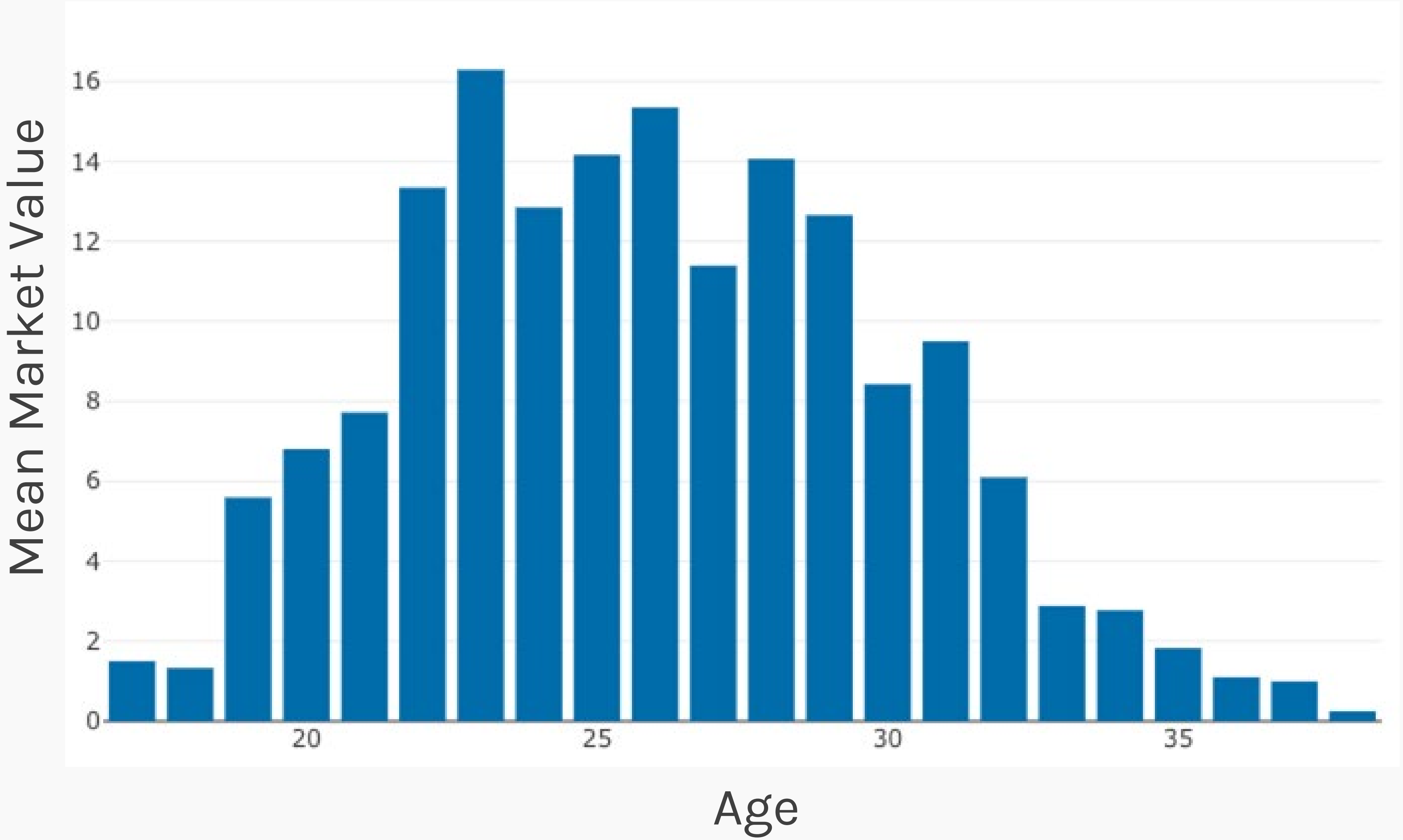
Analyze (Exploratory)

