GEOG 210B Analytical Methods II

Winter 2018

Introduction with R examples

Syllabus

- COMBINED LAB & LECTURES
  - T R 11:00-12:15 ELLSN 3620
  - T 12:30- 2:20 ELLSN 3620
- Pace = F(your progress/interest)
- Format
  - AM basic definitions = lecture material & examples
  - PM work with you in the lab & your questions and discussion but also complete review of lecture material

- Lab = hands on with real data: survey microdata & spatial data

Work and Grades

- Lab reports experience – four parts (80%)
  - Part 1 Data description and hypothesis testing
  - Part 2 Linear and Count Data Regression
  - Part 3 Spatial statistics
  - Part 4 Categorical data analysis
- Class Participation (20%) Class discussion & questions

Project Report

- The lab reports should contain the following.

- **Background**: A general description of the topic. In this section you can also include the hypotheses you want to test.
- **Model Overview**: To describe the models: list the dependent variable for each, show key defining equations of each model, the types of variables each model is designed to explain, and discuss regression diagnostics available to select a good model.
- **Estimated Model**: In this section describe one model for each type: linear regression, poisson, logit/discrete choice. Provide a behavioral interpretation of the model results and discuss its goodness of fit and any other diagnostics that help you select a model.

- **Guidelines:** Keep each report between 1,000 to 2,000 words. Use 2-3 Tables and 2-3 figures. Style is up to you.
- My recommendation: create tables and figures and then describe them.

Probability and statistics, data analysis, regression methods and some other tools

Textbook Definition

- Statistics is the methodology used in studies that collect, organize, and summarize data through graphical and numerical methods, analyze the data, and ultimately draw conclusions
- It is also a “language”
- In this course we will review some of the basic ideas and use stats methods

Introductions

- All (name, specialty, data analysis needs)
- Probability & Statistics background

- What do you expect from this class?

- What did you study in other stat courses?

Other Resources

(checked December 21, 2017)

- Electronic textbooks
  - http://wiki.stat.ucla.edu/socr/index.php/EBook
  - http://www.itl.nist.gov/div898/handbook/
  - http://www.statsoft.com/textbook/stathome.html
  - http://statlink.tripod.com/id6.html
  - http://www.socialresearchmethods.net/kb/sampstat.htm
- Free statistics software (some are excellent!)
  - http://freestatistics.altervista.org/stat.php
- R
  - https://www.r-project.org
  - https://stats.idre.ucla.edu/r/
- My favorite statistical software support site
  - https://stats.idre.ucla.edu/other/dae/

BASIC IDEAS

Statistical Population

- A set of entities (objects, persons, locations, cities, regions, etc) about which we want to infer something

- Example 1: Travel in California – define statistical population
  - Qualifiers: All persons that live in California? All the persons that actually travelled?
- Example 2:Bicycle use in California

Population

Characteristic

- A measurable attribute of a population entity
  - Sick or healthy
  - Salary
  - Number of songs owned

Variable

- A population characteristic that takes on different values for the elements comprising the population
  - Y=0 if sick, Y=1 if healthy
  - Y=$X/month
  - Y=1,2,…1billion

Census vs Sample

- Census or Register or Inventory = Complete enumeration and tabulation of the population characteristics
  - Feasible?
  - Examples?

Sometimes ok = example the NETS data

- Sample = A subset of the population used to make inferences about population characteristics
  - Ideal properties?
  - Issues we know?

Errors

Sampling Error

- The difference between the value of a population characteristic and the value of the same characteristic in the sample
  - Error attributable exclusively to sampling

Non-sampling Error

- Any error that is due to imperfect measurement
  - Selection bias
  - Wrongly worded questions
  - Instrument problems

