Objectives

- Use basic analytics methods such as distributions, statistical tests and summary operations to investigate a data set.
- Use R as a tool to perform basic data analytics, reporting and basic data visualization.
- Identify a model for your data and define the null and alternative hypothesis.

Putting the Data Analytics Lifecycle into Practice

- From Module 2 you learned a strategy to approach any data analytics problem:
  - **Phase 1: Discovery**
  - **Phase 2: Data Preparation**
  - **Phase 3: Model Planning**
  - **Phase 4: Model Building**
  - **Phase 5: Communicate Results**
  - **Phase 6: Operationalize**

- To begin to analyze the data you need:
  - **A tool that allows you to look at the data – that is "R".**
  - **Skills in basic statistics.**

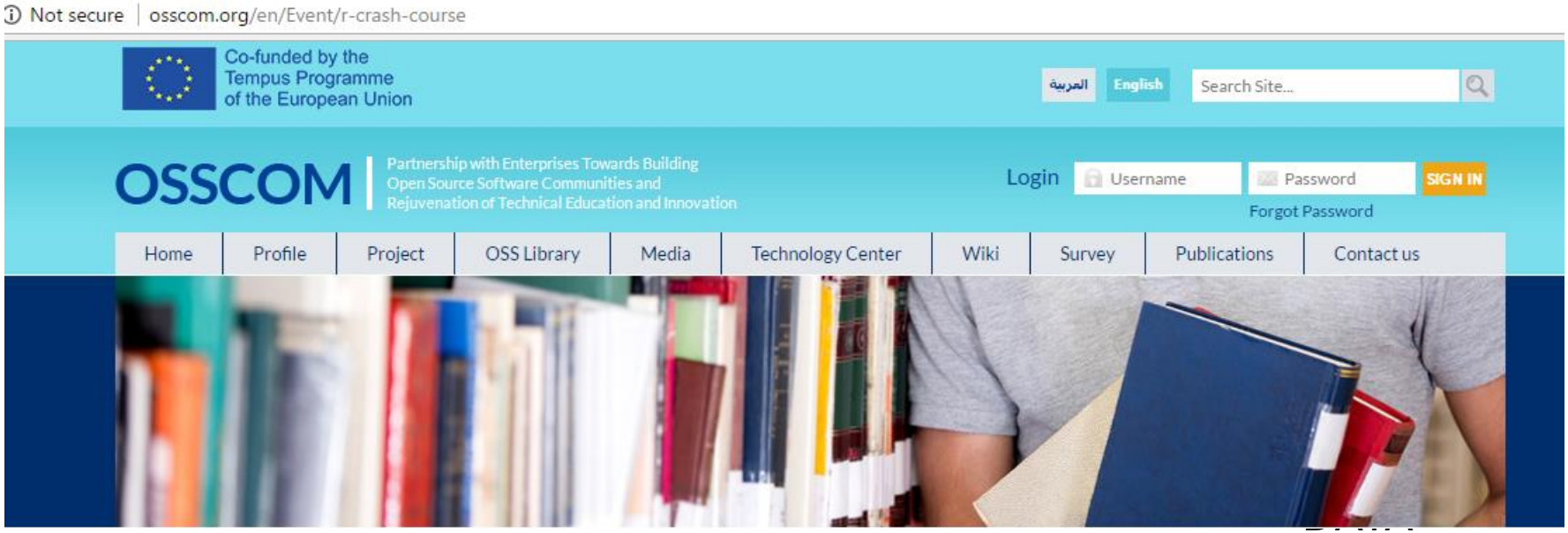
Introduction to R

Using R to Look at Data:

- Using the R Graphical User Interface
- Overview: Getting Data into/out of R
- Data Types Used in R
- Basic R Operations
- Basic Statistics
- Generic Functions

GETTING A HANDLE ON THE DATA

Introduction to R



http://osscom.org/en/content/r-crash-course-materials-0

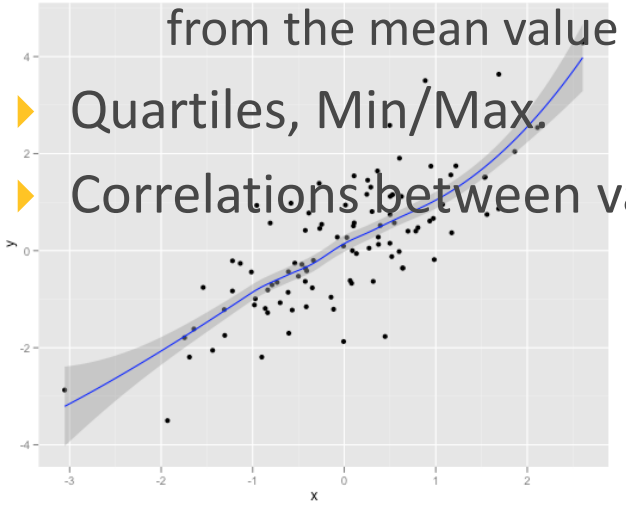
Interactive online course:
https://campus.datacamp.com/courses/free-introduction-to-r

Analyzing and Exploring the Data

Why Visualize?

