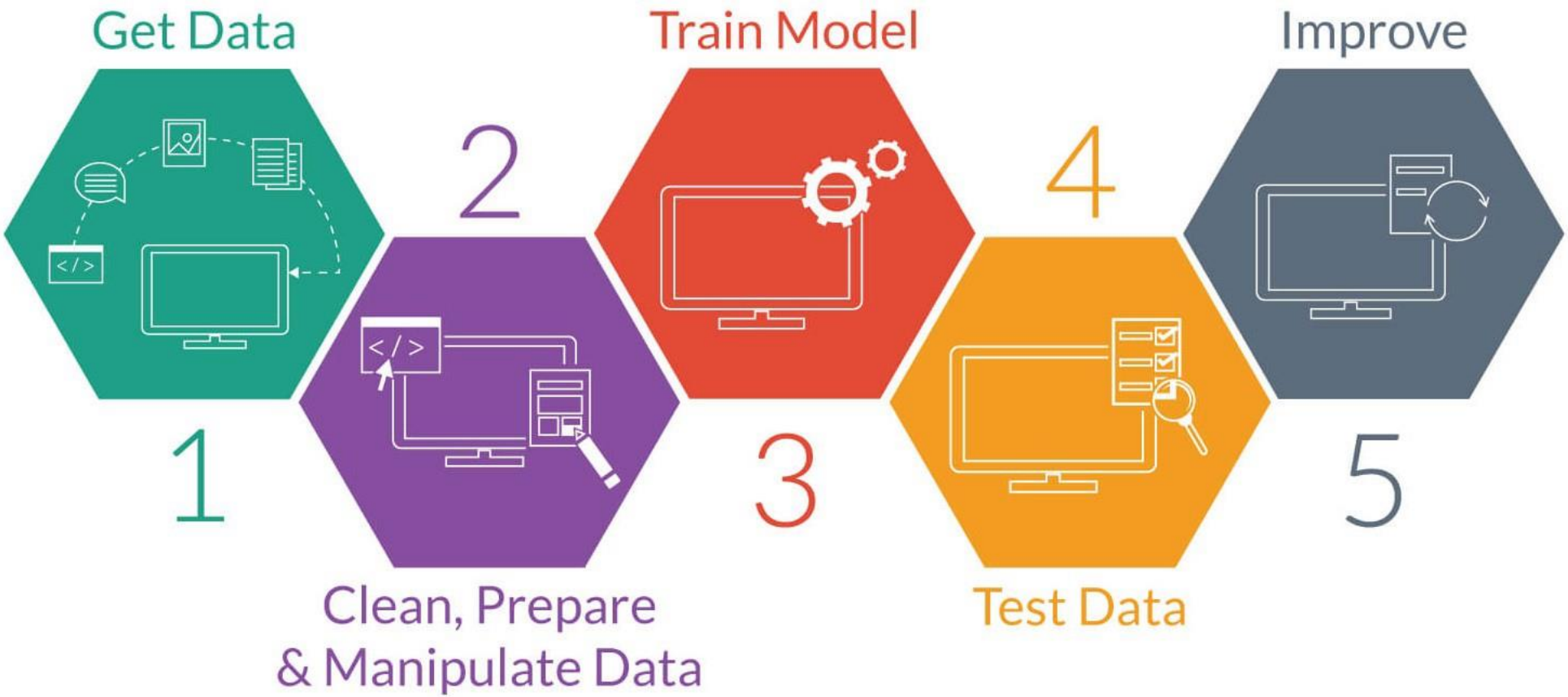
Feature Selection In Machine Learning

dataaspirant.com

Building a Machine Learning model

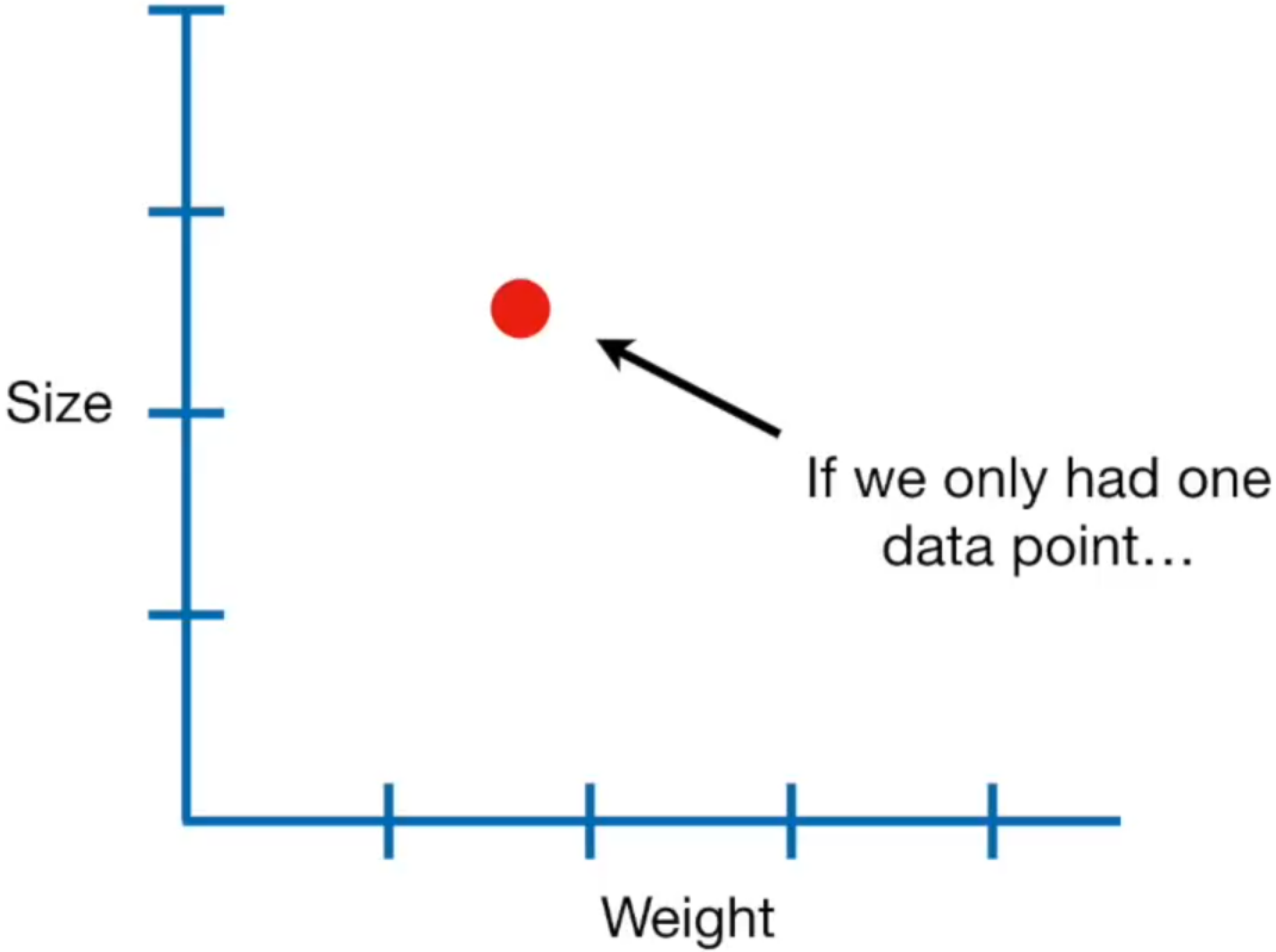


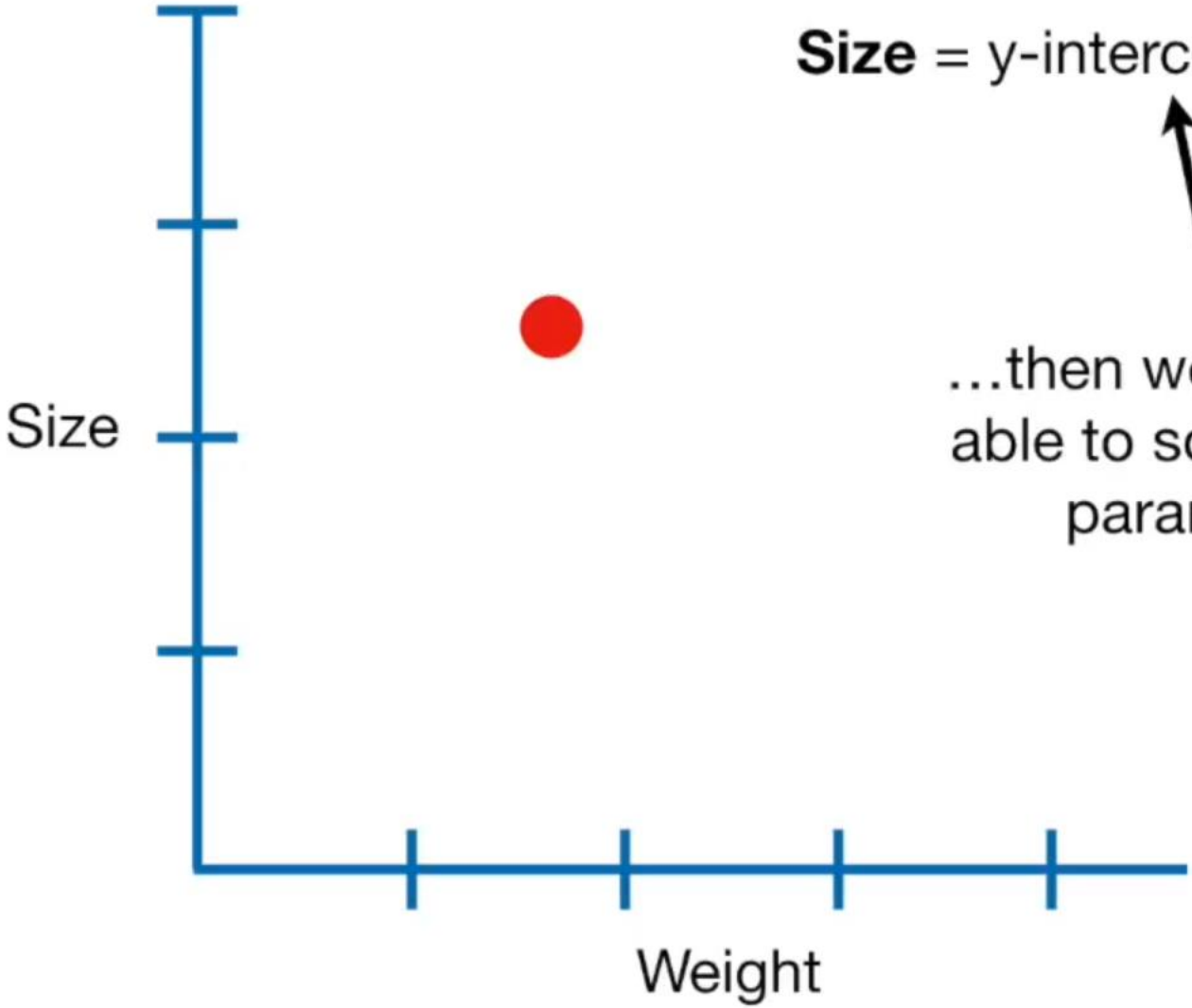
Data Preparation

- Combine data from various sources
- Check data schema:
  - Usually, the source of the data has schema describing the data
- Clean your Data: identify and handle errors in your data.
  - Handle Missing data
  - Handle outliers
- Data Transformation: change the scale of some/all variables. Why?? We will discuss it later.
- Feature Selection: select these features that are most relevant to your task.
- Feature Engineering: combine features, derive new variables, dimensionality reduction, etc.

