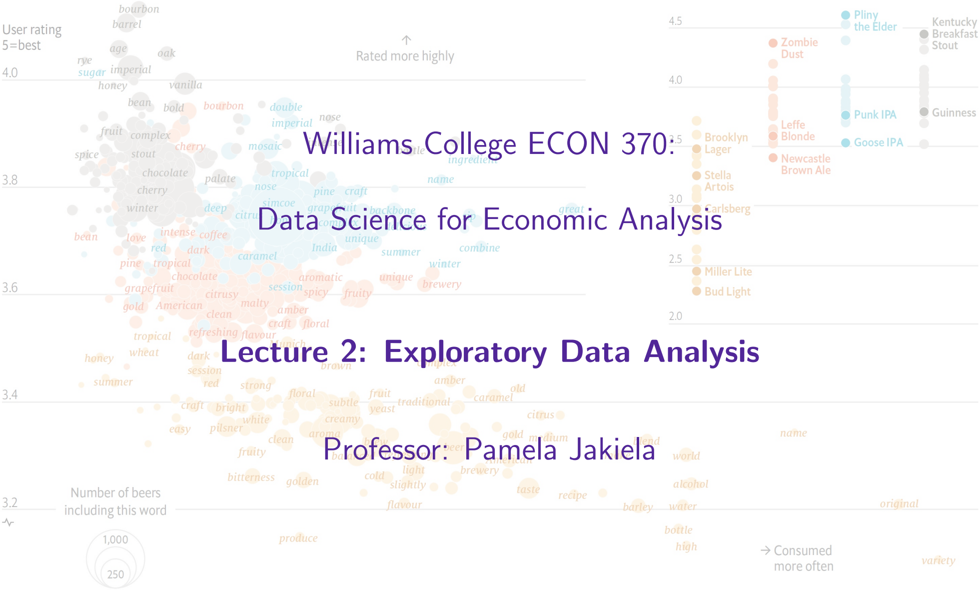


Williams College ECON 370:

Data Science for Economic Analysis

Lecture 2: Exploratory Data Analysis

Professor: Pamela Jakiela

Outline

- The distribution of one variable: summary statistics, histograms, kernel densities
- The relationship between two variables: scatter plots, local polynomial regression

The Most Important Step in Data Analysis

The last step in the data preparation pipeline and the first step in analysis is looking at the data

- Tabulating the values, looking at summary statistics, visualizing the distribution

Exploraty data analysis serves two purposes:

- Detecting errors, problems, outliers, etc.
- Looking for patterns, regularities, relationships in the data

"Measure twice, cut once" – but for data

- There is almost nothing worse than finding a bug in your cleaning/preparation code after you've analyzed the data, written up your results, presented your findings, published, etc.

What's Wrong With This Picture?

Statistic	N	Mean	St. Dev.	Min	Max
Female	812	1.49	0.50	1	2
Age	812	35.72	22.70	−99	60
Education	812	3.00	1.41	1	5
Married	812	0.84	0.37	0	1
Income	683	55.98	26.42	20.18	180.58

Not All Data Issues Appear in Summary Statistics Tables

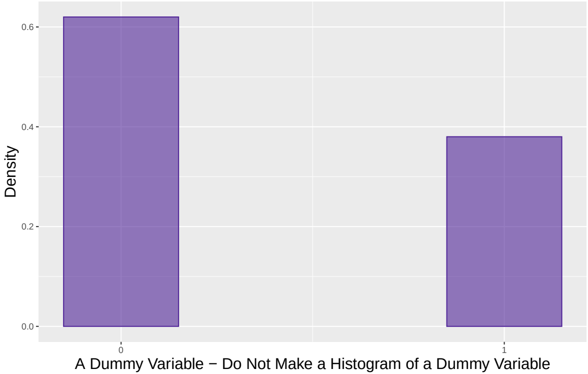
A **histogram** is a bar graph that plots the distribution of a variable X by:

- Partitioning the support of X into equally-spaced bins
- Counting the number of observations in each bin
- Using bars to plot the relationship between the range of X value(s) included in each bin and the number (or the proportion/density) of observations that fall within that bin

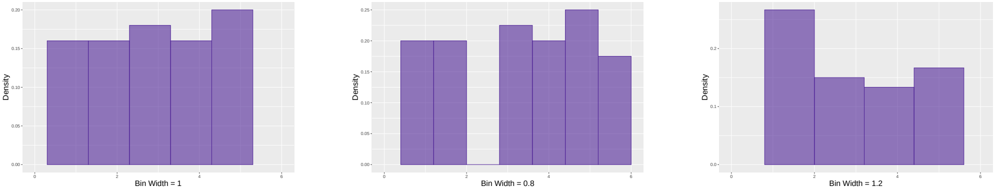
With histograms, there is only one statistical decision to be made: how many bins?

- How many bins is also one of many aesthetic decisions

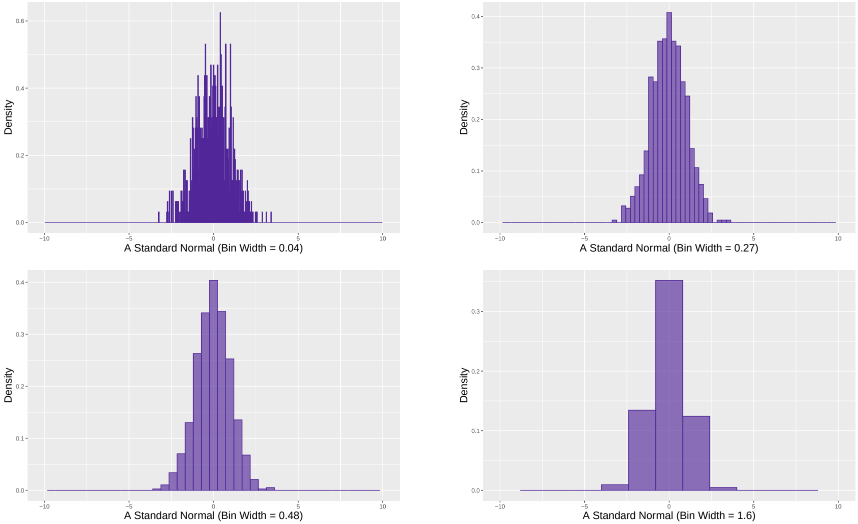
Histograms: The Good, the Bad, and the Ugly



Histograms: The Good, the Bad, and the Ugly



Histograms: The Good, the Bad, and the Ugly



Kernel Density Estimation

Histogram can depend on bin width and the starting point for the first/lowest bin

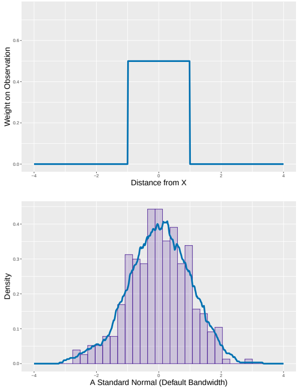
- An alternative would be to define a function $f(x)$ that counted up the number of observations "near" x (i.e. within $h > 0$ of x) for all values in the support of x
- We could then scale the function $f(x)$ so that the area under the curve sums to one

Kernel density estimation generalizes this approach for different weighting functions (kernels)

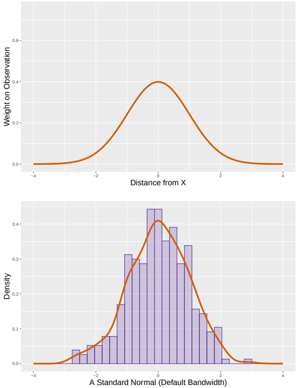
- The example above is kernel density estimation with a rectangular/uniform kernel
  - The rectangular kernel puts equal weight on all data points within bandwidth h of x
- We can instead calculate a weighted count of observations near x
  - Commonly used kernel include: Gaussian (i.e. normal), Epanechnikov (parabolic)

