Data cleaning and transformation

Stephen F Elston | Principle Consultant , Quantia Analytics, LLC

Chapter Overview

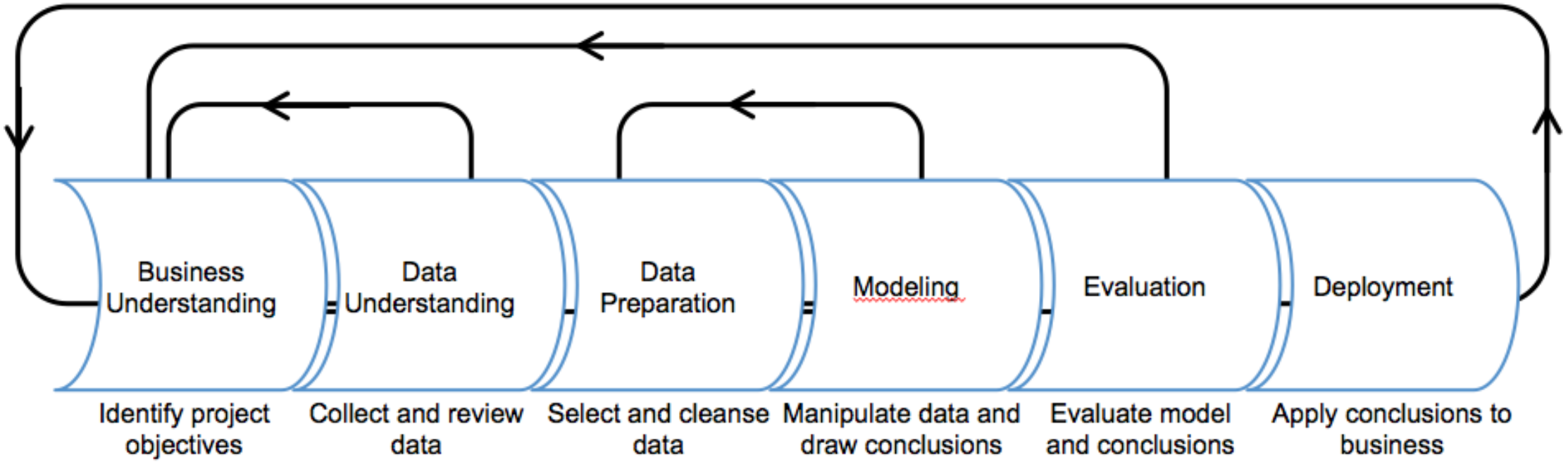
- Data preparation process
- Missing and repeated values
- Outliers and errors
- Scaling

Data Cleaning and Transformation Overview (data munging)

- Data rarely arrives in the form needed for analysis
- Data munging is typically the most time consuming part of a data science project
- Is an iterative process
  - Often discovered with visualization
  - Fix modeling problems

Data Cleaning and Transformation Process

Iterative process



Missing and repeated values

Missing and Repeated Values

- Missing values and repeated values are common
- Many ML algorithms don’t deal with missing values
- Repeated values bias results

Missing Values

Col1	Col2	Col3	Col4	Col5
12456	0.99	Male	43	Small
98567	1.23		55	Medium
34567	9999	Female	NA	Large
67231	0.72	Male	35	?

Treating Missing Values

- Remove rows
- Substitute a specific value
- Interpolate values
- Forward fill
- Backward fill
- Impute

Clean Missing and Repeated Values

- Clean Missing Data module
- With R – is.na()
- With Python – pandas.DataFrame.isnull()

Repeated Values

Key Col	Col2	Col3	Col4	Col5
12456	0.99	Male	43	Small
98567	1.23	Male	55	Medium
34567	1.55	Female	43	Large
34567	1.55	Female	43	Large
34567	1.55	Female	43	Large
34567	.78	Male	43	Large
67231	0.72	Male	35	Small

Clean Missing and Repeated Values

- Clean Repeated Values module
- With R – data.frame[!duplicated(),]
- With Python – DataFrame.drop_duplicates()

