Introduction to Big Data

Data Exploration and Discovery

YORK
UNIVERSITÉ
UNIVERSITY

Review Previous Lesson

Review Concepts from Day 2

The following major topics were introduced last class.

- Business Problem Definition
- Analytics Problem Definition
- Influence Diagramming
- Stakeholder & Analytics Teams
- Data Characteristics
  - Data Structure
  - Data Format
  - Data Granularity
  - Data Latency
  - Data Security
- Variable Types
- SQL Concepts
- Data Visualization Basics

Lesson Review from Day 2

You should be able to answer these questions from class 2

- What are the main items found in a framed business problem?
- What are the main items found in a defined analytics problem?
- What is the purpose of an influence diagram?
- What are the critical roles found in an analytics team?
- What does data granularity refer to?
- What are the different types of data latency?
- What are some examples of ordinal data?
- What are some restrictions placed on interval and ordinal data?
- What criteria could you use when selecting a visualization?
- What are the main building blocks of a SQL Select statement?

Day 3 – New Topics Introduced

The following major topics are discussed this class.

- Data exploration
- Data discovery
- Measurement scales
- Data shape
- Univariate data
- Bivariate data
- Descriptive statistics
- Correlation
- Linear relationships
- Filtering techniques to query from multiple data tables from Pandas/Python

Learning Objectives for Day 3

Day 3 - Learning Objectives

During Day 3 you will learn to:

- Describe and apply Exploratory Data Analysis
  - Concepts, Purpose and Methods
- Describe properties of measurement data
  - By measurement scale
  - By data type
  - By measurement role
  - By analytical application
- Explore and describe univariate data
  - Descriptive statistics
  - Visual methods
- Explore and describe bivariate data
  - Correlations
  - Visual methods
- Discover different types of data relationships
  - Linear correlations
  - Distributions
- Apply filtering techniques and query multiple data tables from Pandas/Python

Exploratory Data Analysis

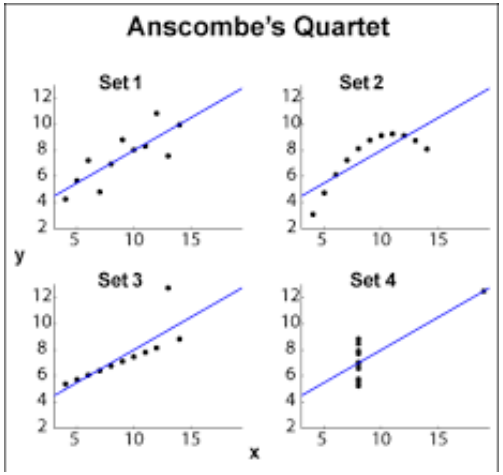
Exploratory Data Analysis Purpose

It has been shown that statistical summaries of data may not always be reliable.

Visual inspection is recommended to avoid misinformation. This example is called Anscombe's Quartet. There are 4 data sets having the same statistics but visual inspection shows 4 distinct patterns

Review the following web site

https://en.wikipedia.org/wiki/Anscombe%27s_quartet



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Four data sets with the same summary statistics but four distinctly different patterns
Visual exploration can find patterns hidden by the statistics

EDA and Visualization

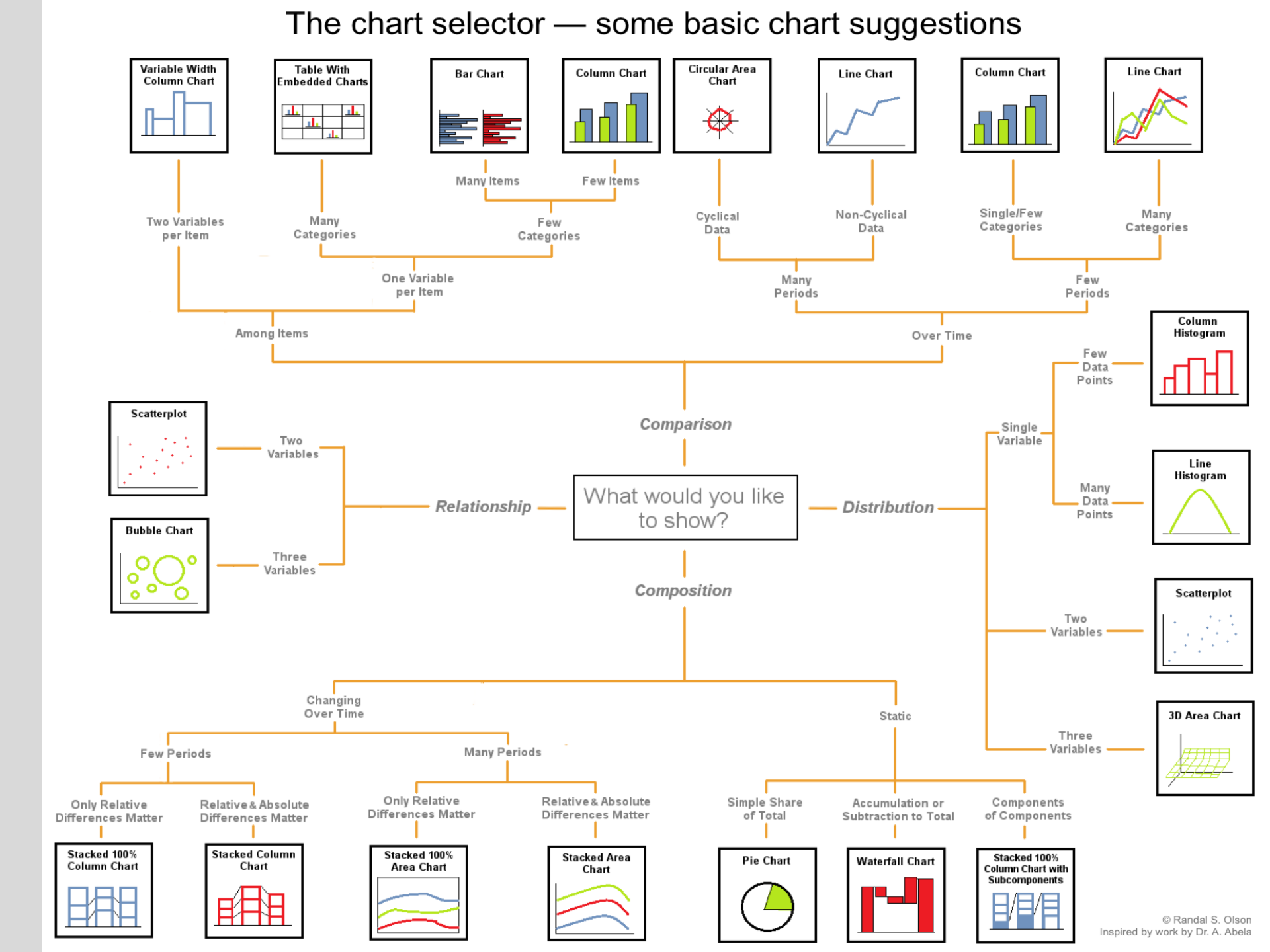
- Exploratory Data Analysis (EDA) and Visualization are very important steps in any analysis task.

- get to know your data!
  - distributions (symmetric, normal, skewed)
  - data quality problems
  - outliers
  - correlations and inter-relationships
  - subsets of interest
  - suggest functional relationships

- Sometimes EDA or viz might be the goal!

school of continuing studies | YORK UNIVERSITÉ UNIVERSITY

The chart selector — some basic chart suggestions



© Randal S. Olson
Inspired by work by Dr. A. Abela

Exploratory Data Analysis (EDA)

- Goal: get a general sense of the data
  - means, medians, quantiles, histograms, boxplots
  - You should always look at every variable - you will learn something!
- data-driven (model-free)
- Think interactive and visual
  - Humans are the best pattern recognizers
  - You can use more than 2 dimensions!
    - x,y,z, space, color, time….

```
> table(porn,spam$spam)

porn     no  yes
  no   1459  685
  yes     2   25
```

- Especially useful in early stages of data mining
  - detect outliers (e.g. assess data quality)
  - test assumptions (e.g. normal distributions or skewed?)
  - identify useful raw data & transforms (e.g. log(x))

- Bottom line: it is always well worth looking at your data!

school of continuing studies | YORK UNIVERSITÉ UNIVERSITY

Exploratory Data Analysis Methods

Measuring Central Tendency

Mean, median, and mode

Mean, median, and mode are different measures of center in a numerical data set. They each try to summarize a dataset with a single number to represent a "typical" data point from the dataset.

Mean: The "average" number; found by adding all data points and dividing by the number of data points.

Median: The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

Mode: The most frequent number—that is, the number that occurs the highest number of times.

Exploratory Data Analysis Methods

Measuring Spread

- Variance
- Standard Deviation
- Inter Quartile Range

If the data distribution is symmetrical all three of these statistics have the same value and any of these statistics can be used to describe the central tendency. The mean is most commonly used for this

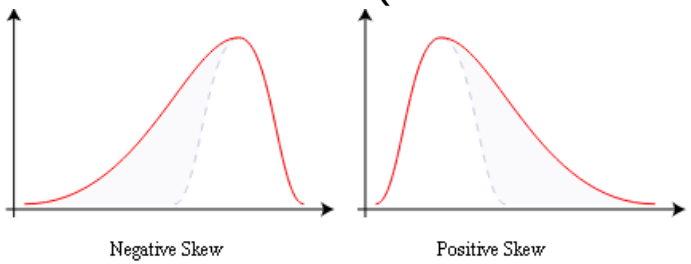
If the data distribution is skewed or unbalanced to the left or to the right, then the median is used to describe the central tendency.

Read the following article on-line for further definitions of descriptive statistics. Following the links within the web site to see definitions for the key terms.

https://en.wikipedia.org/wiki/Descriptive_statistics

Summary Statistics

- *not* visual
- sample statistics of data X
  - mean: $\mu = \sum_i X_i / n$
  - mode: most common value in X
  - median: **X**=sort(X), median = $\mathbf{X}_{n/2}$ (half below, half above)
  - quartiles of sorted **X**: Q1 value = $\mathbf{X}_{0.25n}$, Q3 value = $\mathbf{X}_{0.75\,n}$
    - interquartile range: value(Q3) - value(Q1)
    - range: max(X) - min(X) = $\mathbf{X}_n - \mathbf{X}_1$
  - variance: $\sigma^2 = \sum_i (X_i - \mu)^2 / n$
  - skewness: $\sum_i (X_i - \mu)^3 \;/\; [\,(\sum_i (X_i - \mu)^2)^{3/2}\,]$
    - zero if symmetric; right-skewed more common (what kind of data is right skewed?)