Exploratory Data Analysis

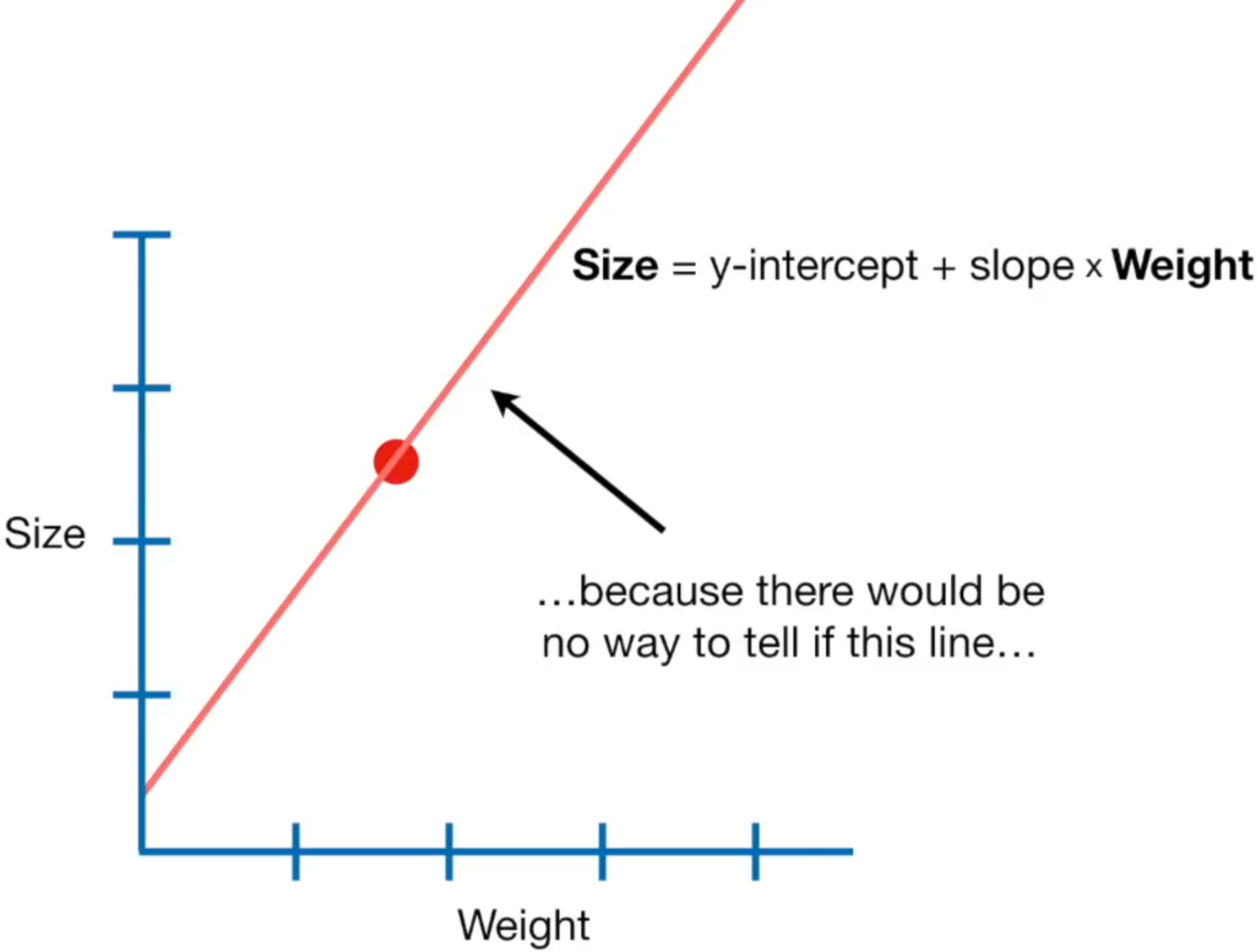
- We can’t prepare the data without first exploring the data.
- Quantify missing data
- Identify numerical and categorical variables
- Determine unique values (cardinality) in categorical features
- Check rare/ dominant categories in categorical features
- Highlight outliers
- Identify linear relationships
- Identify a normal distribution
- Check histograms

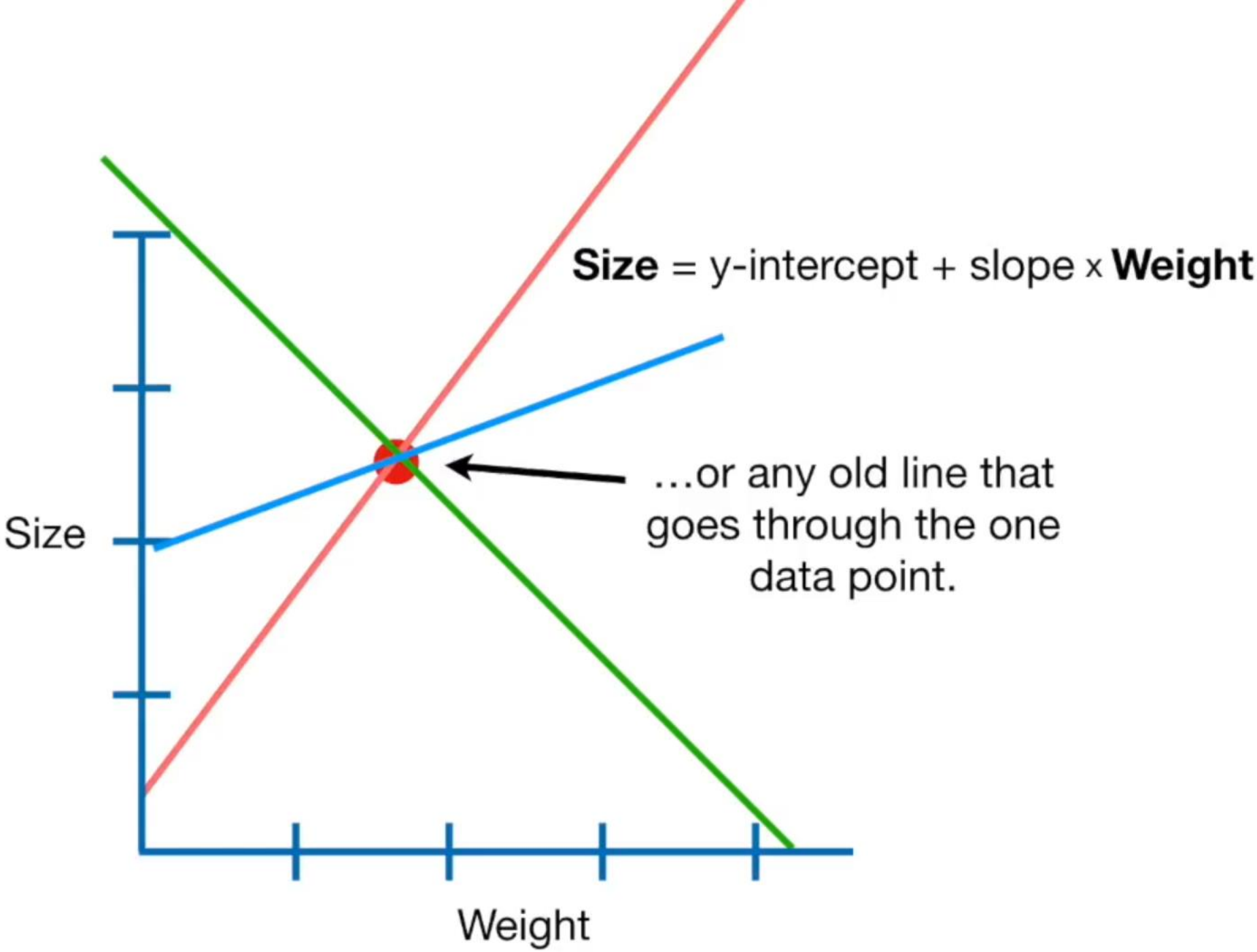

Size
If we only had one
data point…
Weight

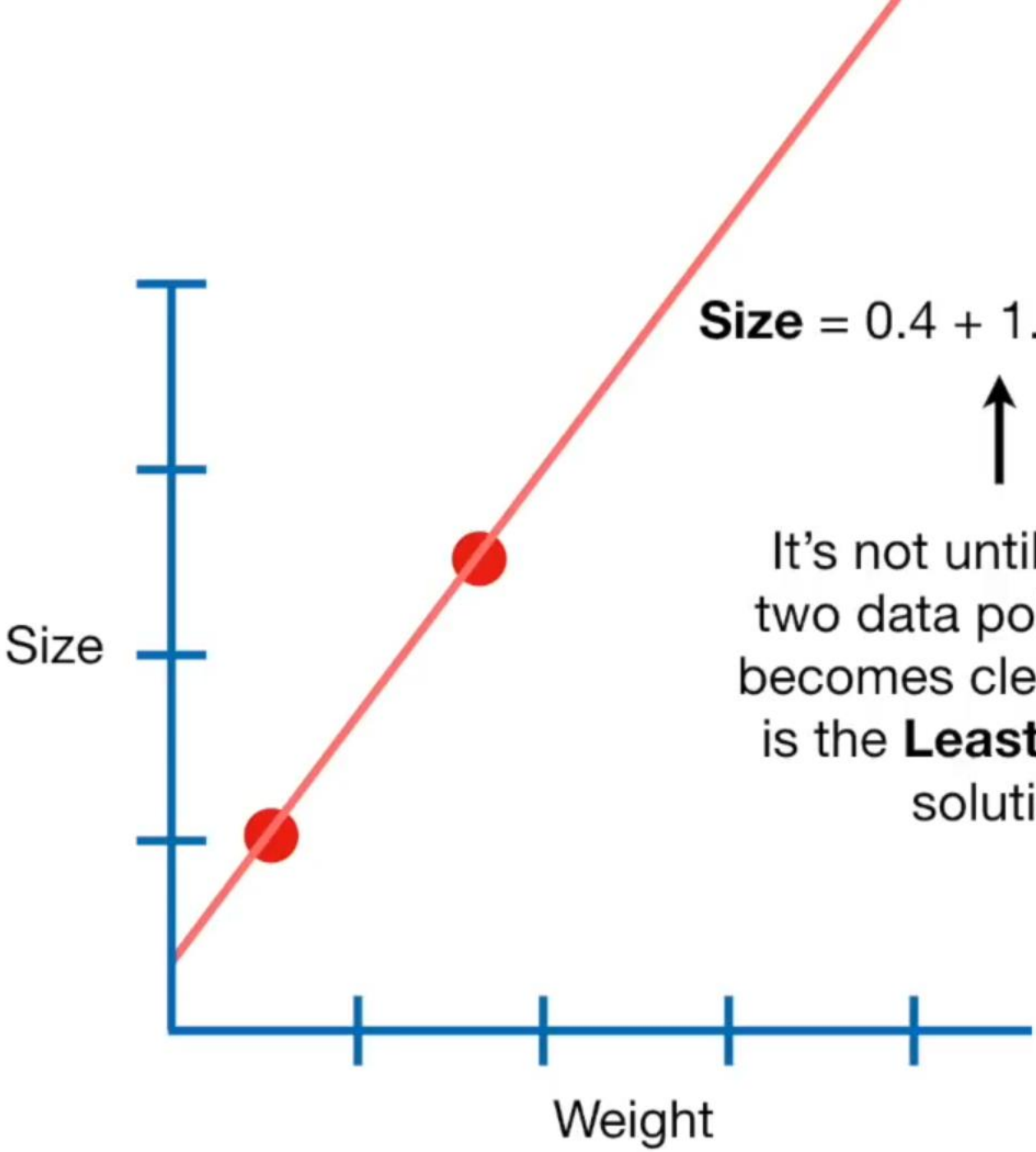


Size = y-intercept + slope x **Weight**

…then we wouldn't be able to solve for these parameters…


Size = y-intercept + slope x **Weight**
…because there would be
no way to tell if this line…
Size
Weight


