Module 2, Day 1

Alireza Samar

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

MLDS
MACHINE LEARNING
FOR DATA SCIENCE

Who Am I?

Alireza Samar @alirezasmr

Graduate Research Assistant at UTM MLDS

Curator at Machine Learning Weekly - mlweekly.com

UTM MLDS MACHINE LEARNING FOR DATA SCIENCE

ML

MLDS

What We've Learned in Module 1

- Anaconda Package Manager
- Install and run code on Jupyter
- Version Controlling Concept
- Git
- GitHub and GitHub Desktop

MLDS

What We've Learned in Module 1

- Statistics
- Linear Algebra
- Optimization
- Bayes Rule
- Maximum Likelihood, Gradient Descent, ...

MLDS

What We've Learned in Module 1

- Python Basics (Syntax, Arrays, Loops, Functions and etc)
- Numpy
- Matplotlib

MLDS

Workshop Materials

GitHub: **UTMMLDS**

https://github.com/utmmlds

MLDS

https://youtu.be/14VYnFhBKcY

MLDS

The E[DA]

Exploratory Data Analysis

Edward Tufte (2006). *Beautiful Evidence*, Graphics Press LLC. www.edwardtufte.com

UTM UNIVERSITI TEKNOLOGI MALAYSIA

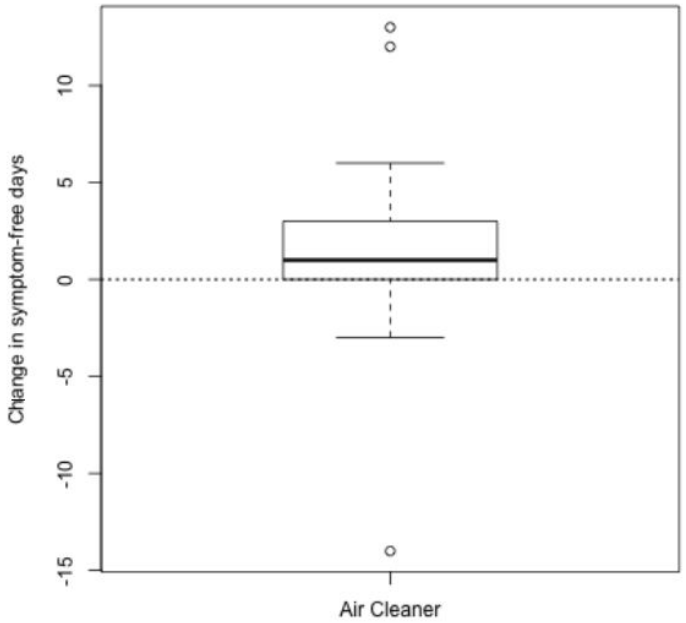
MLDS MACHINE LEARNING FOR DATA SCIENCE

Principles of Analytic Graphics

- **Principle 1**: Show comparisons
  - Evidence for a hypothesis is always ***relative*** to another competing hypothesis.
  - Always ask "Compared to What?"

MLDS

Show Comparisons



A boxplot which looks at the effect of an air cleaner on the asthma symptoms of children.

An air cleaner was introduced into a child's home, to reduce indoor air pollution levels.

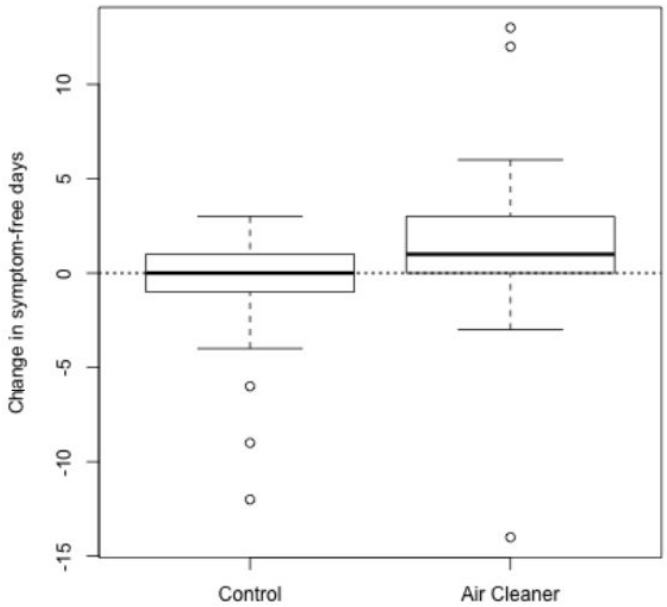
Positive outcome! median increase was about one symptom-free day over 2 weeks.