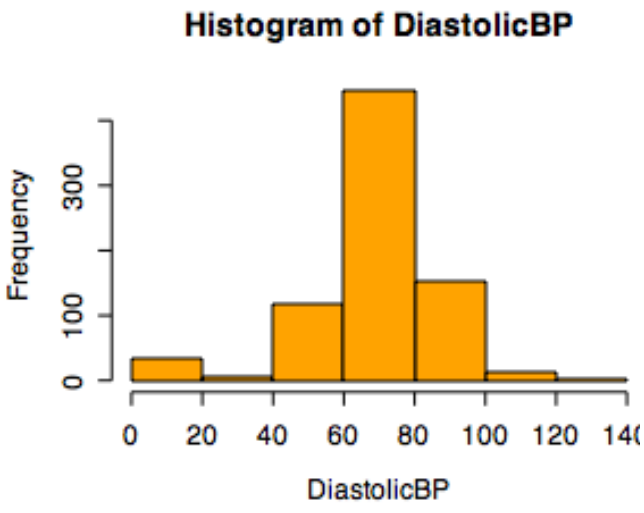


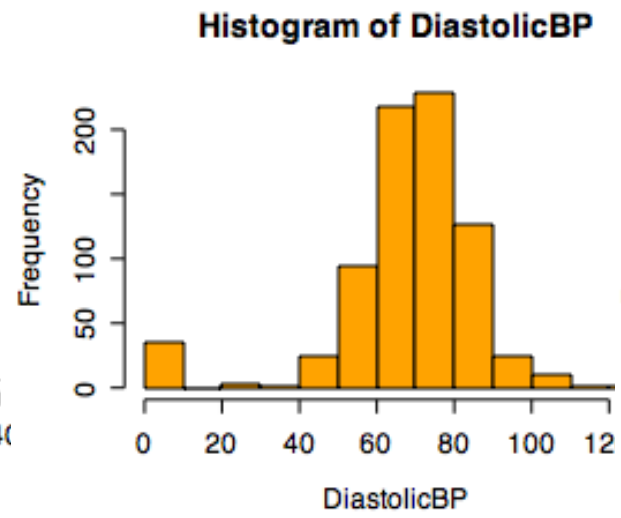
  - number of distinct values for a variable (see unique() in R)
  - Don't need to report all of thses: Bottom line…do these numbers make sense???

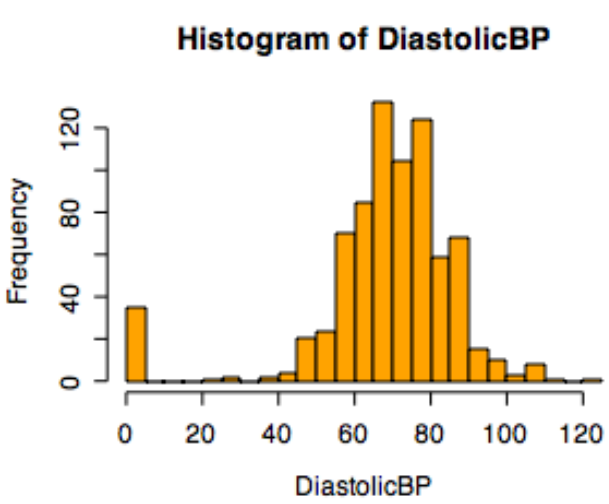
school of continuing studies | YORK UNIVERSITÉ UNIVERSITY

Single Variable Visualization

- Histogram:
  - Shows center, variability, skewness, modality,
  - outliers, or strange patterns.
  - Bin width and position matter
  - Beware of real zeros







Issues with Histograms

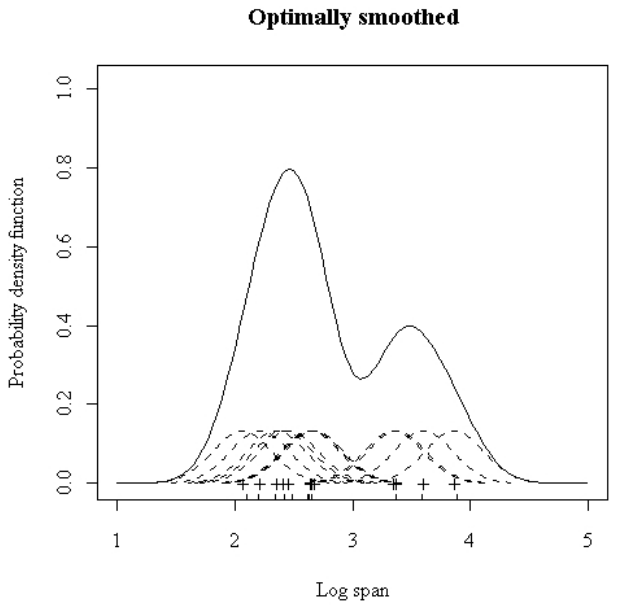
- For small data sets, histograms can be misleading.
  - Small changes in the data, bins, or anchor can deceive
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution.
- Histograms effectively only work with 1 variable at a time
  - But ‘small multiples’ can be effective

Smoothed Histograms - Density Estimates

- Kernel estimates smooth out the contribution of each datapoint over a local neighborhood of that point.

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

h is the kernel width

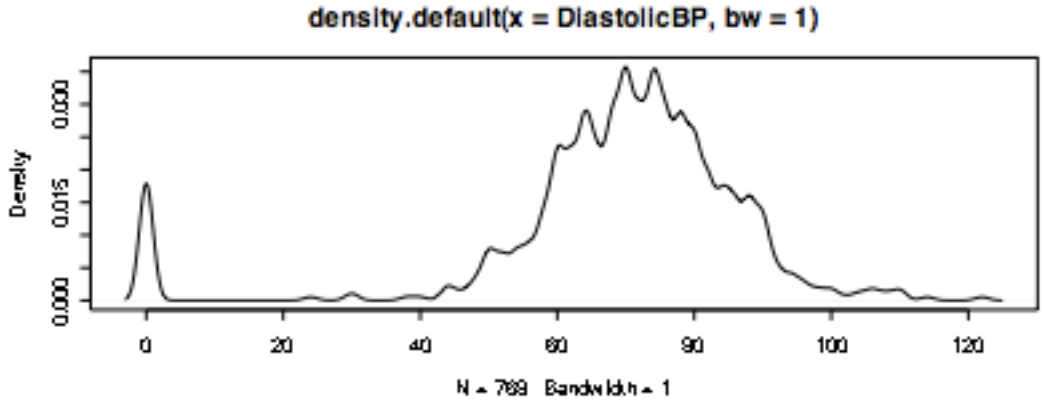


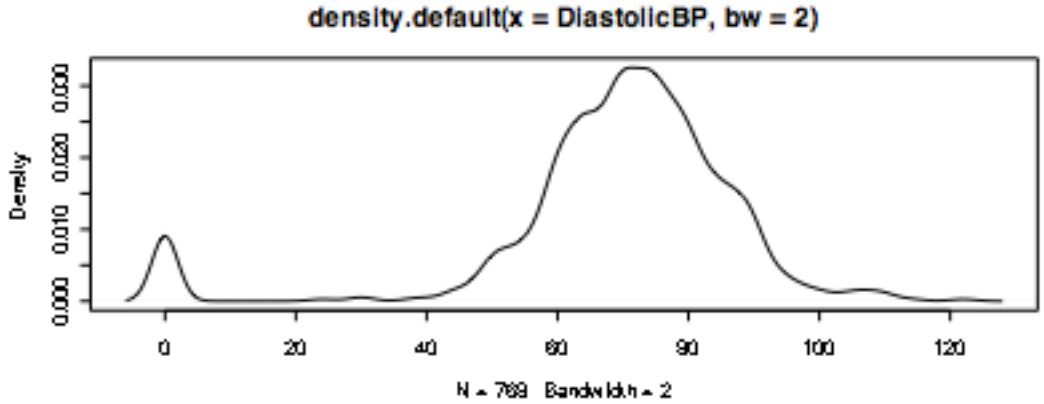
- Gaussian kernel is common:

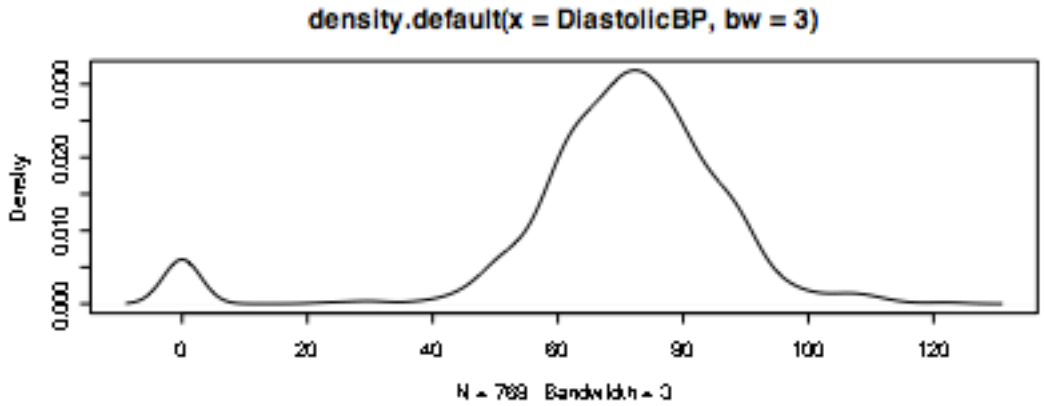
$$Ce^{-\frac{1}{2}\left(\frac{x-x(i)}{h}\right)^2}$$