Size = y-intercept + slope x **Weight**
…or any old line that goes through the one data point.
Size
Weight

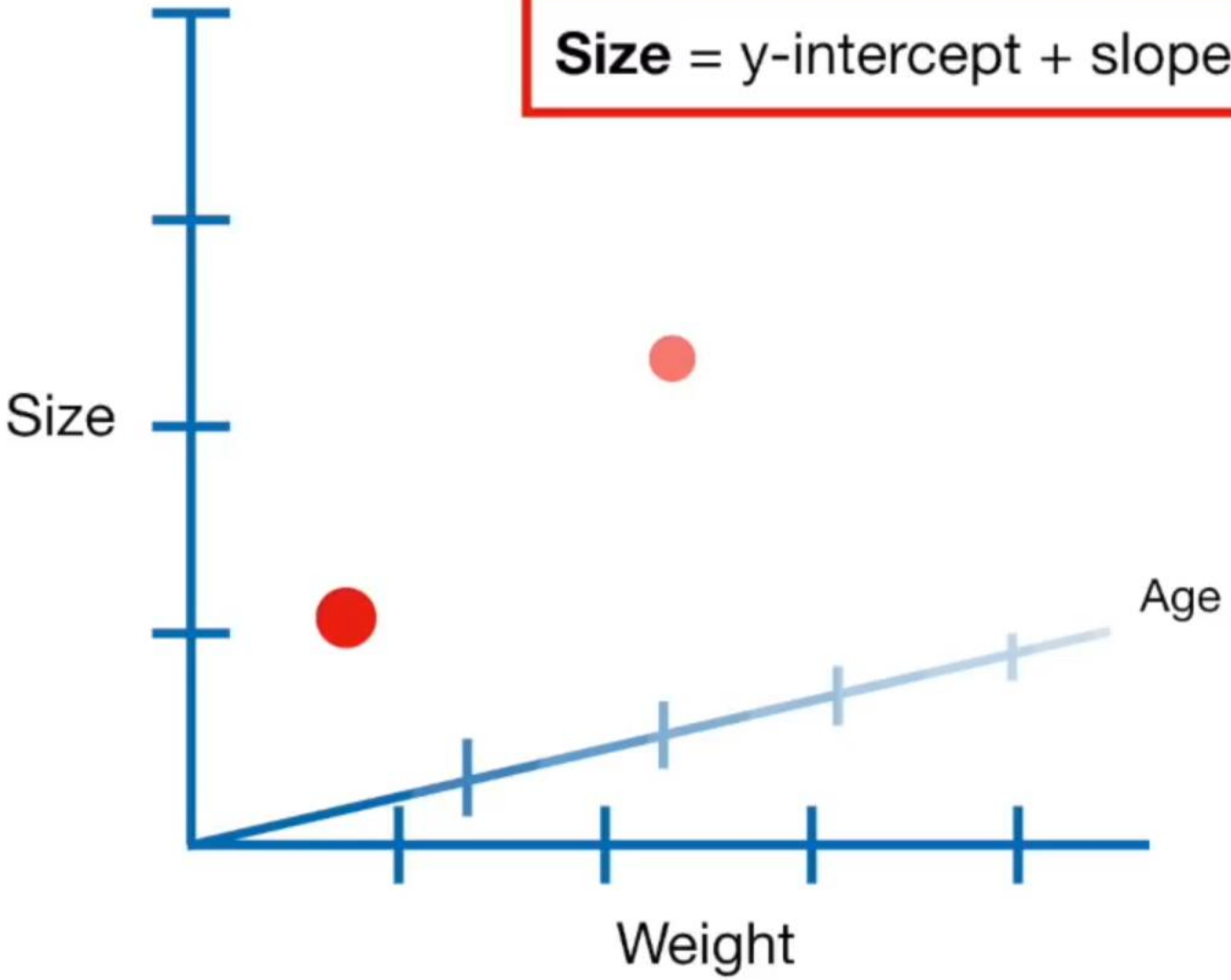


Size = 0.4 + 1.3 × **Weight**

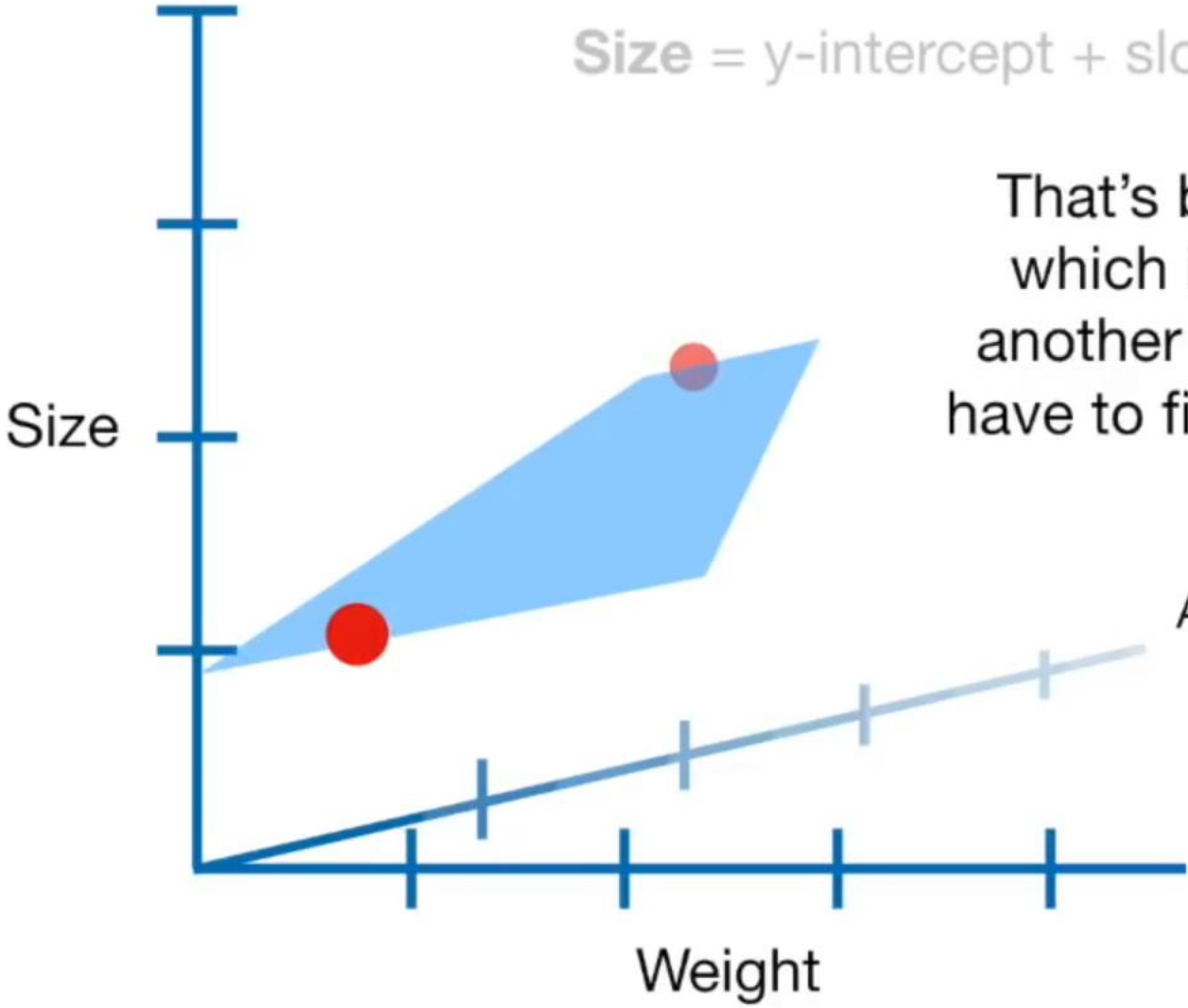
It's not until we have two data points that it becomes clear that this is the **Least Squares** solution.

Now let's look at an equation that has three parameters to estimate.