Compared to what?

Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

MLDS

Show Comparisons



This was a randomized control trial that looked at installing an air cleaner in a child's home.

So in the control homes, the average change in the symptom free case was about 0.

Well relative to doing nothing, the air cleaner is actually a little bit better at showing an improvement in the child's symptoms.

Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

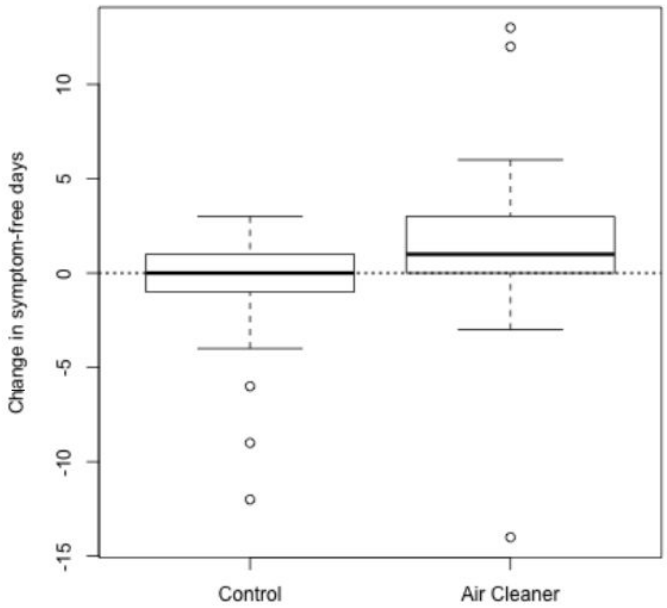
MLDS

Principles of Analytic Graphics

- **Principle 1**: Show comparisons
  - Evidence for a hypothesis is always ***relative*** to another competing hypothesis.
  - Always ask "Compared to What?"
- **Principle 2**: Show causality, mechanism, explanation, systematic structure
  - What is your causal framework for thinking about a question?

MLDS

Show causality, mechanism



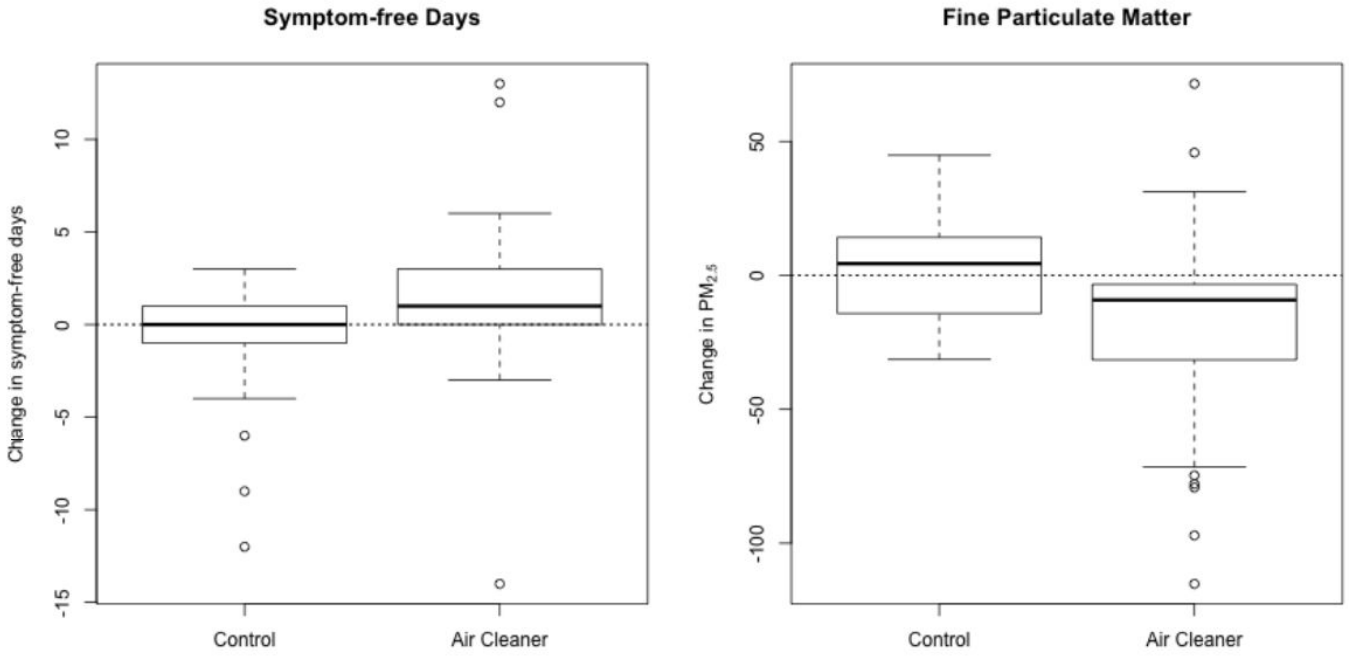
We saw that if you install an air cleaner in a child's home, that on average, they're going to experience a one symptom-free day increase, so a better outcome in their asthma symptoms.