Bandwidth choice is an art

Usually want to try several

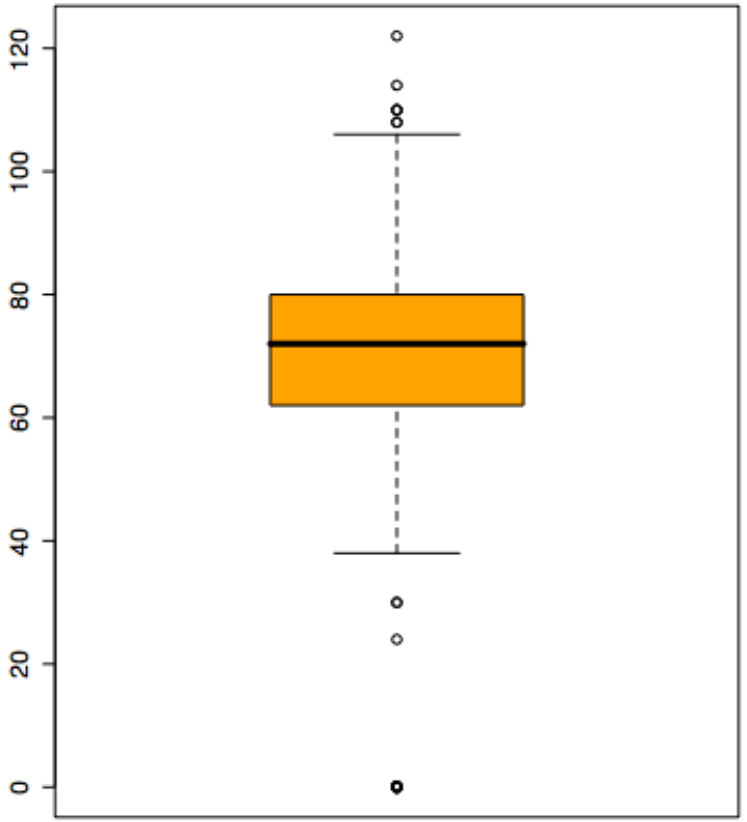






Boxplots

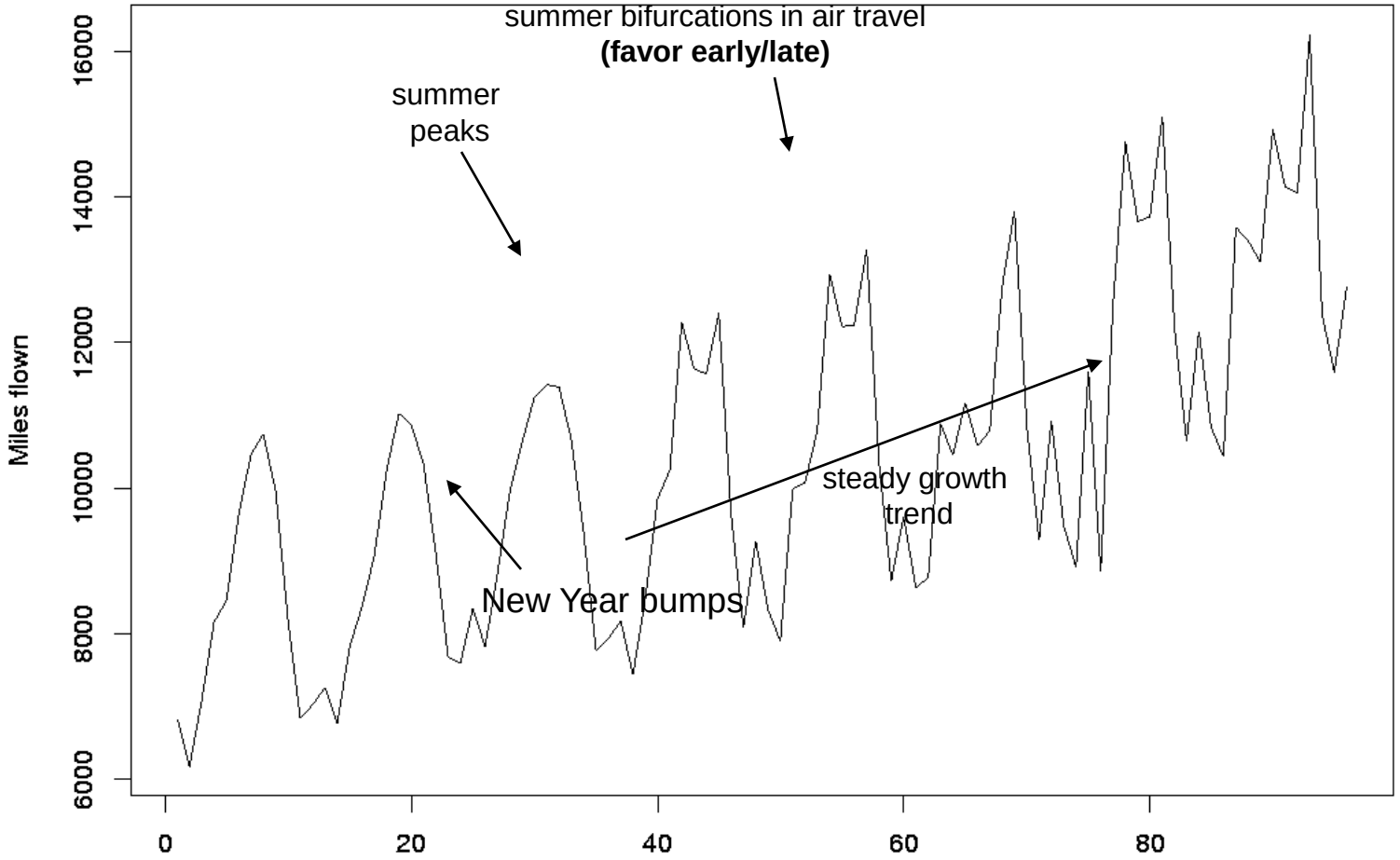
- Shows a lot of information about a variable in one plot
  - Median
  - IQR
  - Outliers
  - Range
  - Skewness
- Negatives
  - Overplotting
  - Hard to tell distributional shape
  - no standard implementation in software (many options for whiskers, outliers)



school of continuing studies | YORK UNIVERSITÉ UNIVERSITY

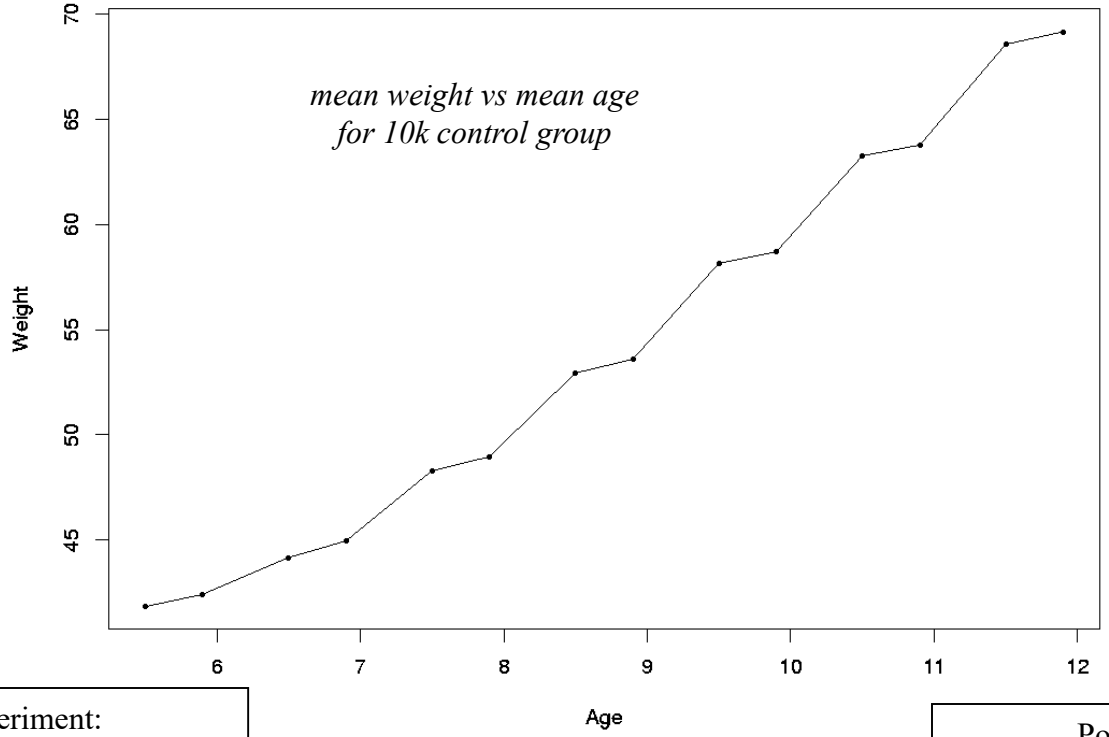
Time Series

If your data has a temporal component, be sure to exploit it



school of continuing studies | YORK UNIVERSITÉ UNIVERSITY

Time-Series Example 3



Scotland experiment:
"↑ milk in kid diet → better health" ?

20,000 kids:
5k raw, 5k pasteurize,
10k control (no supplement)

Would expect smooth weight growth plot.

Visually reveals
<u>unexpected pattern</u> (*steps*),
not apparent from raw data table.

<u>Possible explanations</u>:

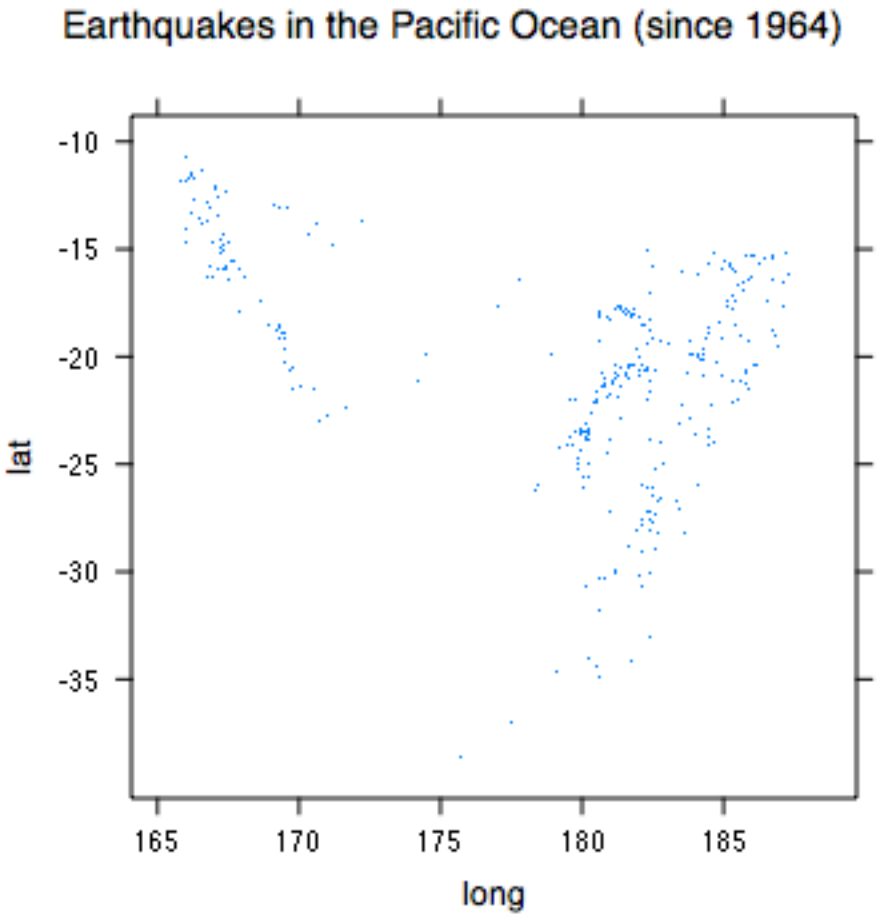
Grow less early in year than later?