Cleaning outliers and errors

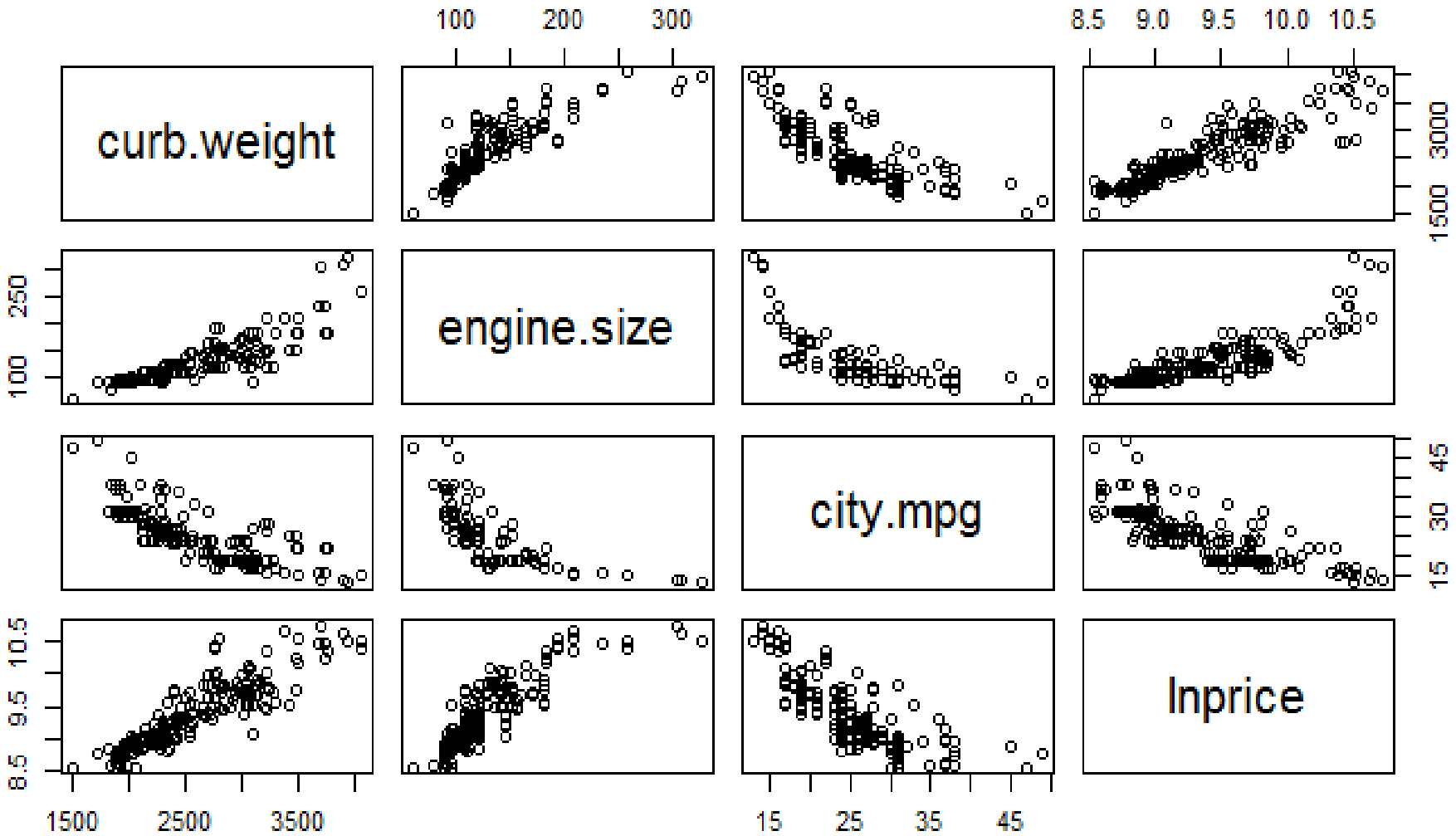
Outliers and Errors

- Errors and outliers can bias model training
- Many possible sources of errors
  - Erroneous measurements
  - Entry errors
  - Transposed values in table
- Discover and evaluate with summary statistics and visualization