Why is it that installing an air cleaner in a child's home, improves their symptoms?

Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

MLDS

Show causality, mechanism



Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

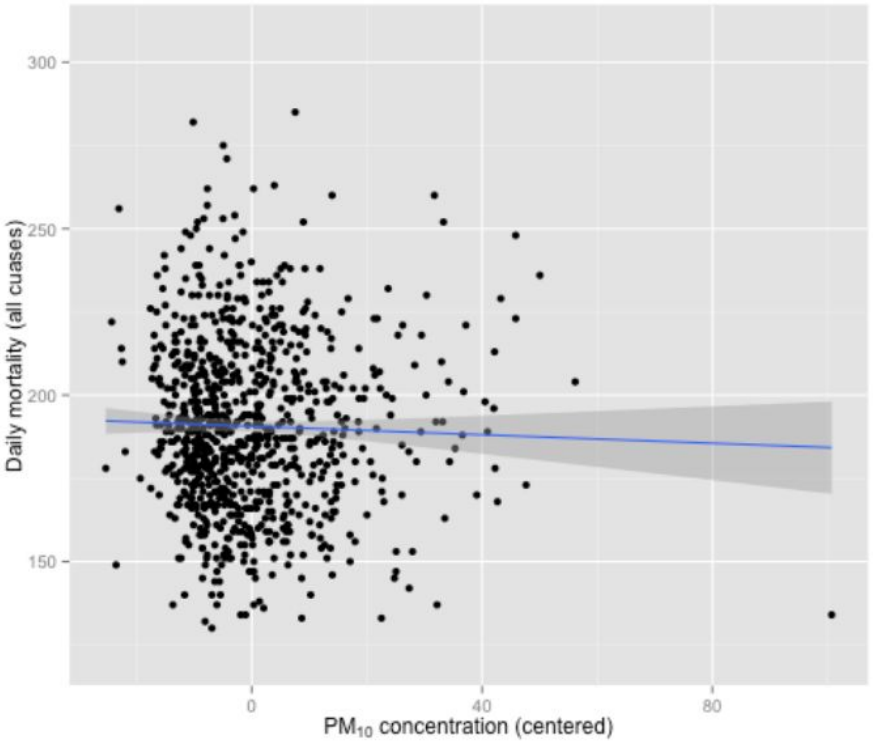
MLDS

Principles of Analytic Graphics

- **Principle 1**: Show comparisons
  - Evidence for a hypothesis is always ***relative*** to another competing hypothesis.
  - Always ask "Compared to What?"
- **Principle 2**: Show causality, mechanism, explanation, systematic structure
  - What is your causal framework for thinking about a question?
- **Principle 3**: Show multivariate data
  - Multivariate = more than 2 variables
  - The real world is multivariate
  - Need to "escape flatland"