Discussion: Exit Polls in Elections

Sampling

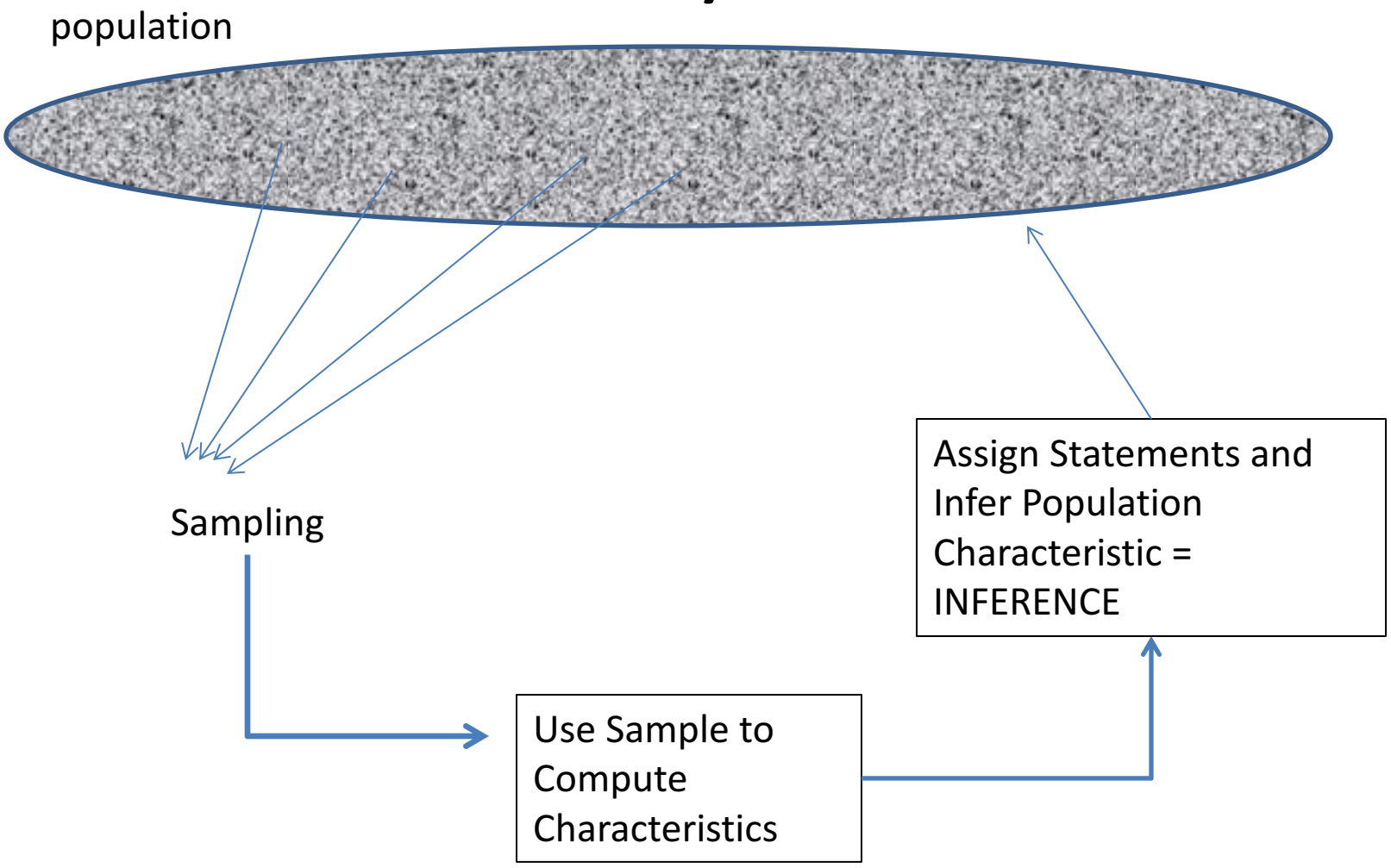
Representative Sample

- The characteristic of the sample closely matches the population characteristic

Equiprobability Random Sample

- Every entity (persons, objects, locations, etc) has the same probability of selection
  - Not the only way to sample!

Two Way Path


population
Sampling
Assign Statements and Infer Population Characteristic = INFERENCE
Use Sample to Compute Characteristics

Inferential Statistics

Statistical Estimation

- Use information in a sample to estimate the value of an unknown population characteristic

Hypothesis Testing

- Procedure(s) in which we decide if the data in a sample support a proposition (Null hypothesis) about the value of a population characteristic

These two go hand in hand in inference

Background

Data analysis

- We have a set of measurements (y_1, y_2, …, y_i, ..y_N)
- We need to understand the data variation
- We need to come up with indicators that describe their "essence"
- In this way we can compare and contrast these measurements with other measurements
- We also want to understand if the variation is just by chance or some kind of underlying process that is not immediately obvious
- Then we want to use our findings to forecast the future
- What would you do?