Summary statistics give us some sense of the data:

- Mean vs. Median
  Median is the middle number in a sorted list of numbers.
- Standard deviation.
  - a quantity expressing by how much the members of a group differ from the mean value for the group.
- Quartiles, Min/Max.
- Correlations between variables.




```
summary(data)
       x                    y
 Min.   :-3.05439   Min.   :-3.50179
 1st Qu.:-0.61055   1st Qu.:-0.75968
 Median : 0.04666   Median : 0.07340
 Mean   :-0.01105   Mean   : 0.09383
 3rd Qu.: 0.56067   3rd Qu.: 0.88114
 Max.   : 2.60614   Max.   : 4.28693
```

Visualization gives us
a more holistic sense

Analyzing and Exploring the Data

Symbol	Names	Definition
Q_1	first quartile lower quartile 25th percentile	splits off the lowest 25% of data from the highest 75%
Q_2	second quartile median 50th percentile	cuts data set in half
Q_3	third quartile upper quartile 75th percentile	splits off the highest 25% of data from the lowest 75%

Analyzing and Exploring the Data

Method 1

1. Use the median to divide the ordered data set into two halves.
    1. If there is an odd number of data points in the original ordered data set, **do not include** the median (the central value in the ordered list) in either half.
    2. If there is an even number of data points in the original ordered data set, split this data set exactly in half.
2. The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.

Example 1

Ordered Data Set: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

	Method 1		
Q_1	15		
Q_2	40		
Q_3	43		

Analyzing and Exploring the Data

Method 1

1.Use the median to divide the ordered data set into two halves.

   1. If there is an odd number of data points in the original ordered data set, **do not include** the median (the central value in the ordered list) in either half.
   2. If there is an even number of data points in the original ordered data set, split this data set exactly in half.

2.The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.

Example 2: Ordered Data Set: 7, 15, 36, 39, 40, 41

	Method 1		
Q_1	15		
Q_2	37.5		
Q_3	40		

A Synthesized Example (Anscombe's Quartet)

4 data sets, characterized by the following. Are they the same, or are they different? See next slides ☺

Property	Values
Mean of x in each case	9
Exact variance of x in each case	11
Exact mean of y in each case	7.5 (to 2 d.p)
Variance of Y in each case	4.13 (to 2 d.p)
Correlations between x and y in each case	0.816
Linear regression line in each case	Y = 3.00 + 0.500x (to 2 d.p and 3 d.p resp.)

i

X	y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

ii

x	y
10.00	9.14
8.00	8.14
13.00	8.74
9.00	8.77
11.00	9.26
14.00	8.10
6.00	6.13
4.00	3.10
12.00	9.13
7.00	7.26
5.00	4.74

