Exploratory Data Analysis

Part B - Visualizing the data : The drill

Pavlos Protopapas

Visualization

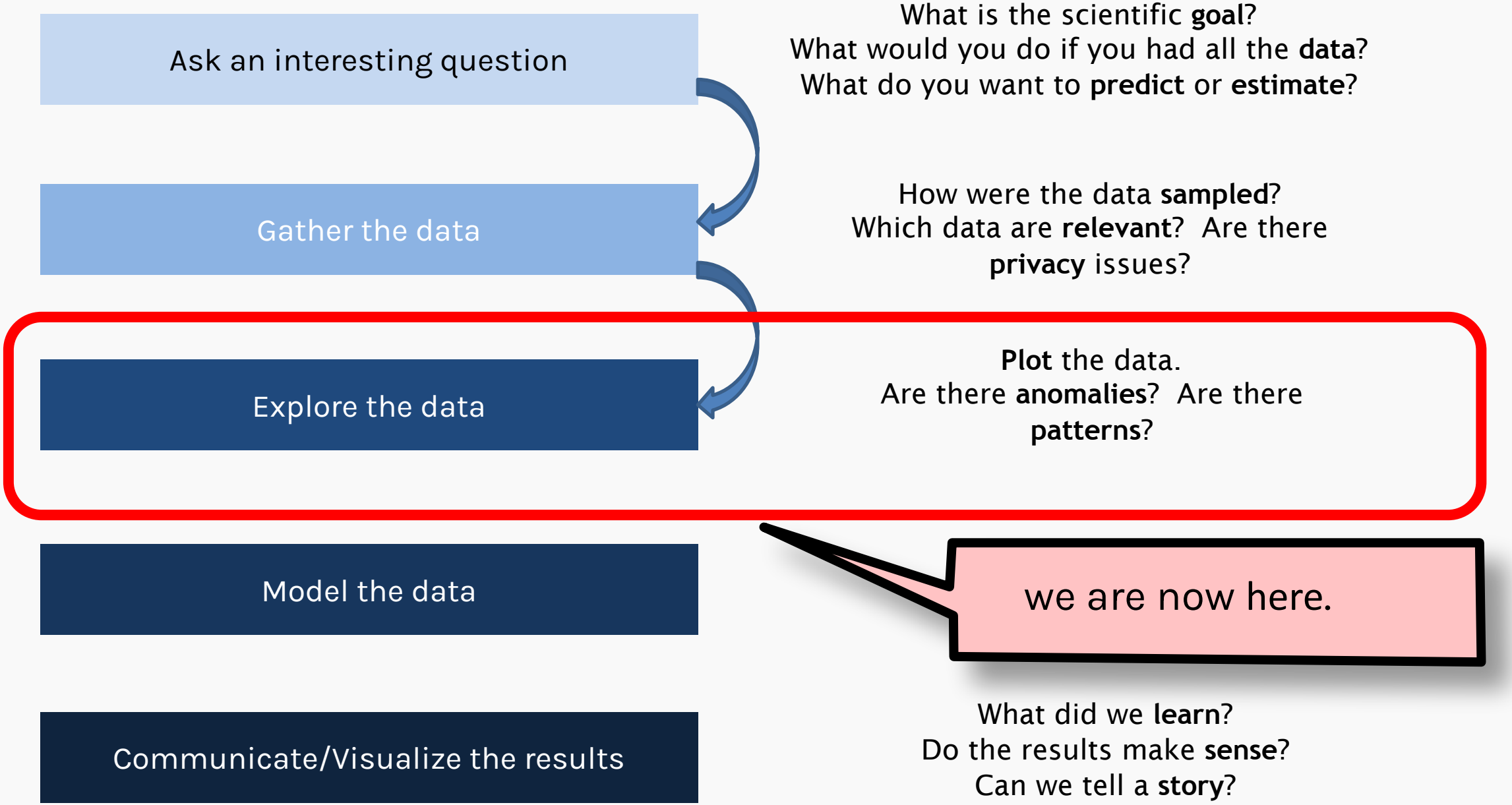
Visualization is incredibly important, both for EDA and for communicating our results to others.

"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.)

Let’s recall the data science process first!

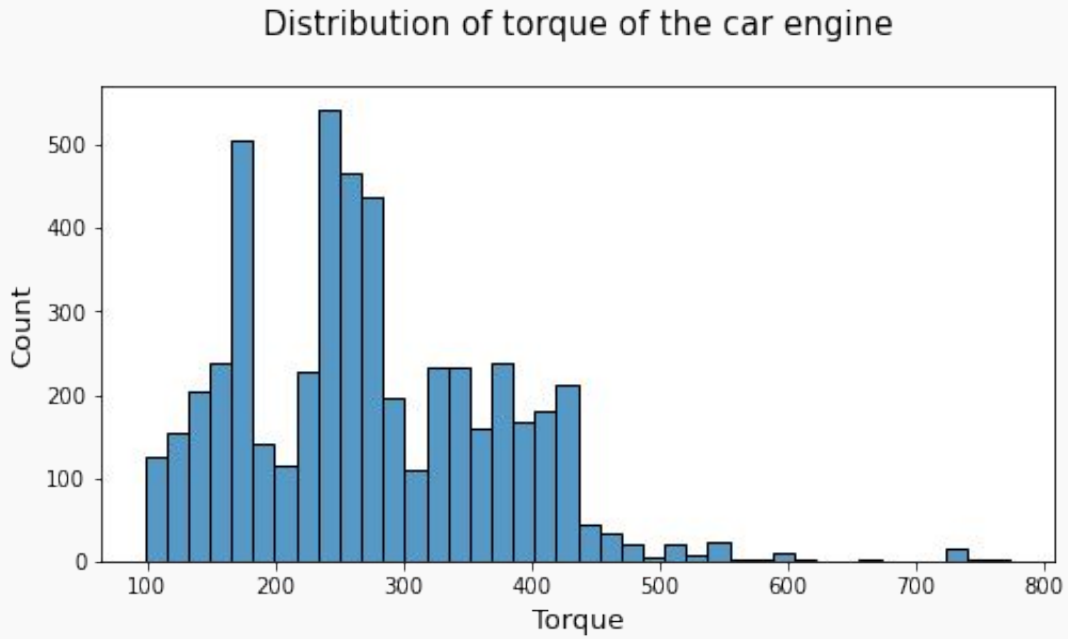

Ask an interesting question
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?
Gather the data
How were the data **sampled**?
Which data are **relevant**? Are there **privacy** issues?
Explore the data
Plot the data.
Are there **anomalies**? Are there **patterns**?
Model the data
we are now here.
Communicate/Visualize the results
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Graphical EDA, aka Visualization

Univariate graphical EDA

Summary statistics does not provide the full picture therefore graphical methods are required

Histograms



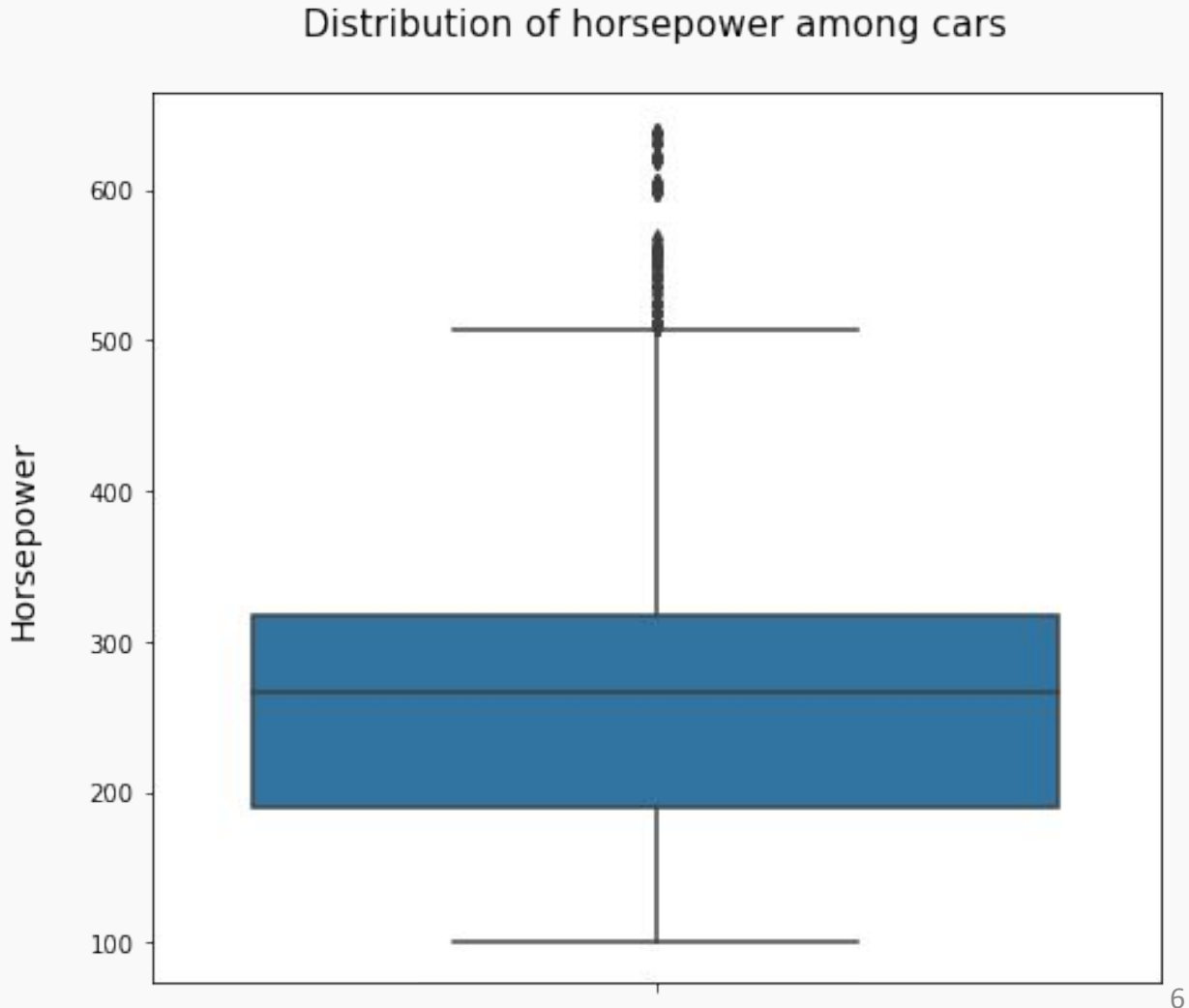
Histograms are one of the best ways to quickly learn a lot about the variable, including central tendency, spread, modality, shape and outliers.