Kernel Density Estimation in Practice

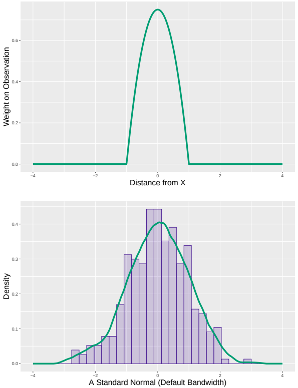
Rectangular Kernel



Gaussian Kernel

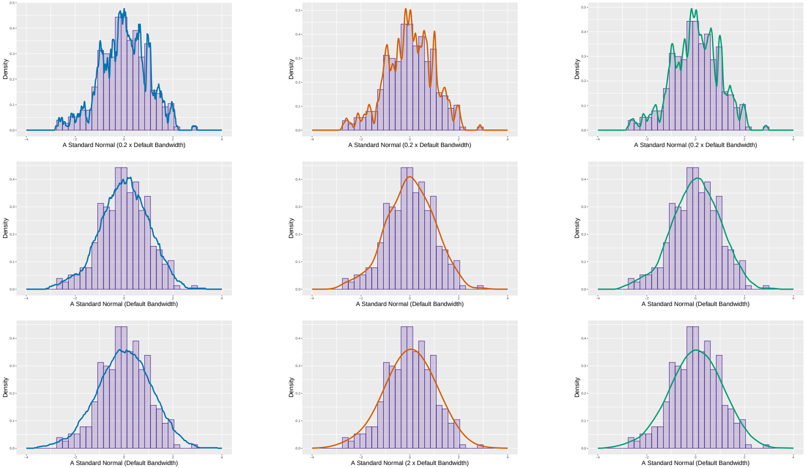


Epanechnikov Kernel

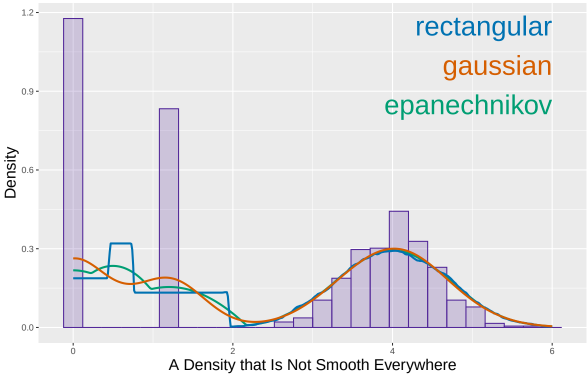


Kernel Density Estimation in Practice

Rectangular Kernel　　Gaussian Kernel　　Epanechnikov Kernel



Q: When Shouldn't You Use a Kernel Density?

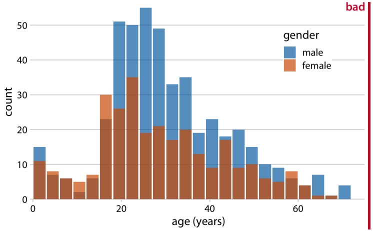


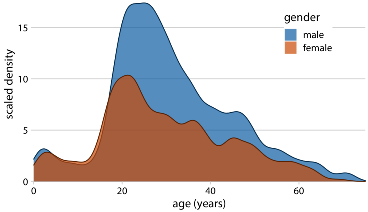
Q: When Shouldn't You Use a Histogram?

There is usually nothing wrong with using a histogram as long as you choose the size and placement of the bins carefully (though see the example on Slide 8 for how this can go wrong)

- However, it is usually better to use a kernel density plot if (you believe) the underlying density is smooth, as the bins add little to our understanding (see Slide 9 for an example)