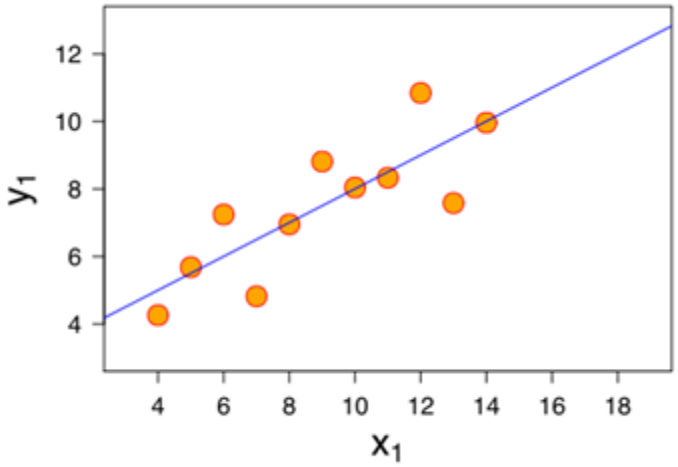
iii

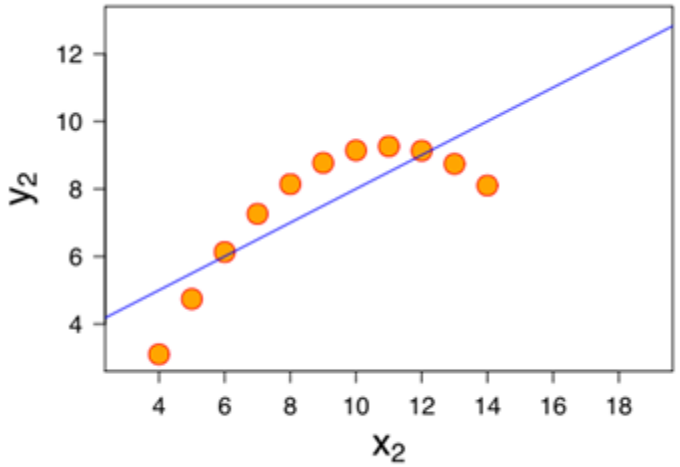
x	y
10.00	7.46
8.00	6.77
13.00	12.74
9.00	7.11
11.00	7.81
14.00	8.84
6.00	6.08
4.00	5.39
12.00	8.15
7.00	6.42
5.00	5.73

iv

x	y
8.00	6.58
8.00	5.76
8.00	7.71
8.00	8.84
8.00	8.47
8.00	7.04
8.00	5.25
19.00	12.50
8.00	5.56
8.00	7.91
8.00	6.89

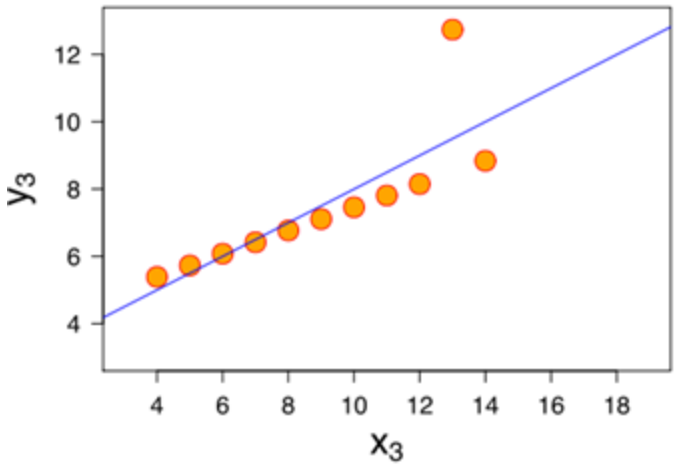
Moral: Visualize Before Analyzing!

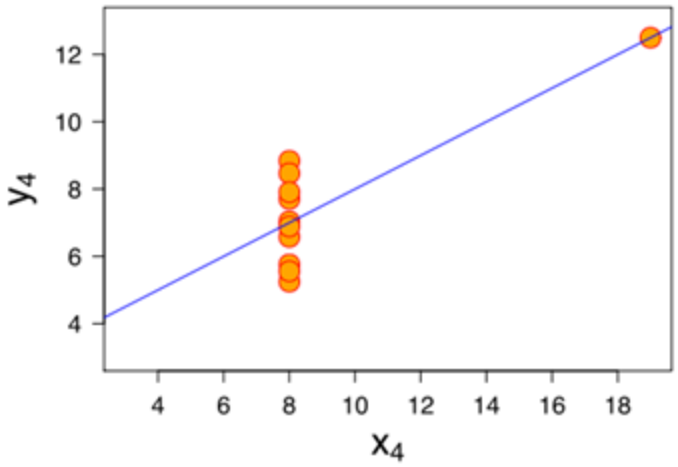




However, if we visualize each data set using a scatterplot and a regression line superimposed over each plot, the datasets appear quite different. Dataset 1 is the best candidate for a regression line, although there is a lot of variation. Dataset 2 is definitely non-linear.

Moral: Visualize Before Analyzing!





Dataset 3 is a close match, but over predicts at higher value of x and has an extreme outlier. And Dataset 4 isn't captured at all by a simple regression line.