Univariate graphical EDA

Summary statistics does not provide the full picture therefore graphical methods are required

Box plots

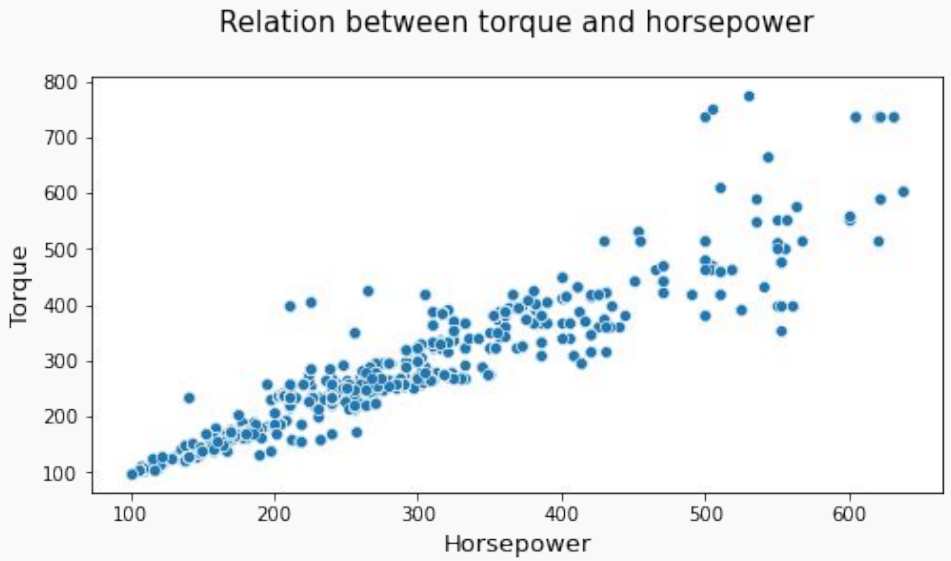
Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers.



Multivariate graphical EDA

There are several ways to graphically represent a relation between two or more variables

Scatter plot

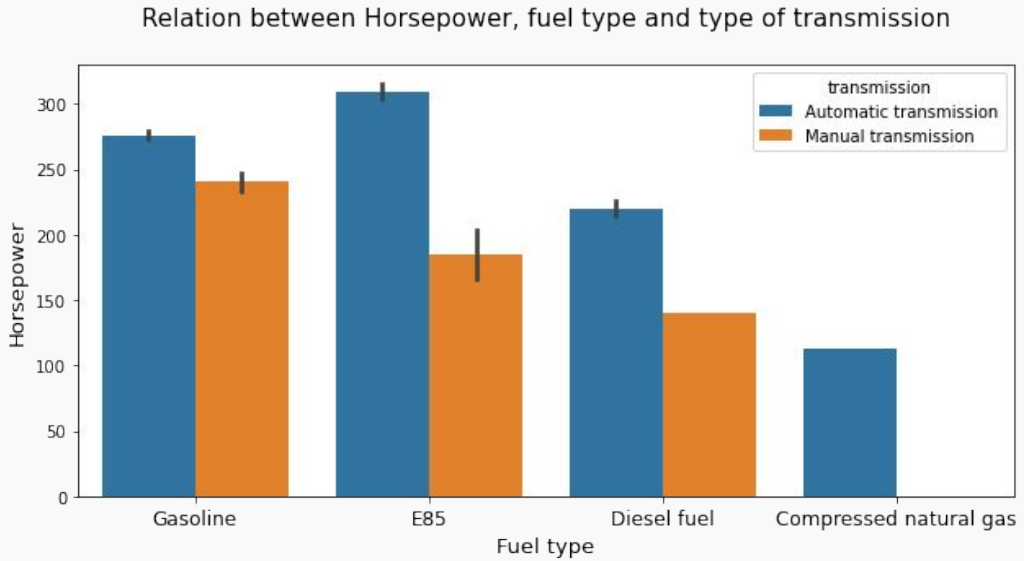


Scatter plot is the best graphical representation to show relation between two quantitative variables

Multivariate graphical EDA

There are several ways to graphically represent a relation between two or more variables

Grouped barplot



Grouped barplots helps us understand relation between three variables when two of them are categorical.

What's the need to visualize?

Anscombe's Quartet

Set A		Set B		Set C		Set D	
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	12	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	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!

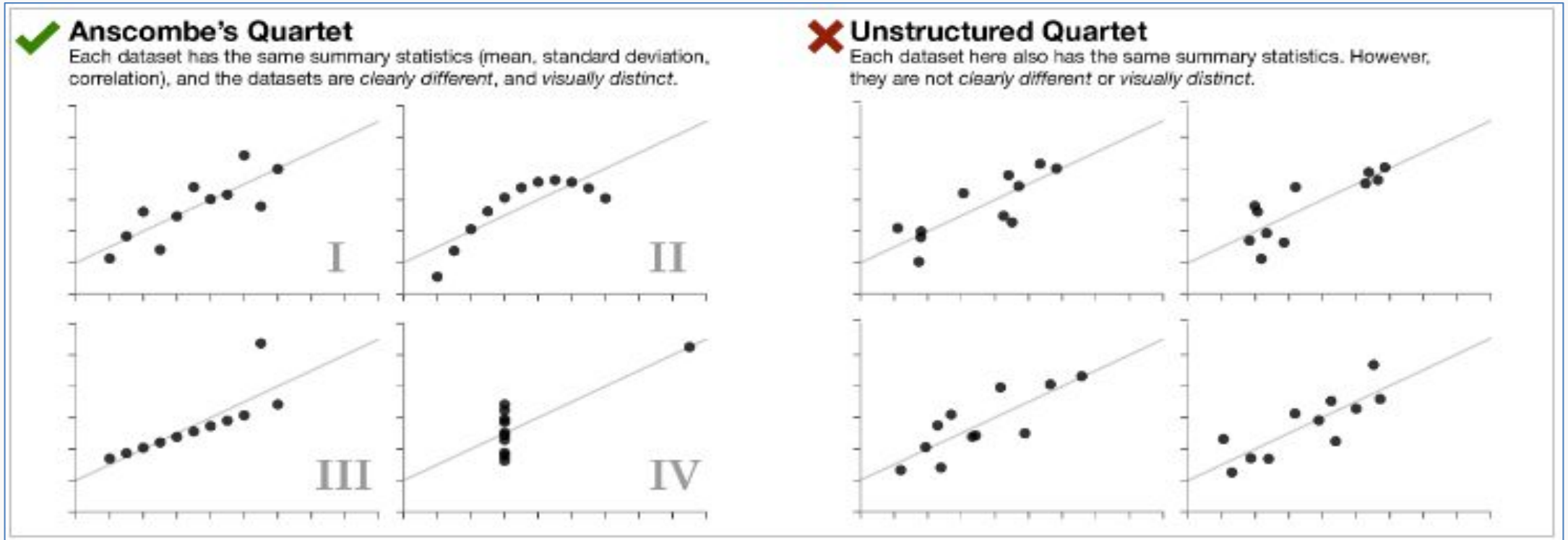


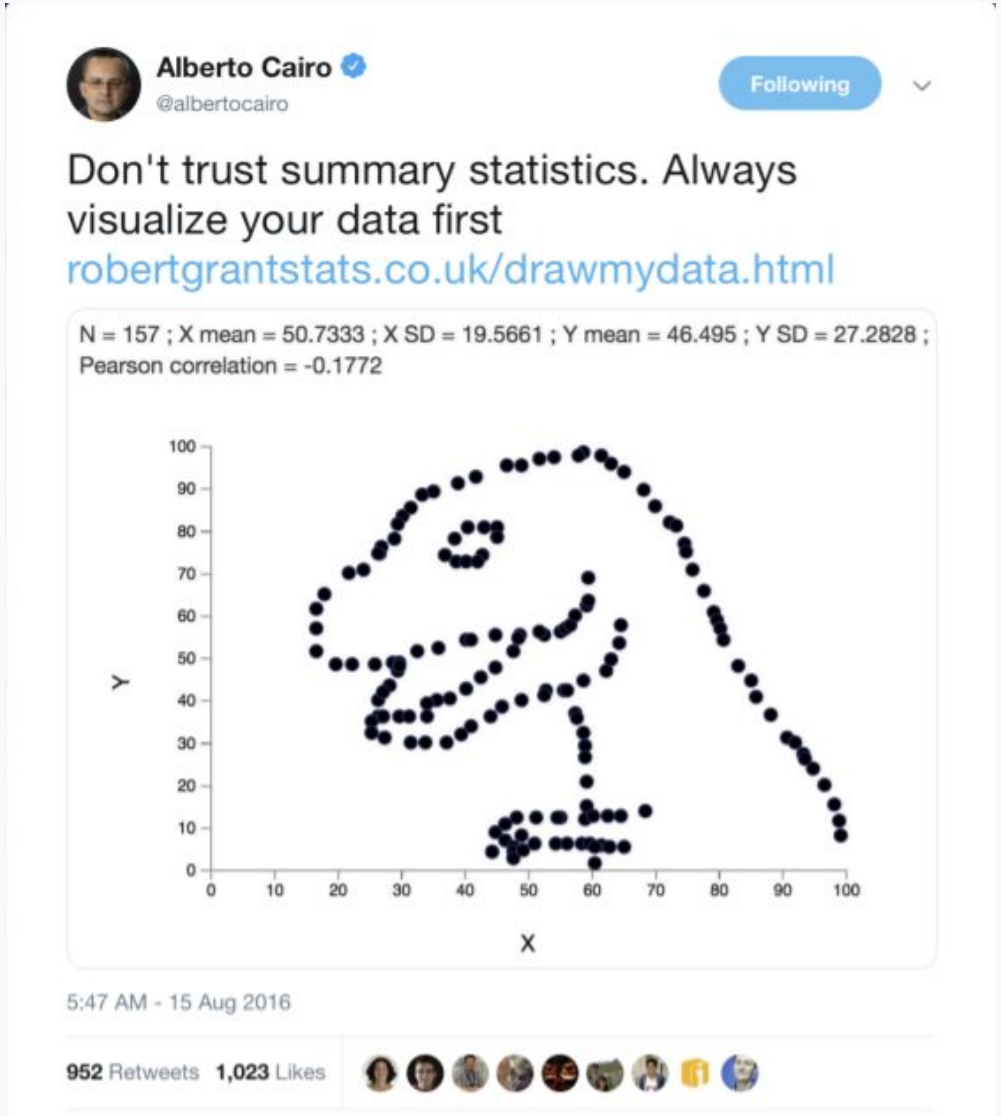
Fig 1. Anscombe's Quartet (left), and a "Unstructured Quartet" on the right, where the datasets have the same summary statistics as those in Anscombe's Quartet, but lack underlying structure or visual distinction.