Different ways

- Data description using tabulations and graphs
- Descriptive statistics (data summaries)
- Frequency analysis and fitting of distributions
- Regression
  - Single equation regression methods
  - Multiple equation regression methods
  - Single equation latent variable models
  - Multiple equation latent variable models
- Clustering, grouping, classification models
- Data mining and neural networks
- Spatial correlations
- Spatial dependency and trends
- Comparison of models with theory!!!!!!

In the Laboratory

- Examine databases (a survey sample & a spatial database)
- Create summaries of data
- Learn how to use R using RStudio
- Build models
- Interpret and report findings

FOUNDATION

Statistical Inference

- Infer = derive a conclusion from facts or premises
- Infer in stat = infer about large groups using information from a small group
- Large group = statistical **population**
- Small group = **sample**
- Central purpose of contemporary statistics
  - Estimation of population parameters
  - Tests of hypotheses
  - Building models

Probability vs Statistics

- http://wiki.stat.ucla.edu/socr/index.php/EBook
- Probability: Theoretical & mathematical framework for statistical inference
- Statistics: The tools to collect data and make sense from them about the world
- Tools to be used in helping you find answers to substantive questions!

Quick list A (variable types)

- Level of measurement
  - **Nominal (categorical) =** names or labels (operation: same or not same).
  - **Ordinal** = nominal with an order (good-bad, agree-disagree, love-hate).
  - **Interval** = ordinal but also separated by same interval (the year date).
  - **Ratio** = all the features of interval and meaningful ratios between pairs. Zero not arbitrary. Add, substract, multiply, divide are ok

Level of Measurement of Data	Definition	Example & Model
Nominal (categorical)	Categorical in nature and observations are recorded in discrete units	Names, colors, means of travel (car driver, car passenger, bus, bicycle, walk)
Ordinal	Responses/observations with a inherent or constructed order	good-bad, strongly agree-agree - disagree, love-hate
Interval	Measurements along a scale that possesses a fixed but arbitrary interval and a arbitrary origin. Addition or multiplication by a constant will not alter the nature of the observed data	Continuous: Temperature in degrees Farheneit, average height of populations, average number of miles driven Discrete: Number of occurrences, Trips a person makes in a day
Ratio data	Similar to interval but the scale has a true zero origin	Continuous: Number of miles driven per day Discrete: Number trips a person makes in a day

Quick list B (variable types)

- **Continuous** = (theoretically) an infinite number of gradations between two values – your weight or height, gasoline you consume

- **Discrete** = definite gap between values and cannot be expressed in portions - #cars you own, houses you lived, places you visited

A + B = type of analysis

- If we have a continuous variable that is a ratio => **linear regression** is a good method
  - If we have large amount of observations at zero: Zero Inflated regression
- If you have a discrete variable that is categorical (red car vs. blue car) => a **Logit** or **Probit** regression is a good method
  - When two categories: called **Binomial/binary**
  - When many categories: called **Multinomial**
- If you have counts of events (trips)
  - Count data models (Poisson, Negative Binomial)
  - If we have large amount of data at zero: Zero inflated Poisson, NegBin etc.
- If you have options from which people select one => **discrete choice** is a good method

At the end of this course you will be able to identify proper use of different methods