Visualizing Your Data

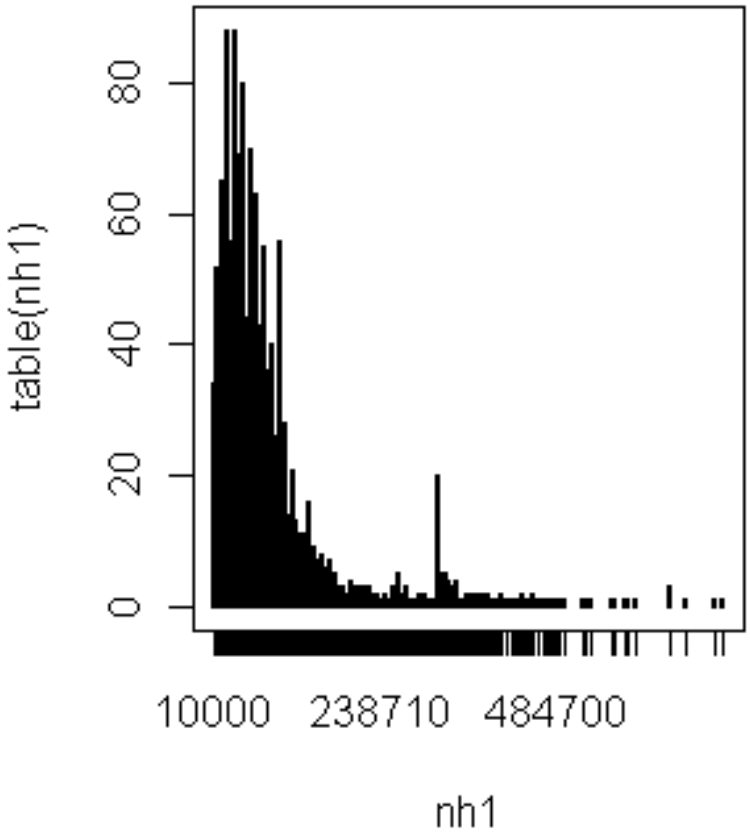
- Examining the distribution of a single variable
- Analyzing the relationship between two variables
- Establishing multiple pair wise relationships between variables
- Analyzing a single variable over time
- Data exploration versus data presentation

Examining the Distribution of a Single Variable Example 1

Graphing a single variable

- plot(sort(.)) – for low volume data
- hist(.) – a histogram
- plot(density(.)) – densityplot
  - A "continuous histogram“

- Example
  - Frequency table of household income



Analyzing a Single Variable over Time

What?

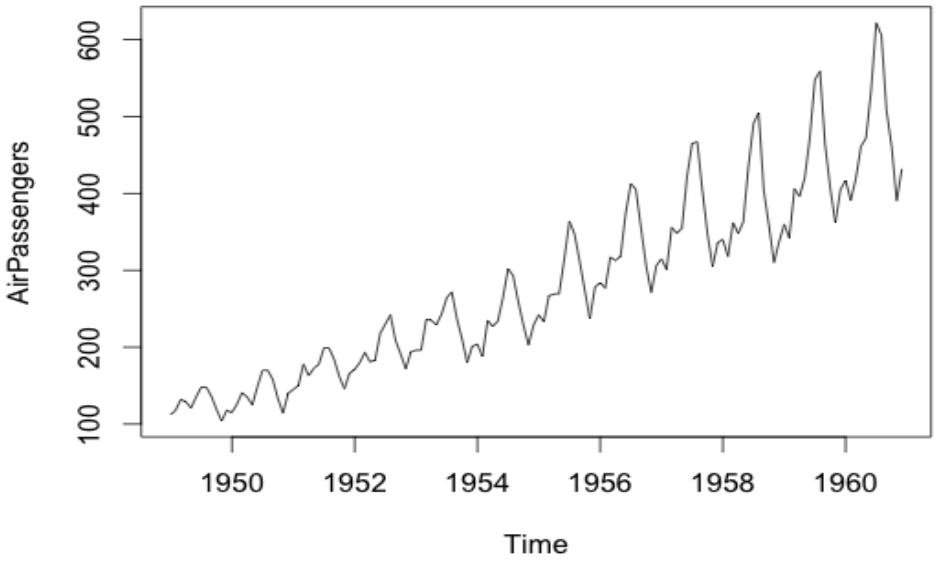
- Looking for …
  - Data range
  - Trends
  - Seasonality

How?

- Use time series plot

Example:

- International air travel (1949-1960)
- Upward trend: growth appears superlinear
- Seasonality
  - Peak air travel around Nov. with smaller peaks near Mar. and June



What are we looking for?