Same stats do not imply same graphs

Same graphs do not imply same stats

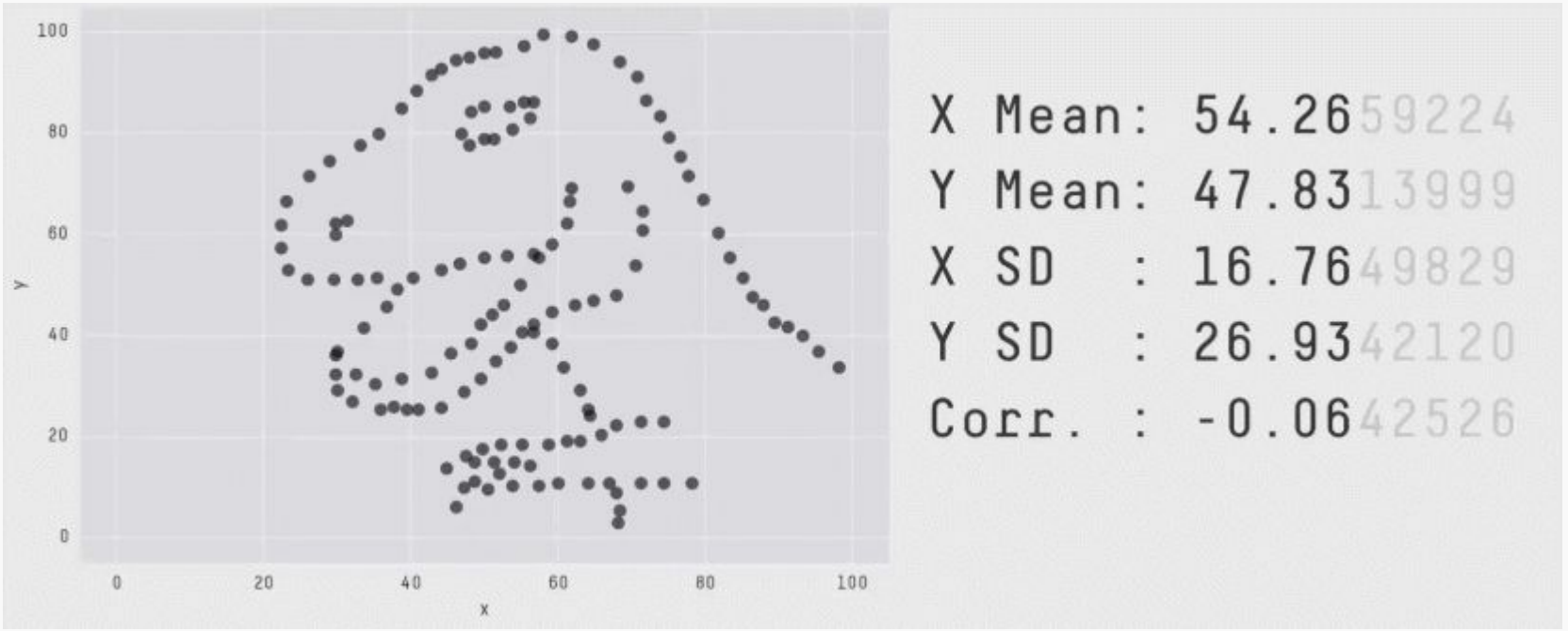
Credits: Same Stats, Different Graphs. (Autodesk Research)

Same stats, different graphs!



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.



THE DATASAURUS DOZEN

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

Visualization Goals

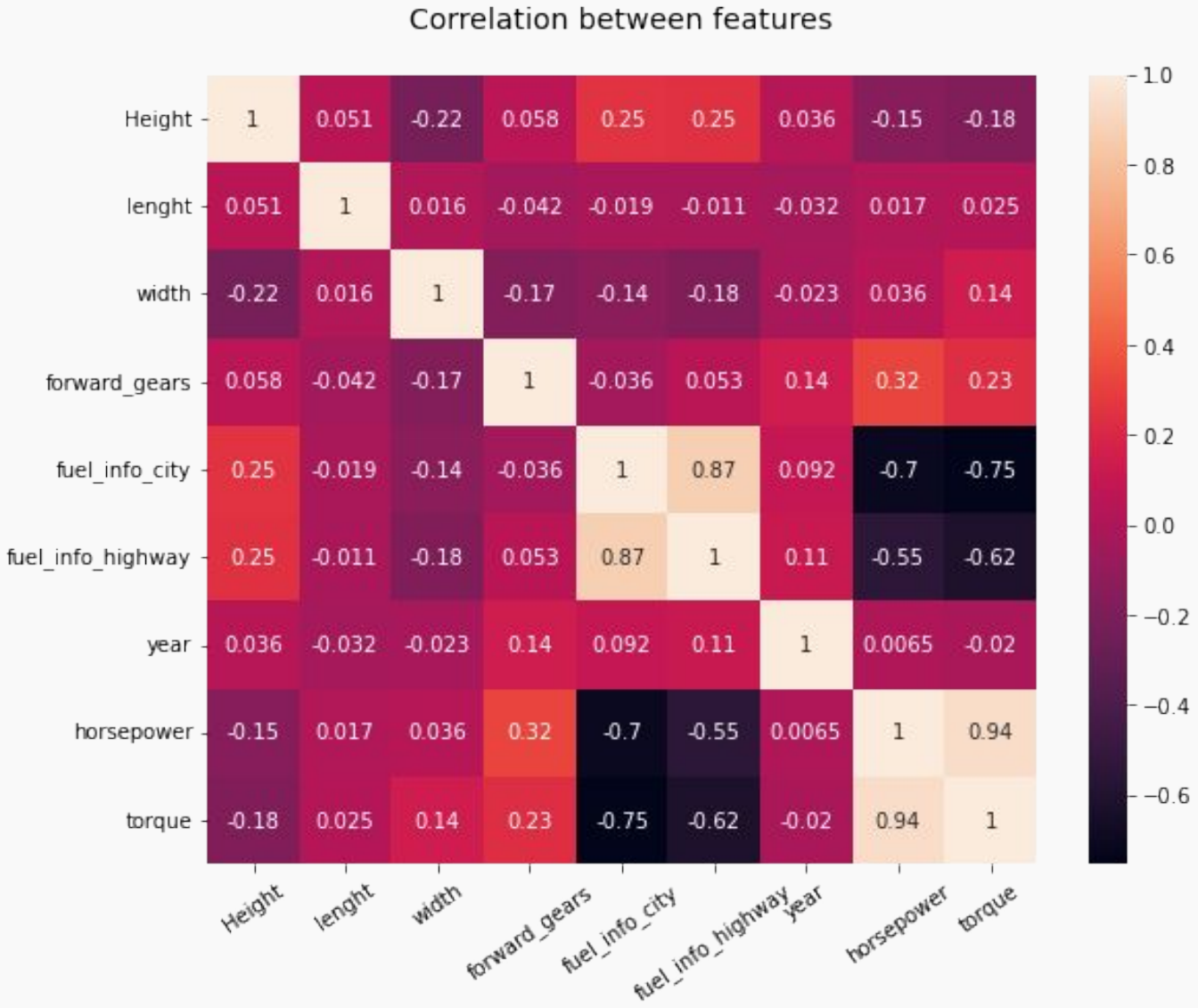
Analyze (Exploratory)

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 while building a classification model to predict customer satisfaction using flight and facilities data.