MLDS

Show Multivariate Data



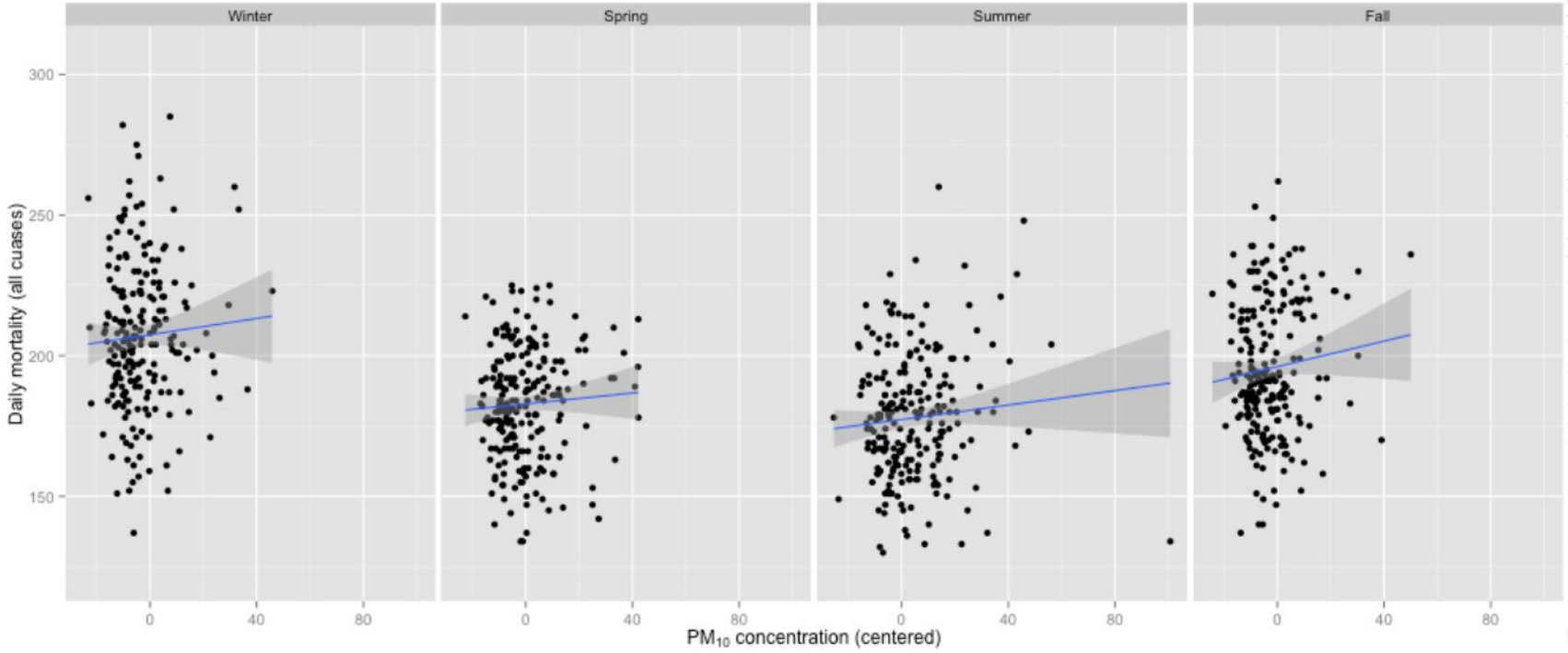
The X axis we have, particulate matter less than 10 microns in aerodynamic diameter, the concentrations of those, from day to day.
Every circle → a daily concentration

The y axis, we have the daily mortality in New York City. This is for the time period 1987 to, 2000.

How does this relationship change across different seasons?

MLDS

Show Multivariate Data



MLDS

Simpson's Paradox

Google it!