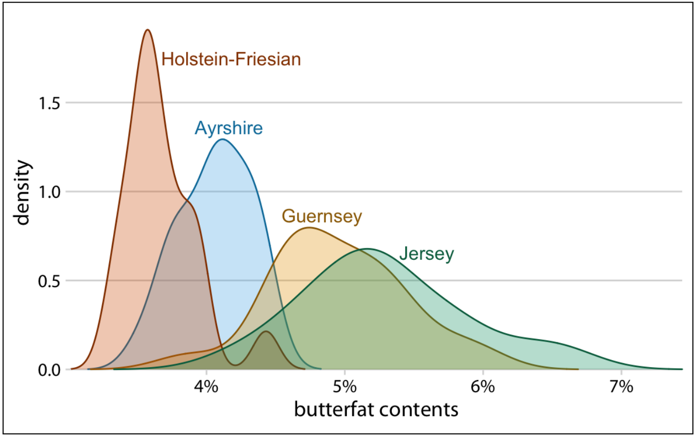
Kernel density plots also work much better when you want to show more than one distribution





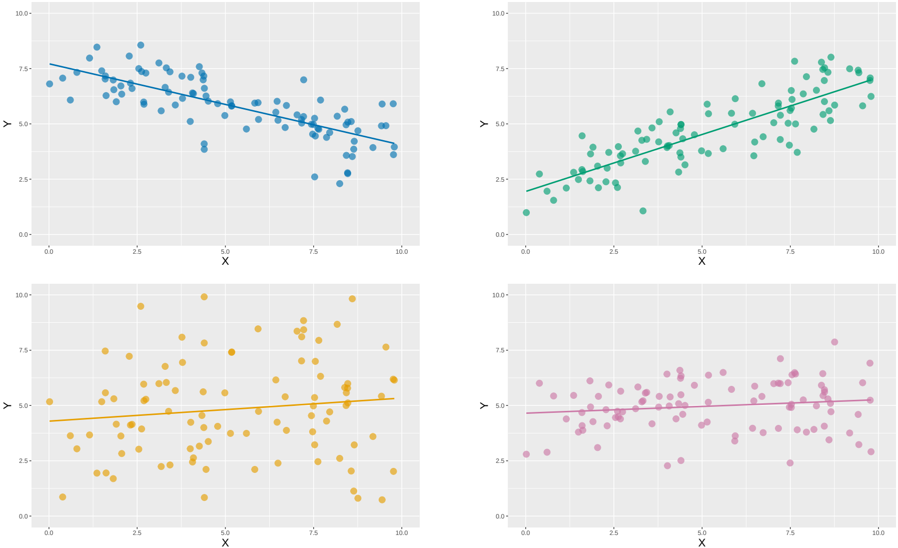
source: Wilke (2019)

Q: When Shouldn't You Use a Histogram?

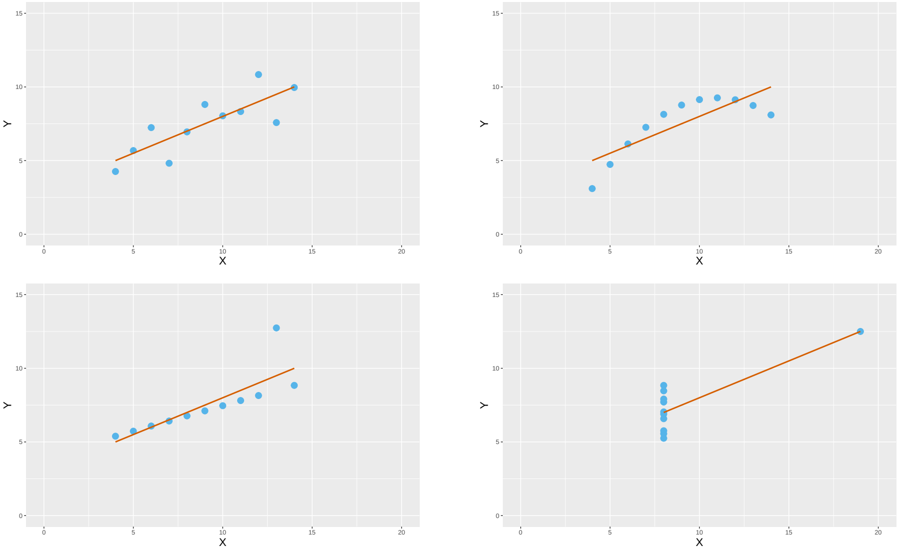


source: Wilke (2019)