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

Visualization: The English Premier League

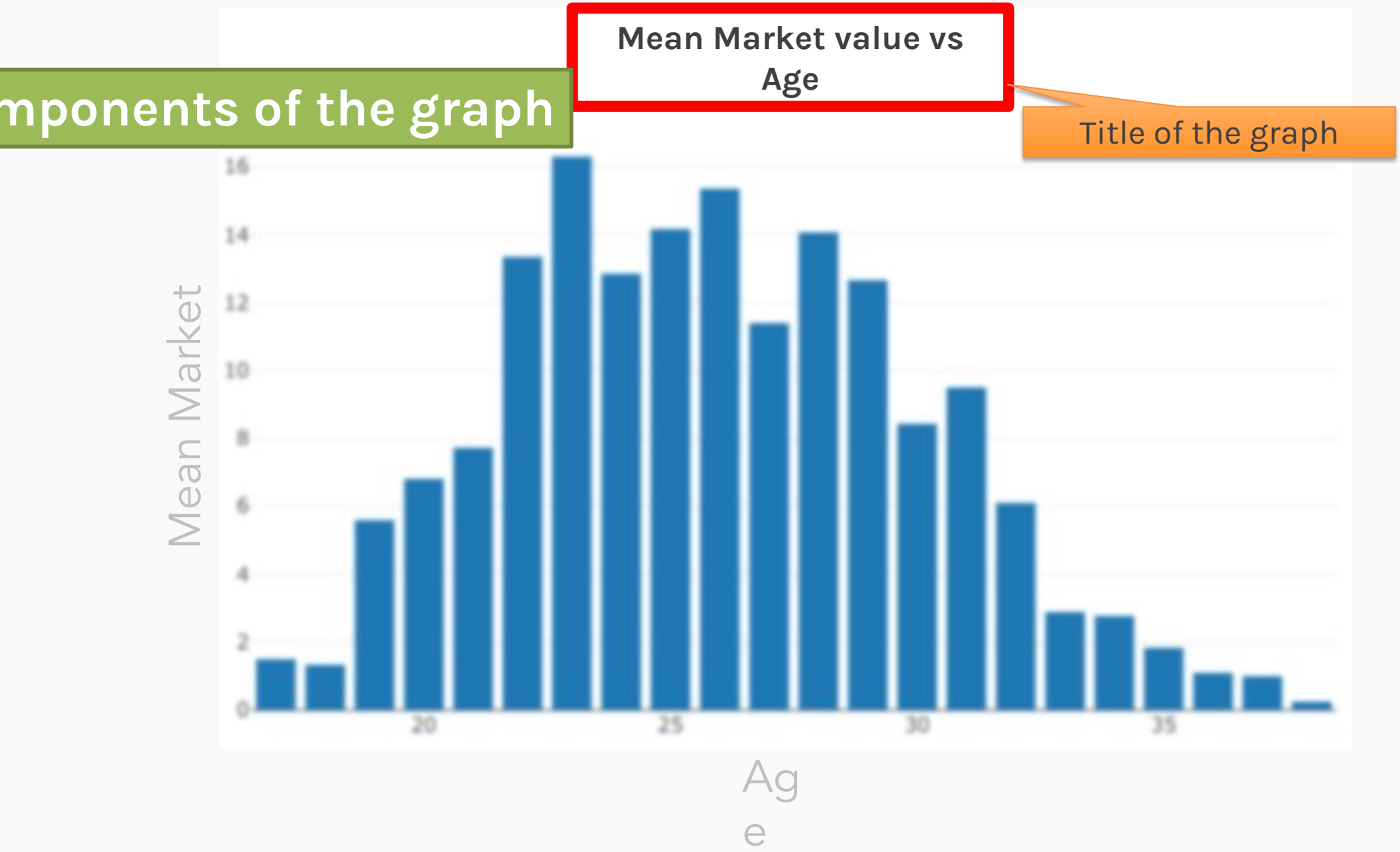
Components of the graph

Visualization: The English Premier League

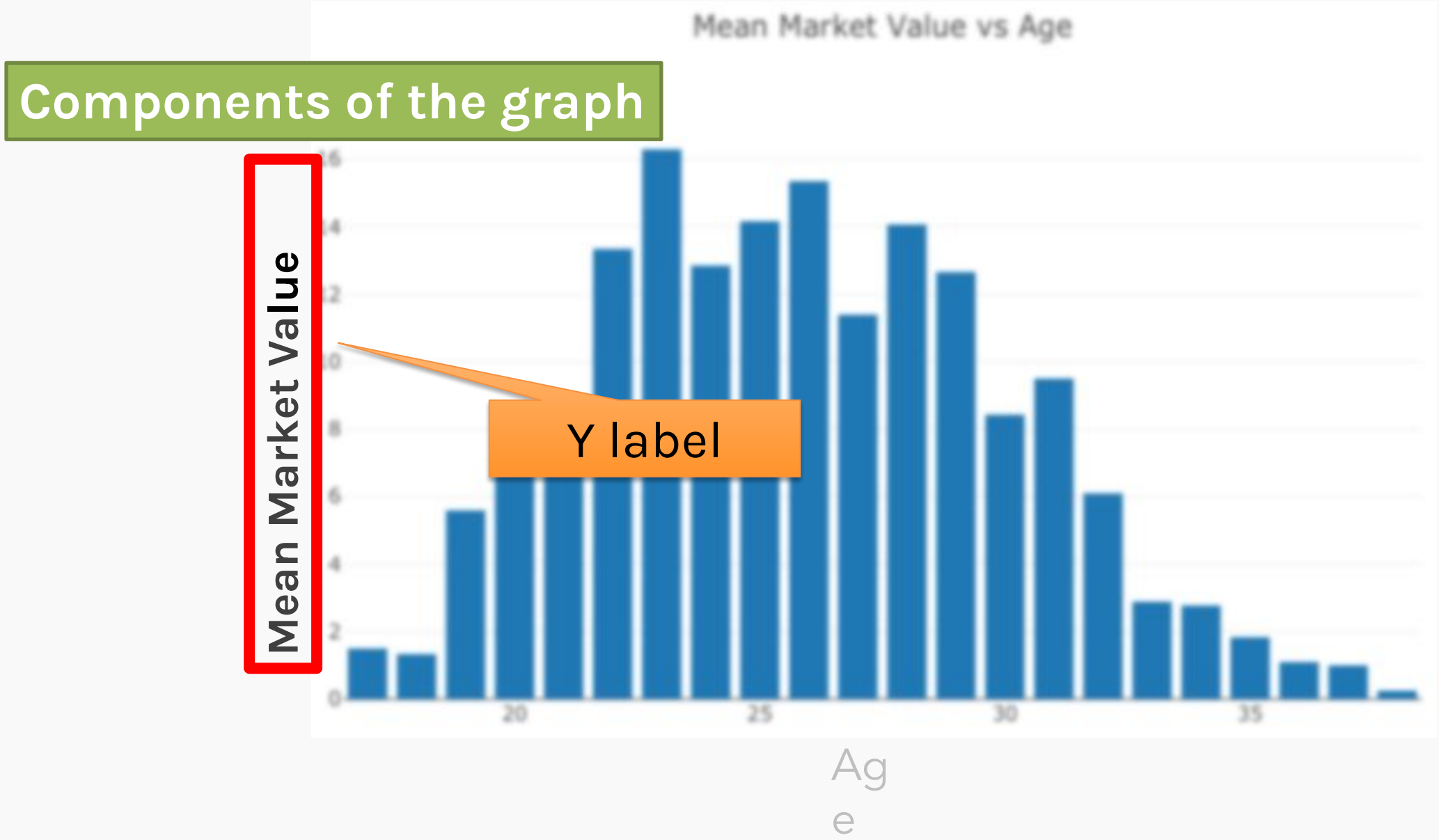