No steps in height plots; so why height ↑ uniformly, weight ↑ spurts?

Kids weighed in clothes: summer garb lighter than winter?

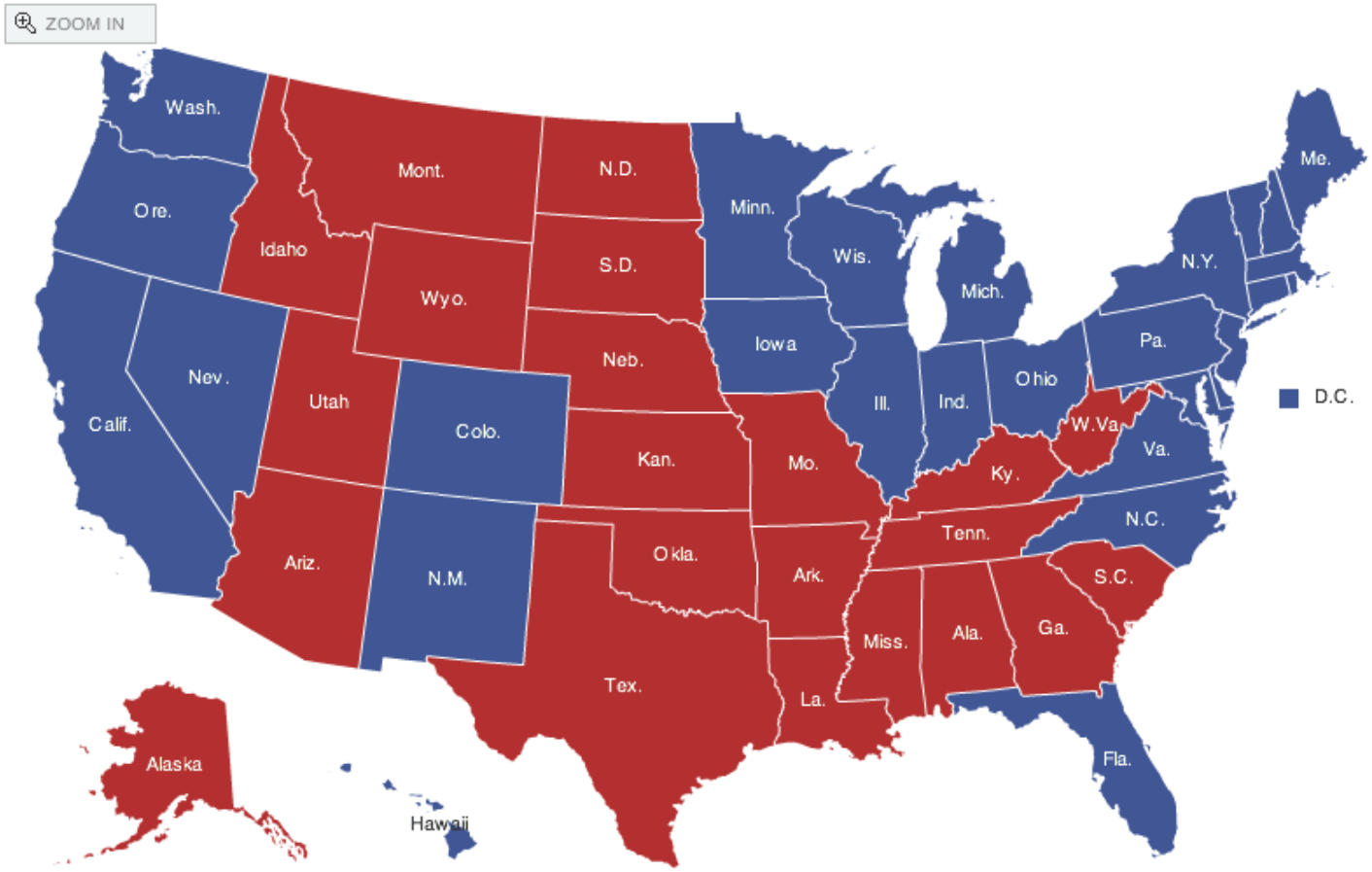
Spatial Data

- If your data has a geographic component, be sure to exploit it
- Data from cities/states/zip cods – easy to get lat/long
- Can plot as scatterplot



school of continuing studies | YORK UNIVERSITÉ UNIVERSITY

Spatial data: choropleth Maps

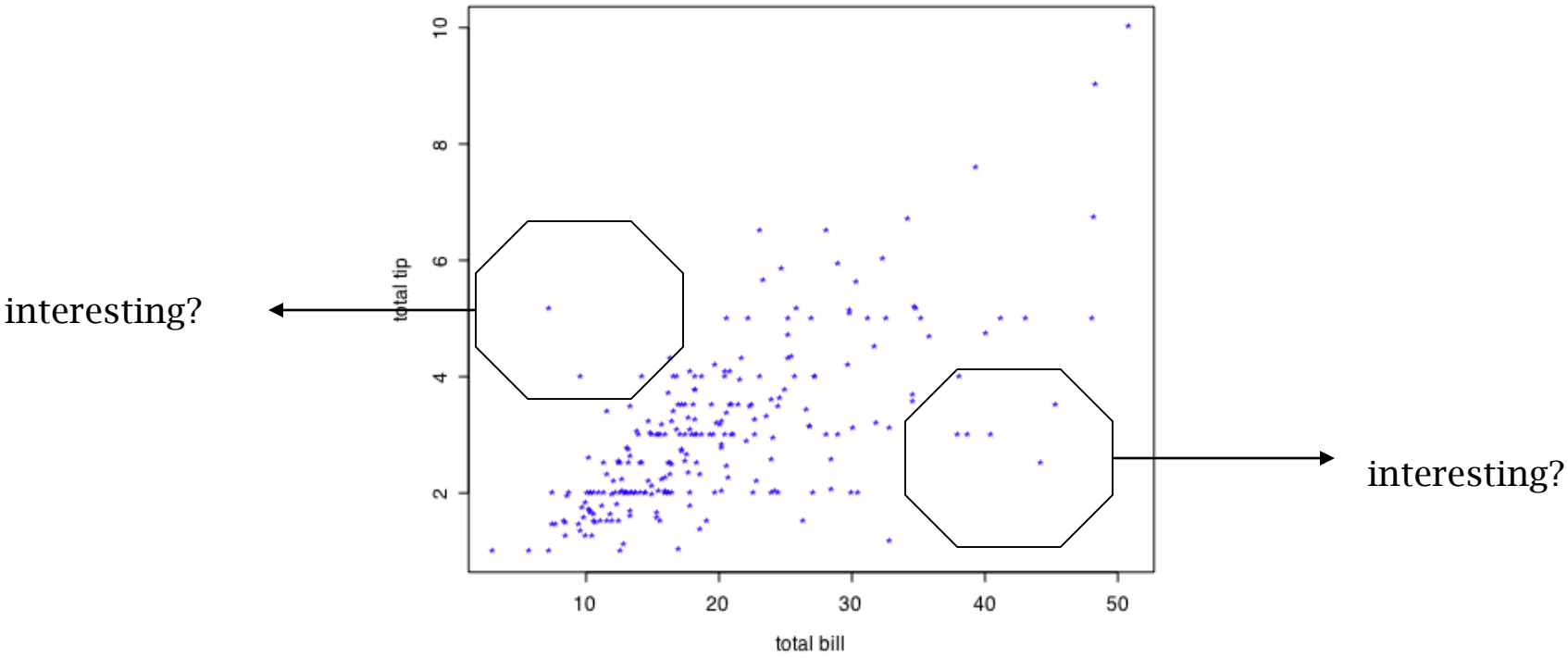


- Maps using color shadings to represent numerical values are called choropleth maps
- http://elections.nytimes.com/2008/results/president/map.html

school of continuing studies | YORK UNIVERSITÉ UNIVERSITY

Two Continuous Variables

- For two numeric variables, the scatterplot is the obvious choice

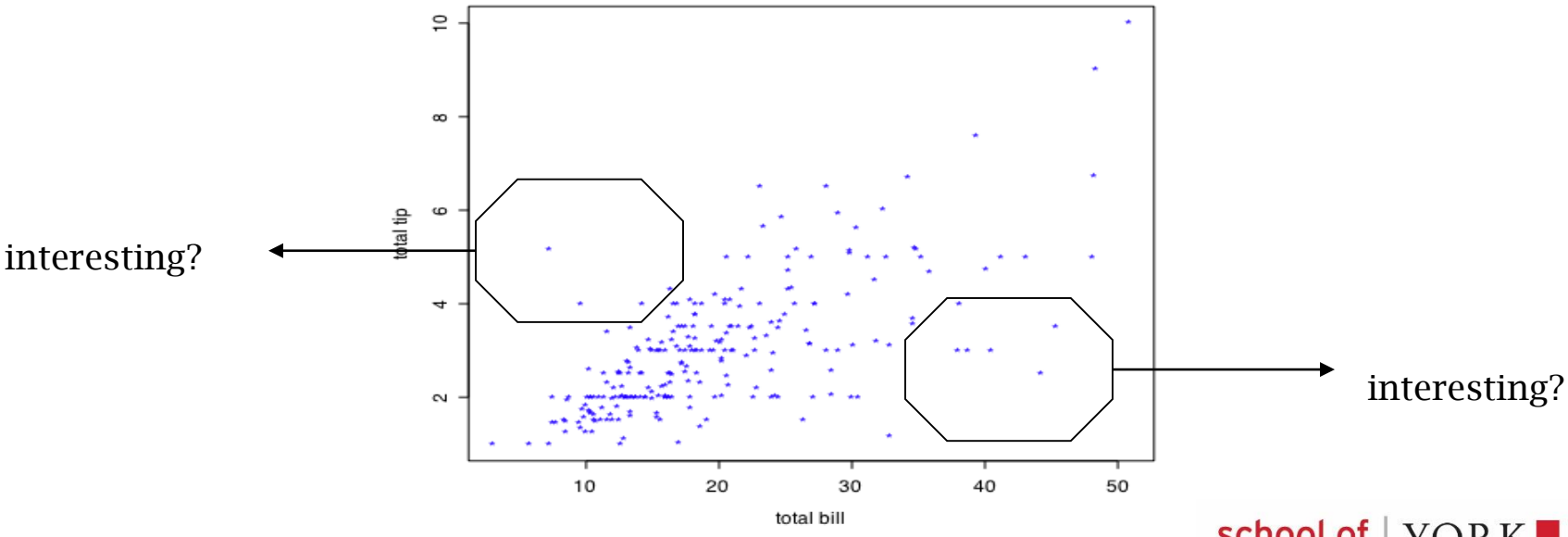


school of continuing studies | YORK UNIVERSITÉ UNIVERSITY

2D Scatterplots

- standard tool to display relation between 2 variables
  - e.g. y-axis = response, x-axis = suspected indicator

- useful to answer:
  - x,y related?
    - linear
    - quadratic
    - other
  - variance(y) depend on x?
  - outliers present?