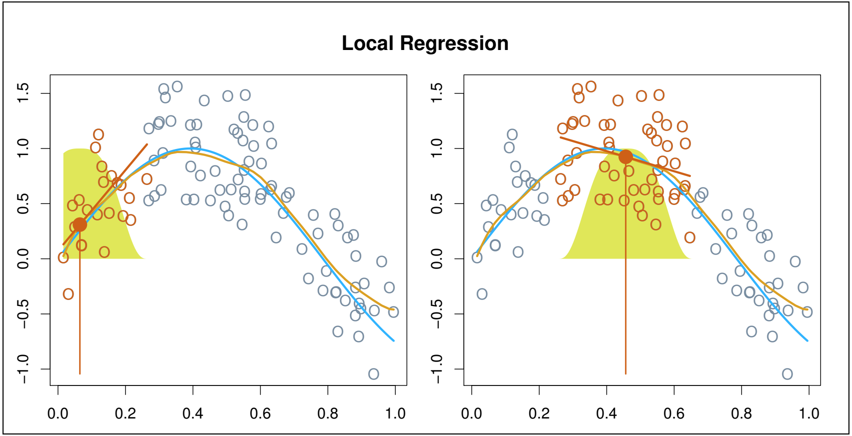
A Scatter Plot Is Worth a Thousand Words



Anscombe's Quartet

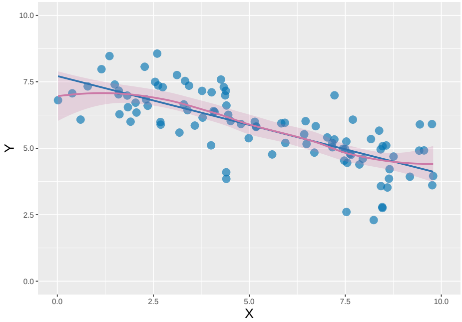


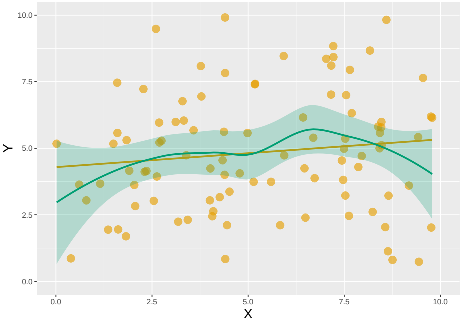
Local Linear Regression



source: James et al. (2021)

Your Workhorse Exploratory Scatter Plot





Summary

- Summary statistics table: mean, standard deviation, minimum, maximum, count
- One variable: histogram, kernel density, or both
- Two variables: scatter plot with a linear and/or local polynomial fit

Provisions Data

Provisions Williamstown has graciously shared 18 months of transactions data with us

- `ECON370-provisions.zip`, which was emailed to you, contains all relevant files

The file `ECON370-provisions-transactions.csv` contains data on $N = 16{,}003$ (almost all) cash register sales transactions that took place between January 1, 2023, and June 30, 2024

- Mostly clean, `prep-transactions-2024-09-12.R` is the cleaning file

The file `ECON370-provisions-items.csv` contains data on all items sold at Provisions

- Data is as it was when it was downloaded from square (so you get to clean it)

Lab #2

You're going to conduct exploratory data analysis on the Provisions transaction data

- Cleaning, summary statistics table, several histograms, and a scatter plot
- You will need to choose a measure of sales and/or revenue and justify it
- You will also need to aggregate the data up to the daily, weekly, or monthly level
- You will need to convert your summary statistics table into a pdf or an html file (the R package `stargazer` can output formatted tables, latex and pdf templates provided)

The file `ECON370-lab2.text` contains an outline of the program you need to write

- Look carefully at the R and Python file path code at the top of the file