Size = y-intercept + slope1 x **Weight** + slope2 x **Age**

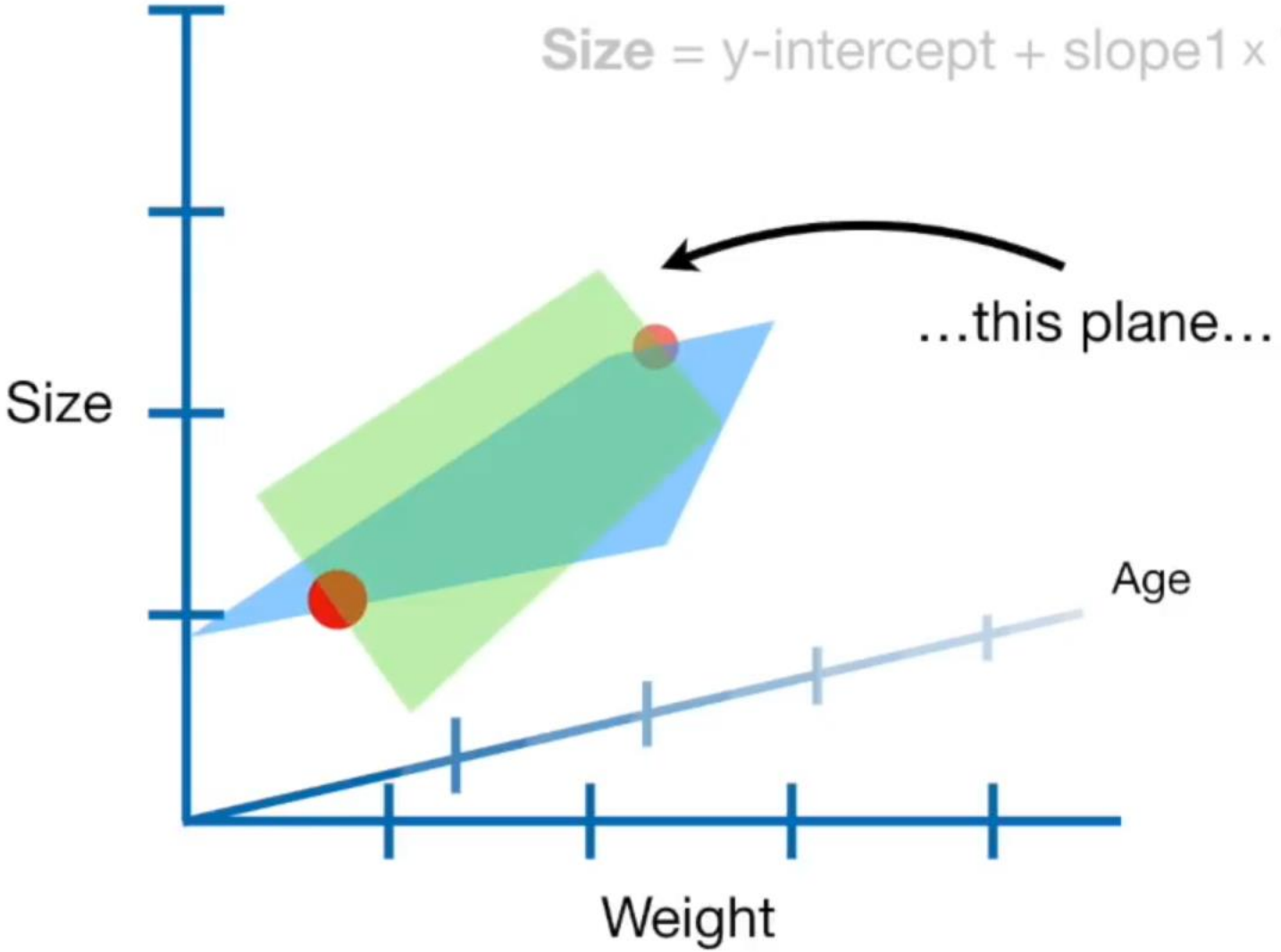


Size = y-intercept + slope1 × **Weight** + slope2 × **Age**

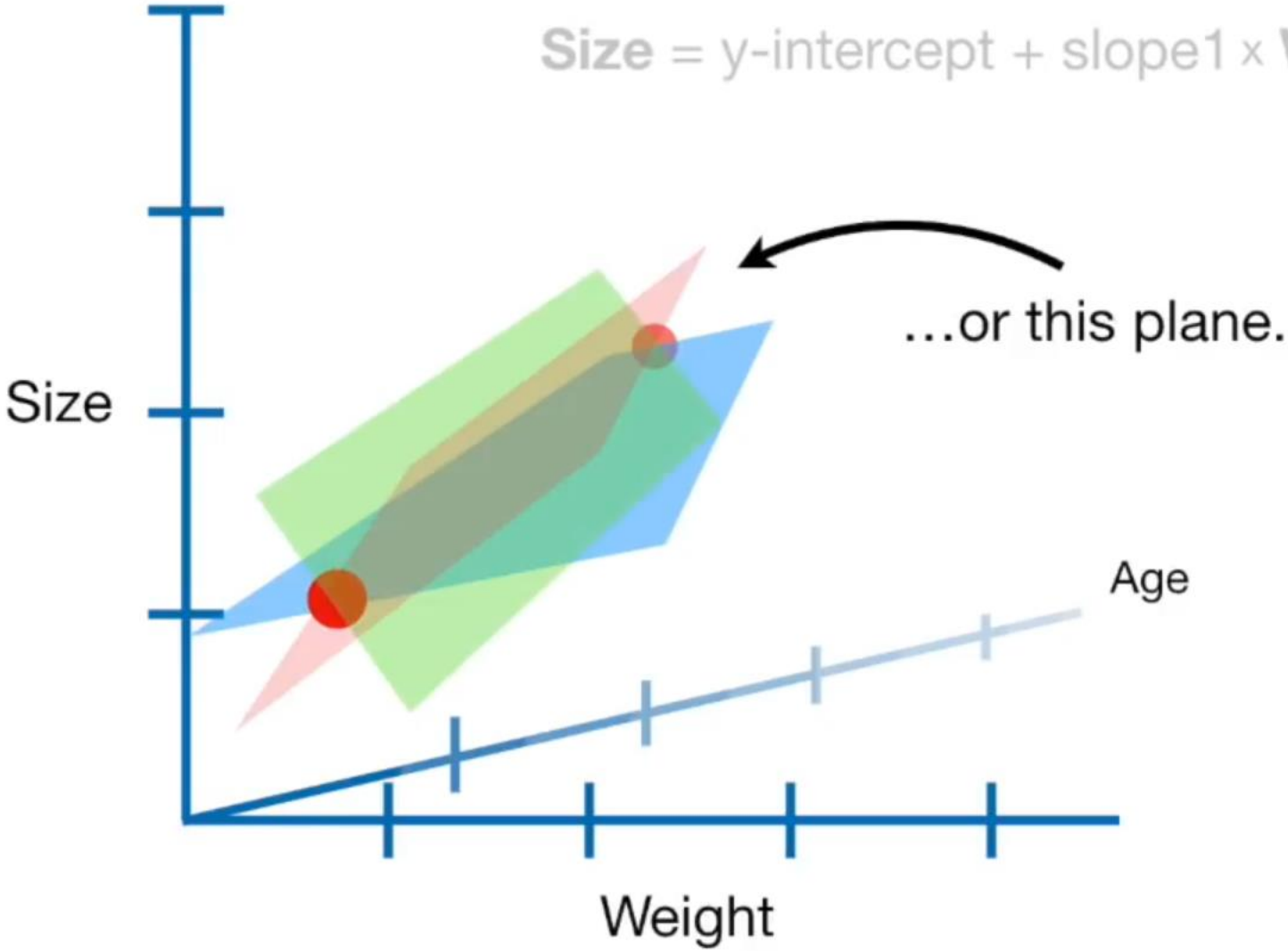


That's because in three dimensions, which is what we get when we add another axis to our graph for **Age**, we have to fit a plane to the data instead of just a line…

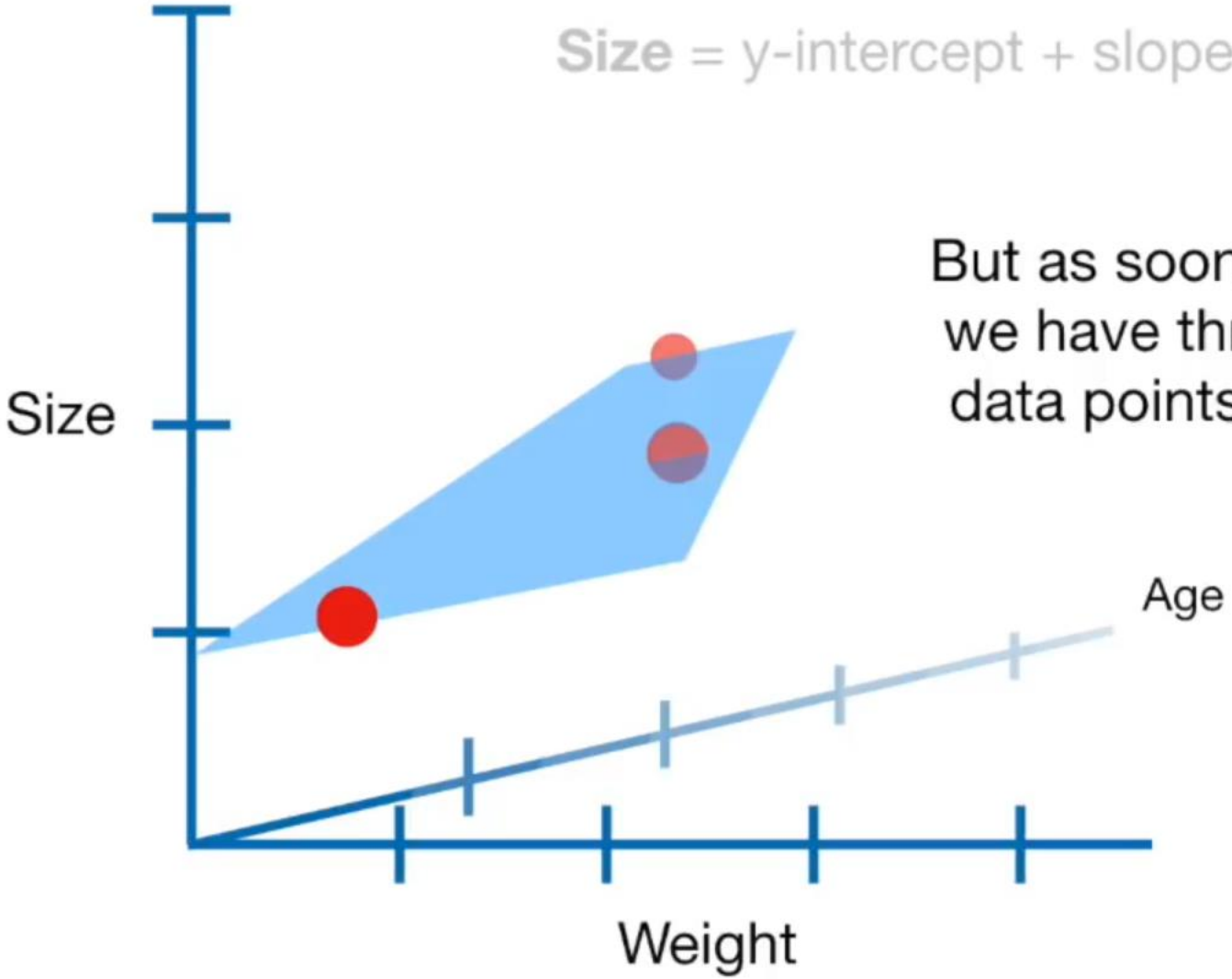
Size = y-intercept + slope1 × **Weight** + slope2 × **Age**