Exploratory Data Analysis Methods

Univariate non-graphical methods

The goal of this method to answer 2 basic questions about the values of a single numerical variable. The data set is called a sample data distribution

1. How similar are the values to each other?
2. How different are the values from each other?

The first question is answered with an understanding of the "central tendency" of the data. How the values cluster around a common value is key to understanding the "similarity" property in the data.

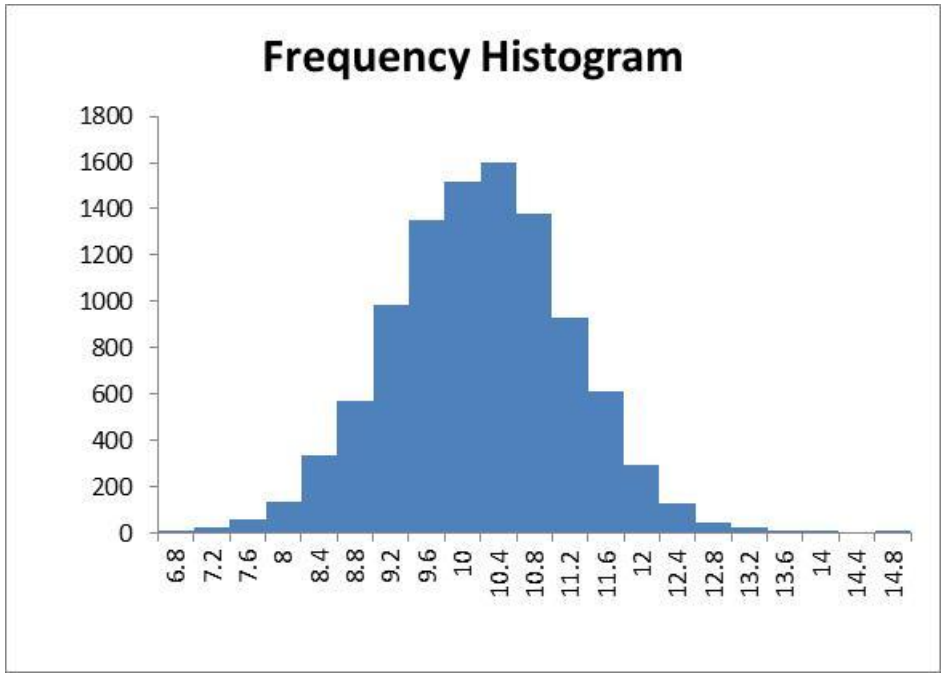
The second question is answered with an understanding of the "spread" or "dispersion of the data. How much the values differ the "central tendency" value provides an understanding of the "difference" property in the data.

Common descriptive values calculated include

- Center (central tendency)
- Spread
- Modality (number of peaks in the data)
- Shape
- Outliers.

Exploratory Data Analysis Examples

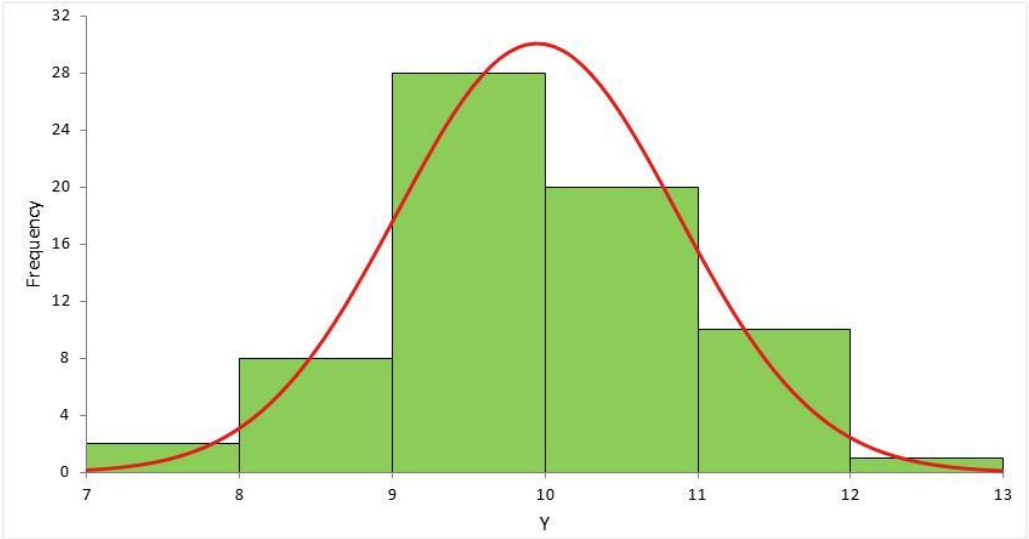
Example 1 – Symmetrical Sample Distribution



Mean	**9.9901**
Number of Trials	10000
Standard error	0.0099
Minimum	6.4387
Maximum	14.623
Median	10.006
Range	8.1842
Standard Deviation	**0.9903**
Variance	0.9808
Skewness	-0.02
Kurtosis	3.05

Exploratory Data Analysis Examples

Example 2 – Symmetrical Distribution



N	69

	Mean	Mean SE	SD	Variance	Skewness	Kurtosis
Y	9.948499167	0.110071116	0.914319358	0.835979889	0.0	0.12

	Minimum	1st quartile	Median	3rd quartile	Maximum	IQR
Y	7.49256138	9.385248010	9.814749751	10.519968900	12.06988867	1.134720890

	Mode
Y	-

Measurement Data

Properties of Measurement Data

Overview

Measurements are data elements that quantify a property of interest to us.

The four types of measurement scales (and related variables) were introduced last Day.

The following link provides a tutorial on measurement scales. Watch the video and read the content on the website that describes the four measurement scales.

http://stattrek.com/statistics/measurement-scales.aspx?Tutorial=AP

Measurement Data Scales

Ratio Measurement Scales

The zero point on the ratio scale is meaningful and not arbitrary. It means the "absence" of something.

Values plotted on a ratio scale are continuous and numerical. The implication of this scale is that 2 properties can be determined.

Proportion and Difference Example:

- comparing a=5.0 and b=10.0
- conclusion 1 – b is twice as large as a
- conclusion 2– b is 5.0 units greater than a

Measurement Data Scales

Interval Measurement Scales

The zero point on the interval scale is not meaningful and is arbitrary. It does not mean the "absence" of something. An example is a temperature scale. Zero on the Fahrenheit or Celsius scale does NOT mean the absence of energy.

Values plotted on an interval scale are continuous and numerical. The implication of this scale is that only 1 property can be determined.

Difference example:

- comparing a=5.0 and b=10.0
- conclusion 1 – b is twice as large as a
- it is incorrect to conclude anything about proportion on an interval scale

Measurement Data Scales

Ordinal Measurement Scales

The ordinal scale is used to rank items in terms of some measured criteria. An example is the five star hotel rating scheme. The only conclusion that is valid relates to the order. Conclusions about difference or proportion are incorrect.

Values plotted on an ordinal scale are rank values, scores or counts. The implication of this scale is that only 1 property can be determined.

Rank example:

- comparing a=5.0 and b=10.0
- conclusion 1 – b is better than a
- it is incorrect to conclude anything about proportion or difference on an ordinal scale

Measurement Data Scales

Nominal Measurement Scales

The nominal scale is used to count and aggregate items in defined categories.

Values plotted on an nominal scale are whole to part measurements. The only conclusion that is valid states how many items are in a defined category

Nominal example:

- comparing a=5.0 and b=10.0
- conclusion 1 – b has 10 units and a has 5 units
- it is also correct to make conclusion about differences and proportions

Univariate Data

Univariate Data Concepts

Numerical Data

- Variables that measure the properties of things
- Depending on the measurement technique, the can be based on ratio, interval, ordinal or nominal scales.