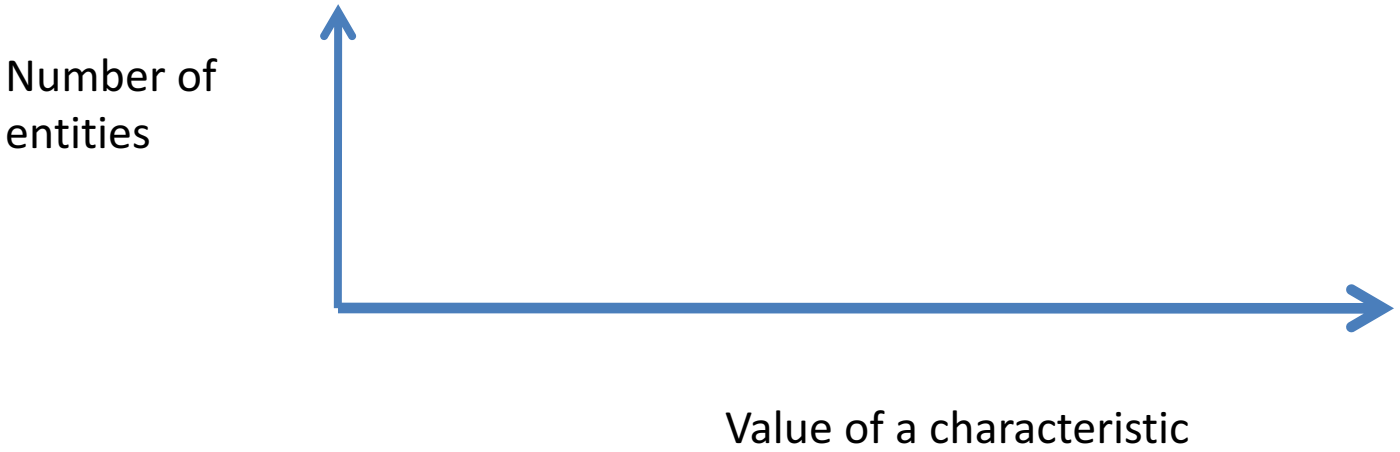
Statistical Measures & Data

- **Statistic** = Characteristic or measure obtained from a sample
- **Parameter** = Characteristic or measure obtained from a population
- Data
  - Measures of central tendency
  - Measures of dispersion (spread)
  - Measures of association (relationship)

