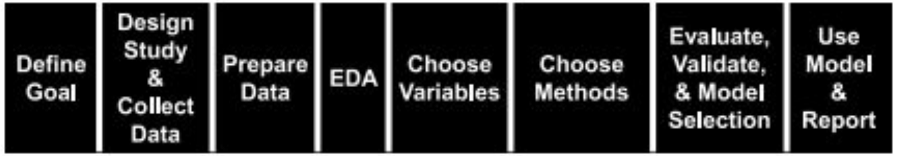
Plan for the week

- M: Fall Break
- W: Exploratory Data Analysis (EDA)
  - Mini-project
- F: Section
  - Data Cleaning

Exploratory Data Analysis

Data Science in five (easy) steps

1. State problem or research question
2. Collect and clean data
3. EDA (before trying fancy math!)
4. Quantitative analysis and evaluation
5. Present results, including visualizations



Define Goal

Design Study & Collect Data

Prepare Data

EDA

Choose Variables

Choose Methods

Evaluate, Validate, & Model Selection

Use Model & Report

Image Source

EDA is a preliminary approach to data analysis that:

- Aims to find patterns and structure in data
- Uncovers outliers and anomalies in the data
- Can help to quickly assess relationships between explanatory (e.g., bill) and response (e.g., tip) variables
  - Explanatory variables are also called independent variables
  - Response variables are also called dependent variables

EDA (John Tukey)

Exploratory data analysis focuses on exploring data to:

- understand the data’s underlying structure
- develop intuition about the data set
- consider how the data were collected (to aid in cleaning)
- decide how to further investigate with more formal statistical methods

Source: Udacity

EDA is any initial investigation of data

- Visualizing one’s data is the most useful way to get an initial understanding of underlying patterns and anomalies
- John Snow’s mapping of Cholera used EDA, as did Florence Nightingale’s mapping of causes of death