1. Explore the data.
2. Assess the situation.
3. Determine how to proceed.
4. Decide what to do.

Now, let's comeback to visualizing the data from English Premier League

Visualization: The English Premier League

Mean Market Value vs Age

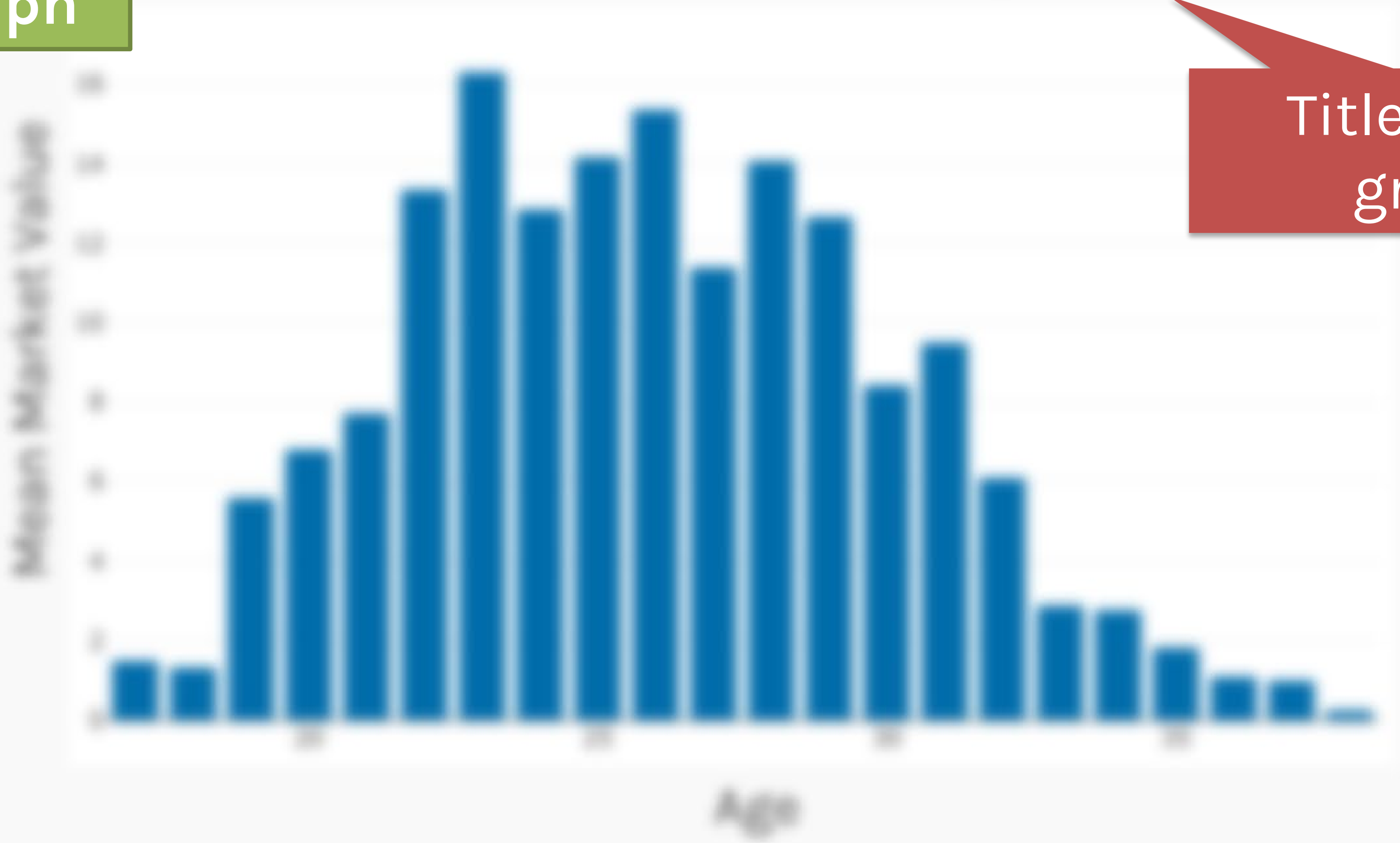


Visualization: The English Premier League

Components of the graph

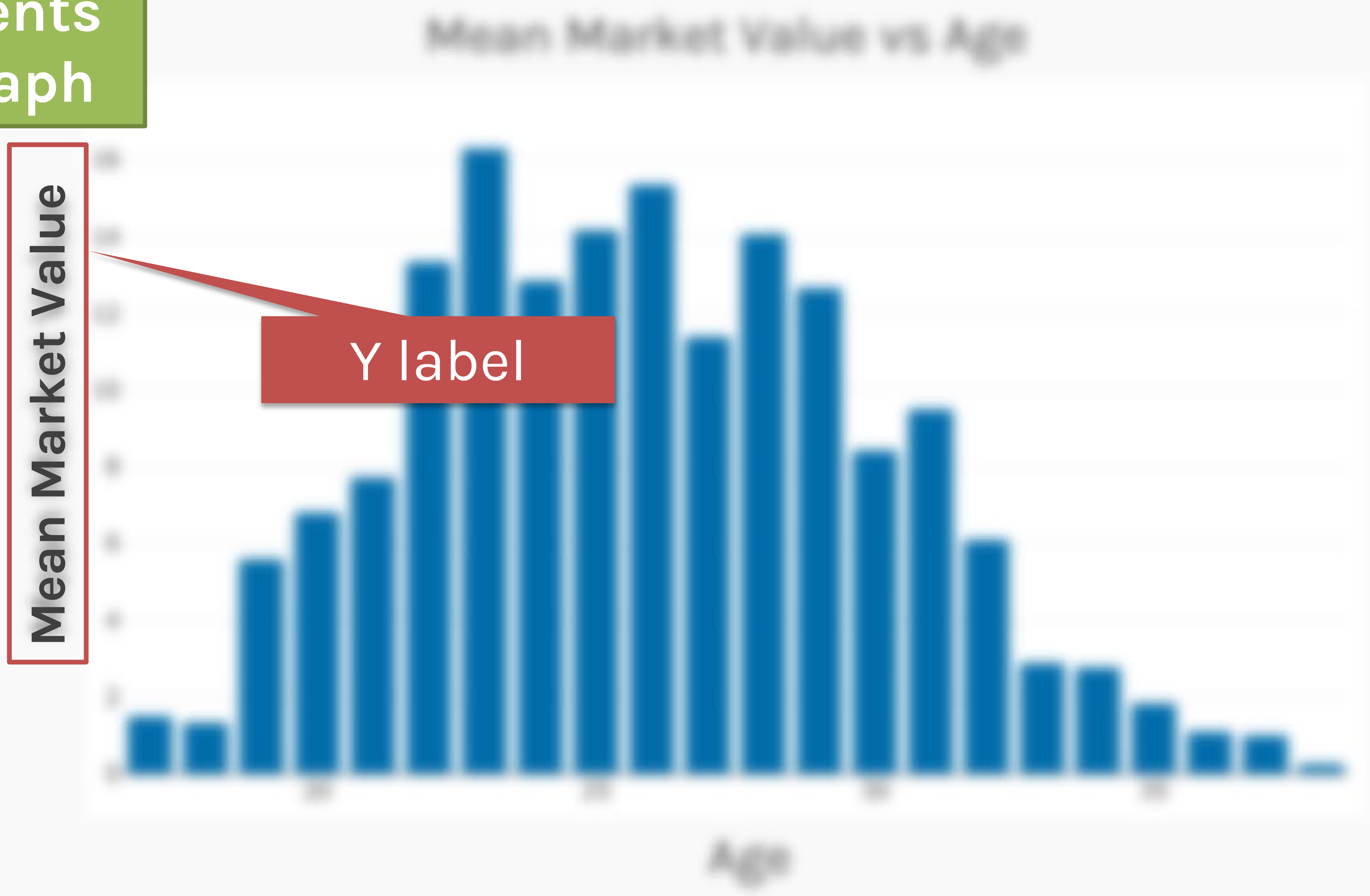
Mean Market Value vs Age

Title of the graph



Visualization: The English Premier League

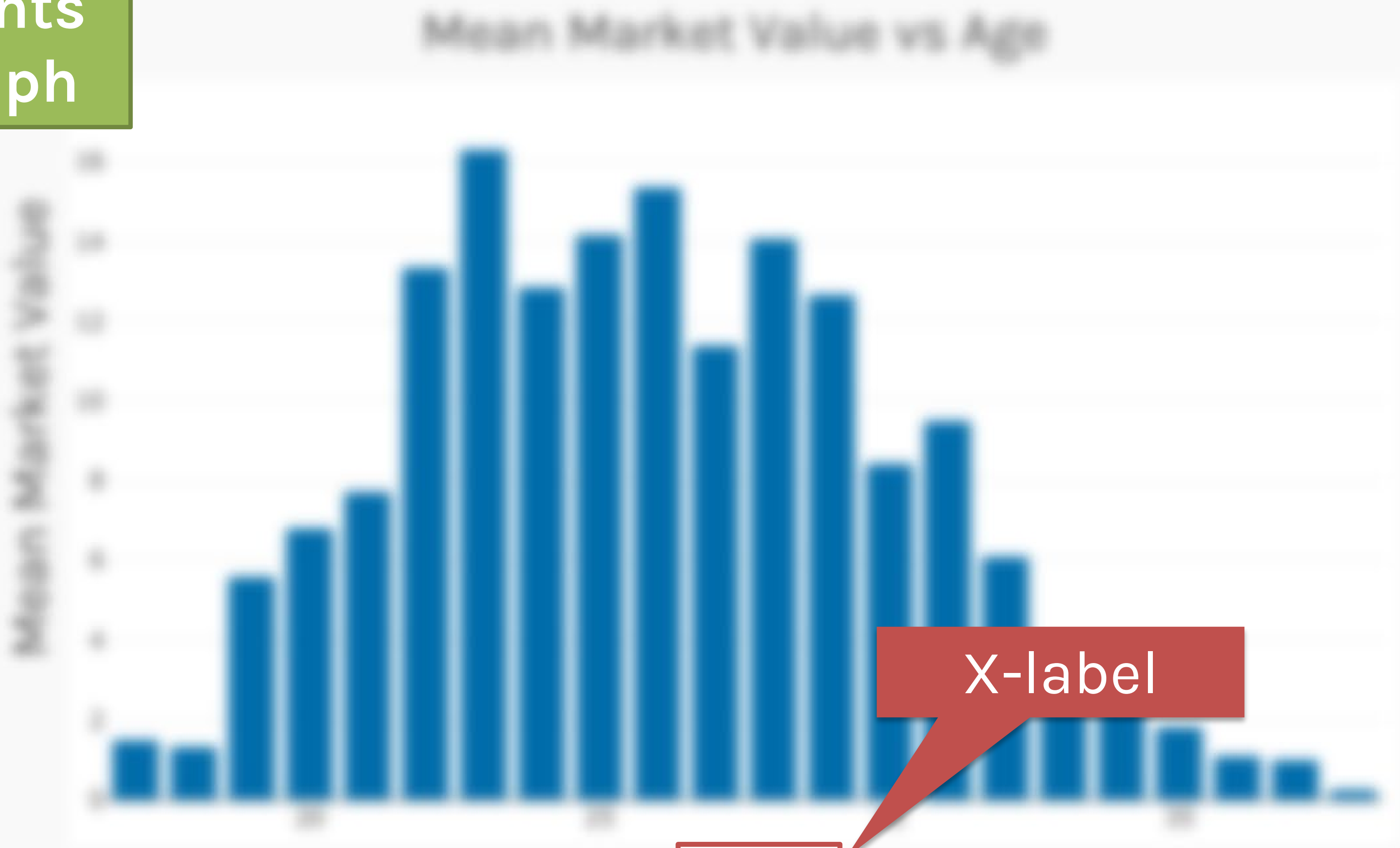