Components of the graph

Visualization: The English Premier League

Components of the graph

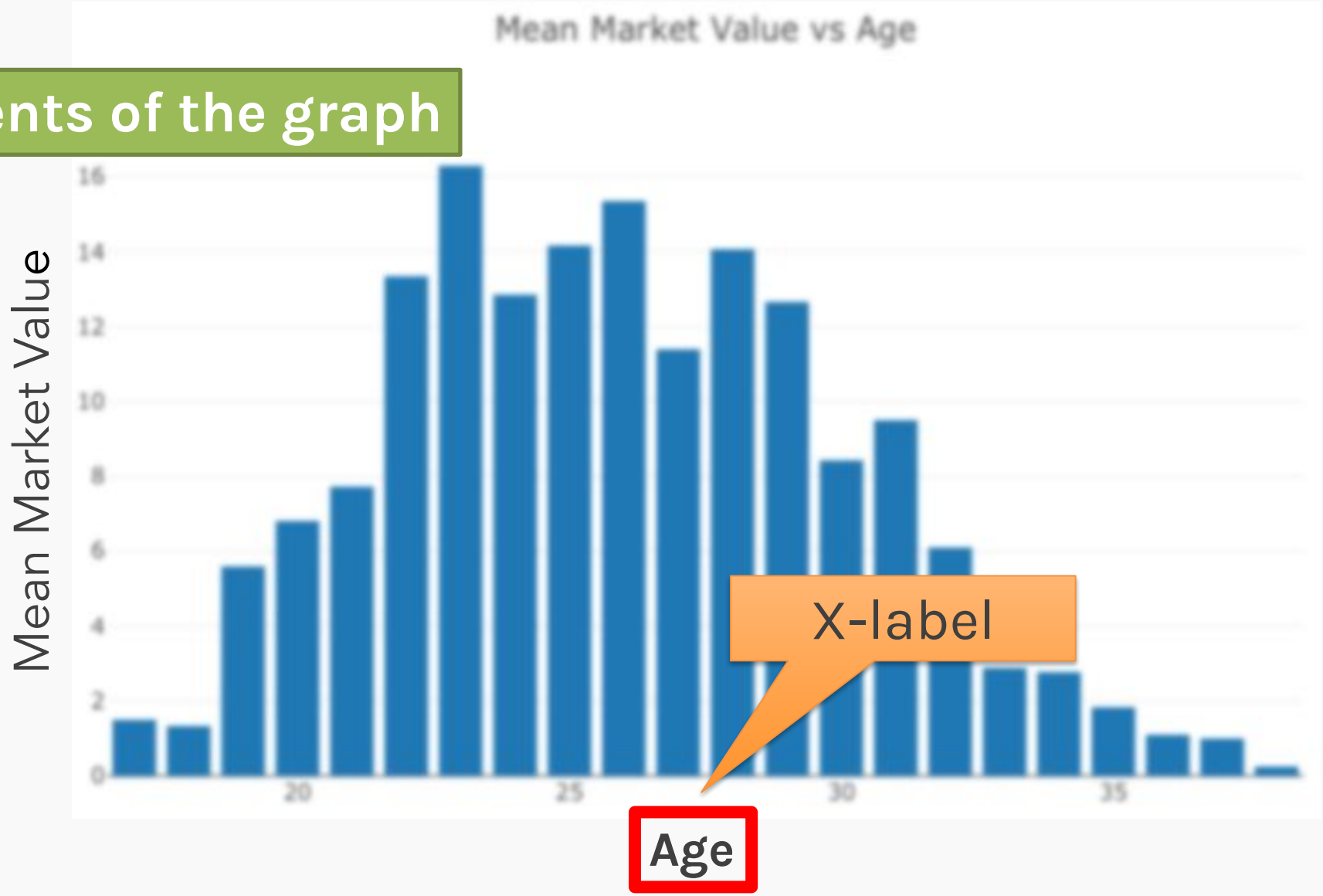


Visualization: The English Premier League

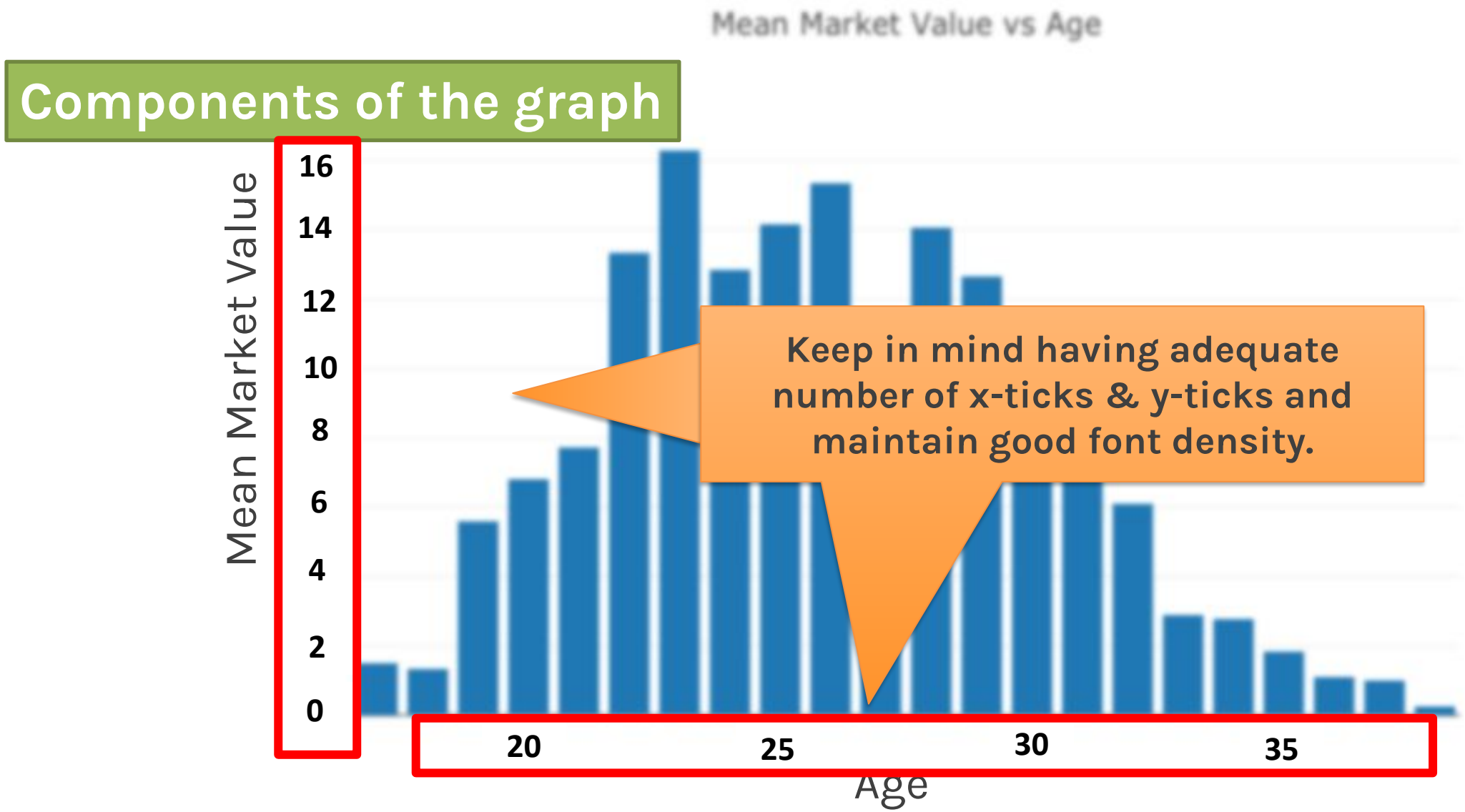


Visualization: The English Premier League