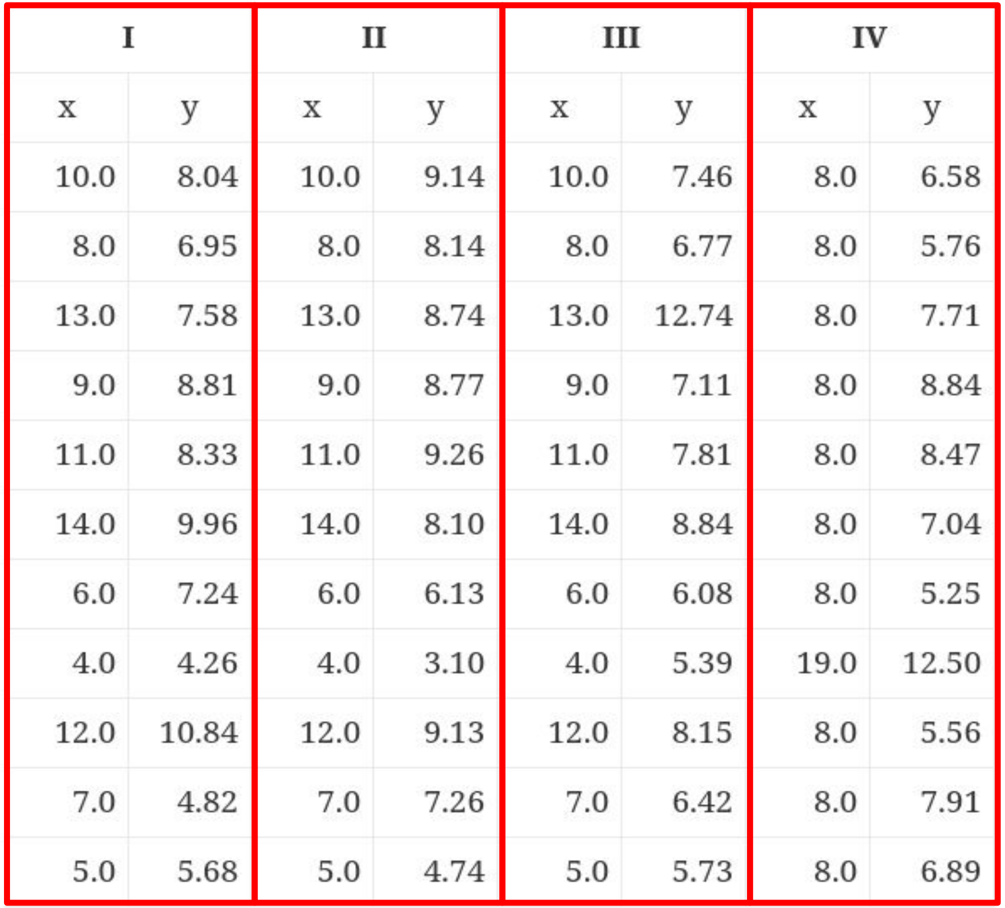
A small, synthetic dataset

Anscombe’s quartet

- *Graphs in Statistical Analysis* (Francis J. Anscombe 1973)
- Anscombe advocated for graphing data before analyzing it
- He wanted to challenge the mainstream sentiment that “numeric calculations are exact, but graphs are rough”

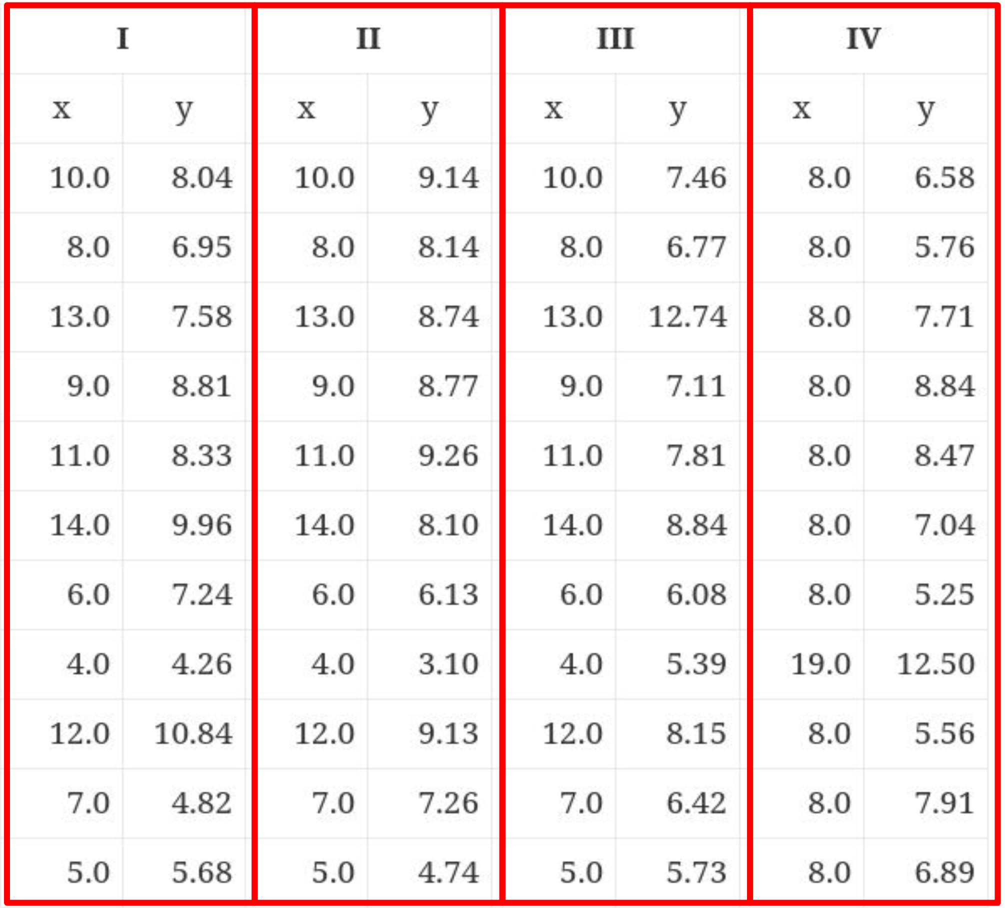
- (He also wanted to emphasize the effects of outliers)