Components of the graph



Y label

Visualization: The English Premier League

Components of the graph

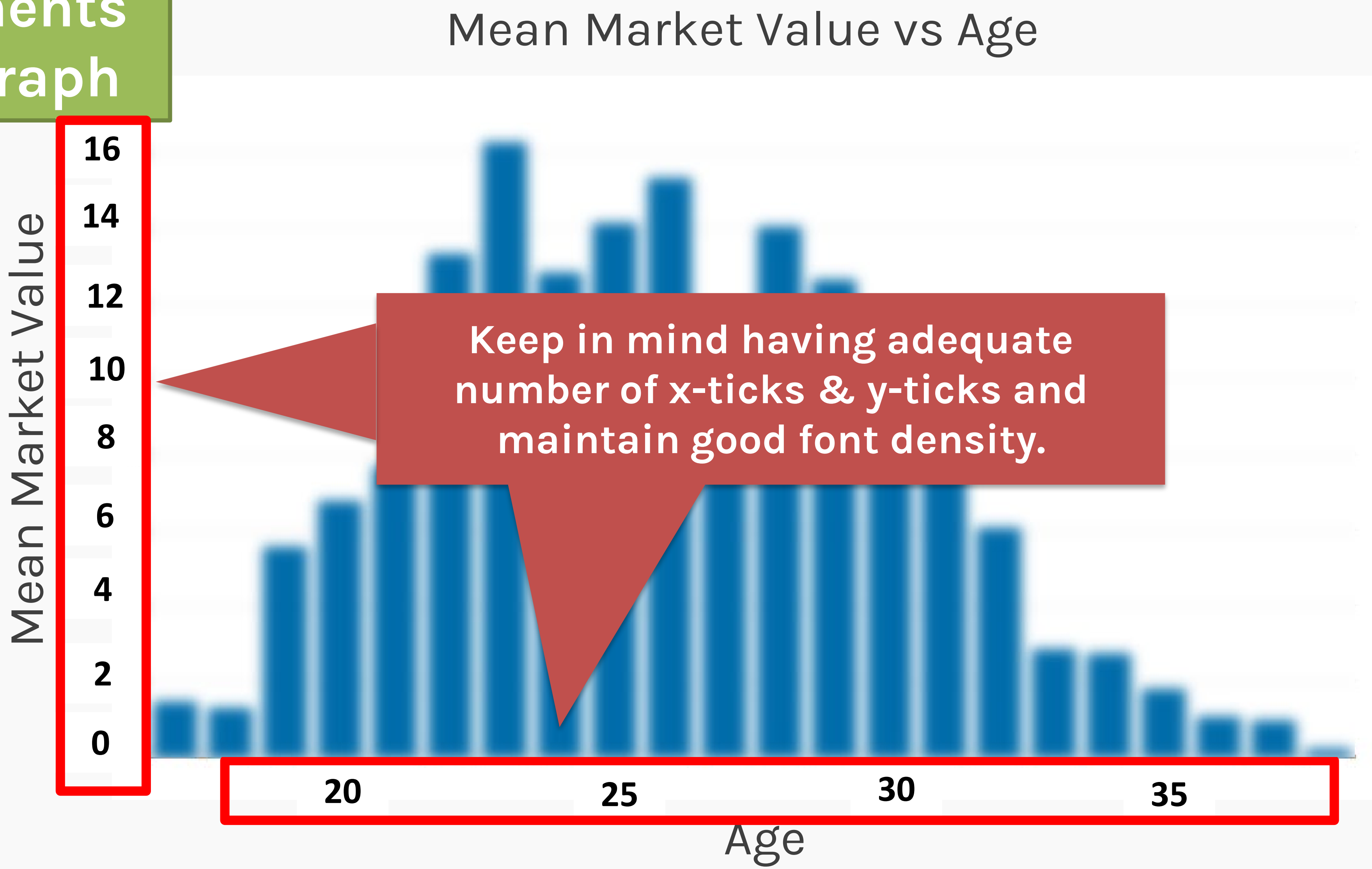


X-label

Age

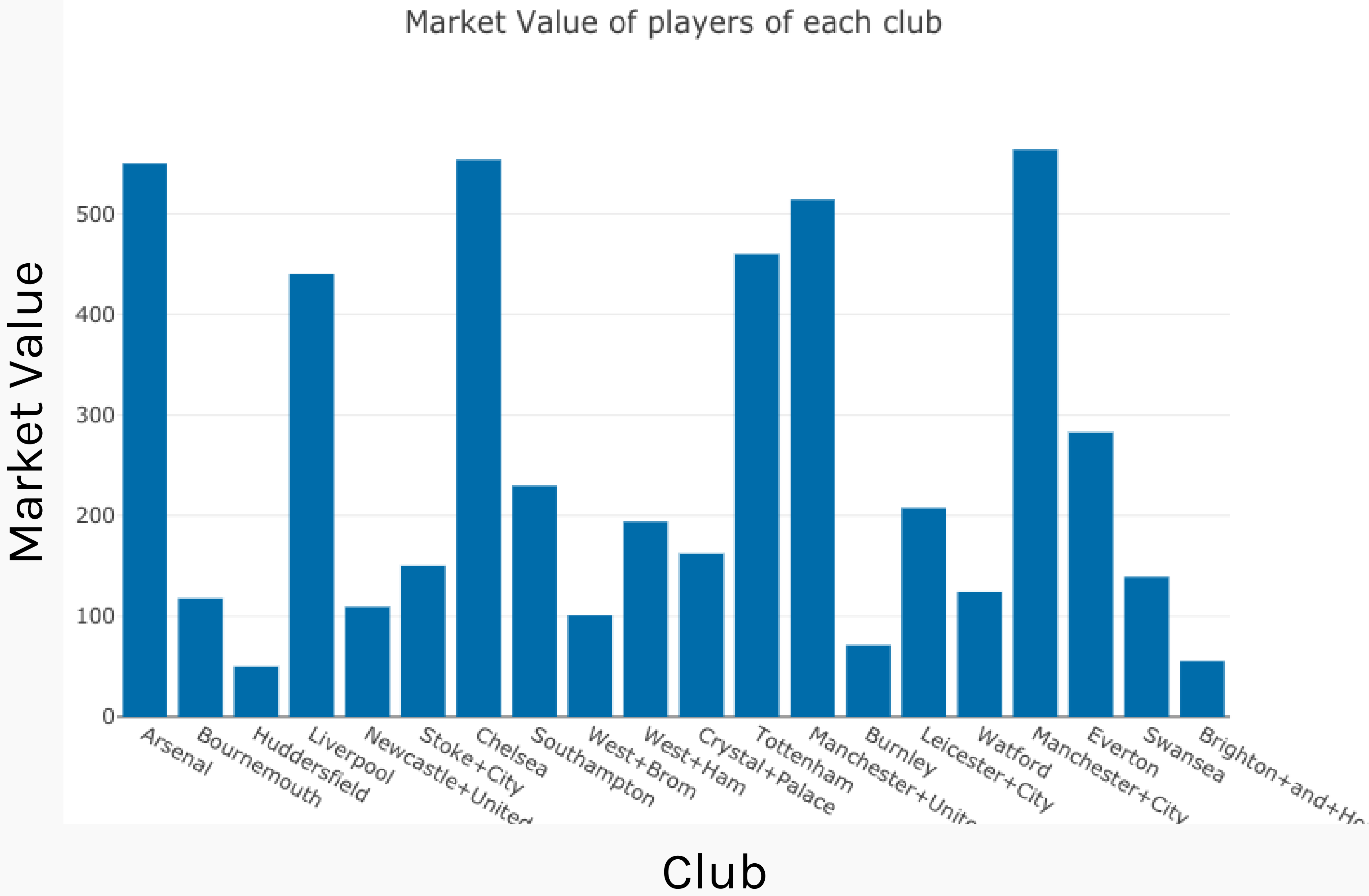
Visualization: The English Premier League