Size = y-intercept + slope1 × **Weight** + slope2 × **Age**



Size = y-intercept + slope1 × **Weight** + slope2 × **Age**

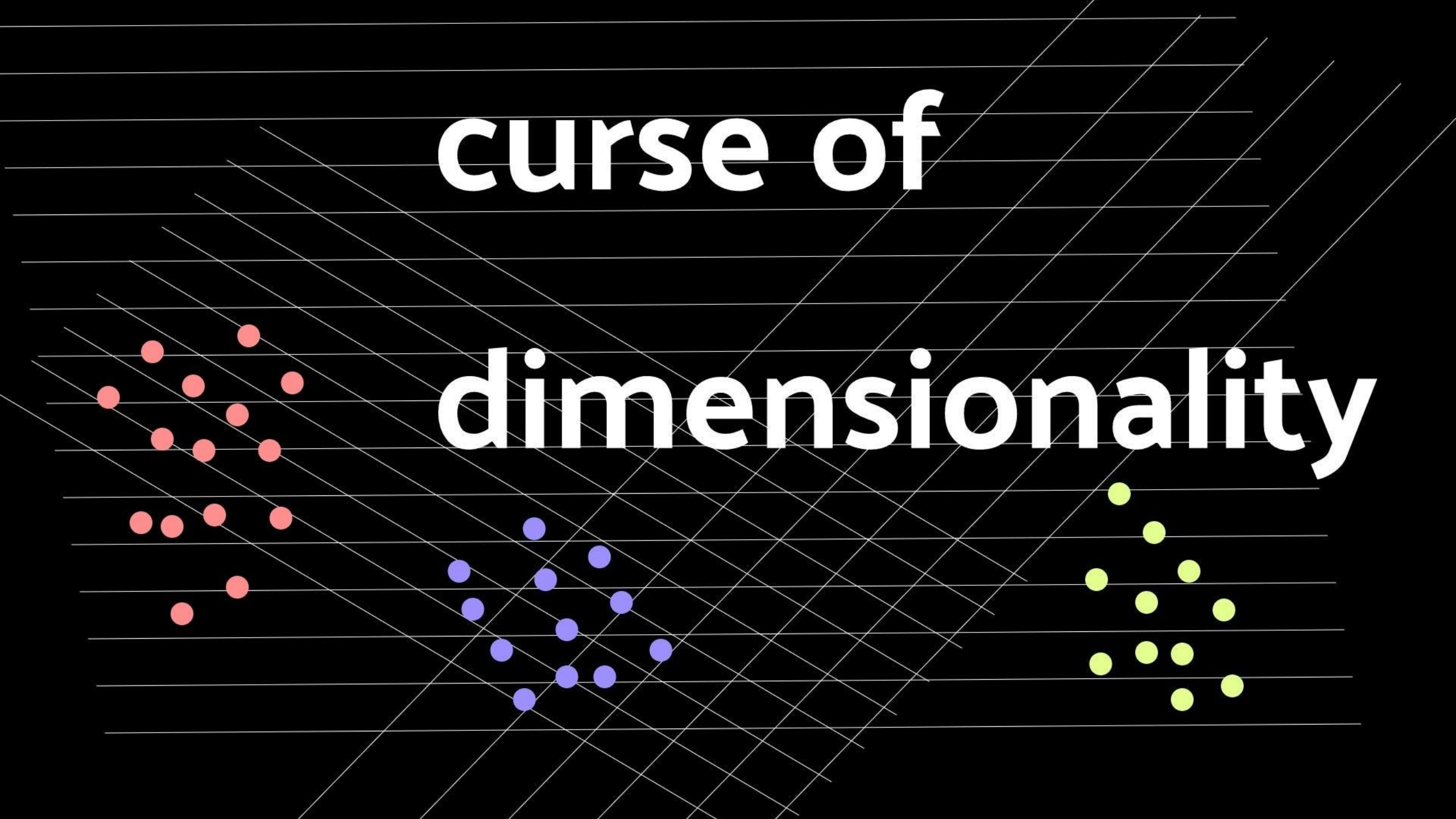


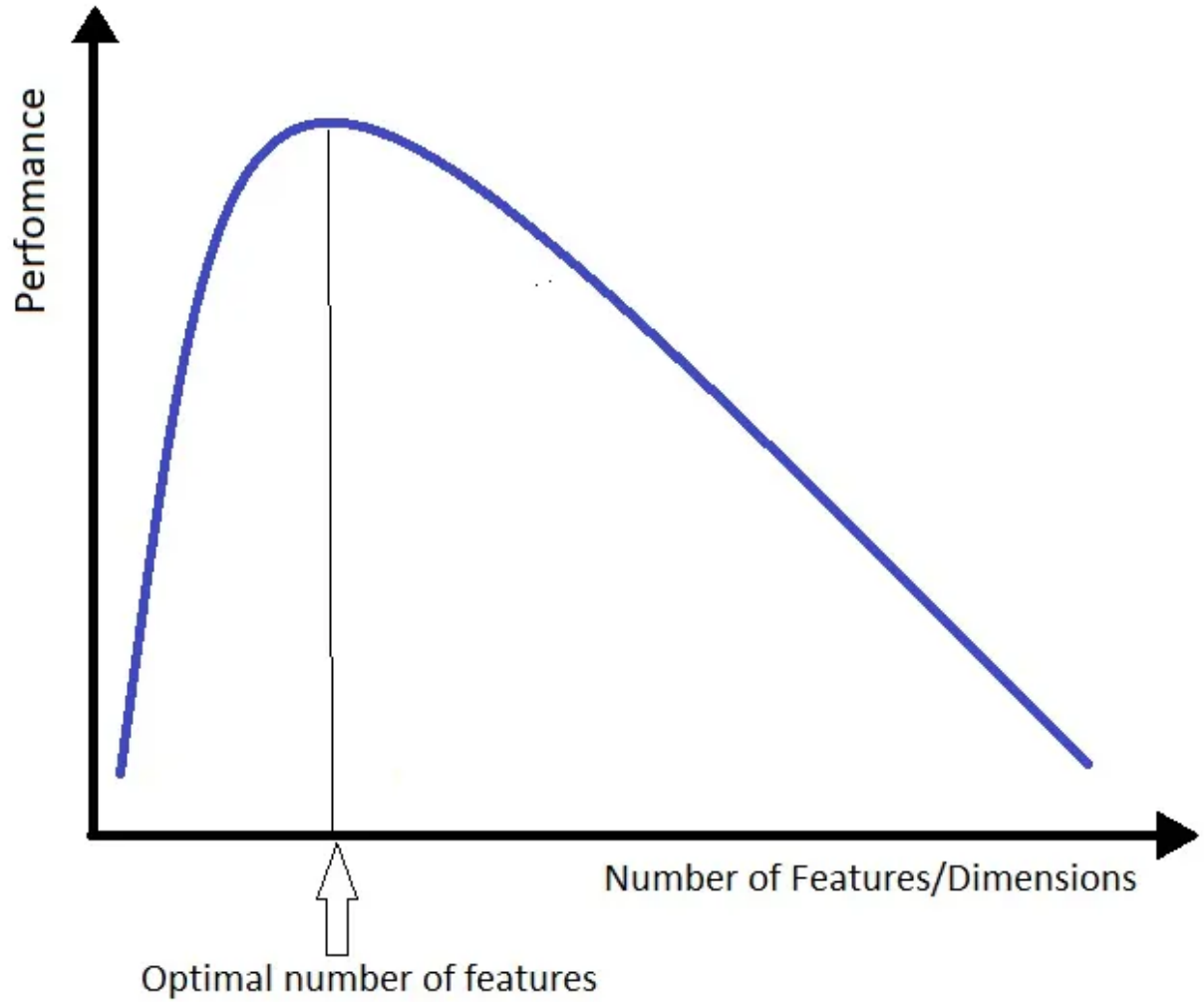
Size = y-intercept + slope1 x **Weight** + slope2 x **Age** + slope3 x **fur color**

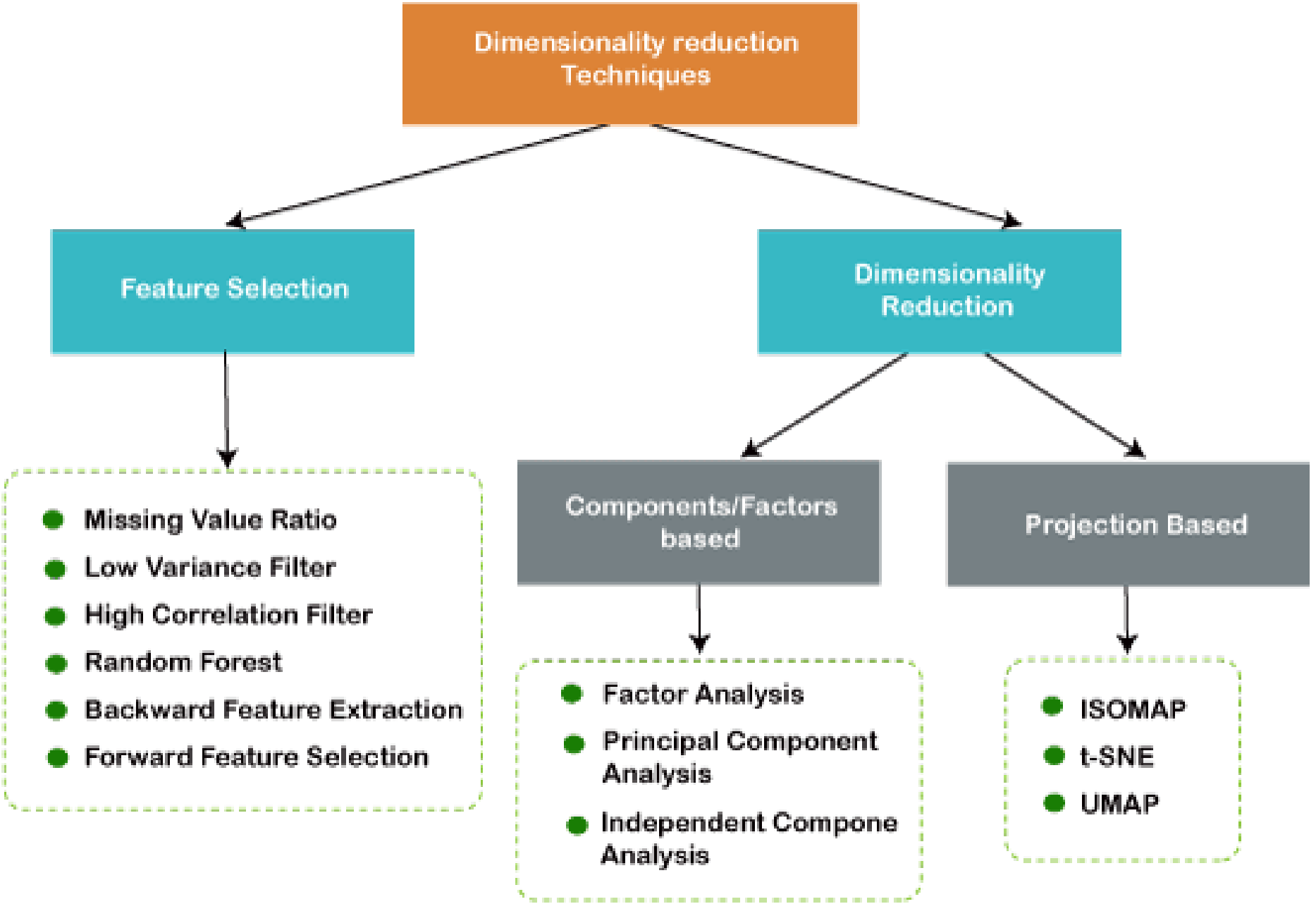
If we have an equation with four parameters…

…then **Least Squares** needs at least four data points to estimate all four parameters.