MLDS

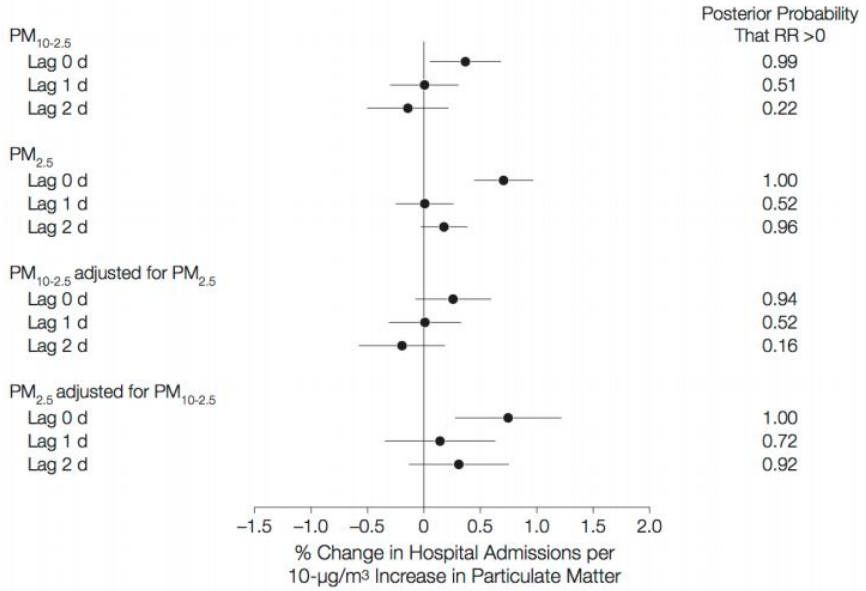
Principles of Analytic Graphics

- **Principle 1**: Show comparisons
  - Evidence for a hypothesis is always ***relative*** to another competing hypothesis.
  - Always ask "Compared to What?"
- **Principle 2**: Show causality, mechanism, explanation, systematic structure
  - What is your causal framework for thinking about a question?
- **Principle 3**: Show multivariate data
  - Multivariate = more than 2 variables
  - The real world is multivariate
  - Need to "escape flatland"
- **Principal 4**: Integration of evidence
  - Completely integrate words, numbers, images, diagrams
  - Data graphics should make use of many modes of data presentation
  - **Don't let the tool drive the analysis**

MLDS

Integrate Different Modes of Evidence

Figure 2. Percentage Change in Emergency Hospital Admissions Rate for Cardiovascular Diseases per a 10-µg/m³ Increase in Particulate Matter



Estimates are on average across 108 counties. $PM_{2.5}$ indicates particulate matter is 2.5 µm or less in aerodynamic diameter; PM_{10}, particulate matter is 10 µm or less in aerodynamic diameter; $PM_{10-2.5}$, particulate matter is greater than 2.5 µm and 10 µm or less in aerodynamic diameter; RR, relative risk. Error bars indicate 95% posterior intervals.

MLDS

Principles of Analytic Graphics

- **Principle 5**: Describe and document the evidence with appropriate labels, scales, sources, etc.
  - A data graphic should tell a complete story that is credible
- **Principle 6**: Content is king
  - Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content

MLDS

Why do we use graphs in data analysis?

MLDS

Why do we use graphs in data analysis?

- To understand data properties
- To find patterns in data
- To suggest modeling strategies
- To "debug" analyses
- To communicate results

MLDS

Why do we use graphs in data analysis?

- To understand data properties
- To find patterns in data
- To suggest modeling strategies
- To "debug" analyses
- To communicate results

Exploratory graphs

MLDS

Characteristics of exploratory graphs

- They are made quickly
- A large number are made
- The goal is for personal understanding
- Axes/legends are generally cleaned up
- Color/size are primarily used for information

MLDS

Example: Air Pollution in the United States

- The U.S. Environmental Protection Agency (EPA) sets national ambient air quality standards for outdoor air pollution
  - U.S. National Ambient Air Quality Standards
- For fine particle pollution (PM2.5), the "annual mean, averaged over 3 years" cannot exceed 12 $\mu g/m^3$.
- Data on daily PM2.5 are available from the U.S. EPA web site
  - EPA Air Quality System

MLDS

Example: Air Pollution in the United States

- **Question**: Are there any counties in the U.S. that exceed that national standard for fine particle pollution?

MLDS

Data

Annual average PM2.5 averaged over the period 2008 through 2010

```
##     pm25  fips region longitude latitude
## 1  9.771 01003   east    -87.75    30.59
## 2  9.994 01027   east    -85.84    33.27
## 3 10.689 01033   east    -87.73    34.73
## 4 11.337 01049   east    -85.80    34.46
## 5 12.120 01055   east    -86.03    34.02
## 6 10.828 01069   east    -85.35    31.19
```

Do any counties exceed the standard of 12 $\mu g/m^3$?