Components of the graph

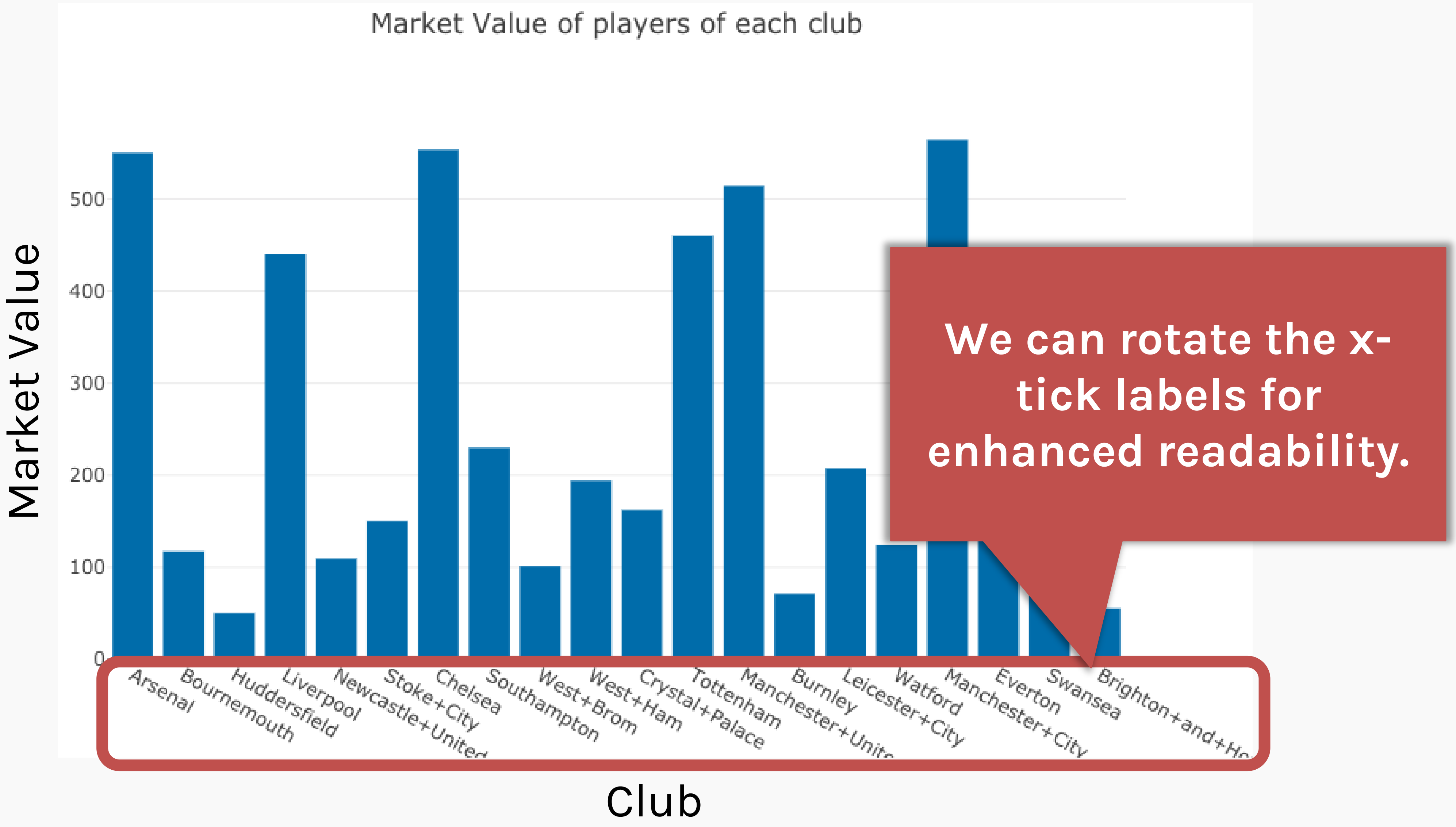


Keep in mind having adequate number of x-ticks & y-ticks and maintain good font density.