A sense of the data range

- If it's very wide, or very skewed, try computing the log

Outliers, anomalies

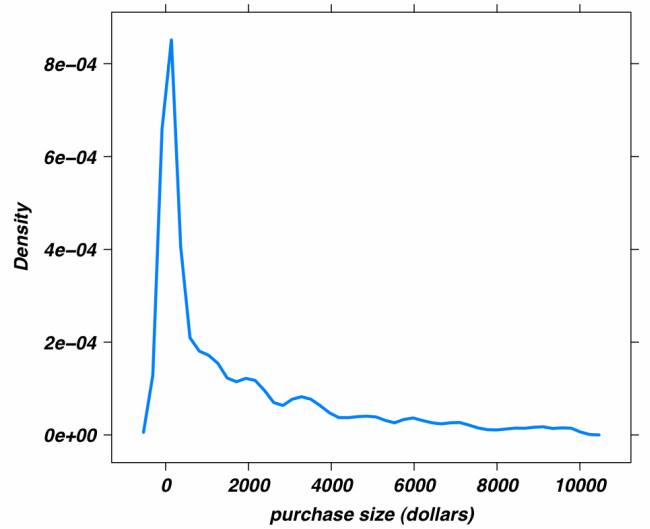
- Possibly evidence of dirty data

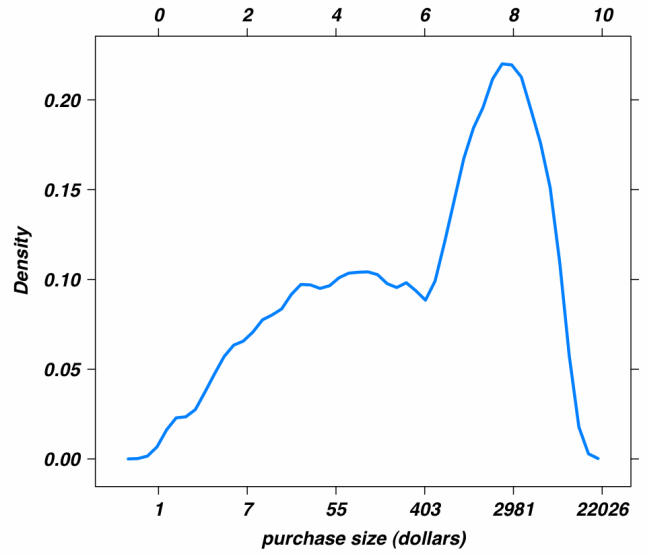
Shape of the Distribution

- Unimodal? Bimodal?
- Skewed to left or right?
- Approximately normal? Approximately lognormal?

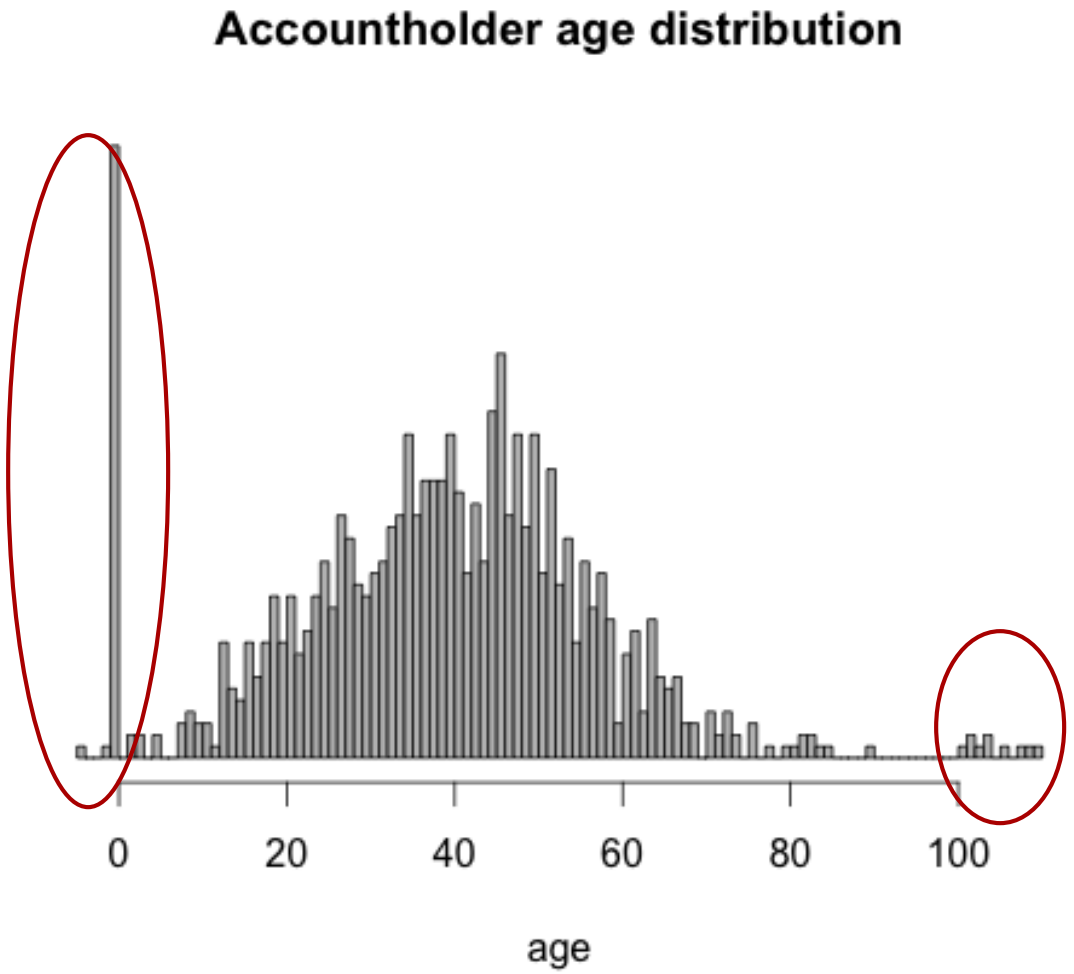
Example - Distribution of purchase size ($)