Anscombe created four numeric data sets



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

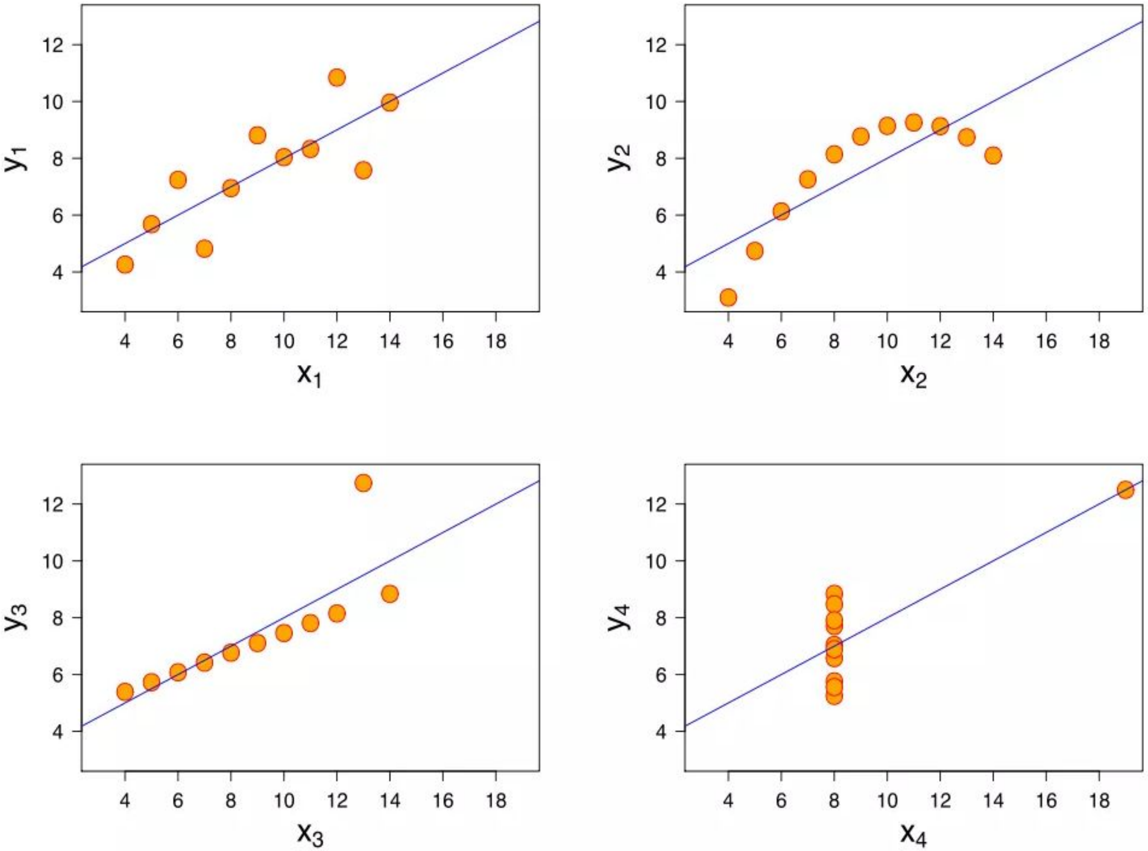
Anscombe calculated statistics for these data sets



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

- The medians of the x’s are all the same and equal to 9.0
- The means of the y’s are all the same and equal to 7.5
- All four data sets are best approximated by the same line: $y = 3 + ½ x$

Anscombe also graphed his data sets



When viewed visually, only the first data set seems to satisfy a relationship anything like *y = 3 + ½ x*