Visualization: The English Premier League

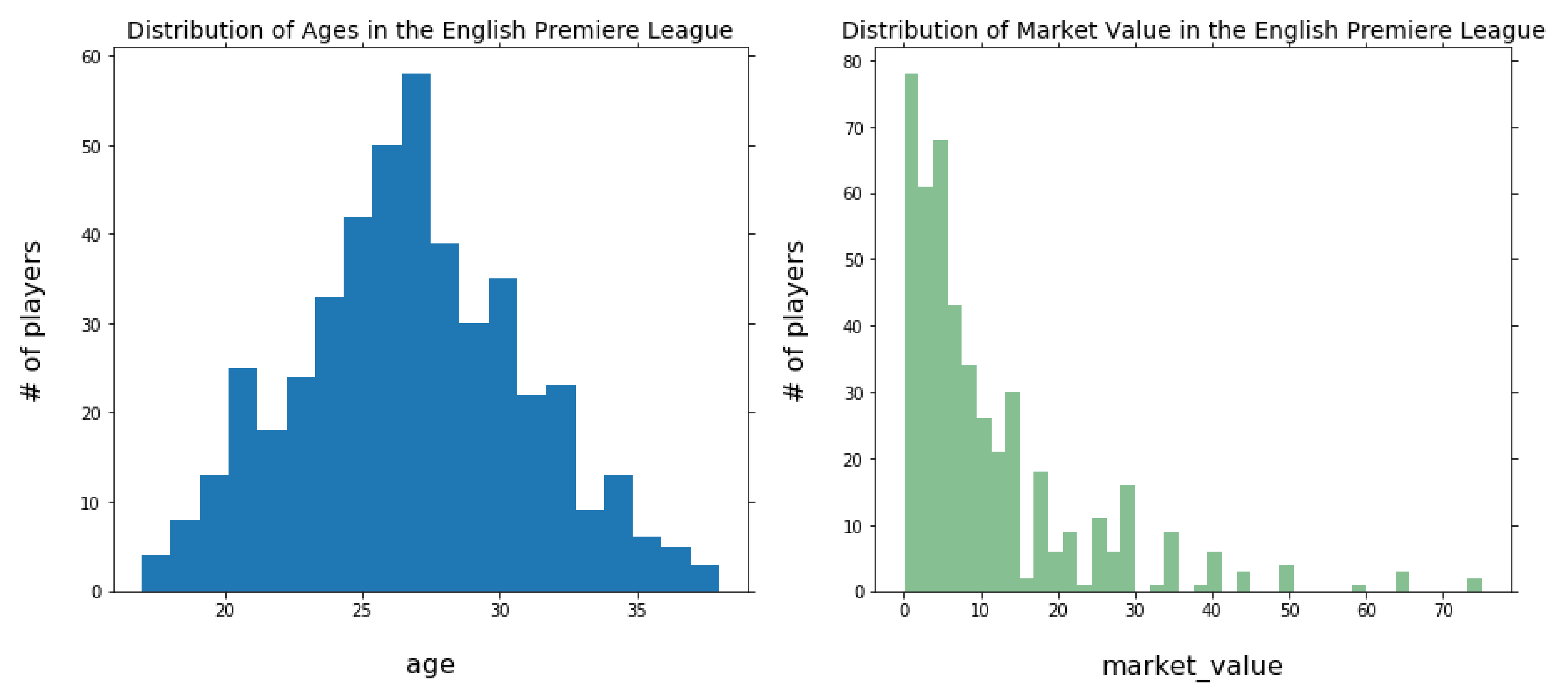


Visualization: The English Premier League



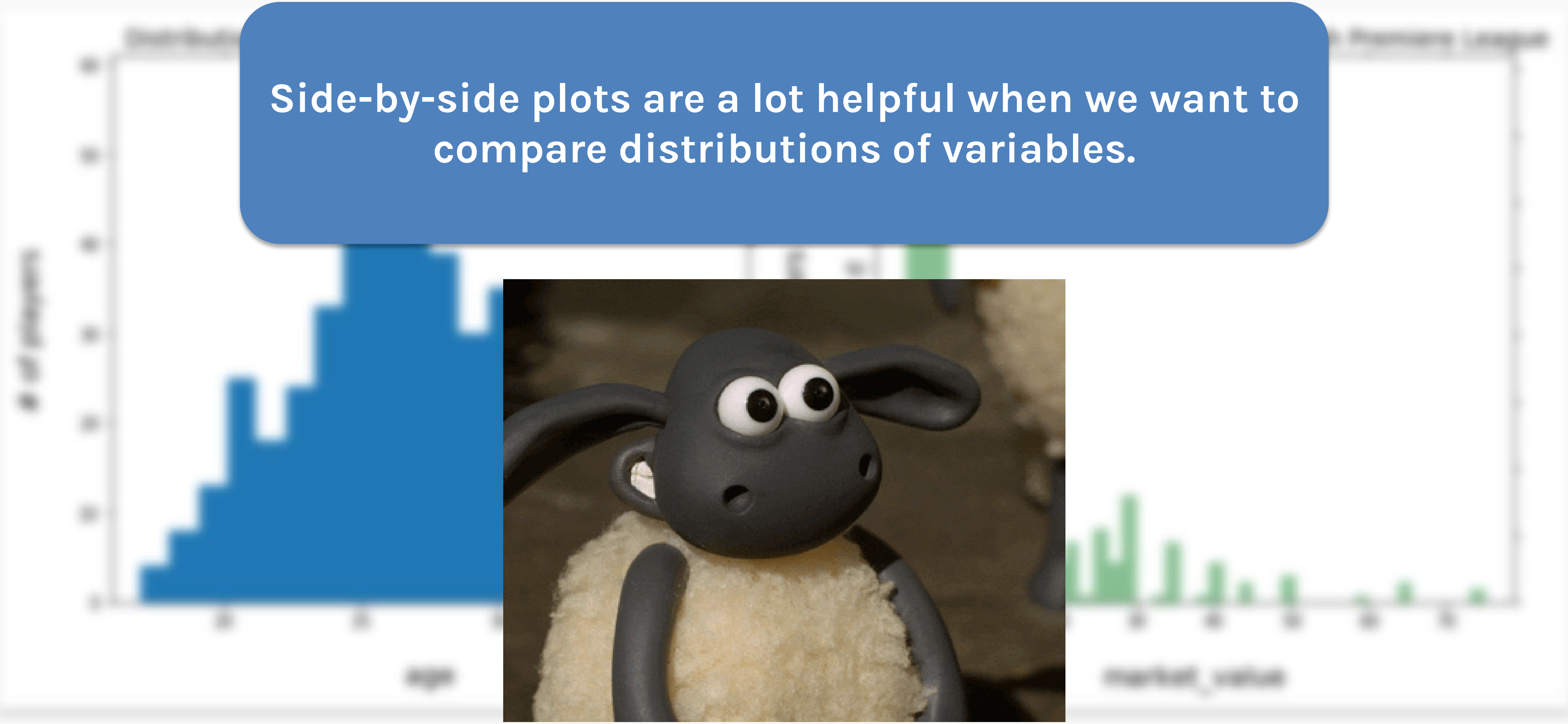
We can rotate the x-tick labels for enhanced readability.