- Range from 0 to < $10K, left skewed
- Typical of monetary data
- Plotting log of data gives better sense of distribution
- Two purchasing distributions
  - ~ $55
  - ~ $2900





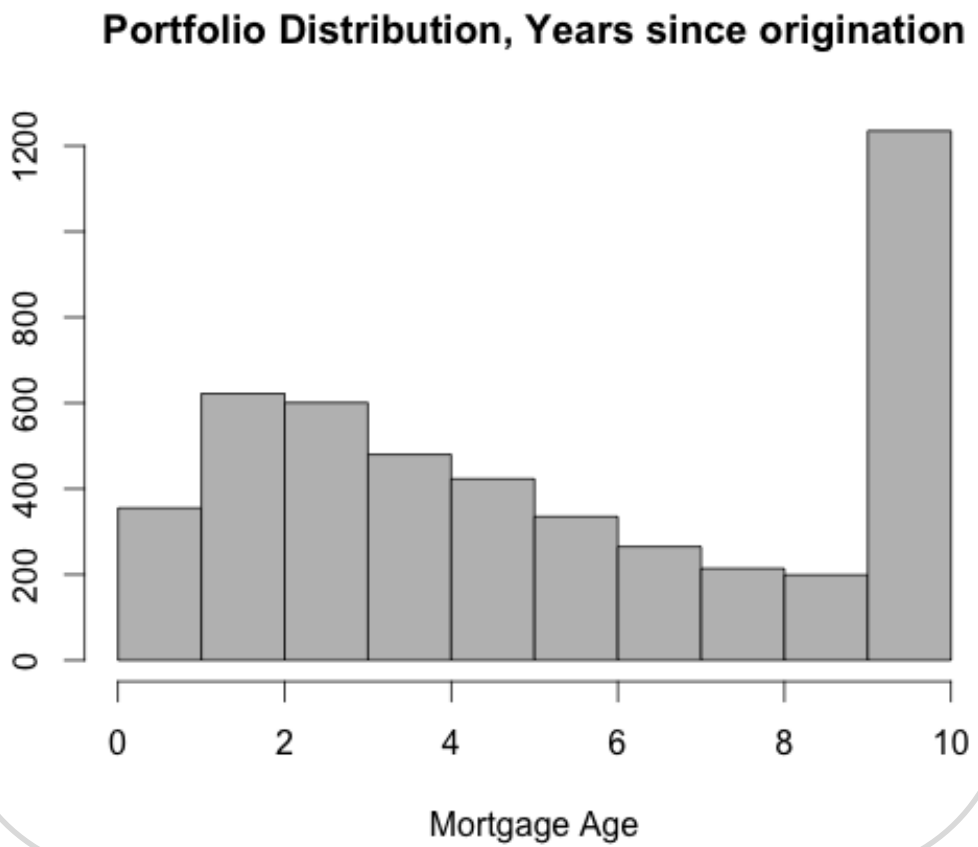
Evidence of Dirty Data



Missing values?

Mis-entered data? Inherited accounts?