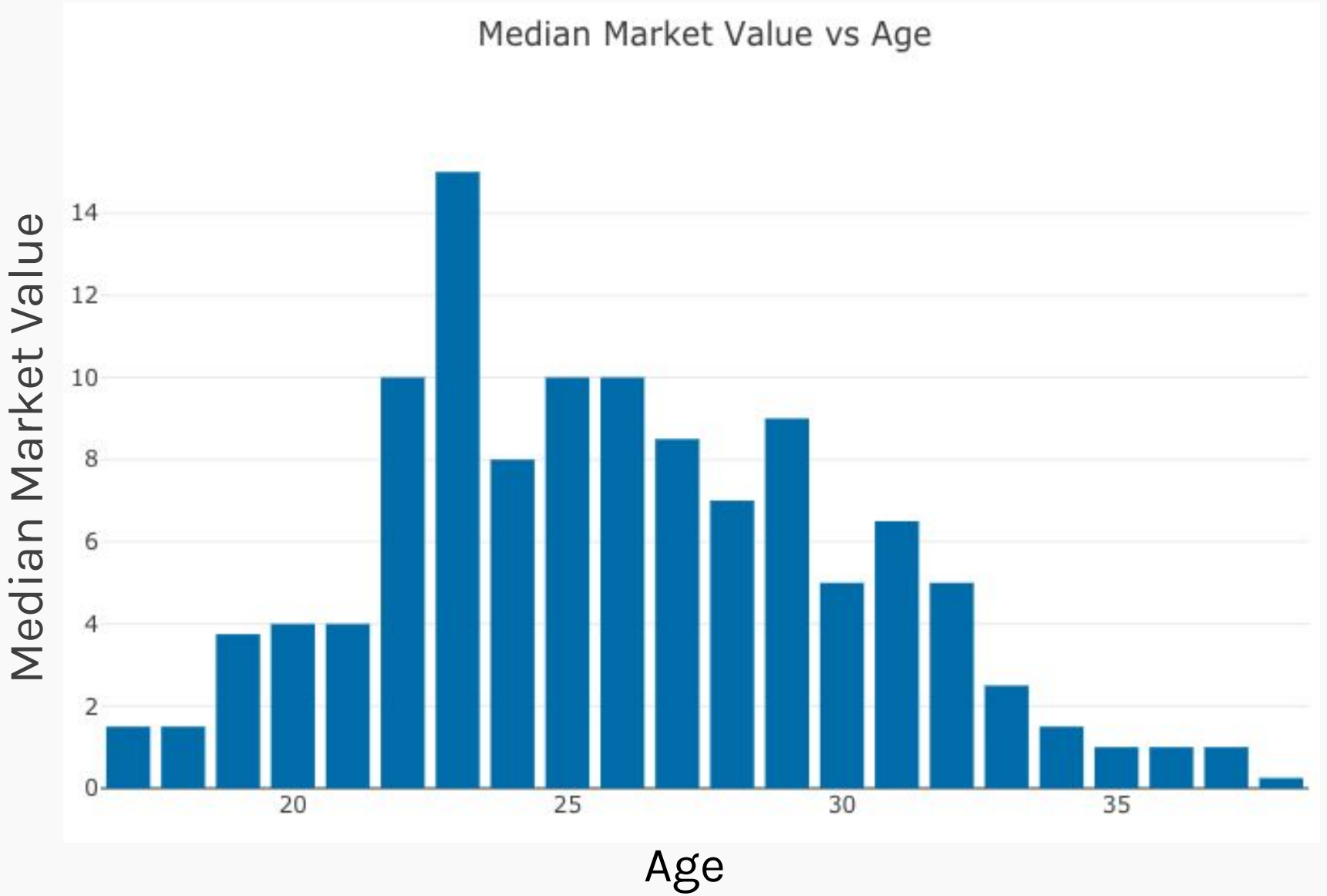
Components of the graph



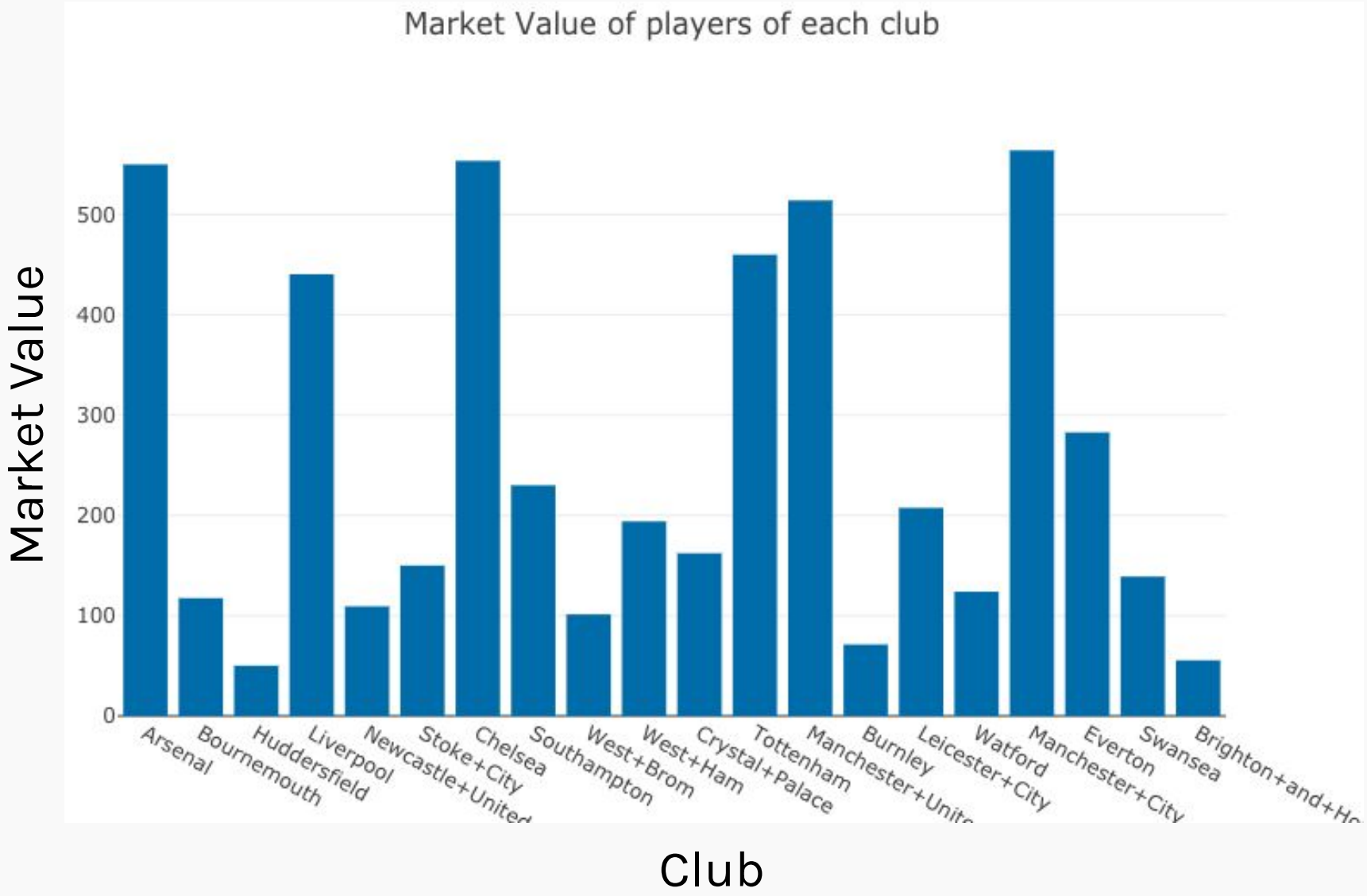
Visualization: The English Premier League



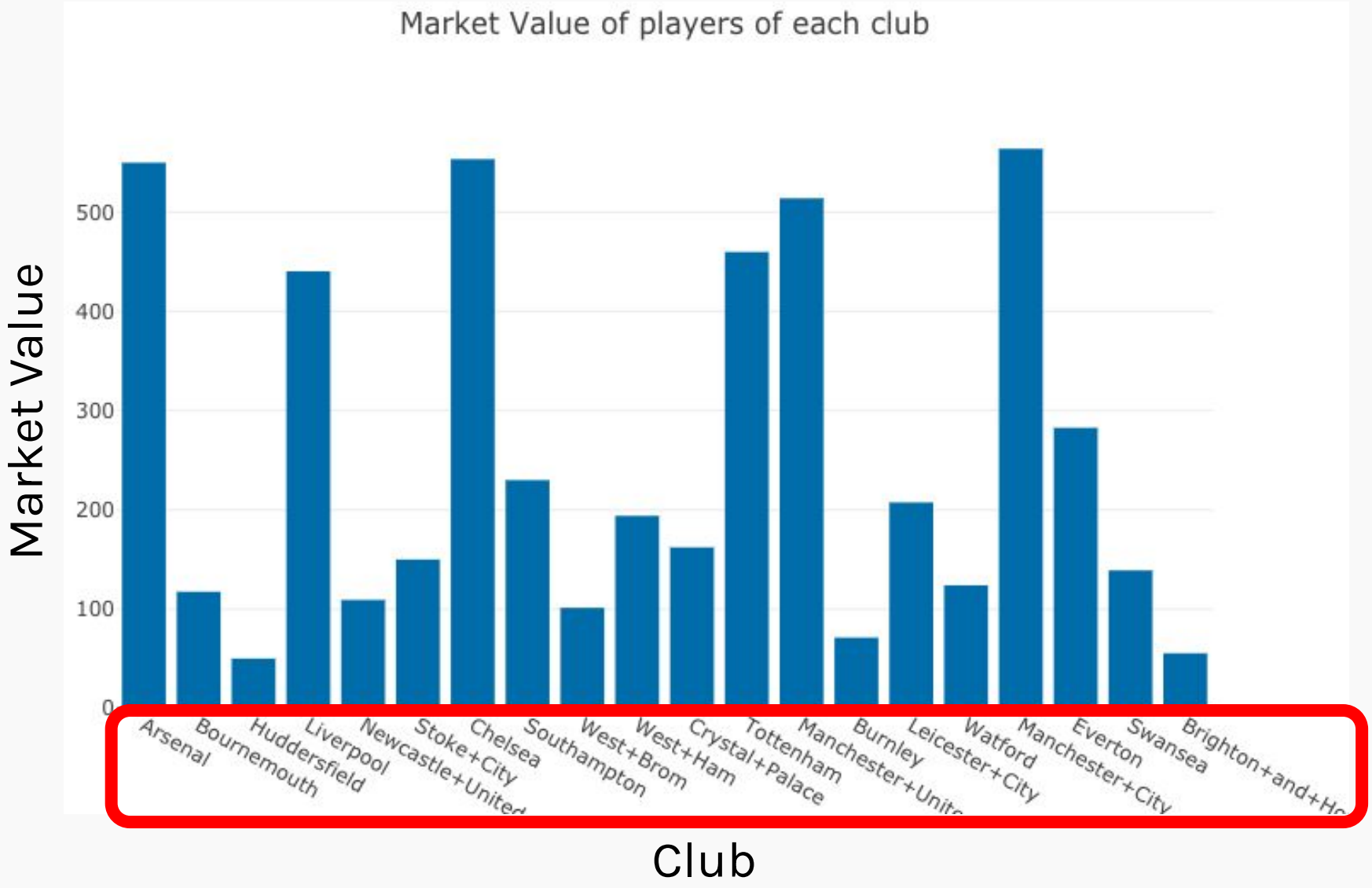
Visualization: The English Premier League