Visualization: The English Premier League

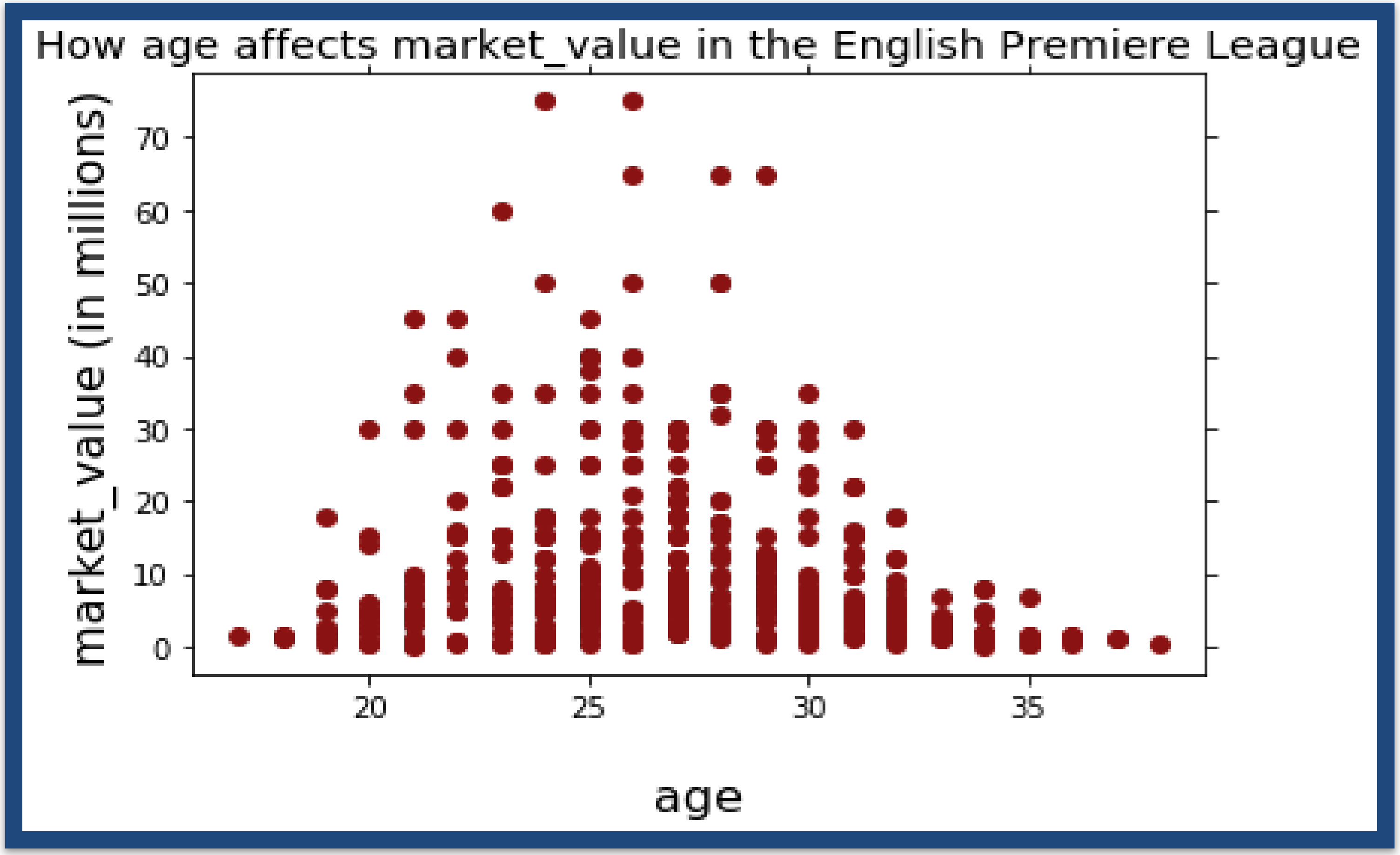


Visualization: The English Premier League

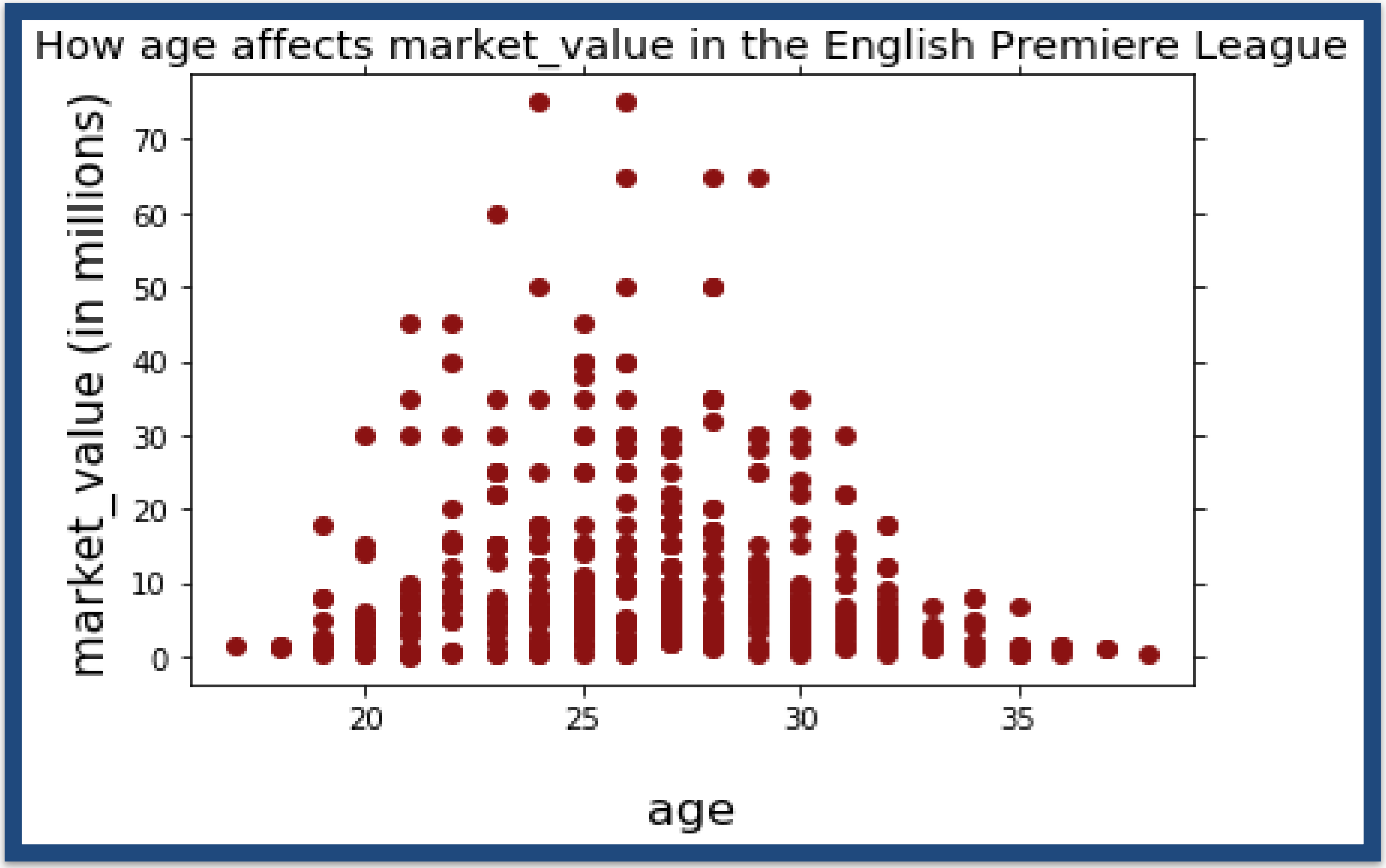
Side-by-side plots are a lot helpful when we want to compare distributions of variables.