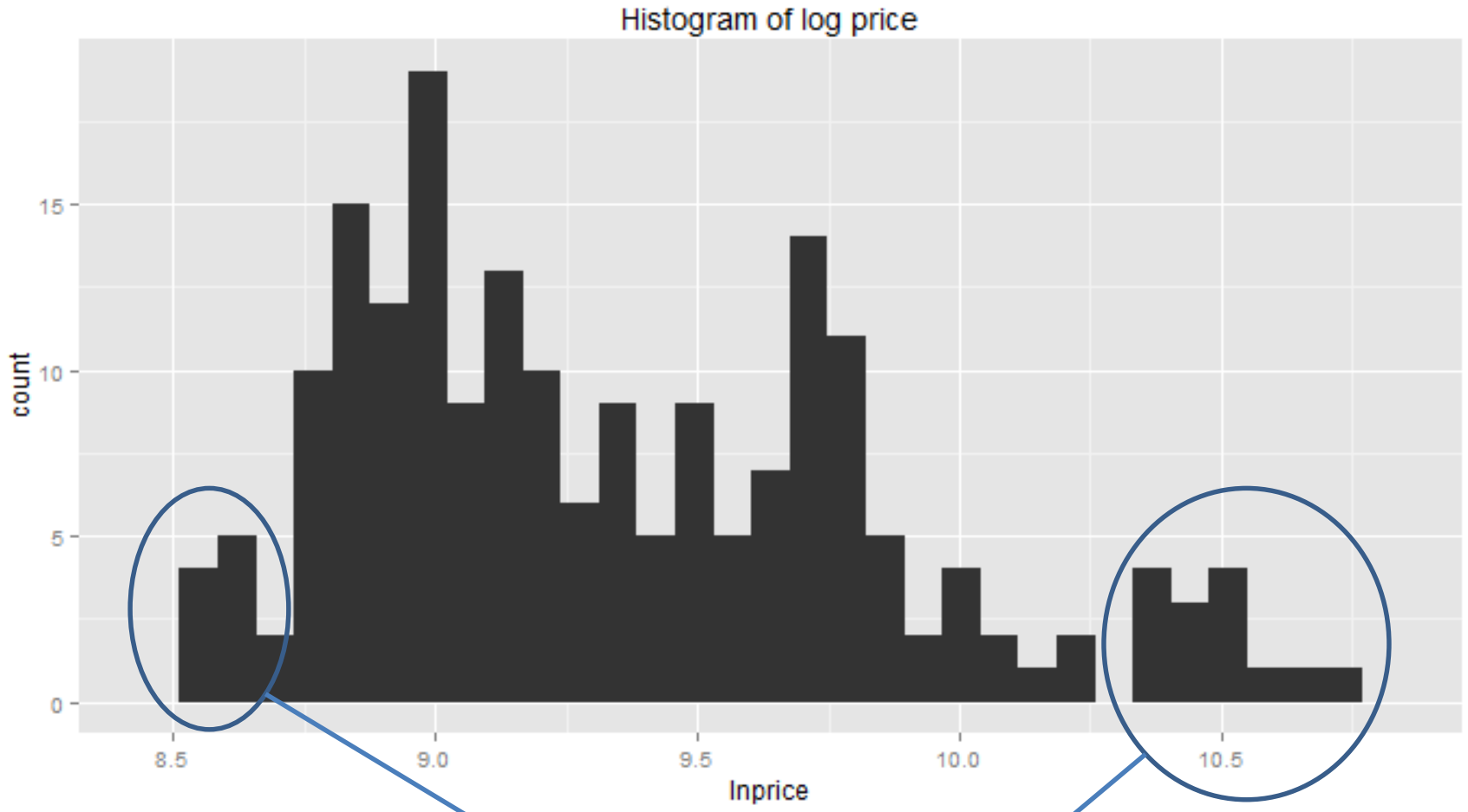
Visualizing Outliers

- Scatter plot matrix helps validate outliers
- R – pairs plot
- Python – pandas.tools.plotting.scatter_matrix