Visualization: The English Premier League

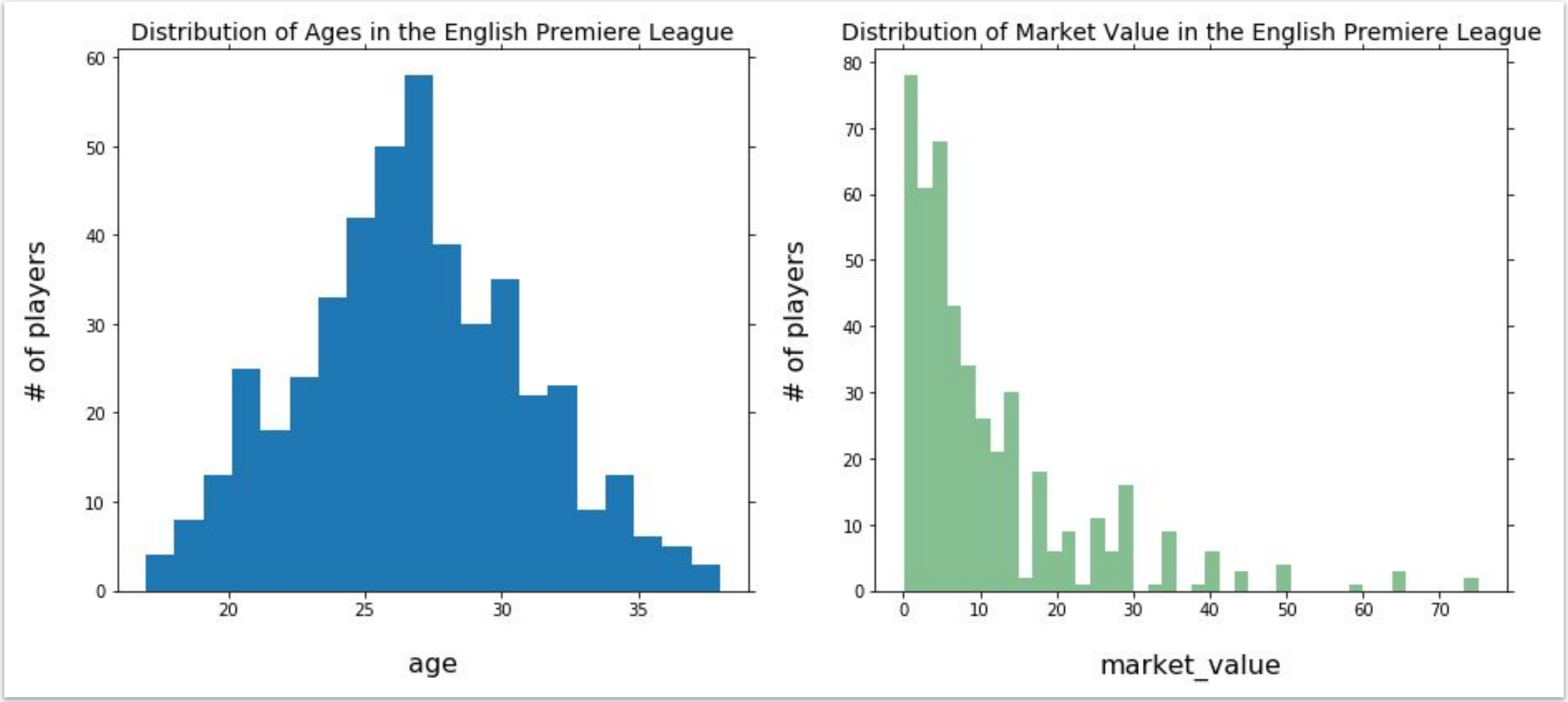


Visualization: The English Premier League

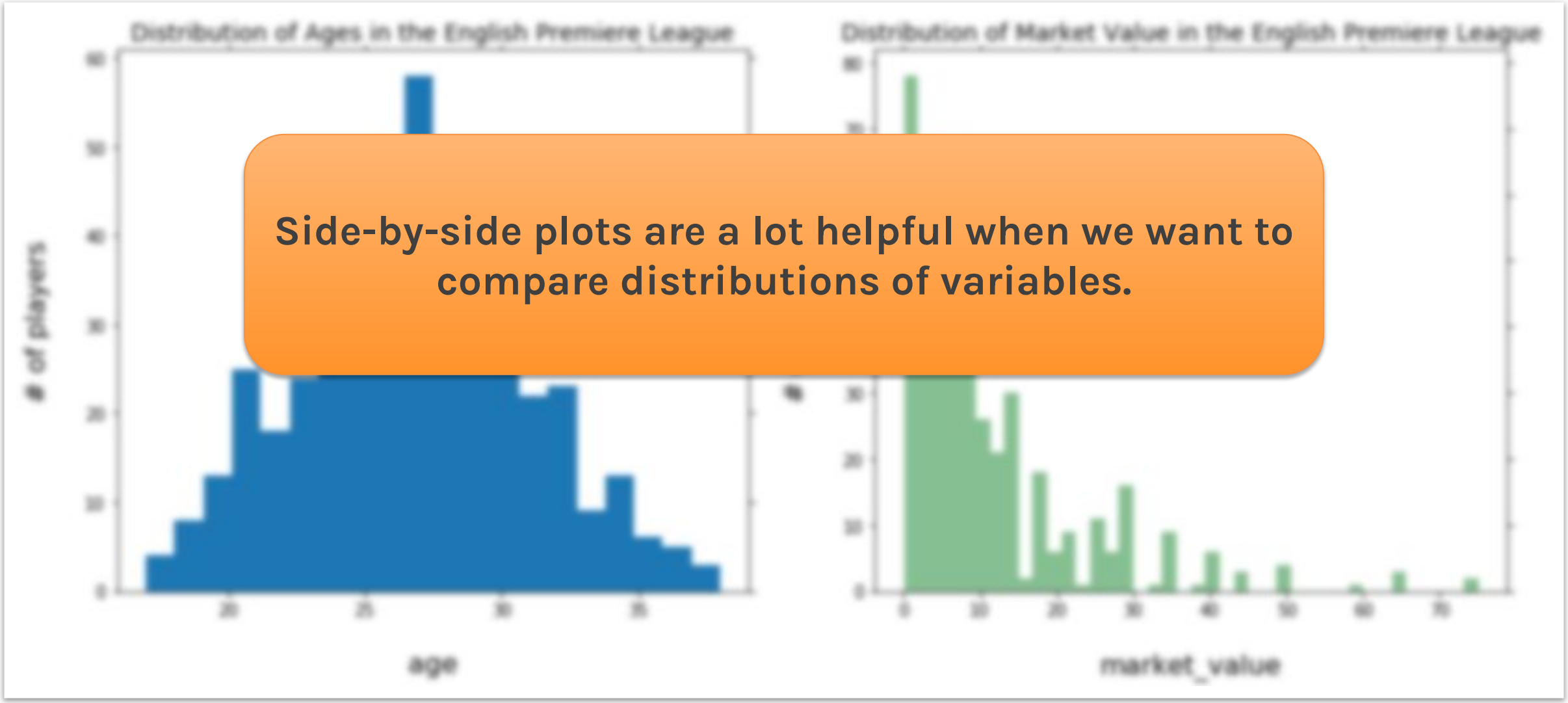


Visualization: The English Premier League

Visualization: The English Premier League

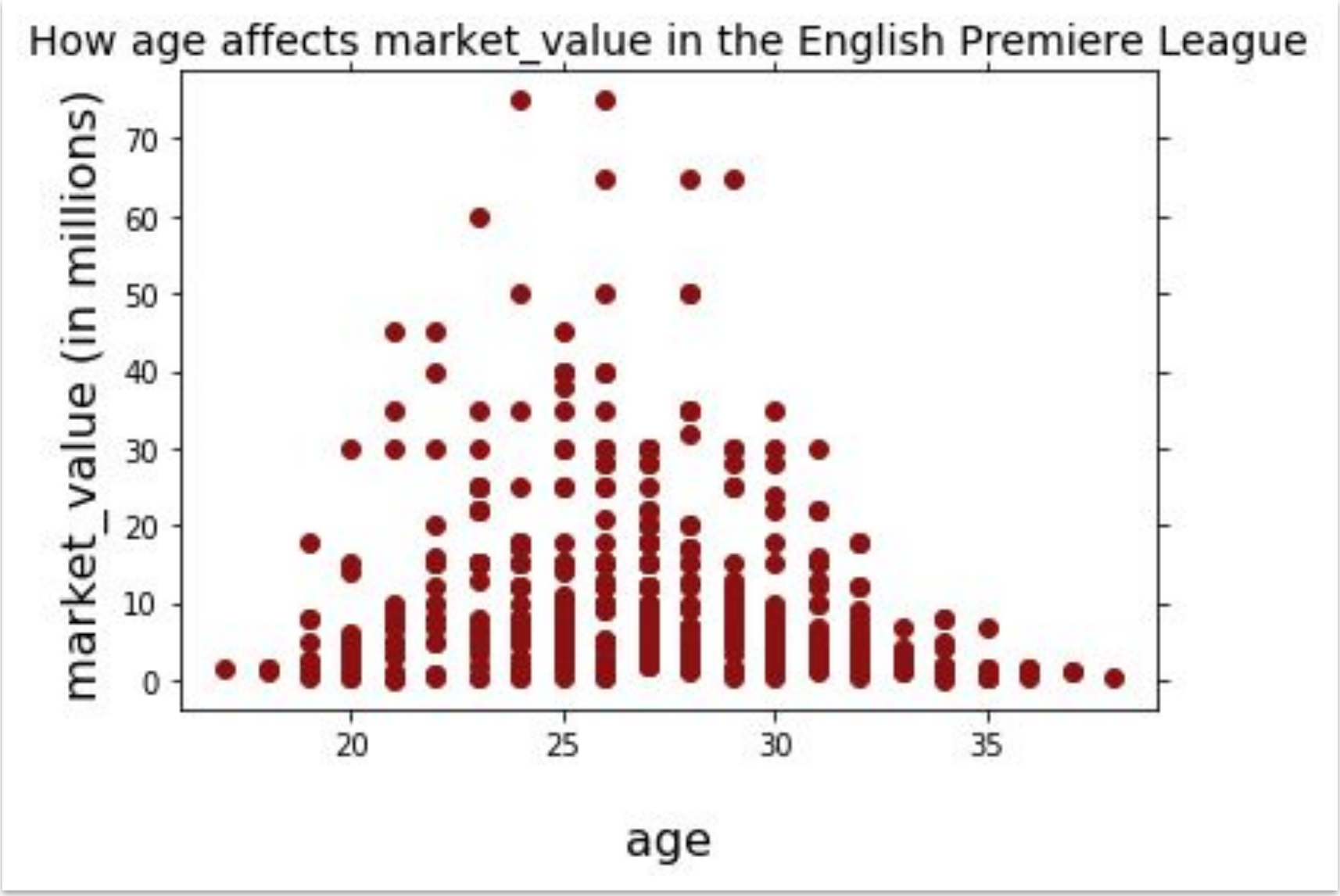


Visualization: The English Premier League

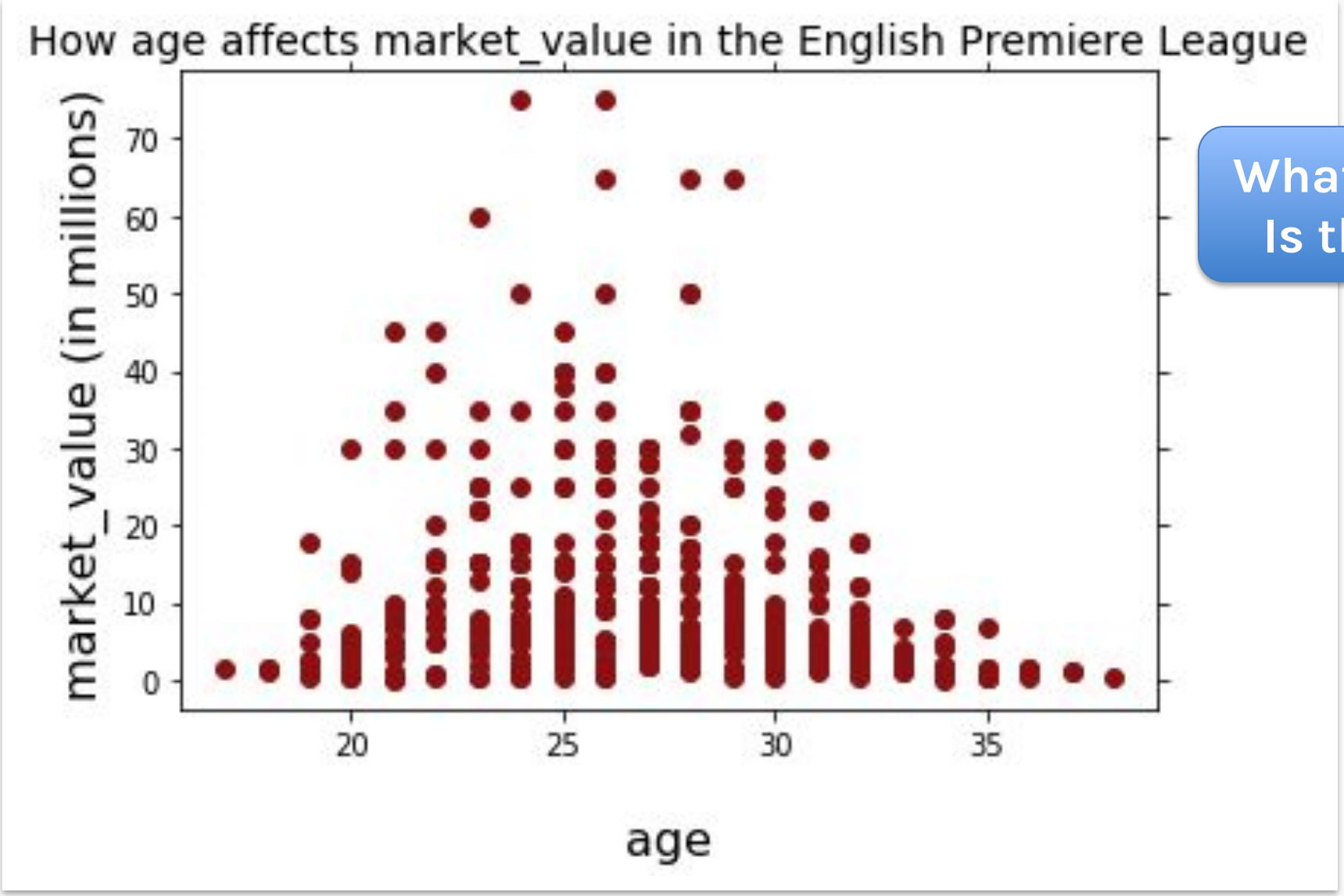


Side-by-side plots are a lot helpful when we want to compare distributions of variables.