Categorical Data

- Variables that describe things
- Categorical data provides the grouping criteria used to create a nominal measurement.
- Also called dimensions

Univariate Data Concepts

Types of Data Exploration Analysis

Numerical Data

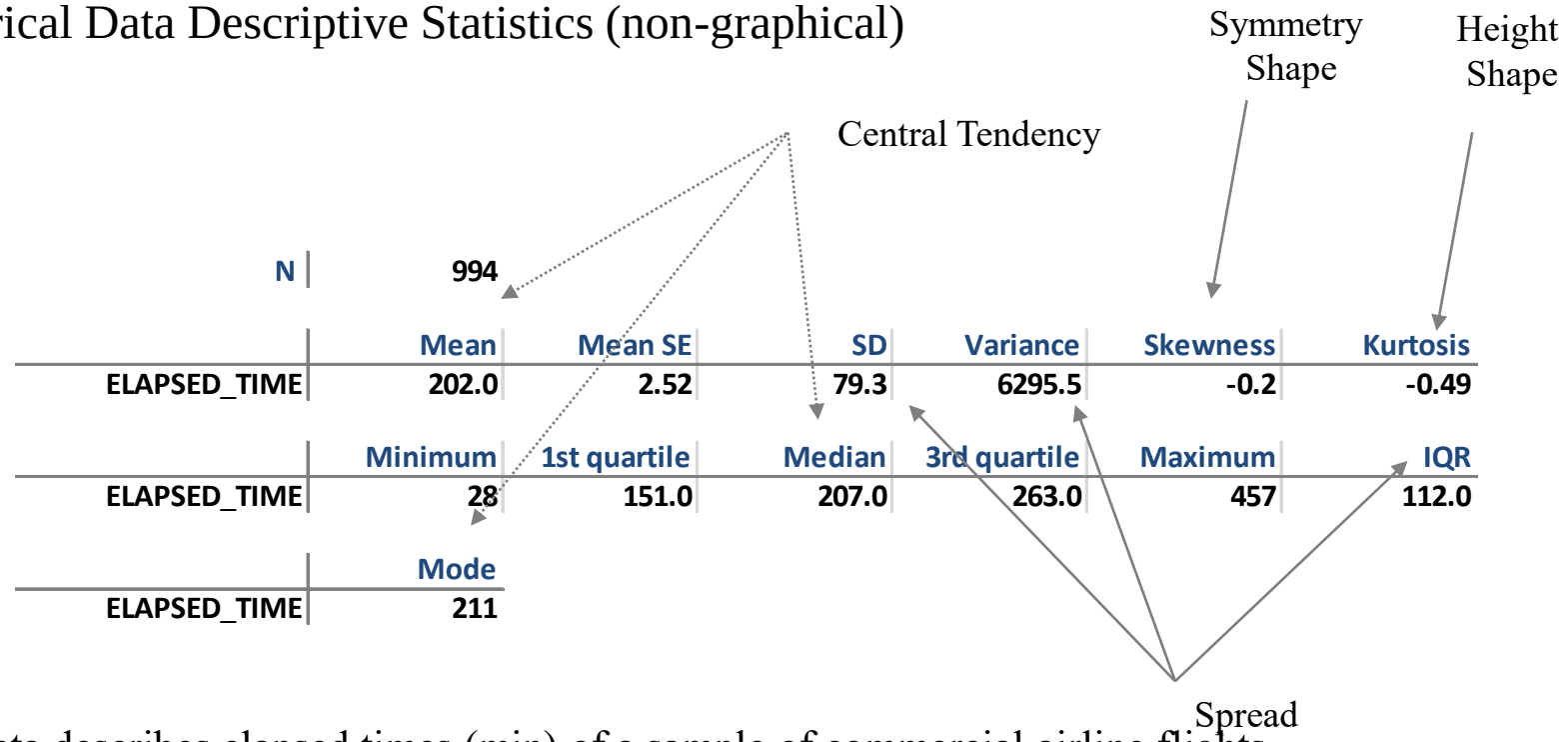
- non-graphical
- graphical

Categorical Data

- non-graphical counts
- graphical

Numerical Data – Non Graphical Technique

Numerical Data Descriptive Statistics (non-graphical)



The data describes elapsed times (min) of a sample of commercial airline flights

Read the following article on-line for further definitions of descriptive statistics. Following the links within the web site to see definitions for the key terms.

https://en.wikipedia.org/wiki/Descriptive_statistics

Numerical Data – Non Graphical Methods

Numerical Data Descriptive Statistics Example:

The data describes departure delay times (min) of a sample of commercial airline flights.

https://www.kaggle.com/usdot/flight-delays

N	999

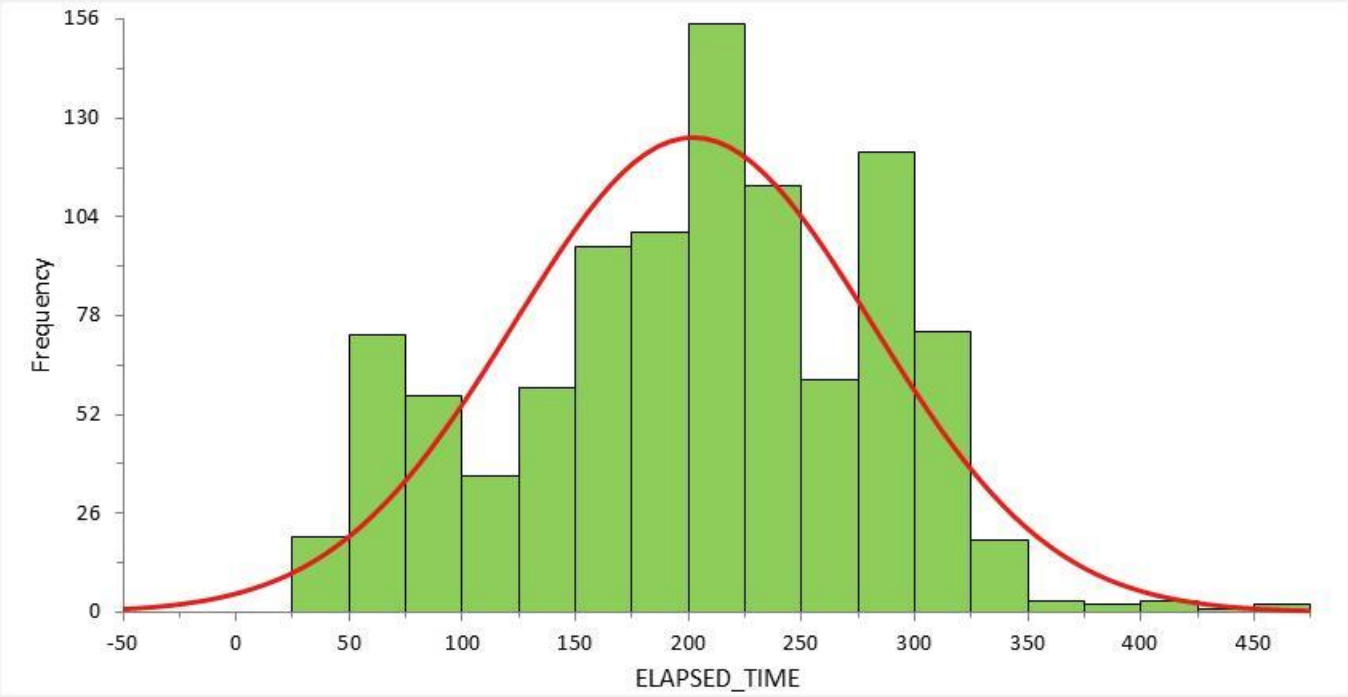
	Mean	Mean SE	SD	Variance	Skewness	Kurtosis
DEPARTURE_DELAY	66.9	3.30	104.4	10900.5	2.9	14.86

	Minimum	1st quartile	Median	3rd quartile	Maximum	IQR
DEPARTURE_DELAY	-16	3.0	19.0	106.0	1058	103.0

	Mode
DEPARTURE_DELAY	-4

Numerical Data – Graphical Technique

Frequency Histogram with an Overlay of a Symmetrical Distribution



The data describes elapsed times (min) of a sample of commercial airline flights.

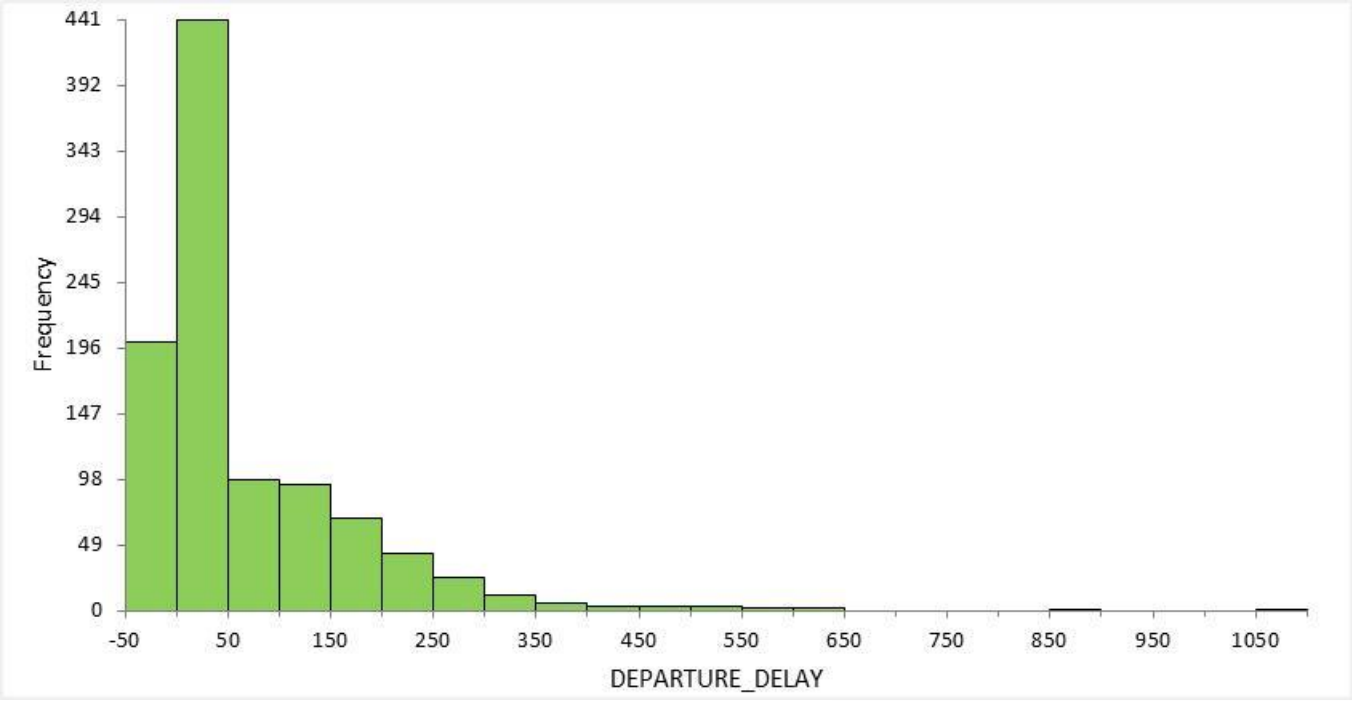
https://www.kaggle.com/usdot/flight-delays

Numerical Data – Graphical Technique

Continuous Data – Frequency Histogram Example

The data describes departure delay times (min) of a sample of commercial airline flights.