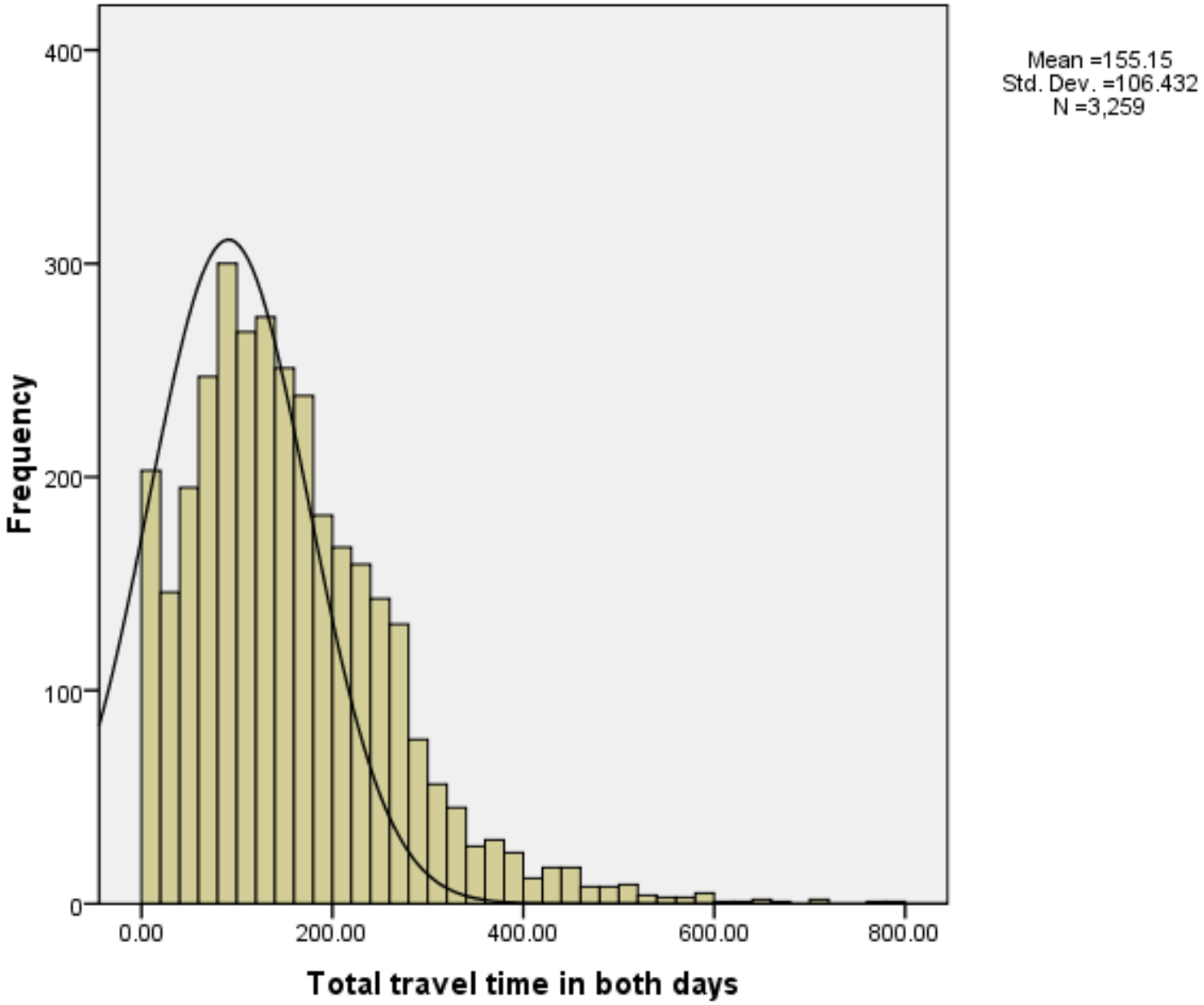
Distributions

- http://www.socr.ucla.edu/htmls/SOCR_Distributions.html

- Mathematical equations representing how values are organized



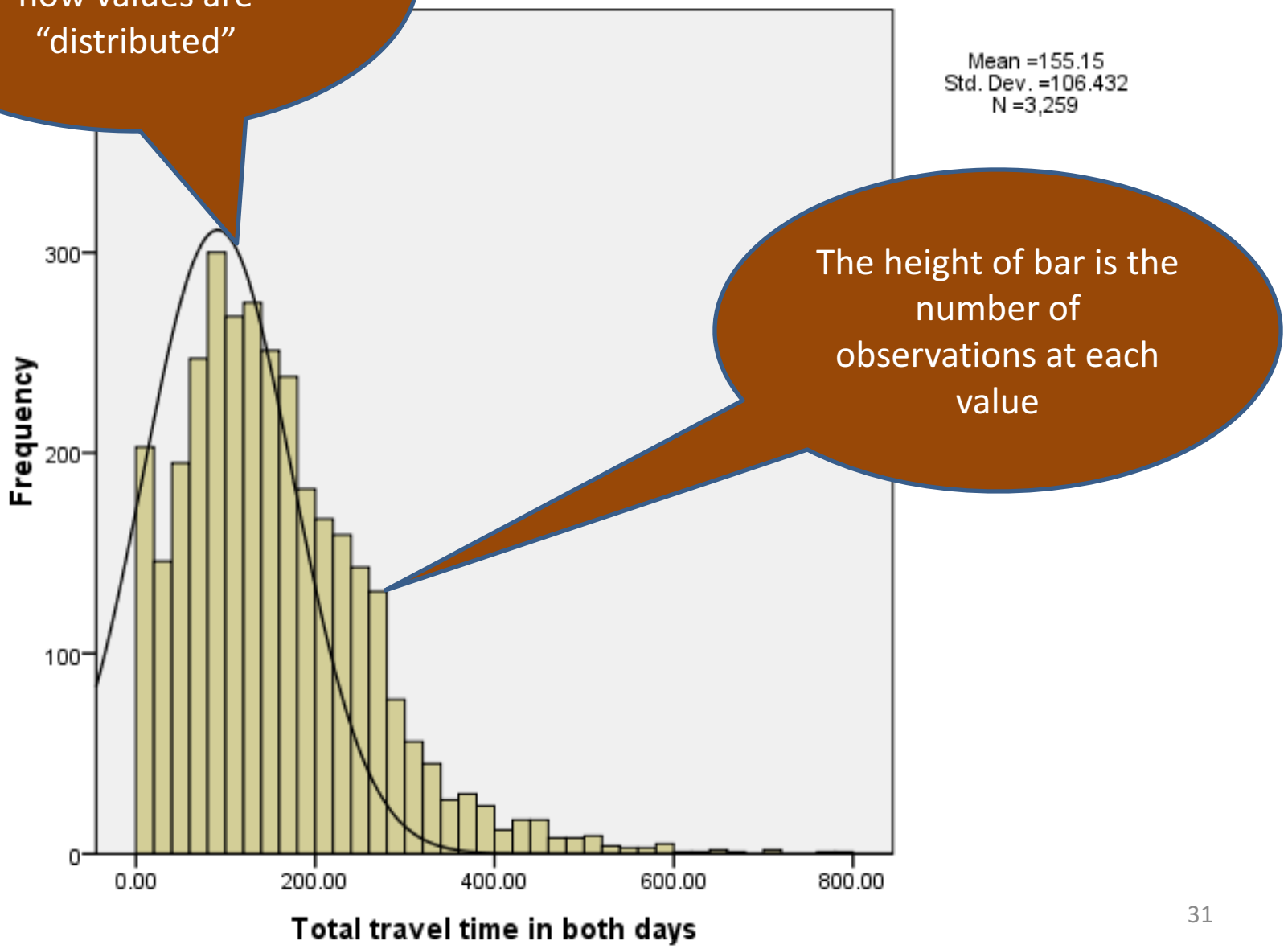
Example
(time traveling in 2 days of a diary