MLDS

Simple Summaries of Data

One dimension

- Five-number summary
- Boxplots
- Histograms
- Density plot
- Barplot

MLDS

Five Number Summary

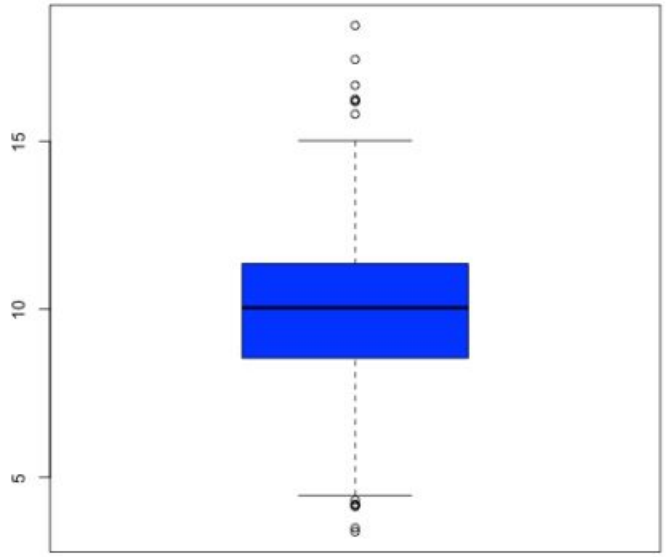
```
summary(pollution$pm25)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.38    8.55   10.00    9.84   11.40   18.40
```

MLDS

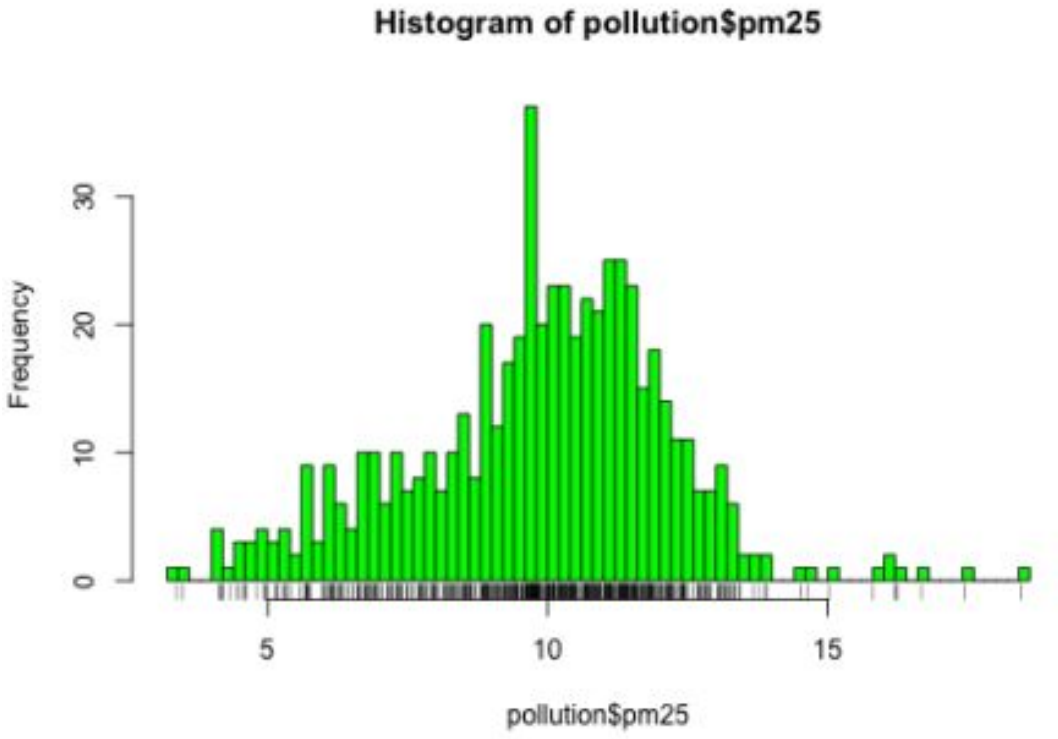
Boxplot

```
boxplot(pollution$pm25, col = "blue")
```