Visualization: The English Premier League



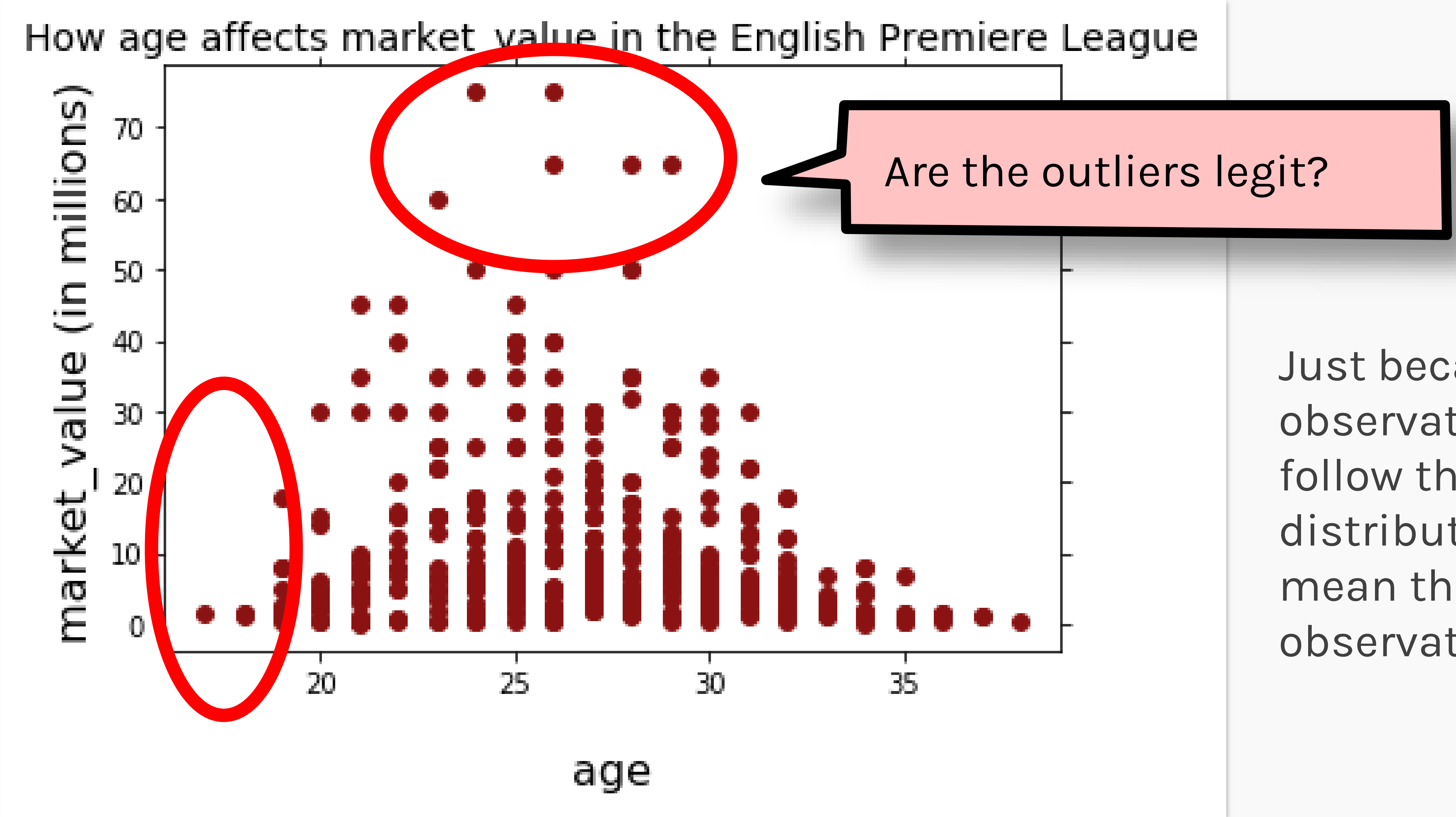
Visualization: The English Premier League



What do see in this graph? Is this sensible enough?

Are there any outliers?

Visualization: The English Premier League



Are the outliers legit?

Just because some observations doesn't follow the usual distribution does not mean that the observation is not real!

Next steps:

- Ensure our data is expected/valid/appropriate for the task.
- Provide insights into the dataset.
- Extract/determine important variables/attributes/features.
- Detect outliers and anomalies.
- Test underlying assumptions.
- Make informed decisions in developing models.

Next steps:

- Ensure our dat [redacted] ate for the task.
- Provide insight
- Extract/determ ibutes/features.
- Detect outliers
- **Test underlying**
- **Make informed dels.**

⚠Fasten your seat belts for the upcoming sessions!