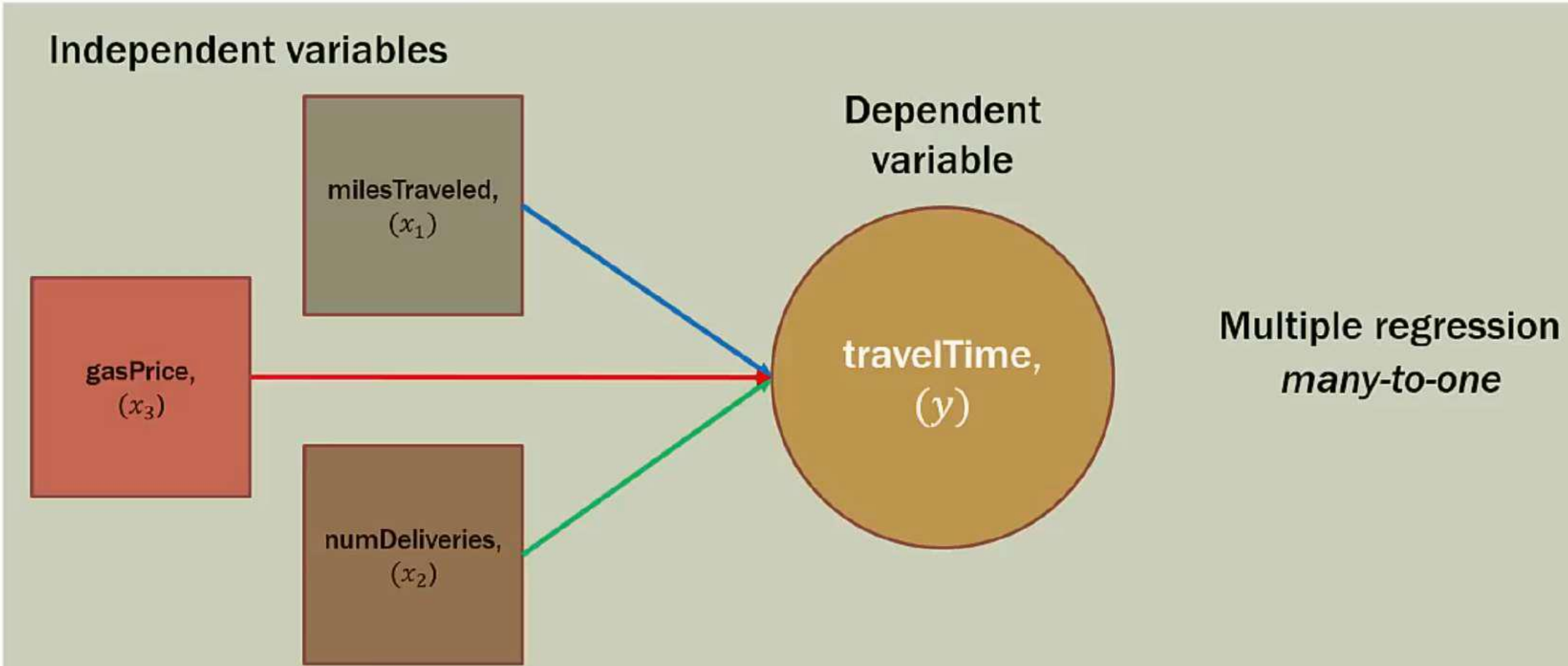
curse of dimensionality




Perfomance
Number of Features/Dimensions
Optimal number of features

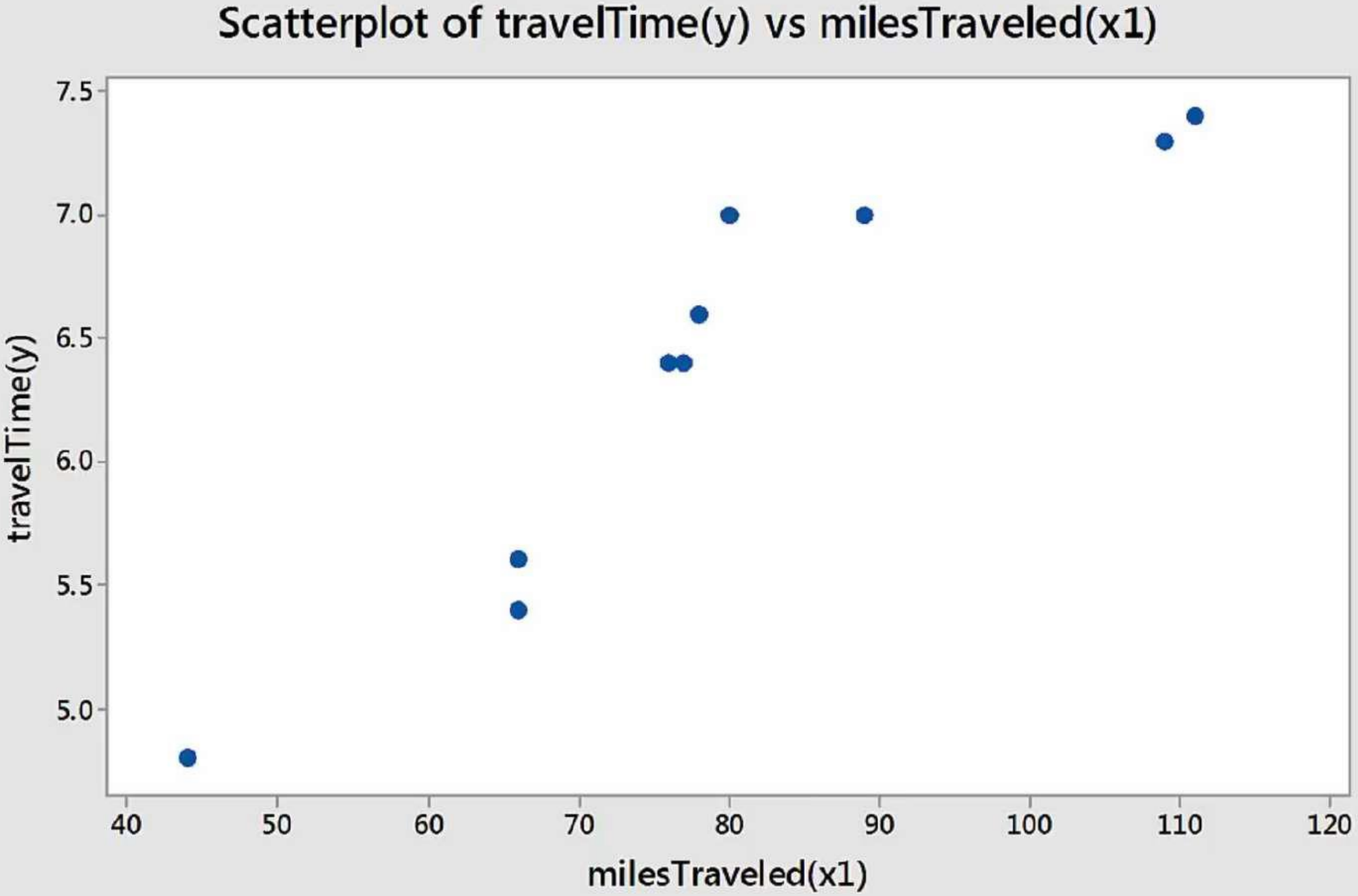

Dimensionality reduction Techniques
Feature Selection
Dimensionality Reduction
Missing Value Ratio
Low Variance Filter
High Correlation Filter
Random Forest
Backward Feature Extraction
Forward Feature Selection
Components/Factors based
Projection Based
Factor Analysis
Principal Component Analysis
Independent Compone Analysis
ISOMAP
t-SNE
UMAP