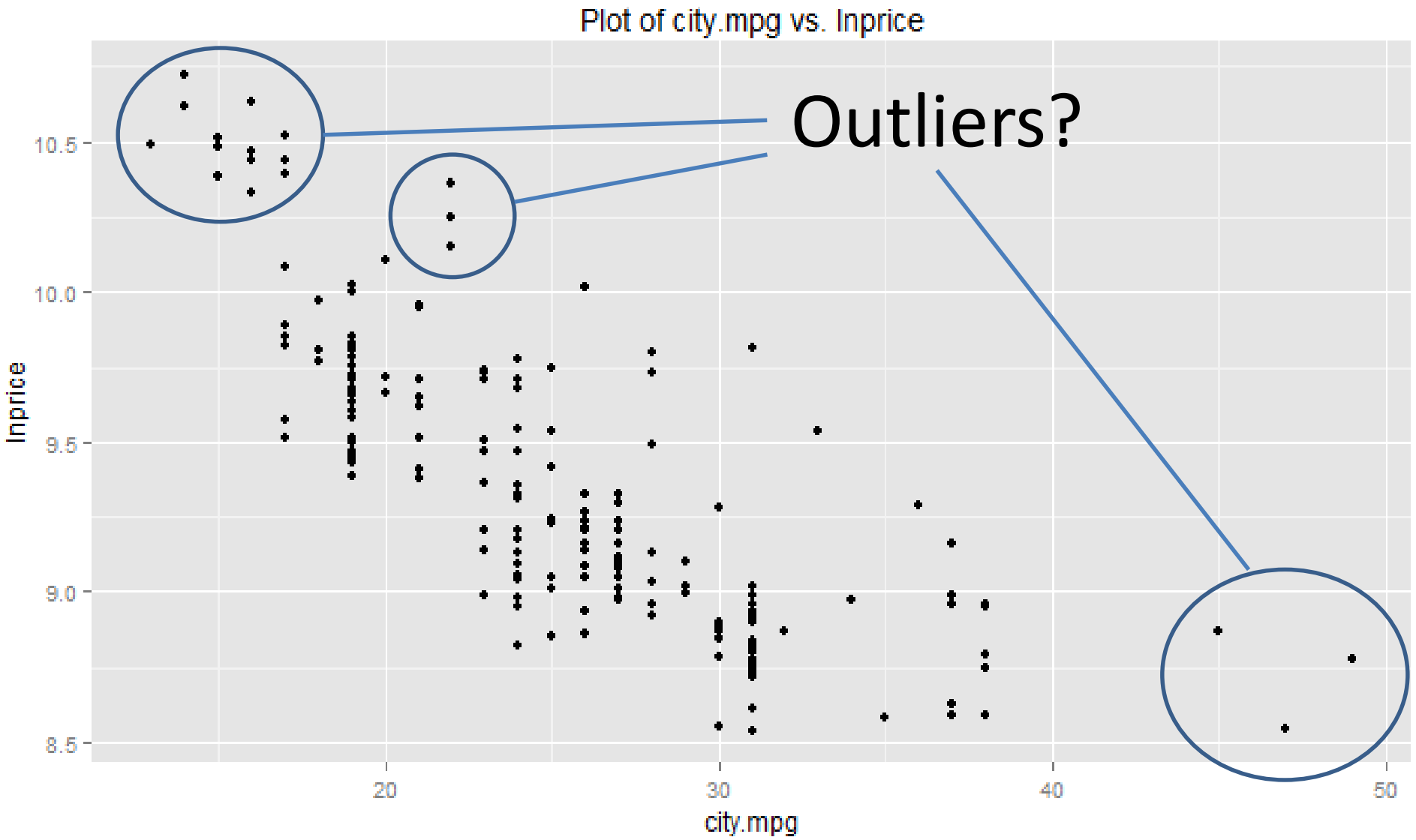
Visualizing Outliers



Identify Outliers and Errors



Identify Outliers and Errors



Clean Outliers and Errors

- Error treatments
  - Censor
  - Trim
  - Interpolate
  - Substitute
- Clip Values module
- With R
- With Python

Removing Outliers

R: data.frame = data.frame[filter.expression,]

```
library(dplyr)
frame1 <- frame1 %>% filter(Col1 > 40) %>%
                     filter(Col2 < 30) %>%
                     filter(Col3 < 3)
```