https://www.kaggle.com/usdot/flight-delays



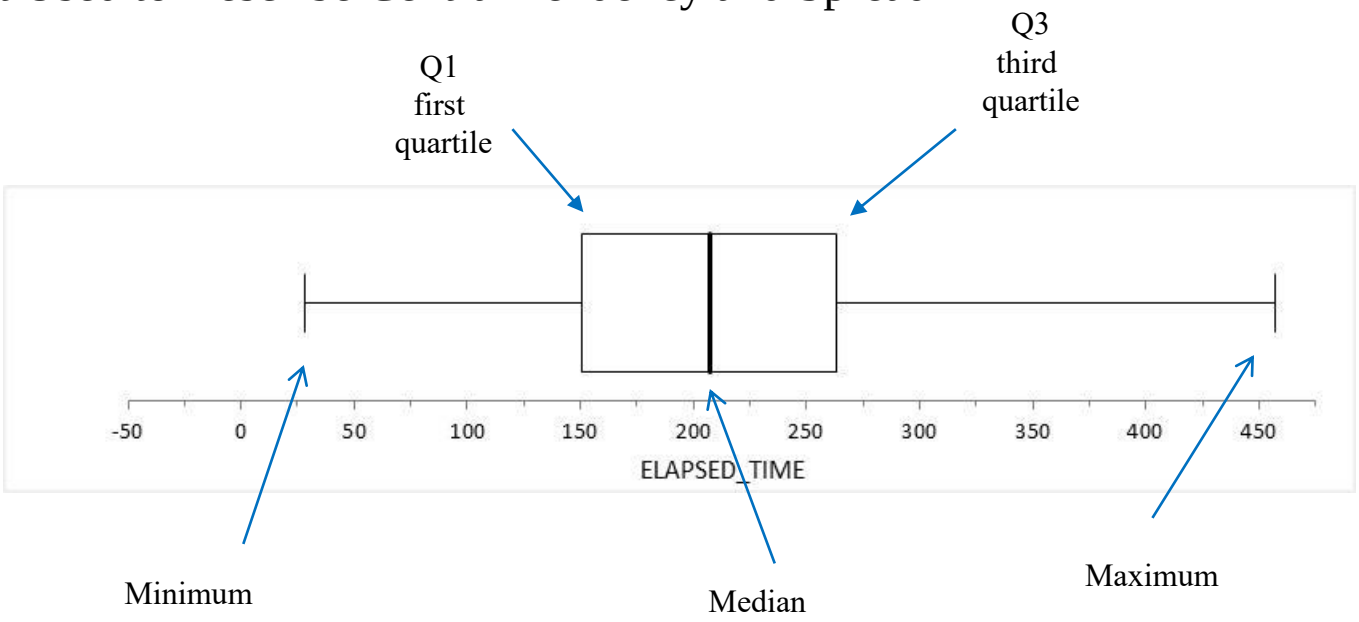
Numerical Data – Box Plot Technique

Box Plots also called Box-Whisker Plots are described in the following video. Please watch the video to gain an introduction to the technique.

https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/v/box-and-whisker-plot-exercise-example

Numerical Data – Box Plot Example

Box Plot Used to Describe Central Tendency and Spread



The data describes elapsed times (min) of a sample of commercial airline flights.

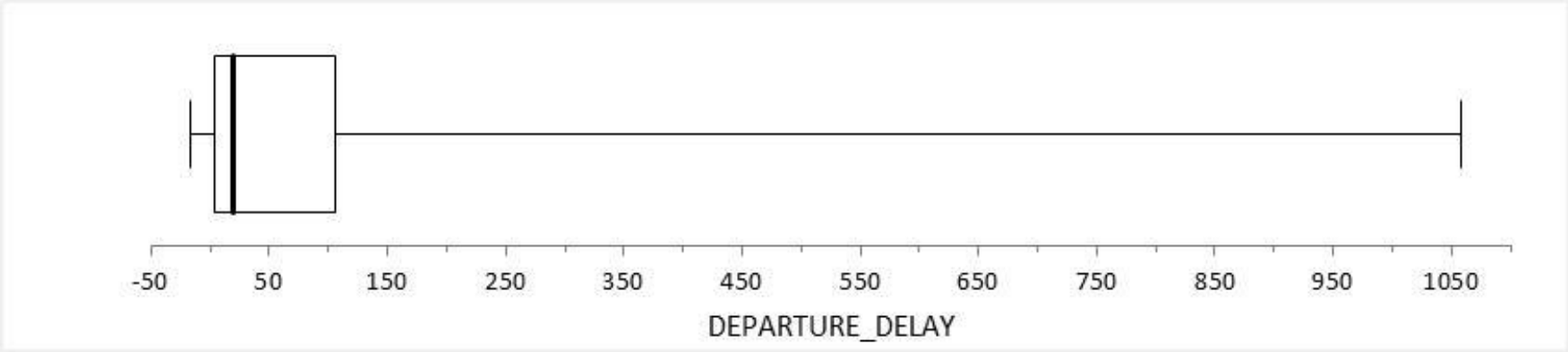
https://www.kaggle.com/usdot/flight-delays

Numerical Data – Box Plot Example

Continuous Data – Box Plot Example Used to Visualize Central Tendency & Spread

The data describes departure delay times (min) of a sample of commercial airline flights.

https://www.kaggle.com/usdot/flight-delays



Numerical Data – Probability Distributions

Distribution functions are created from univariate analysis defined by shape, central tendency and spread.

Please watch the video in this link to gain an introduction to how distribution functions are used to model probabilities.

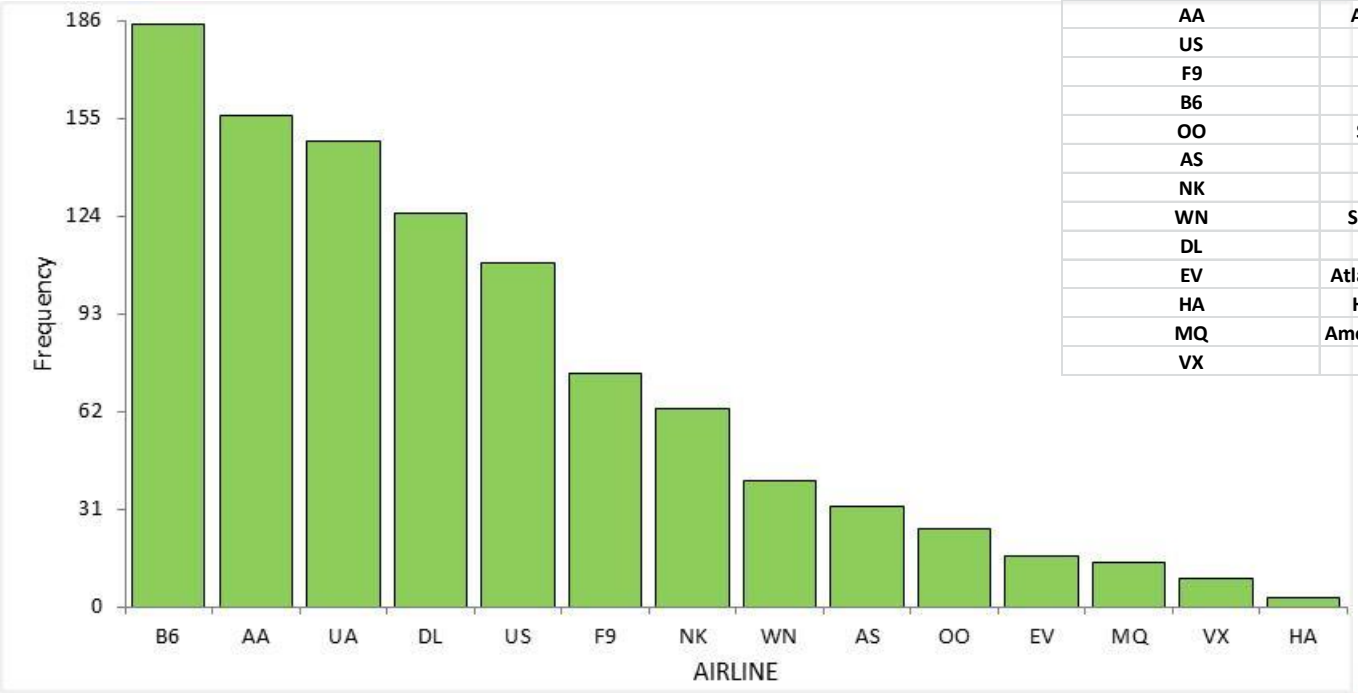
https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/density-curve/v/density-curves

Categorical Data Exploration

Categorical Data – Absolute Frequency Example

The data describes the frequency of different airlines in the flight delay data set.

https://www.kaggle.com/usdot/flight-delays



IATA_CODE	AIRLINE
UA	United Air Lines Inc.
AA	American Airlines Inc.
US	US Airways Inc.
F9	Frontier Airlines Inc.
B6	JetBlue Airways
OO	Skywest Airlines Inc.
AS	Alaska Airlines Inc.
NK	Spirit Air Lines
WN	Southwest Airlines Co.
DL	Delta Air Lines Inc.
EV	Atlantic Southeast Airlines
HA	Hawaiian Airlines Inc.
MQ	American Eagle Airlines Inc.
VX	Virgin America