A small, non-synthetic dataset

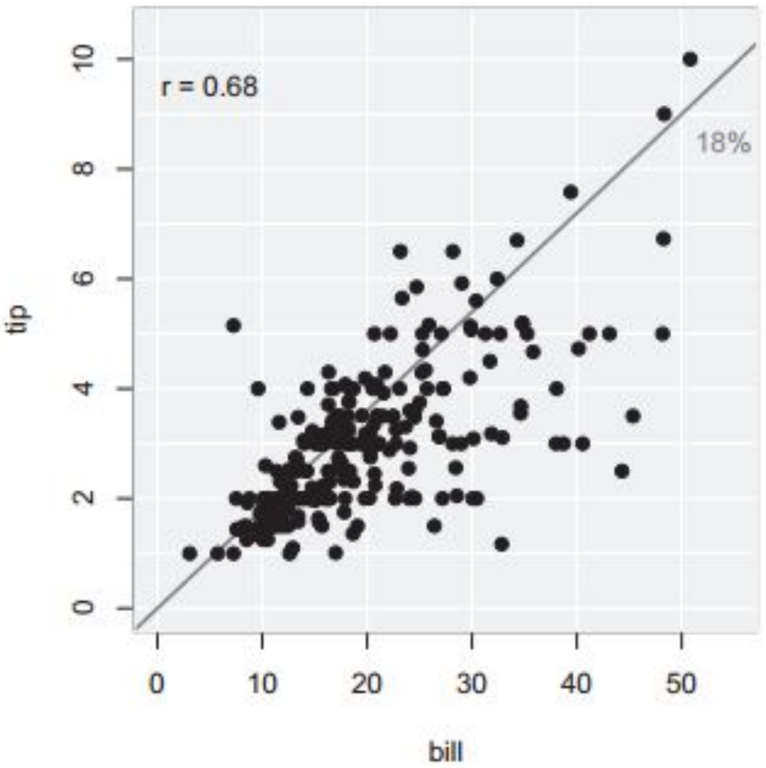
A small data set

- A server had some hunches about what makeup of customers might be the best tippers
- So, he collected some data to investigate this question
- He recorded 244 of his tips over a period of a few months in 1995

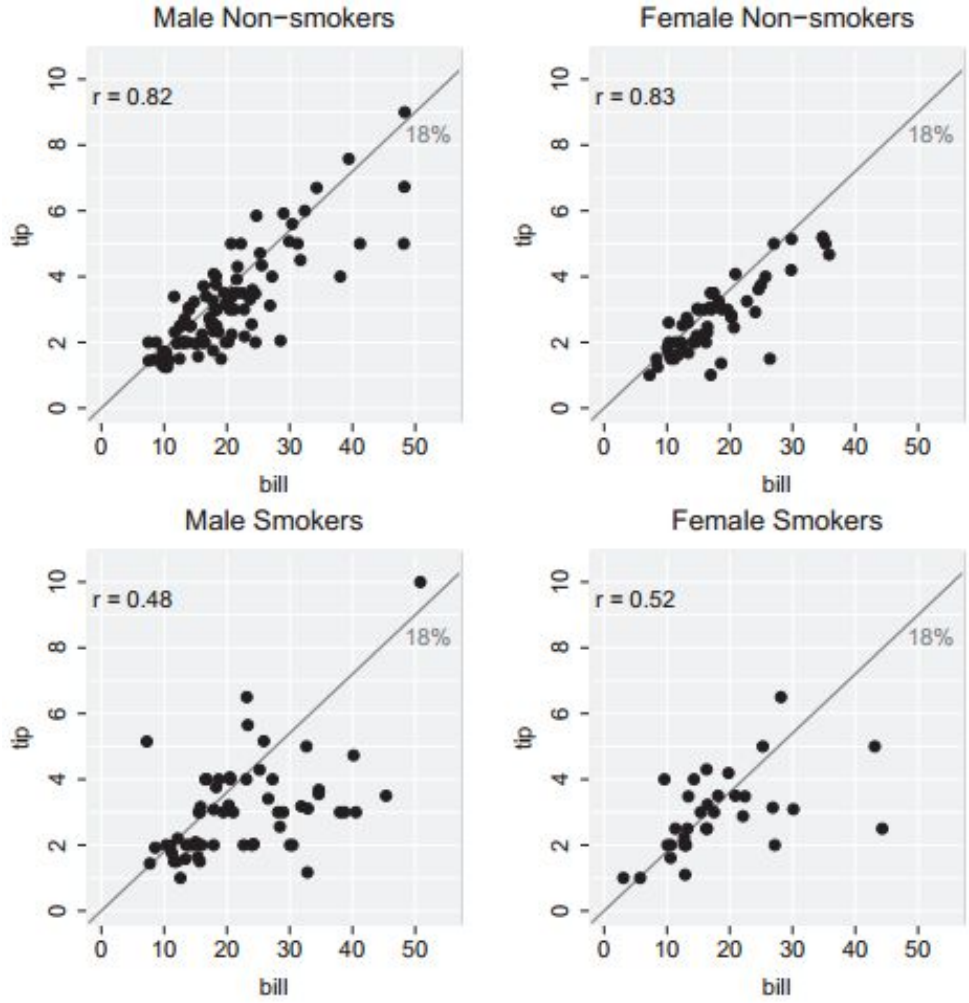
He logged the following information:

- Tip
- Total bill
- Smoking status
- Time of day
- Day of week
- Size of party
- Gender of tipper

Is there a relationship between the bill and the tip?



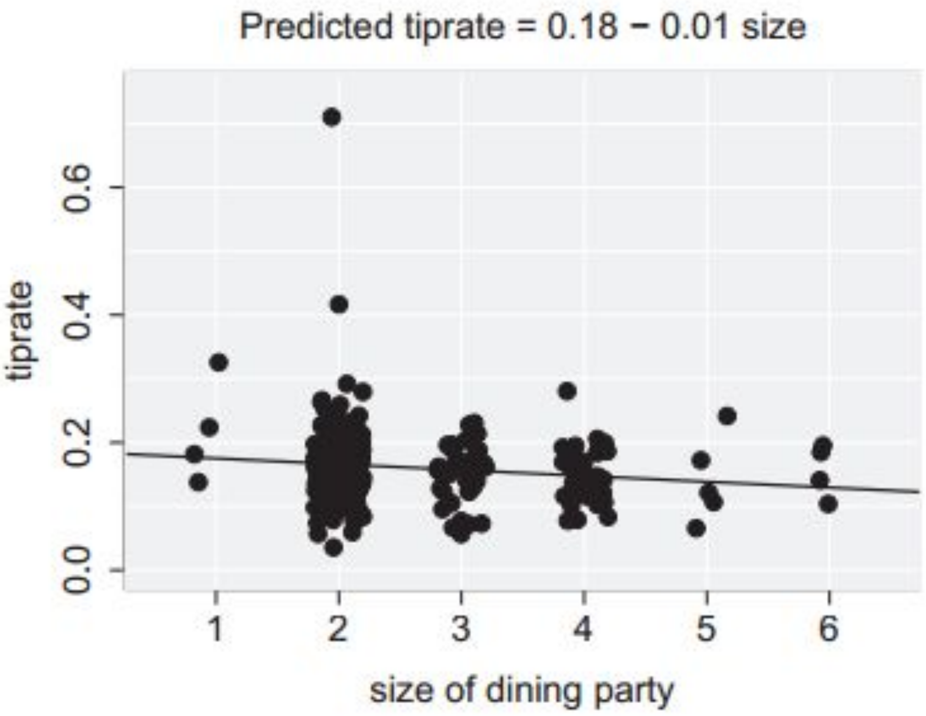
- An equation of a line can be found to best fit the data
- Given no extra information one might assume that this line would be simply 18% of the bill



Conditioning

- Here we drill down into the data some more
- We generate four graphs, each one conditioned on gender and smoking status
- We make more fine-grained observations about the data this way

Adding the size of a group to our model



- Why does the tip rate decline with party size?
- Maybe because the bigger the group, the bigger the bill, and since the tip is more money, the rate goes down?
- Restaurants are onto this phenomenon: Many automatically charge 20% or more for parties of 6 or more!