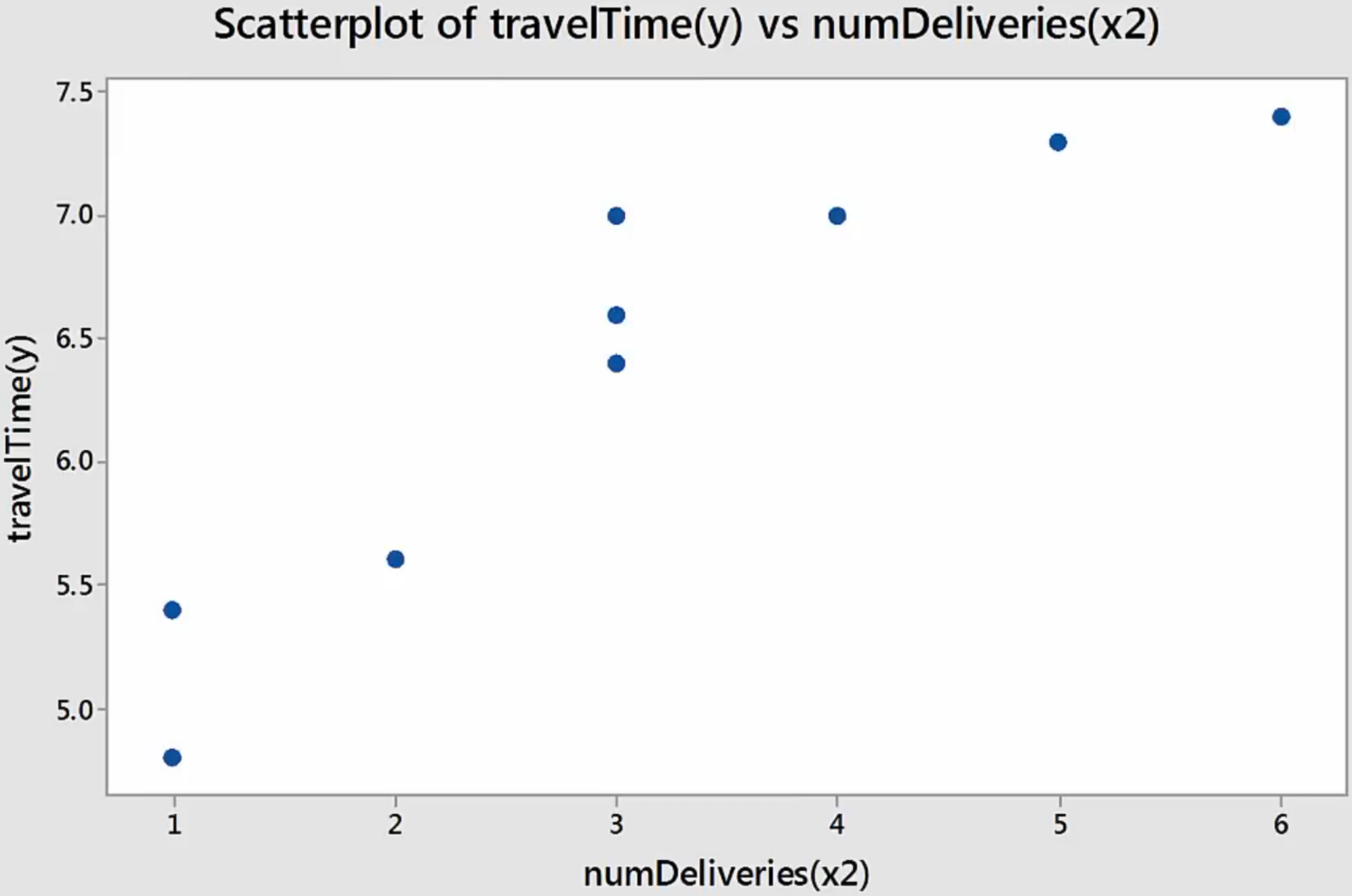
SKETCHING OUT RELATIONSHIPS

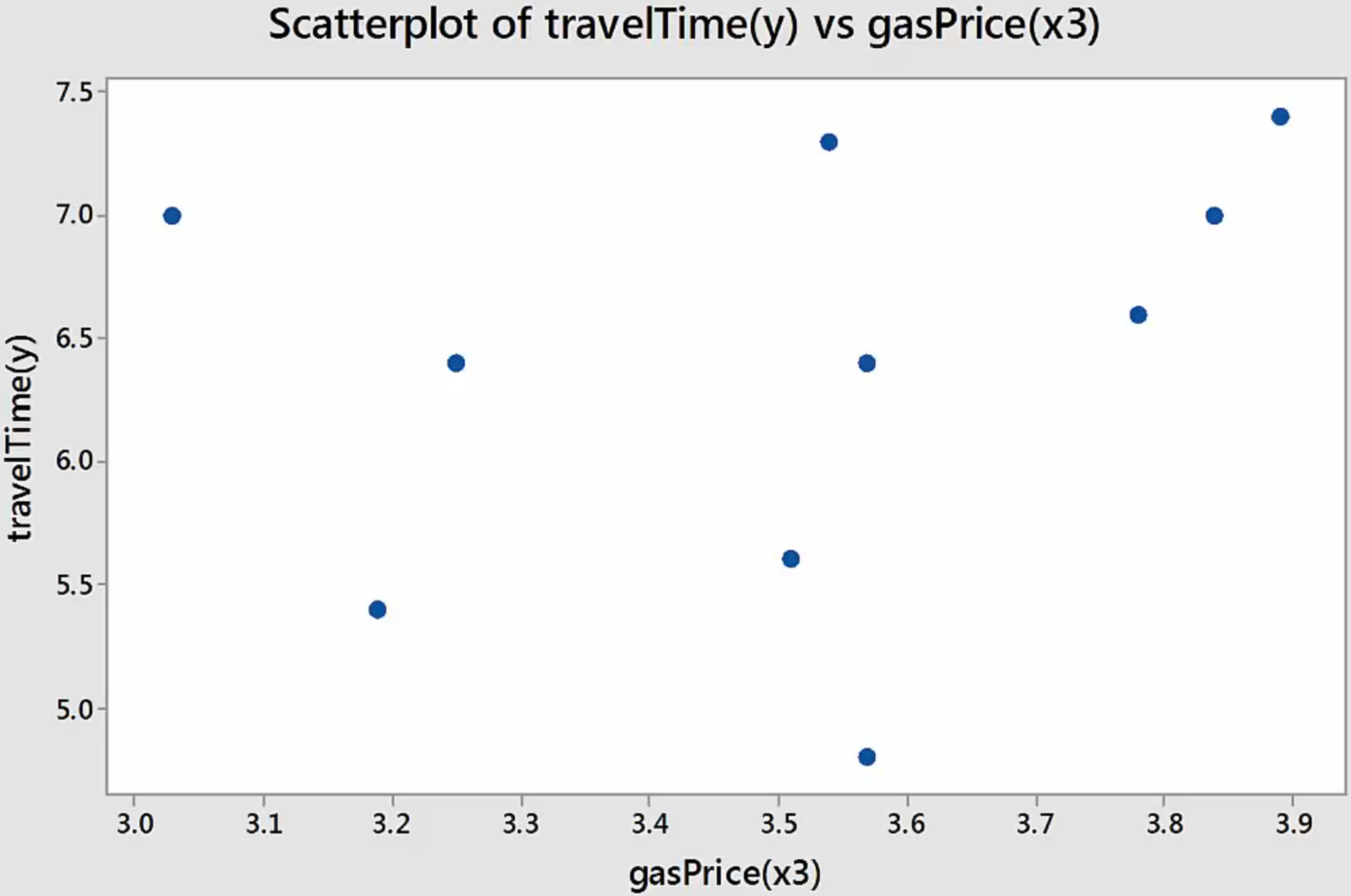


IV TO DV SCATTERPLOTS

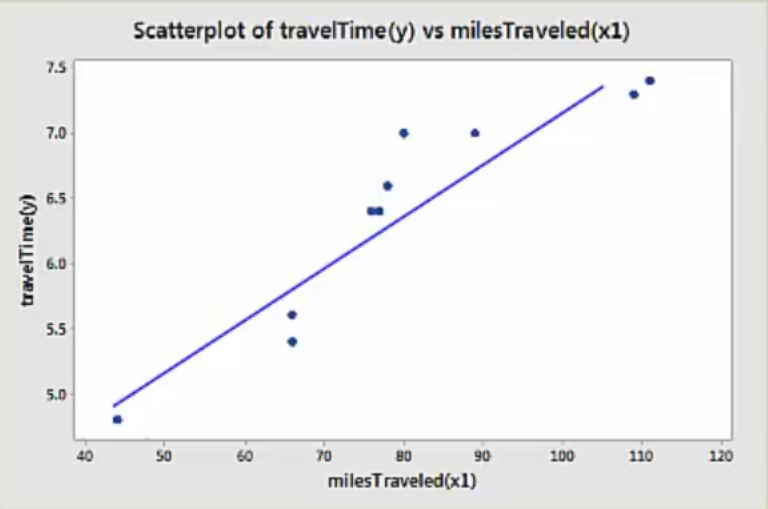
Relevancy Check

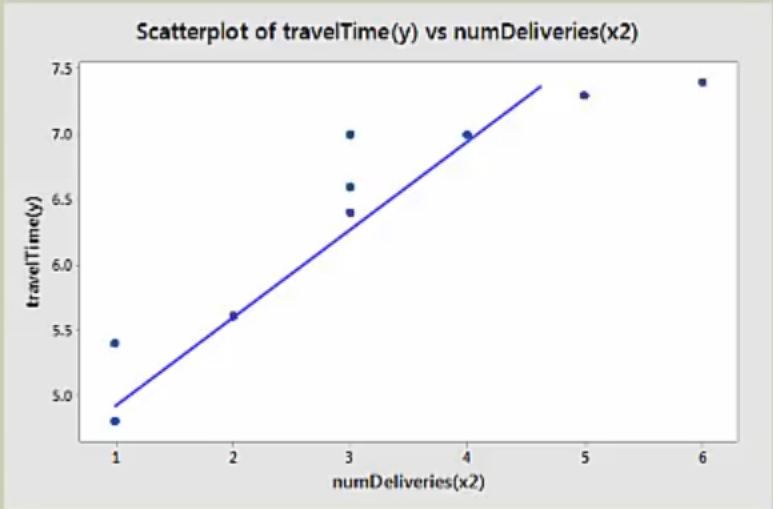

Scatterplot of travelTime(y) vs milesTraveled(x1)
travelTime(y)
7.5
7.0
6.5
6.0
5.5
5.0
40
50
60
70
80
90
100
110
120
milesTraveled(x1)

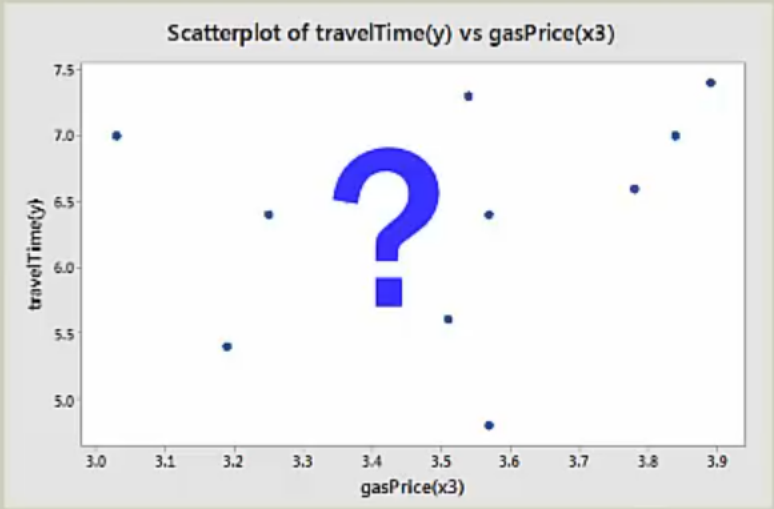

Scatterplot of travelTime(y) vs numDeliveries(x2)
travelTime(y)
7.5
7.0
6.5
6.0
5.5
5.0
1
2
3
4
5
6
numDeliveries(x2)


Scatterplot of travelTime(y) vs gasPrice(x3)
travelTime(y)
7.5
7.0
6.5
6.0
5.5
5.0
3.0
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
3.9
gasPrice(x3)