Python: DataFrame = DataFrame[filter_expression]

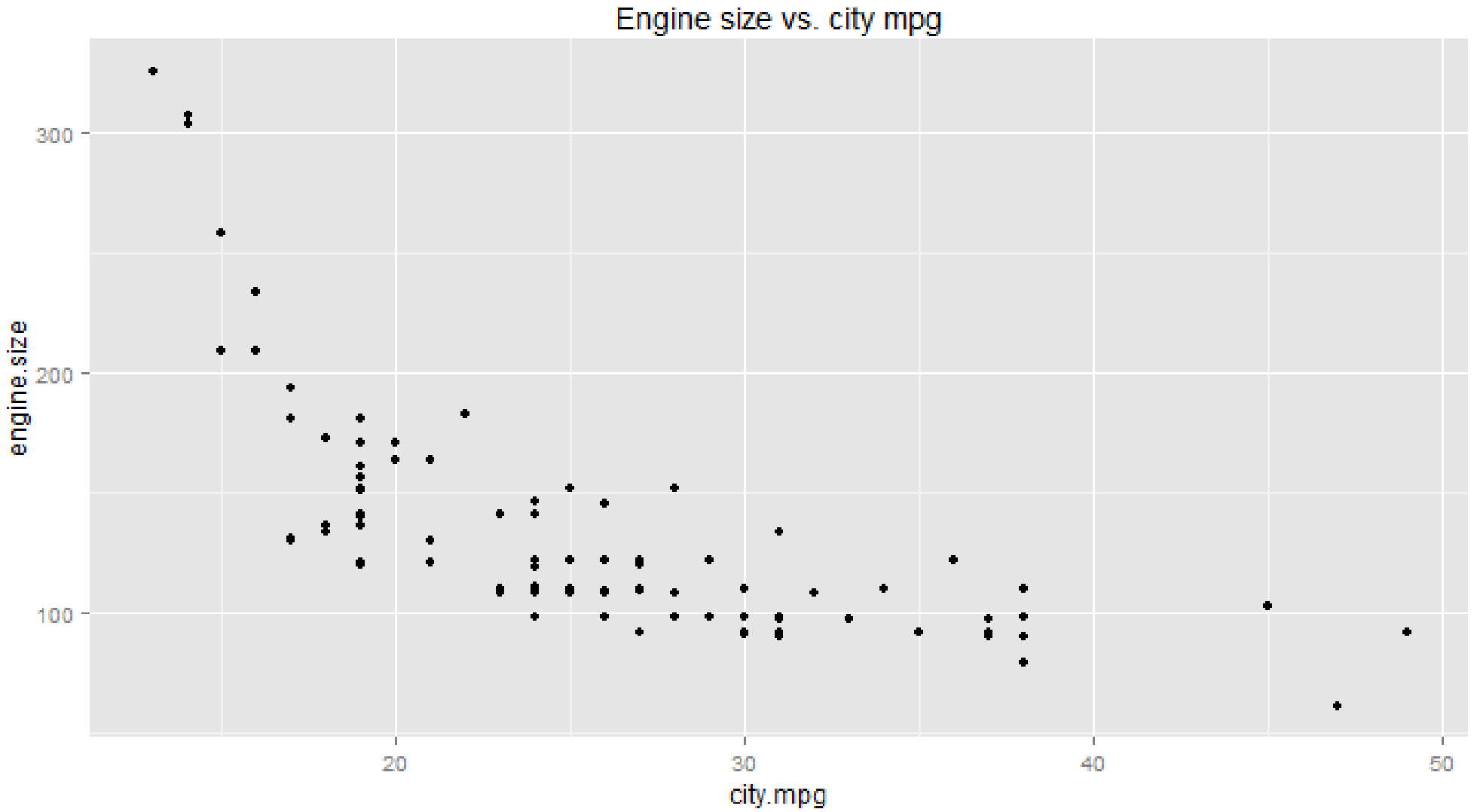
```
frame1 = frame1[(frame1["Col1"] > 40.0) &
                (frame1["Col2"] < 30.0) &
                (frame1["Col3"] < 3.0)]
```