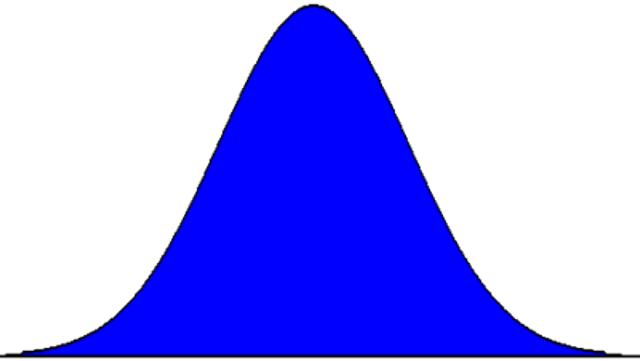
Example

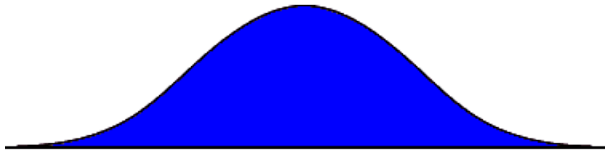
veling in 2 days of a diary

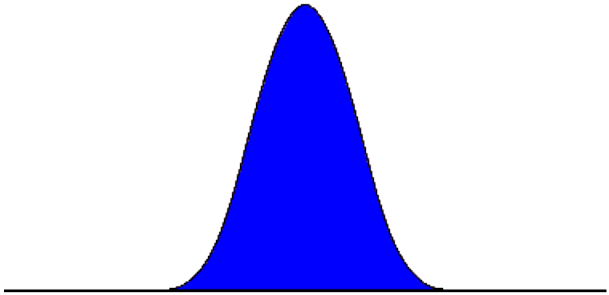
This line is an equation that tries to represent how values are "distributed"