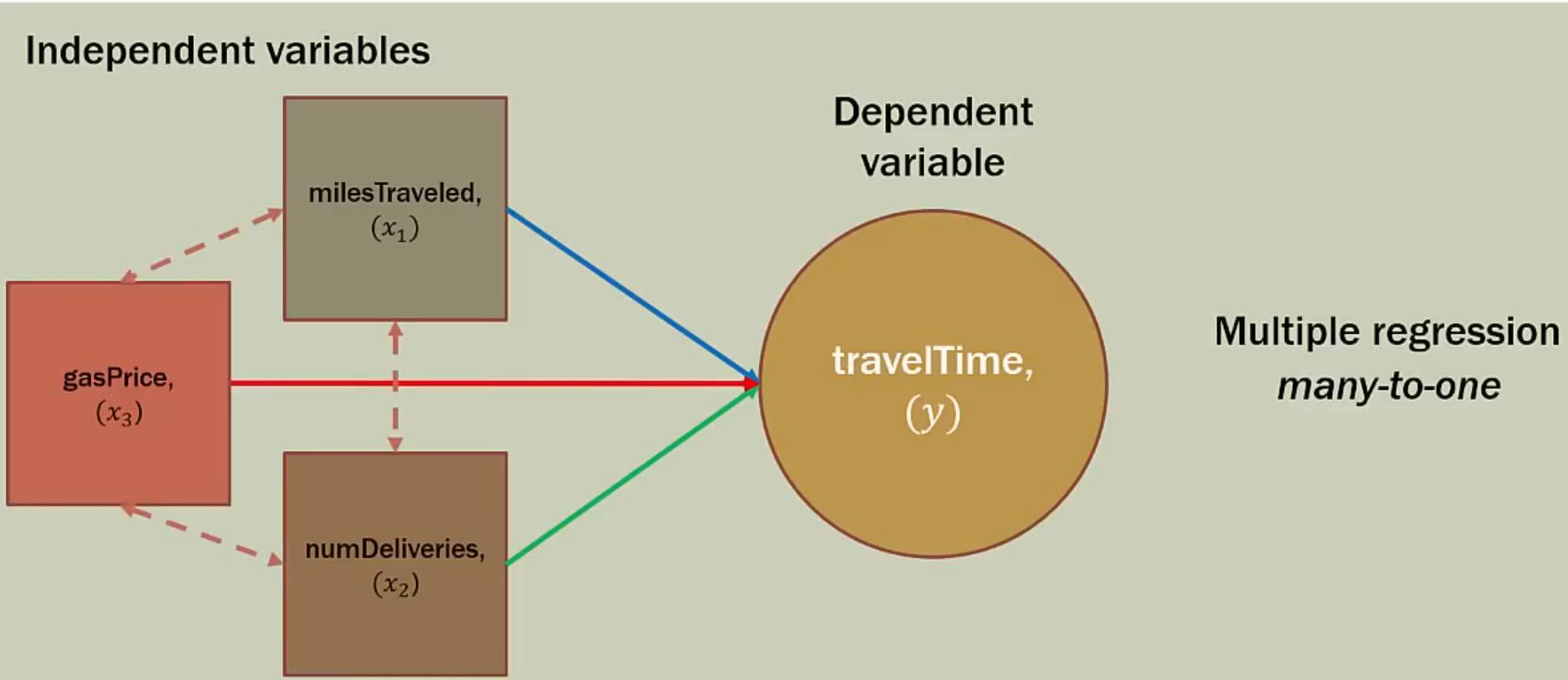
DV VS IV SCATTERPLOTS





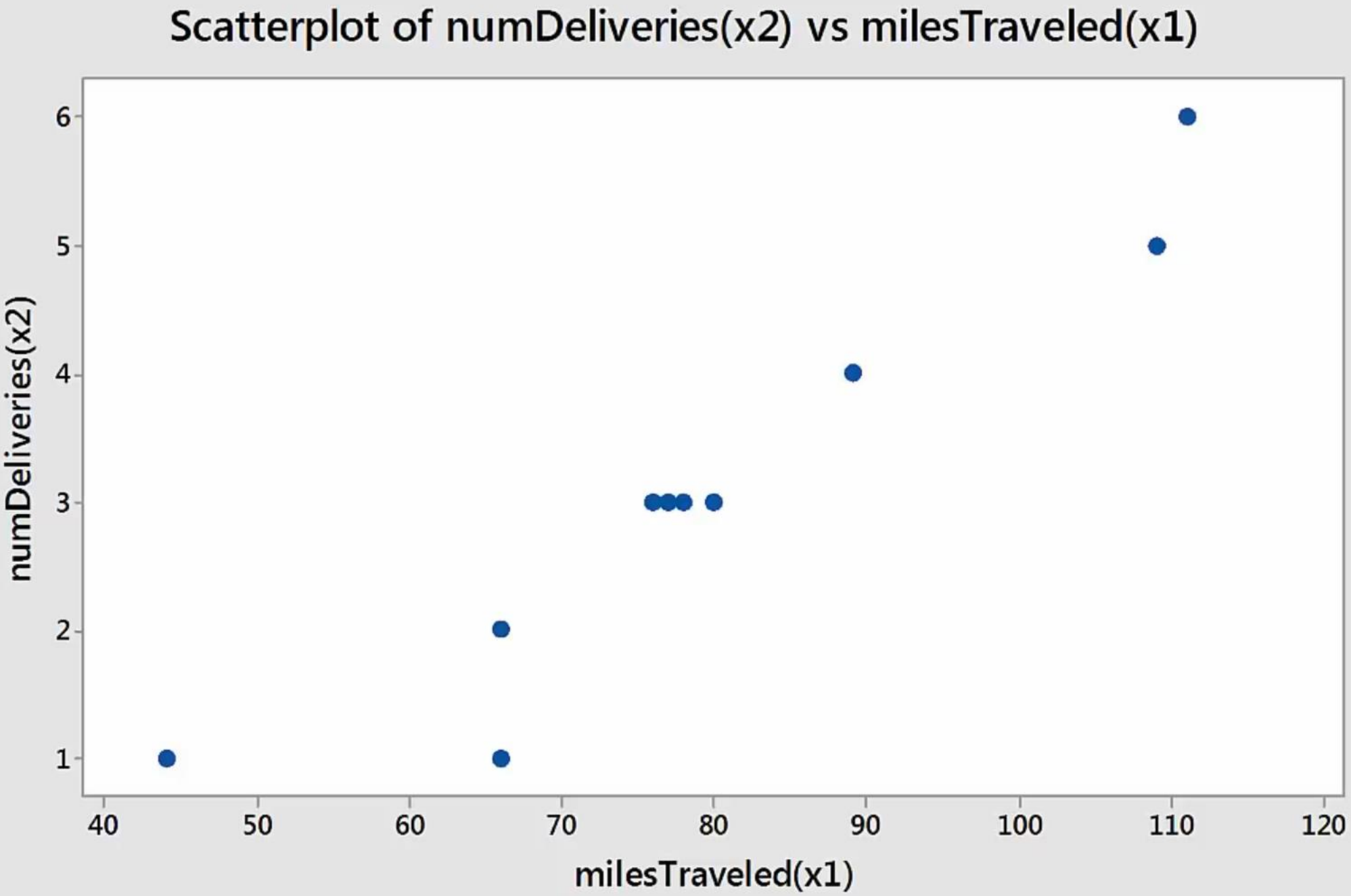


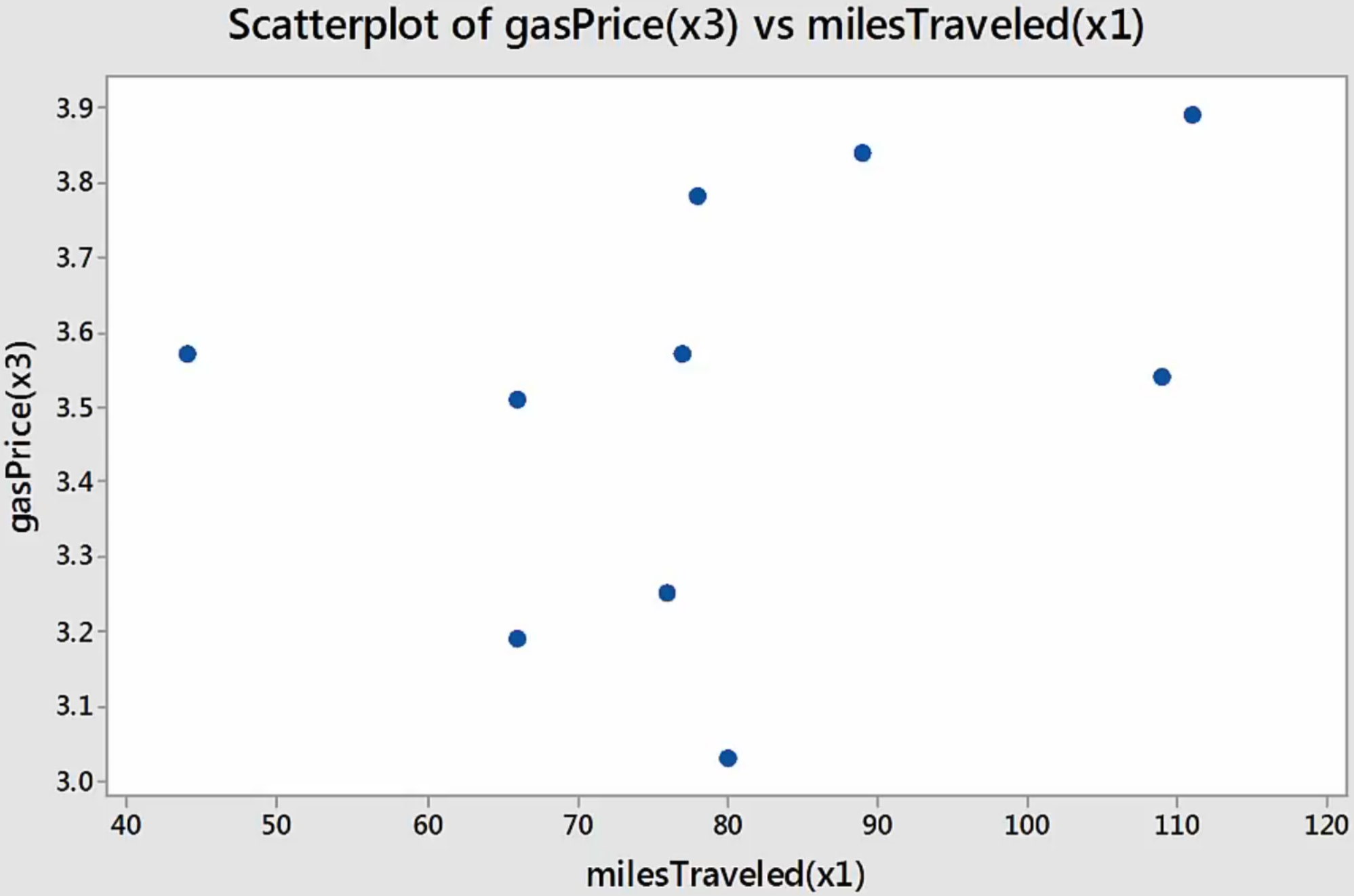
SKETCHING OUT RELATIONSHIPS

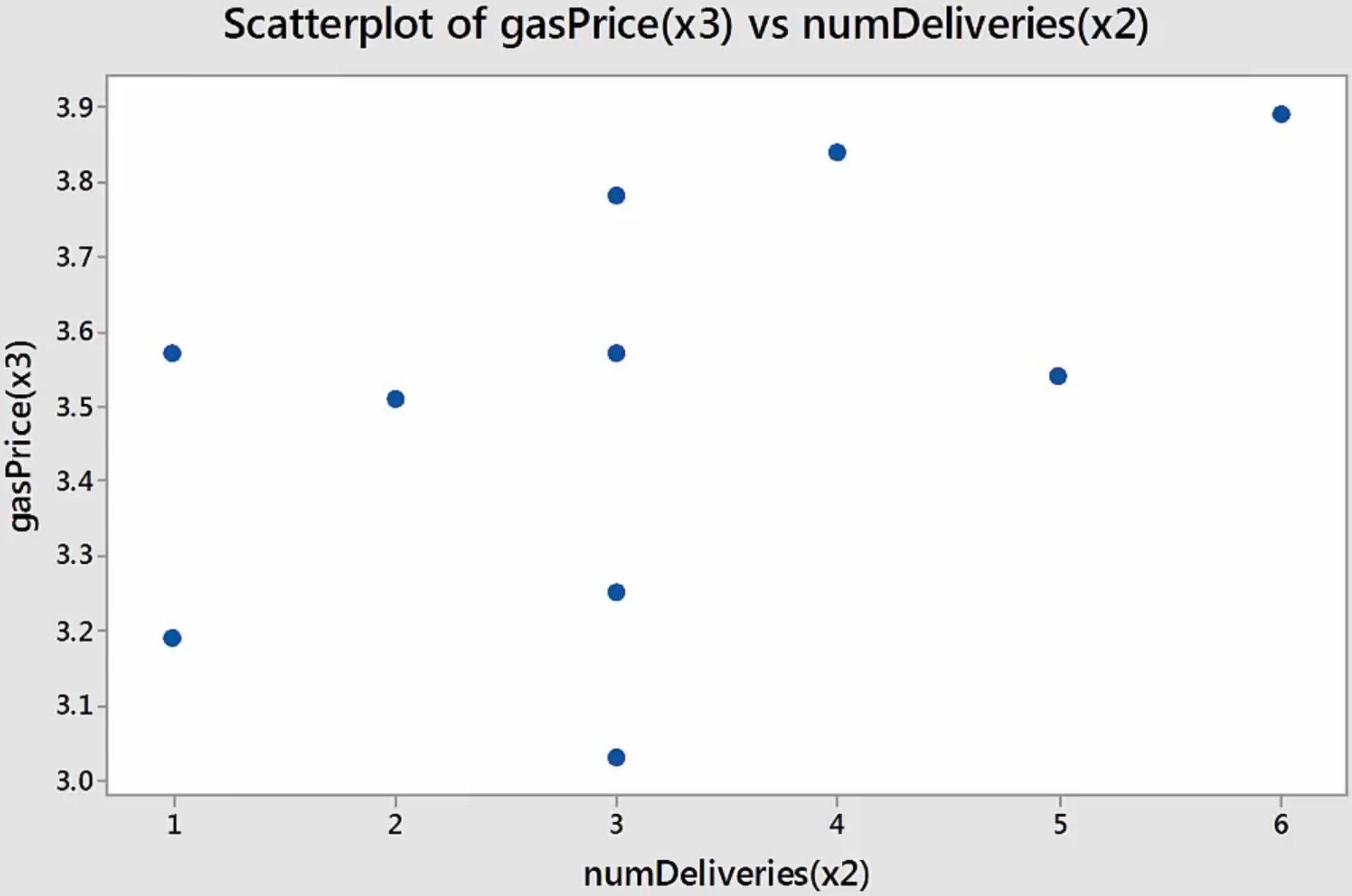


IV TO IV SCATTERPLOTS

Multicollinearity Check


Scatterplot of numDeliveries(x2) vs milesTraveled(x1)
numDeliveries(x2)
6
5
4
3
2
1
40
50
60
70
80
90
100
110
120
milesTraveled(x1)


Scatterplot of gasPrice(x3) vs milesTraveled(x1)
gasPrice(x3)
3.9
3.8
3.7
3.6
3.5
3.4
3.3
3.2
3.1
3.0
40
50
60
70
80
90
100
110
120
milesTraveled(x1)


Scatterplot of gasPrice(x3) vs numDeliveries(x2)
gasPrice(x3)
3.9
3.8
3.7
3.6
3.5
3.4
3.3
3.2
3.1
3.0
1
2
3
4
5
6
numDeliveries(x2)

IV SCATTERPLOTS (MULTICOLLINEARITY)