Scaling Data

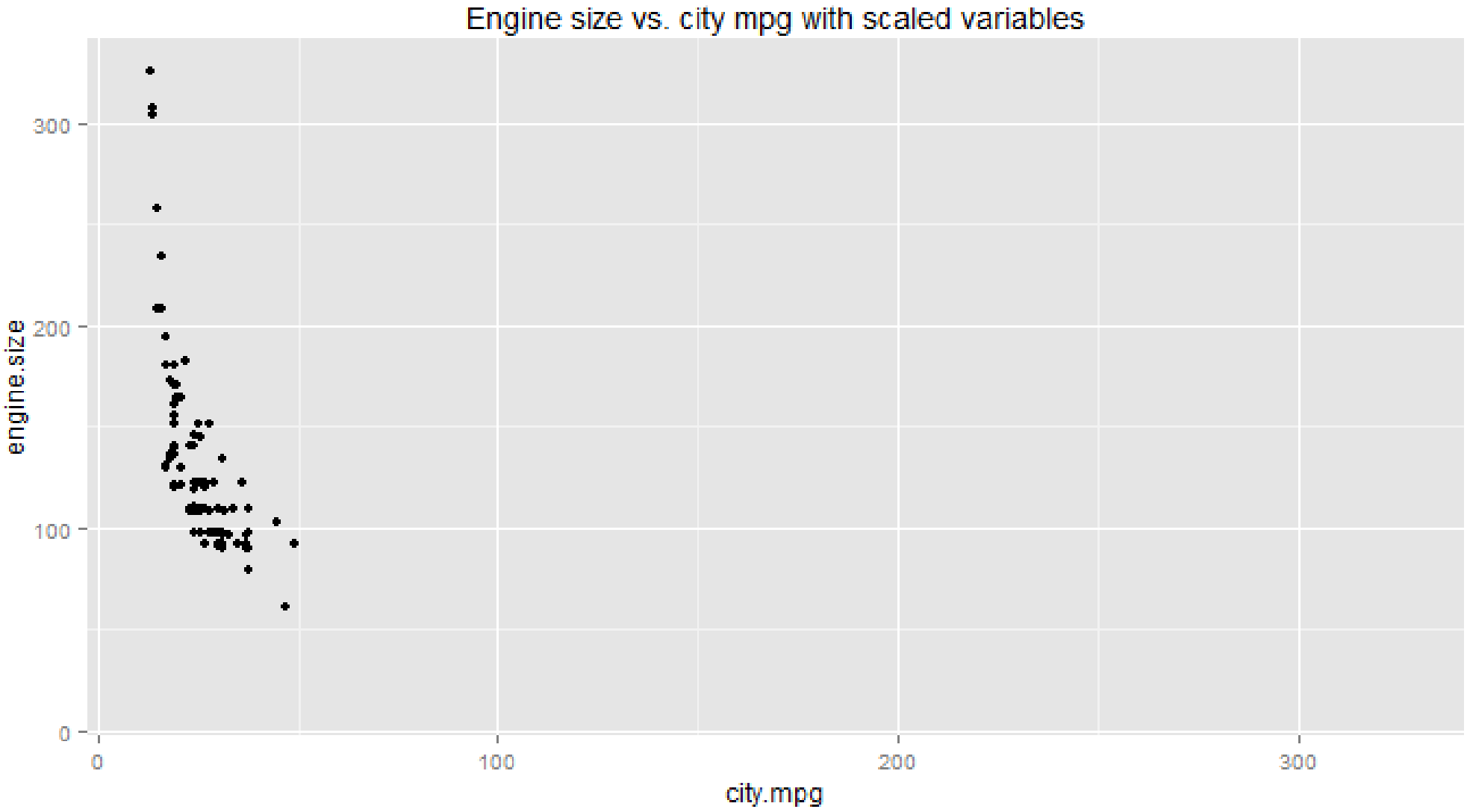
Scaling

- Numeric variables need similar scale
- Often scale to zero mean and unit variance
- May need to de-trend
- Other scaling includes min-max
- Scale after treating outliers