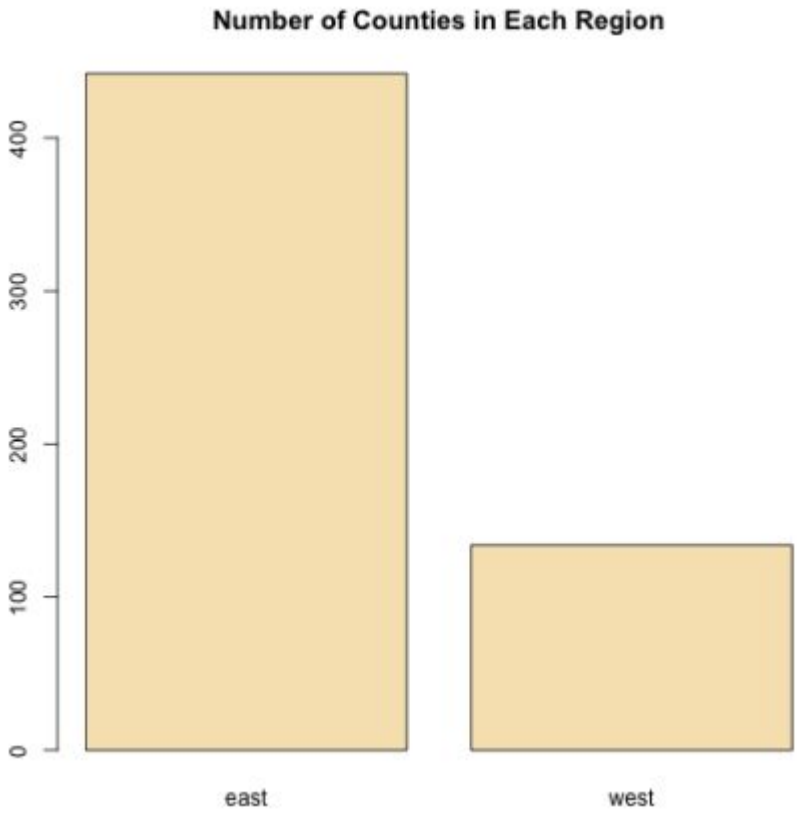
MLDS

Histogram



MLDS

Barplot



MLDS

Simple Summaries of Data

Two dimensions

- Multiple/overlayed 1-D plots (Lattice/ggplot2)
- Scatterplots
- Smooth scatterplots

MLDS

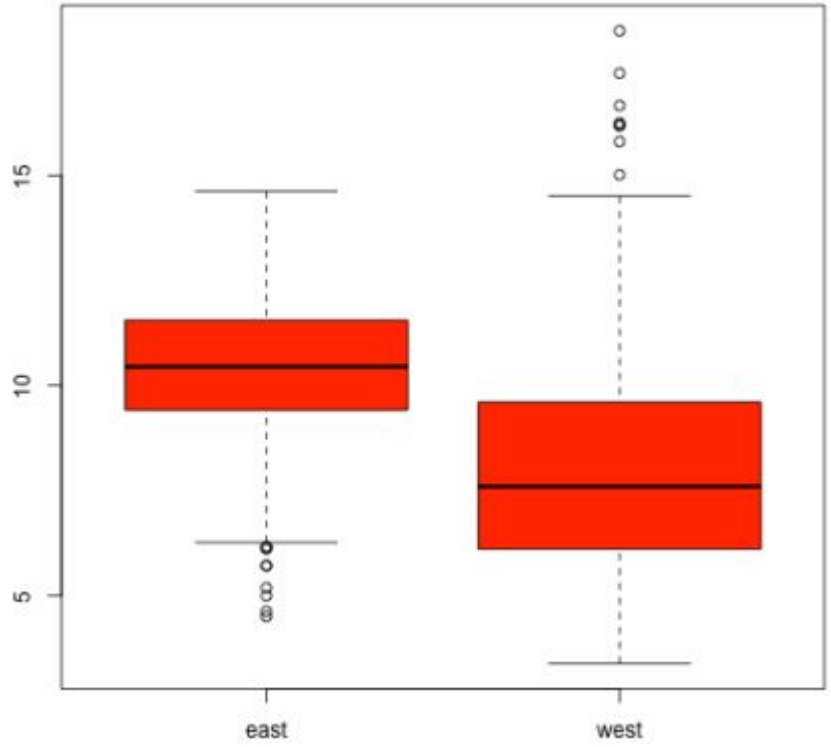
Simple Summaries of Data

> 2 dimensions

- Overlayed/multiple 2-D plots; coplots
- Use color, size, shape to add dimensions
- Spinning plots
- Actual 3-D plots (not that useful)