"Saturated" Data

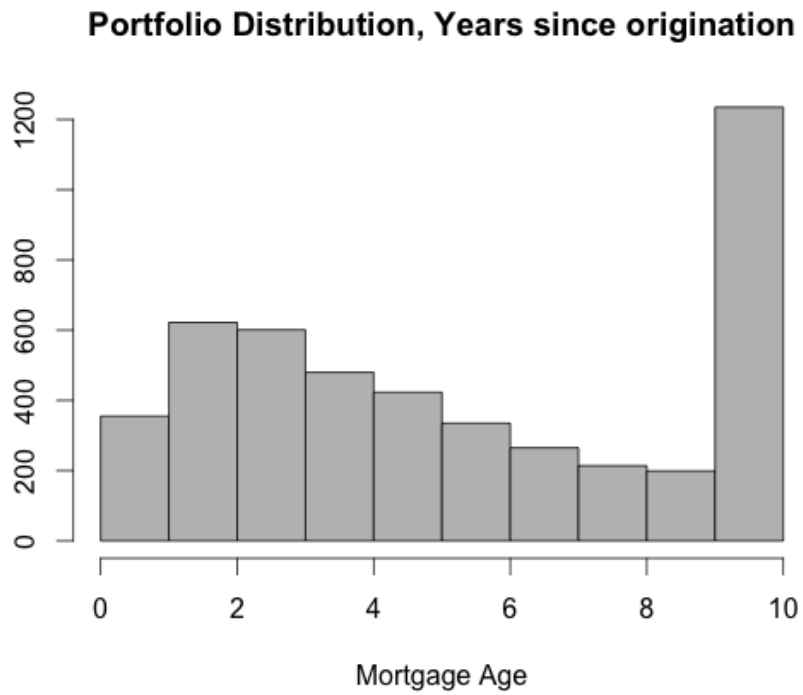


Do we really have no mortgages older than 10 years?

Or does the year 2005 in the origination field mean "2005 or prior"?

"Saturated" Data

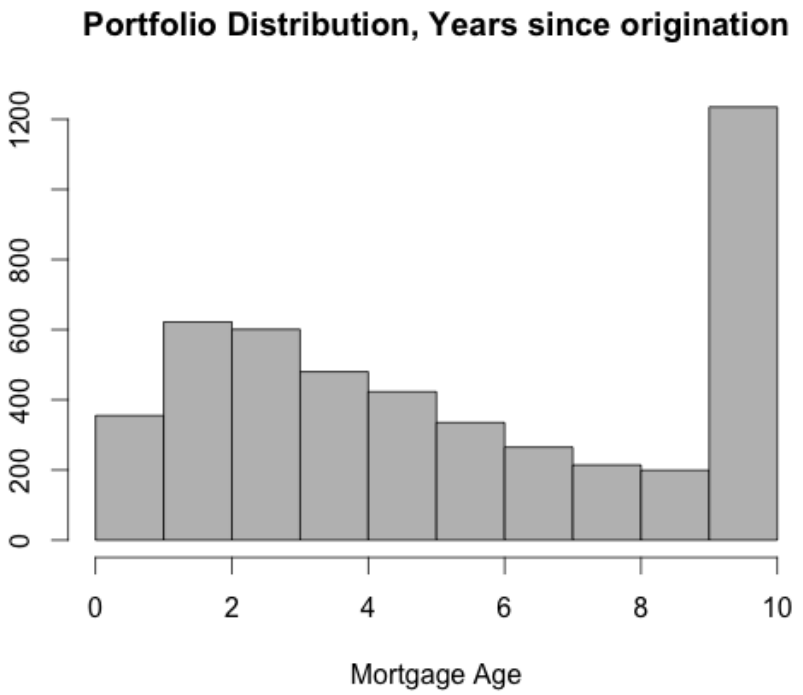
Here's another example of dirty (or at least, "incompletely documented" data). We are looking at the age of mortgages in our bank's home loan portfolio. The age is calculated by subtracting the origination date of the loan from "today" (2011).



The first thing we notice is that we don't seem to have loans older than 10 years old – and we also notice that we have a disproportionate number of ten year old loans, relative to the age distribution of the other loans.

"Saturated" Data

One possible reason for this is that the date field for loan origination may have been "overloaded" so that "2001" is actually a beacon value that means "2001 or prior" rather than literally 2001. (This sometimes happens when data is ported from one system to another, or because someone, somewhere, decided that origination dates prior to 2001 are not relevant).