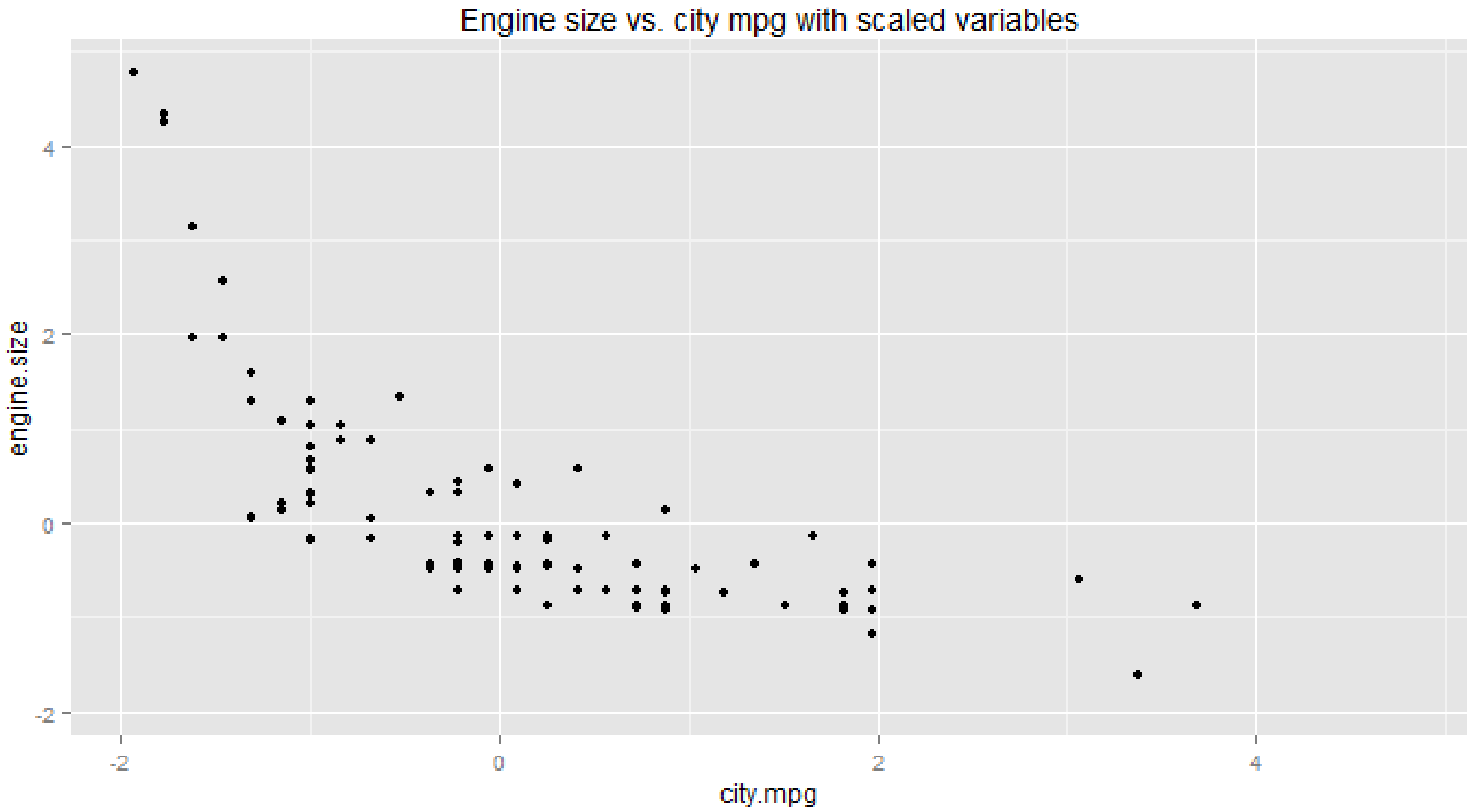
Scatter plot of two numeric columns



Unscaled data biases model construction



Scaled data biases model construction



Scaling

- Normalize Data module
- With R: `scale()`
- With Python:

  e.g. `scikit-learn.preprocessing.Scale()`

Microsoft

©2014 Microsoft Corporation. All rights reserved. Microsoft, Windows, Office, Azure, System Center, Dynamics and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.