Visualization: The English Premier League

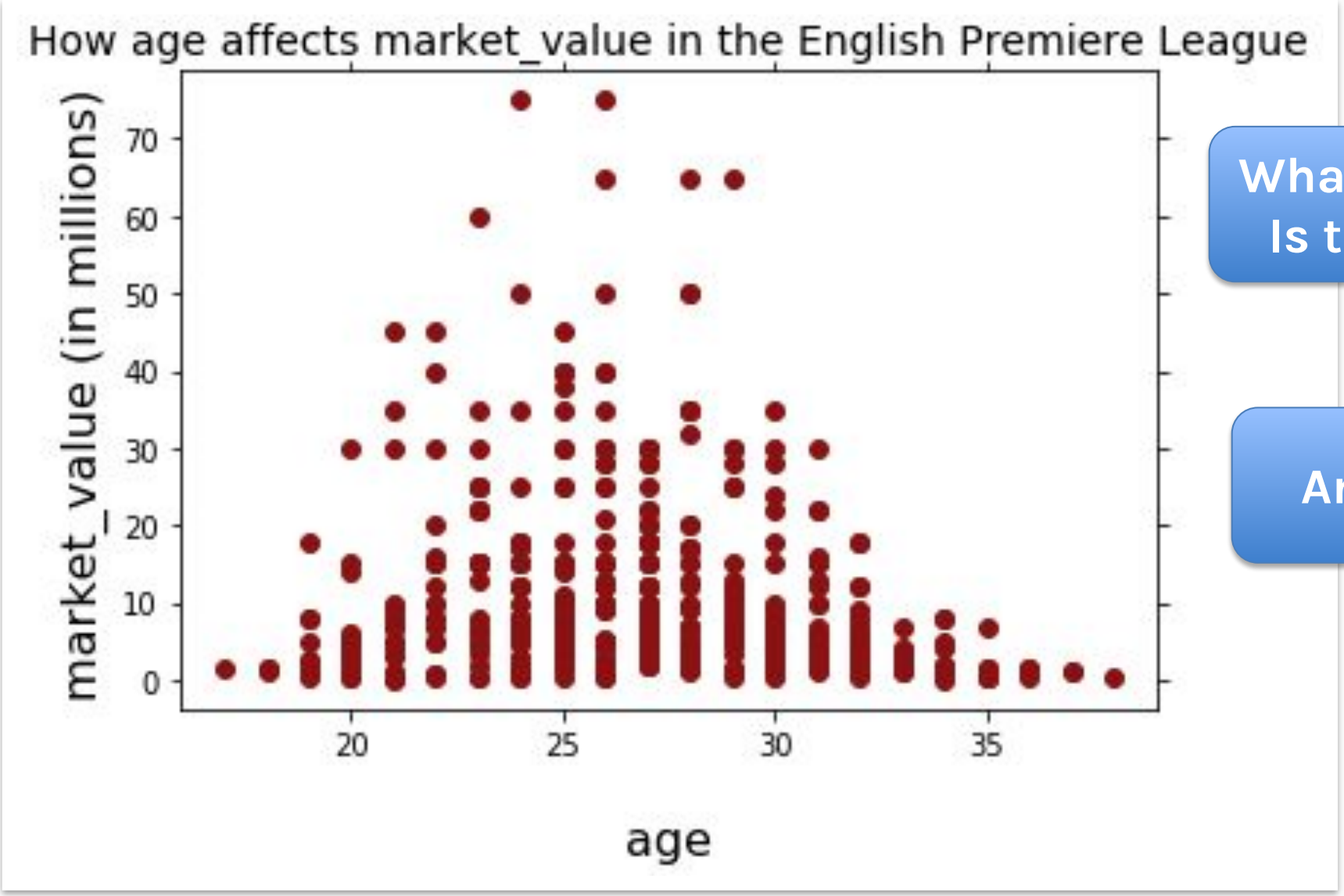


Visualization: The English Premier League



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

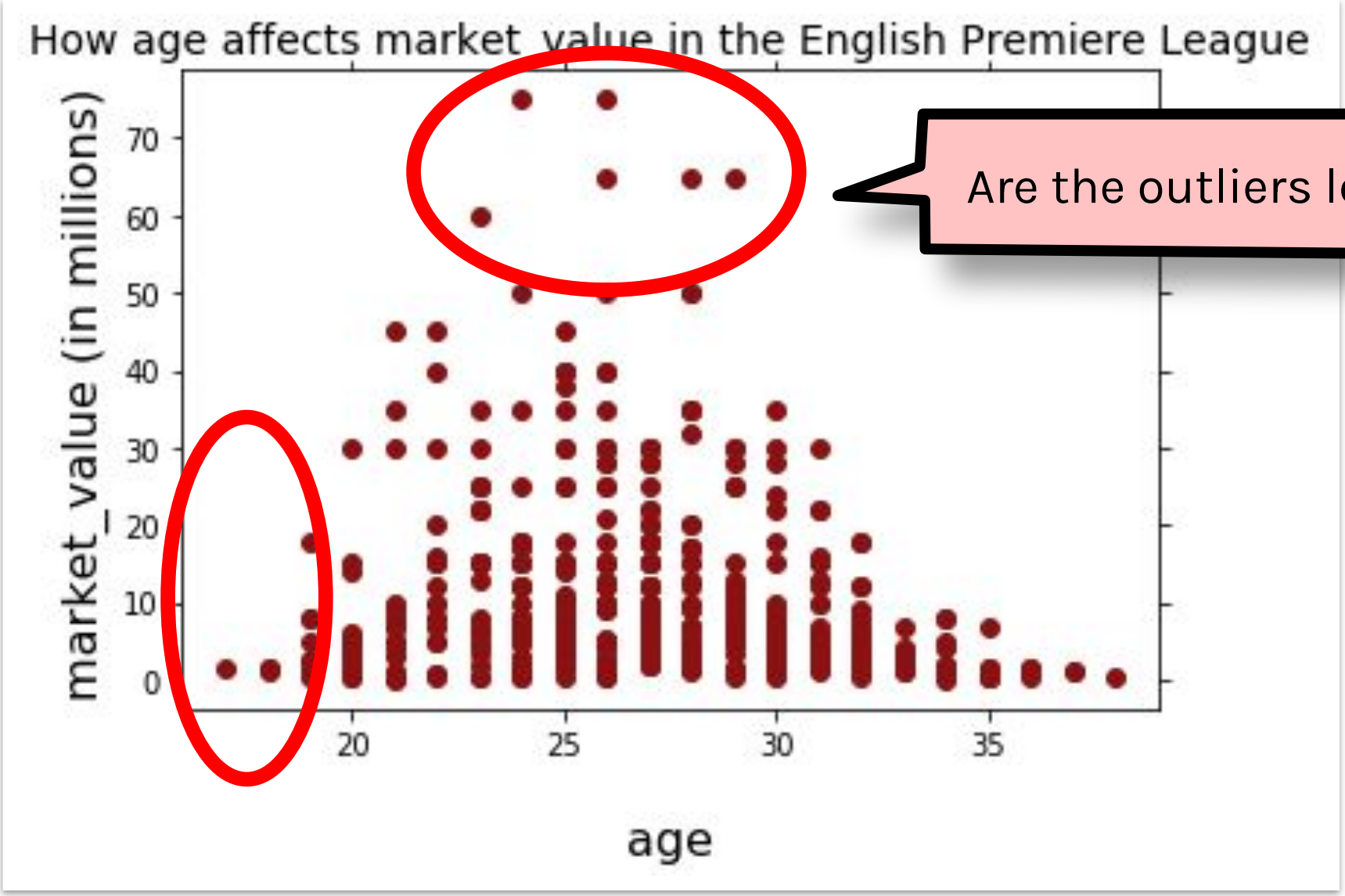
Visualization: The English Premier League



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

Are there any outliers?

Visualization: The English Premier League

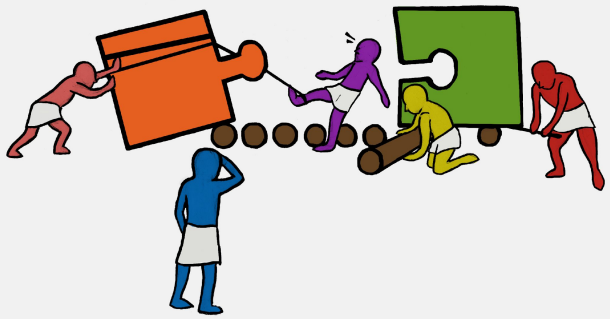


Just because some observations doesn't follow the usual distribution does not mean that the observation is not real

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.

⚠️ **Fasten your seat belts for the upcoming sessions!**



👨‍🏫 Exercise B.1: Effective Visualization (Part A)

The aim of this exercise is to visualize Rossmann Kaggle dataset, you will try different plots to take a look at the data. This is an instructor-led exercise!