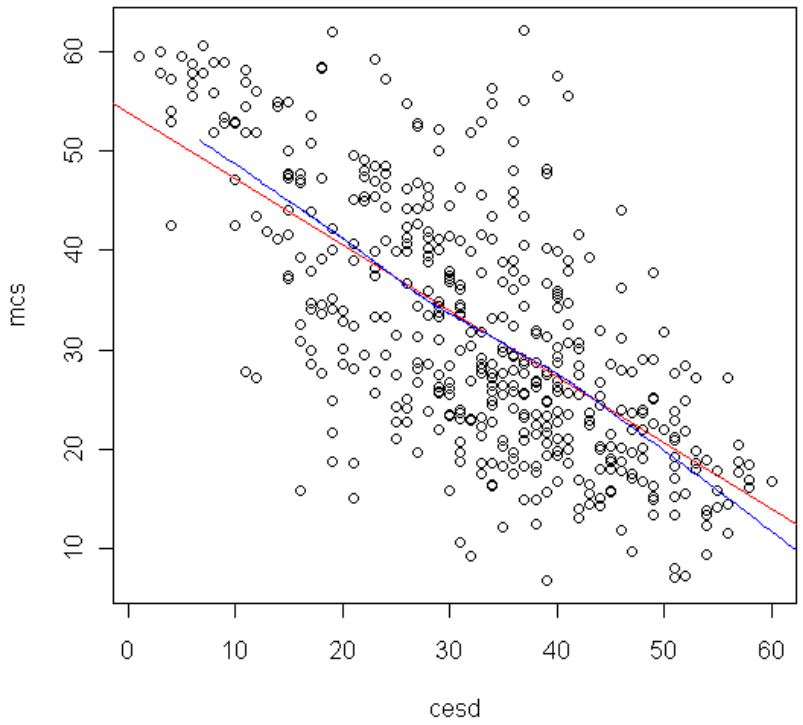


"Saturated" Data

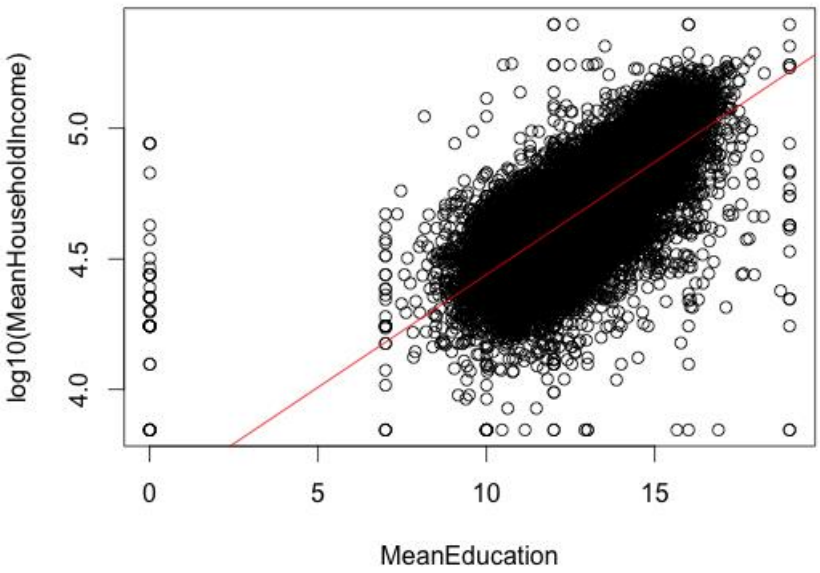
What would we do about this? If we are analyzing probability of default, it is probably safe to eliminate the data (or keep the assumption that the loans are 10 years old), since 10 year old mortgages default quite rarely (most defaults occur before about the 4th year). For different analyses, we may need to search for a source of valid origination dates (if that is possible).

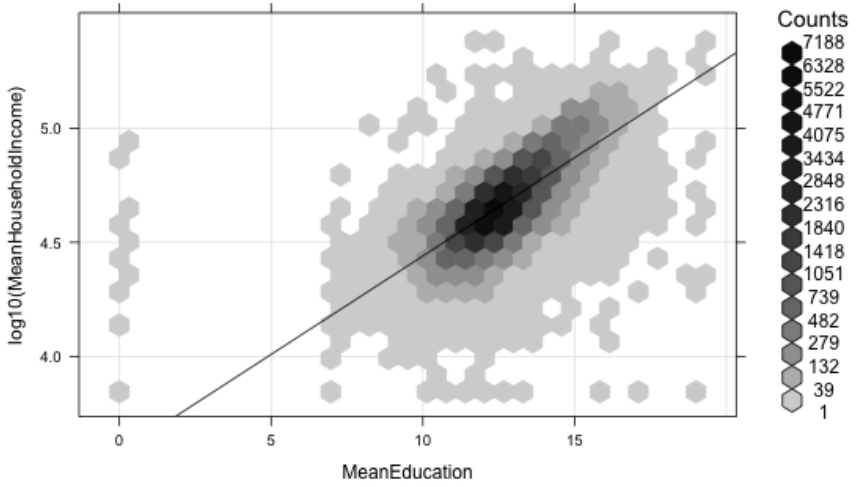
Two Variables: What are we looking for?

- Is there a relationship between the two variables?
  - Linear? Quadratic?
  - Exponential?
    - Try semi-log or log-log plots
  - Is it a cloud?
    - Round? Concentrated? Multiple Clusters?
- How?
  - Scatterplots
- Example
  - linear fit
  - Fairly linear relationship, but with wide variance



Two Variables: High Volume Data - Plotting





Scatterplot:
Overplotting makes it difficult to see structure

Hexbinplot:
Now we see where the data is concentrated.