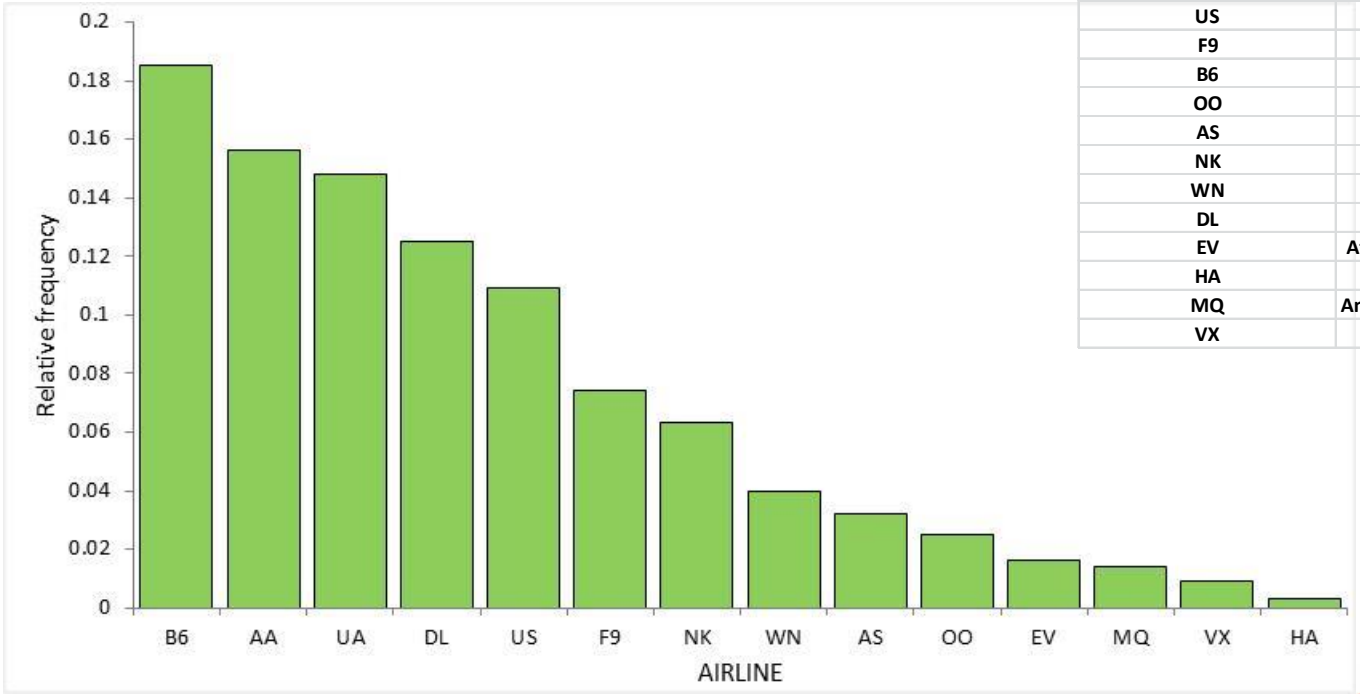
Categorical Data Exploration

Categorical Data – Relative Frequency Example

The data describes the relative frequency of different airlines in the flight delay data set.

https://www.kaggle.com/usdot/flight-delays

IATA_CODE	AIRLINE
UA	United Air Lines Inc.
AA	American Airlines Inc.
US	US Airways Inc.
F9	Frontier Airlines Inc.
B6	JetBlue Airways
OO	Skywest Airlines Inc.
AS	Alaska Airlines Inc.
NK	Spirit Air Lines
WN	Southwest Airlines Co.
DL	Delta Air Lines Inc.
EV	Atlantic Southeast Airlines
HA	Hawaiian Airlines Inc.
MQ	American Eagle Airlines Inc.
VX	Virgin America



Bivariate Data

Bivariate Data – Scatter Plots & Correlations

Scatter plots are a useful method for starting your analysis of bivariate data.

Please watch the following video for an introduction to the technique.

https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/introduction-to-scatterplots/v/constructing-scatter-plot

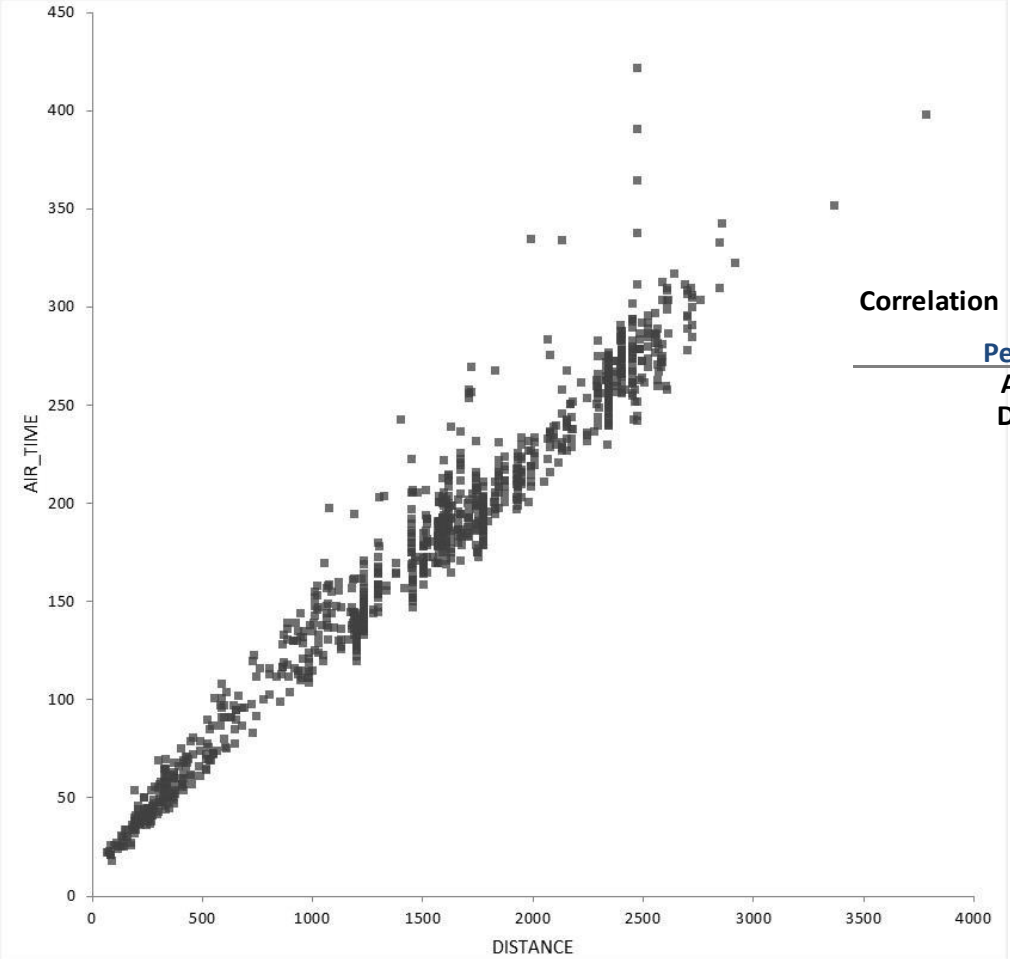
Correlation coefficients measure dependencies between two variables.

Please watch the following video for an introduction to correlations.

https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/scatterplots-and-correlation/v/correlation-coefficient-intuition-examples

Bivariate Data Scatter Plot Example

Scatter Plot with High Positive Correlation

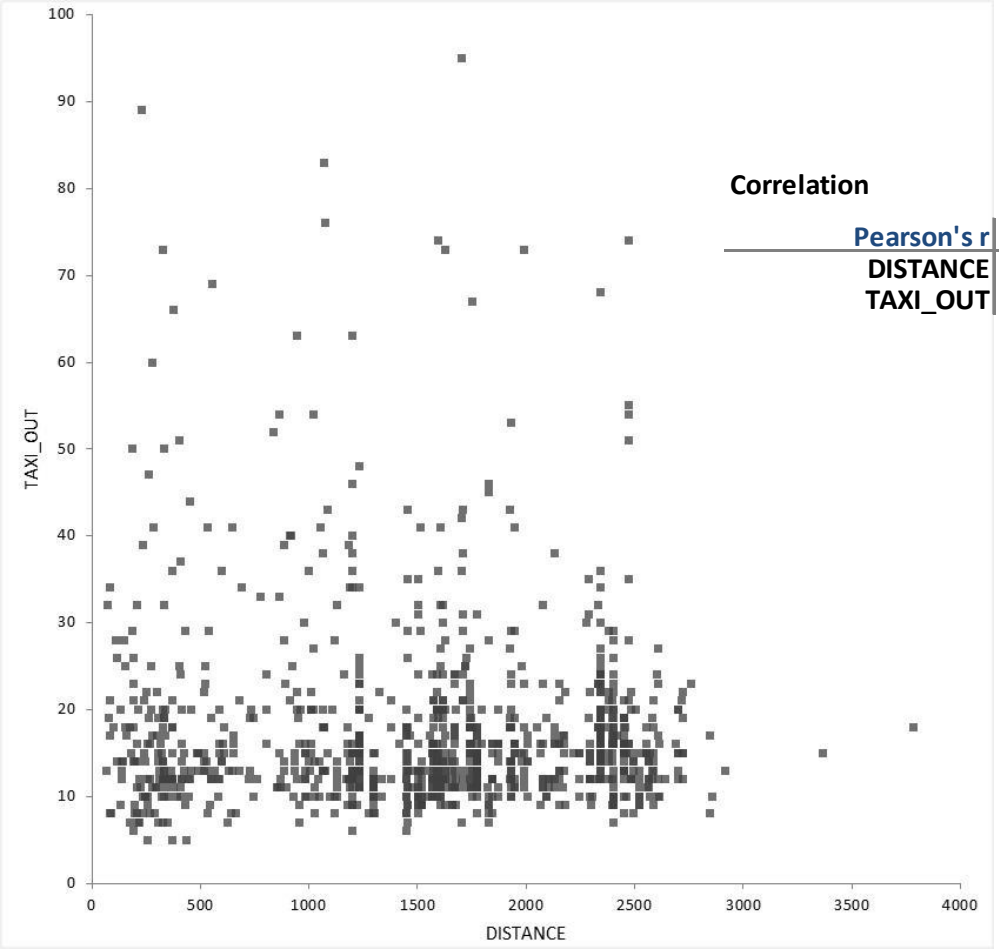


Correlation

Pearson's r	AIR_TIME	DISTANCE
AIR_TIME	-	0.979
DISTANCE	0.979	-

Bivariate Data Scatter Plot Example

Scatter Plot with Low Correlation



Correlation

Pearson's r	DISTANCE	TAXI_OUT
DISTANCE	-	-0.046
TAXI_OUT	-0.046	-

Lesson Review

Consider the following questions that you should be able to answer by completing Day 3.

- What is the scope of data exploration?
- What properties of data are discovered using data exploration?
- What conclusions can be made about two different measurements from a ratio scale?
- What is an example of an invalid conclusion about two different measurements from an interval scale?
- What descriptive statistics are used to describe the shape of data?
- What property of univariate data describes similarity?
- What property of bivariate data is used to describe dependencies?
- What properties of univariate data is expressed in a box plot visual?

Lesson Summary

Day 3 Lesson Summary

- During Day 3 you learned to:
- Identify, interpret and correctly apply categories of measurement data
  - Ordinal
  - Nominal
  - Interval
  - Ratio
- Interpret descriptive statistics to quantitatively describe the shape of a variable
  - Central tendency measures
  - Dispersion measures
  - Skewness and Kurtosis “height” measures
- Understand and describe bivariate data set according to:
  - Scatter plots.
  - Linear relationships
  - Correlations from a scatter plot
  - Correlation coefficient
- Understand basic Data Exploratory methods to develop Probability Distributions
  - Probability Density
  - Data Relationships
  - Data Frequency Distributions
- Apply filtering techniques and query from multiple data tables from Pandas/Python