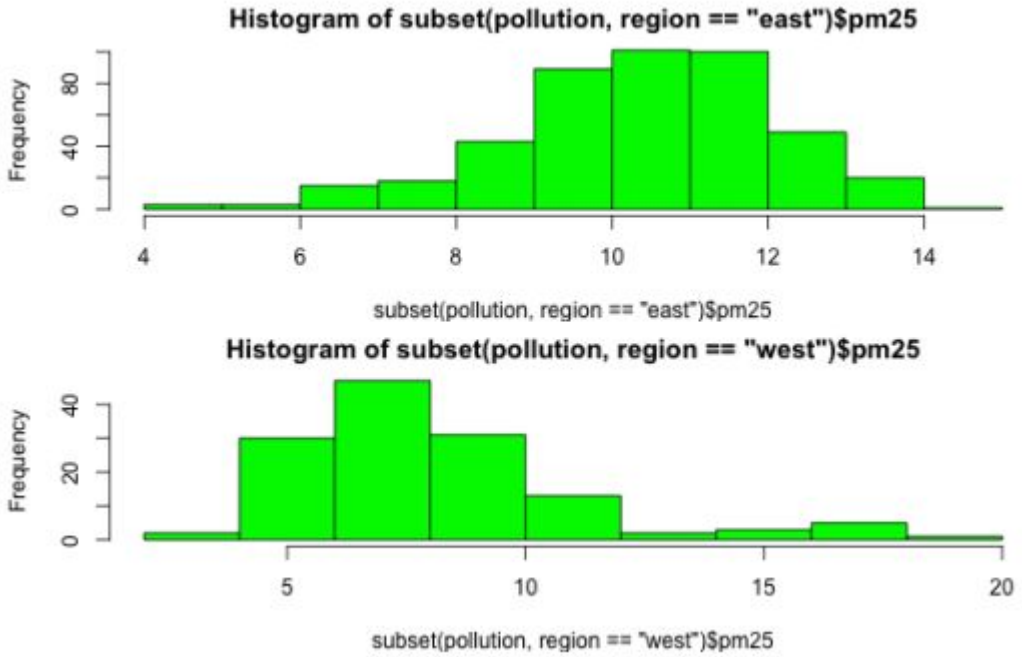
MLDS

Multiple Boxplots



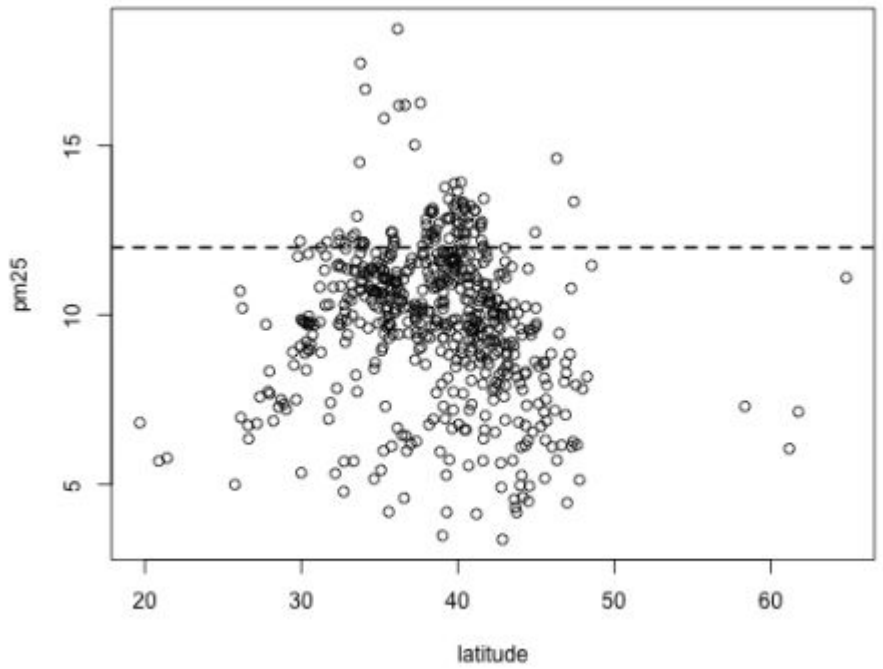
MLDS

Multiple Histograms



MLDS

Scatterplot



MLDS

Summary

- Exploratory plots are "quick and dirty"
- Let you summarize the data (usually graphically) and highlight any broad features
- Explore basic questions and hypotheses (and perhaps rule them out)
- Suggest modeling strategies for the "next step"

MLDS

Thanks!

Machine Learning for Data Science Interest Group
Advanced Informatics School
Universiti Teknologi Malaysia

@utmmlds
ais.utm.my/mlds

April 2017