Vietnam Draft

Vietnam War Draft Procedure

- In 1969, the Vietnam war was at its height. The Selective Service was charged with finding a fair procedure for drafting young men into the U.S. military.
- The aim of the procedure was to be fair in the sense of not favoring any culturally or economically defined subgroup of American men.
- They decided that selecting draftees based solely on a person’s birth date would be fair.

A birthday lottery

- Pieces of paper representing 366 days of the year (including February 29) were placed into plastic capsules, poured into a rotating drum, and then selected one at a time.
- The first number selected was 258, which meant that someone born on the 258th day of the year (September 14th) would be among the first group to be drafted.
- The second number was 115, so someone born on the 115th day (April 24th) was among the second group to be drafted.
- All 366 birth dates were assigned draft numbers in this way.
- Was the draft lottery fair? Let’s ask the data!

Draft rank by month during Vietnam war: raw data

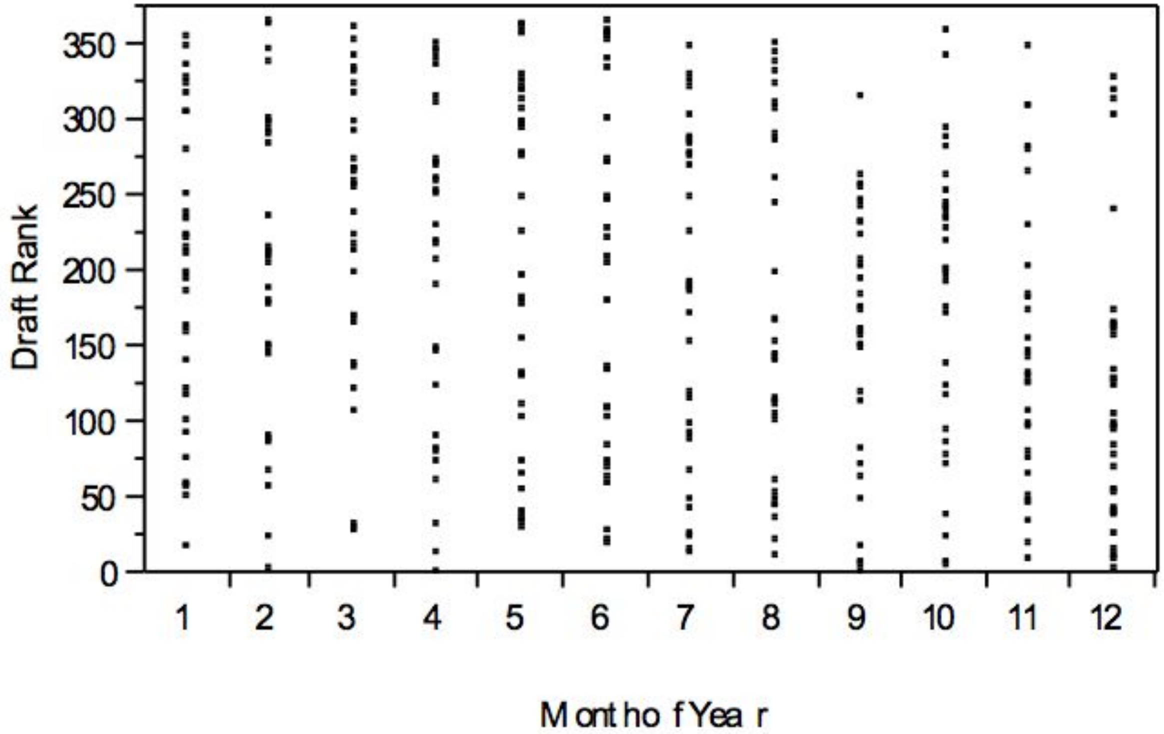


Image Source

Draft rank by month during Vietnam war: box plots

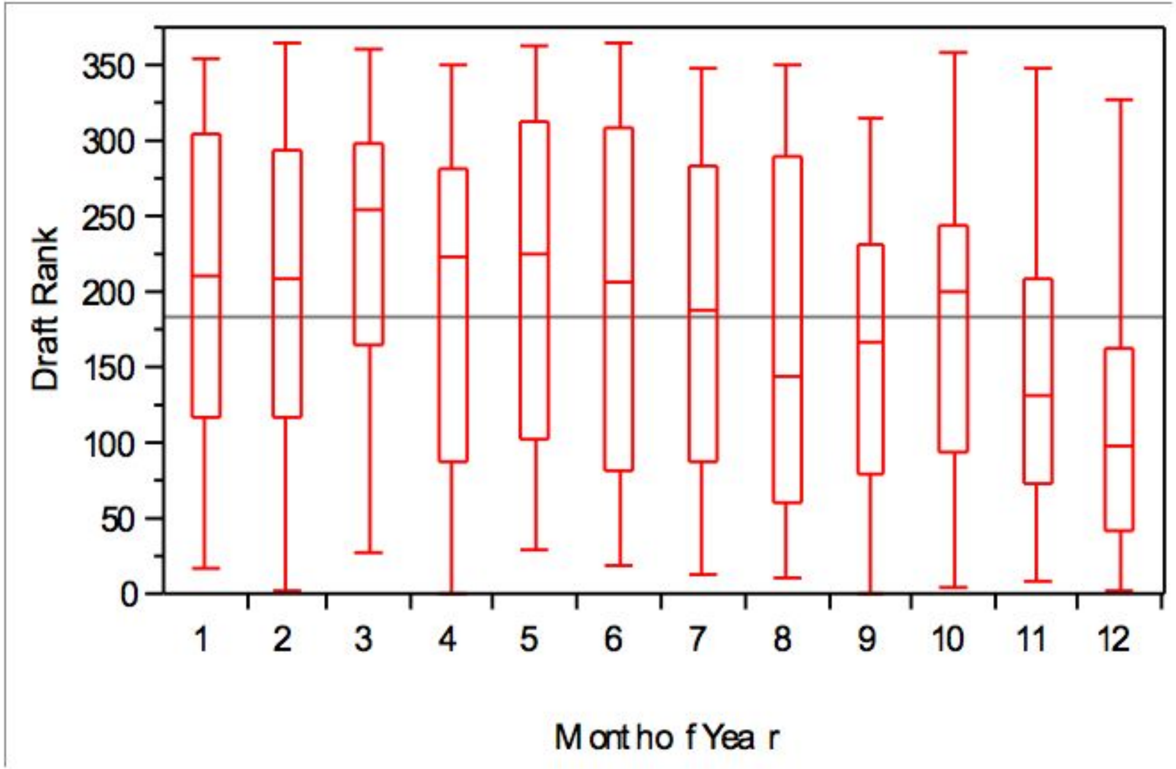
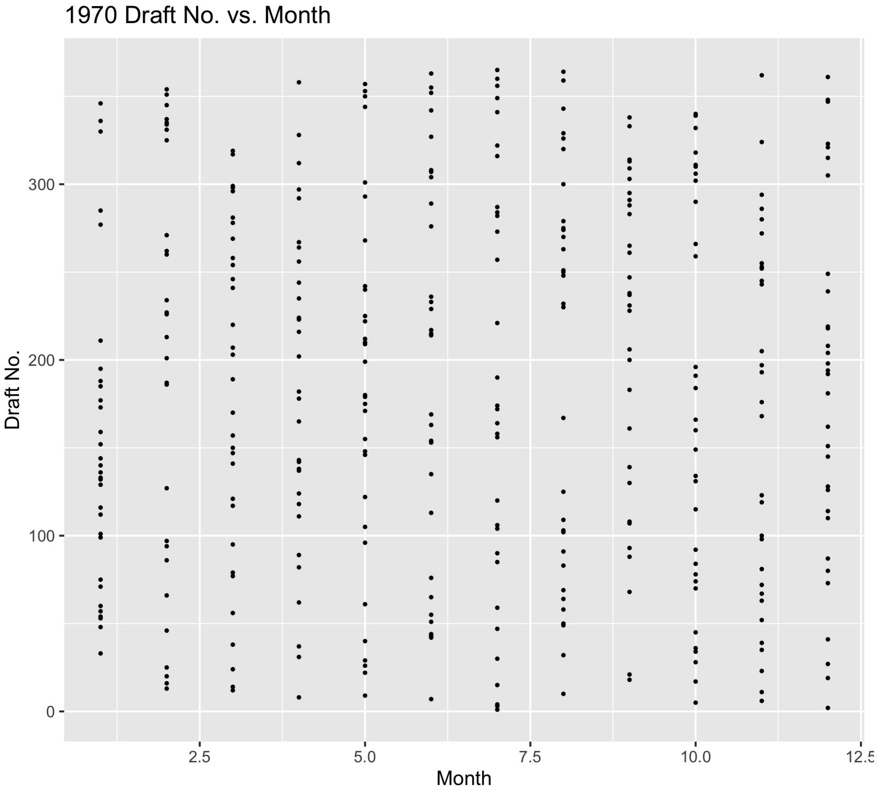
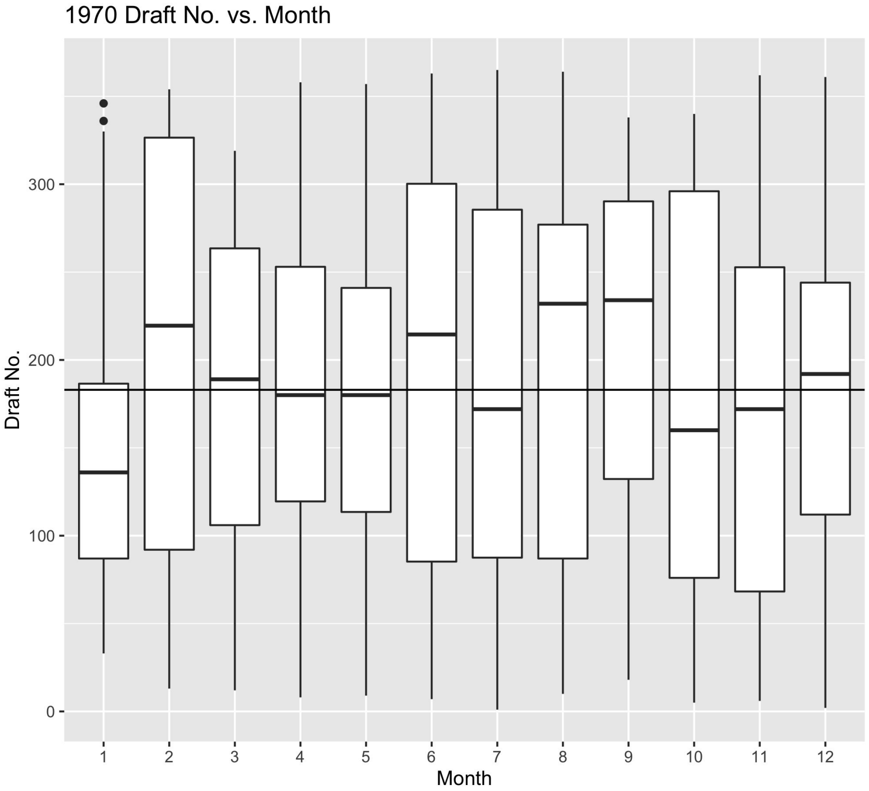


Image Source

Draft rank by month during Vietnam war: 1970





Conclusion

By putting birthdays into the drum in sequence, and by insufficiently mixing them up, later days in the calendar year were more likely to be selected than earlier ones.