The height of bar is the number of observations at each value



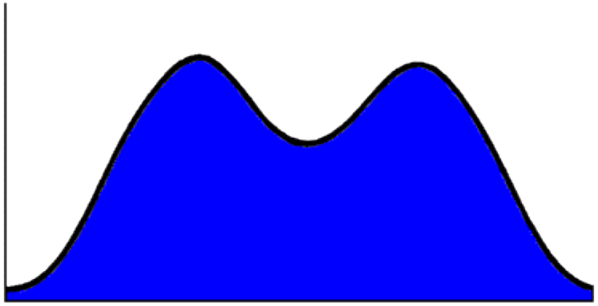




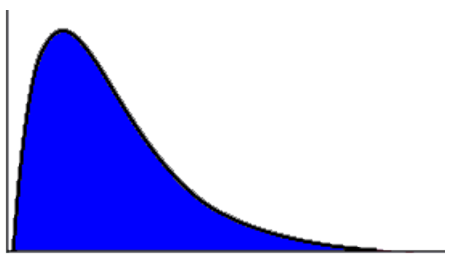
Symmetric:
1) mesokurtic,

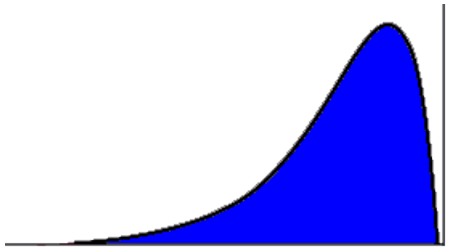
2) platycurtic (flat),

3)leptocurtic (thin)

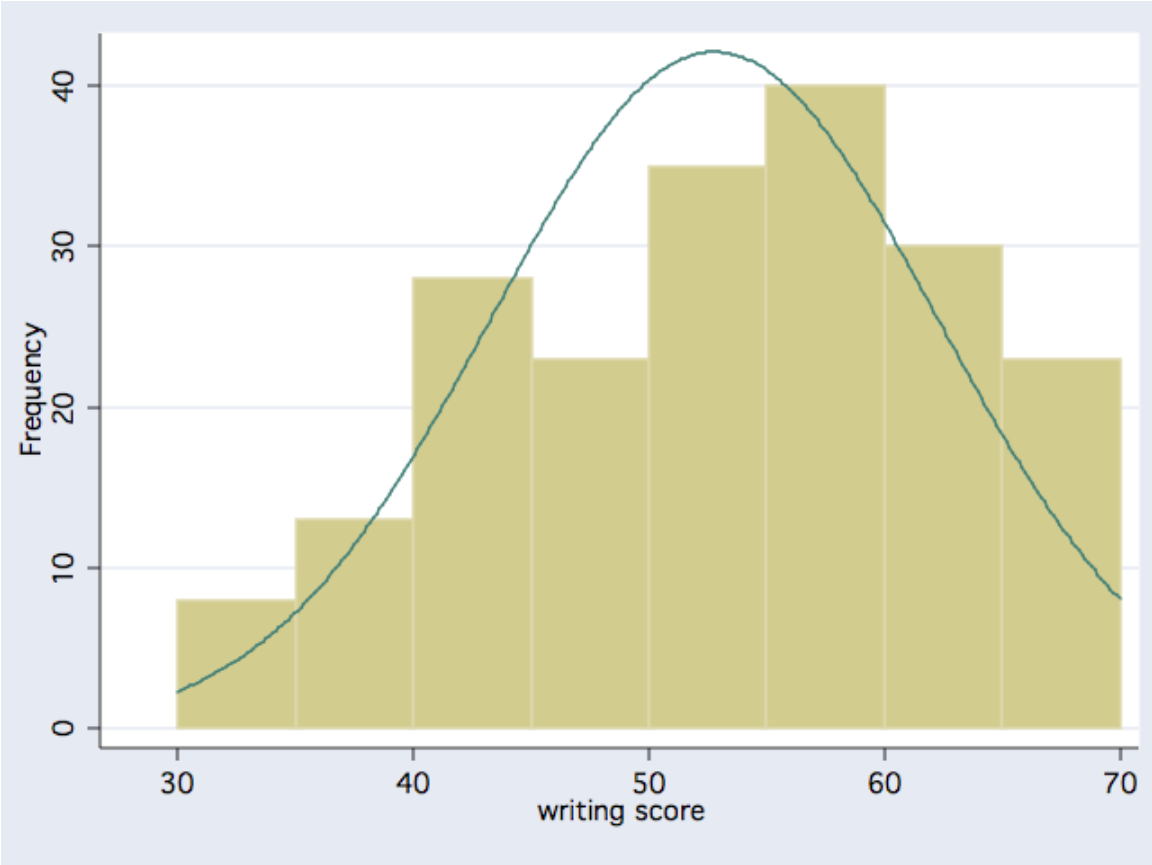


Bimodal





Skewed



Center of the data

CENTRAL TENDENCY

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

population

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

sample

Mean

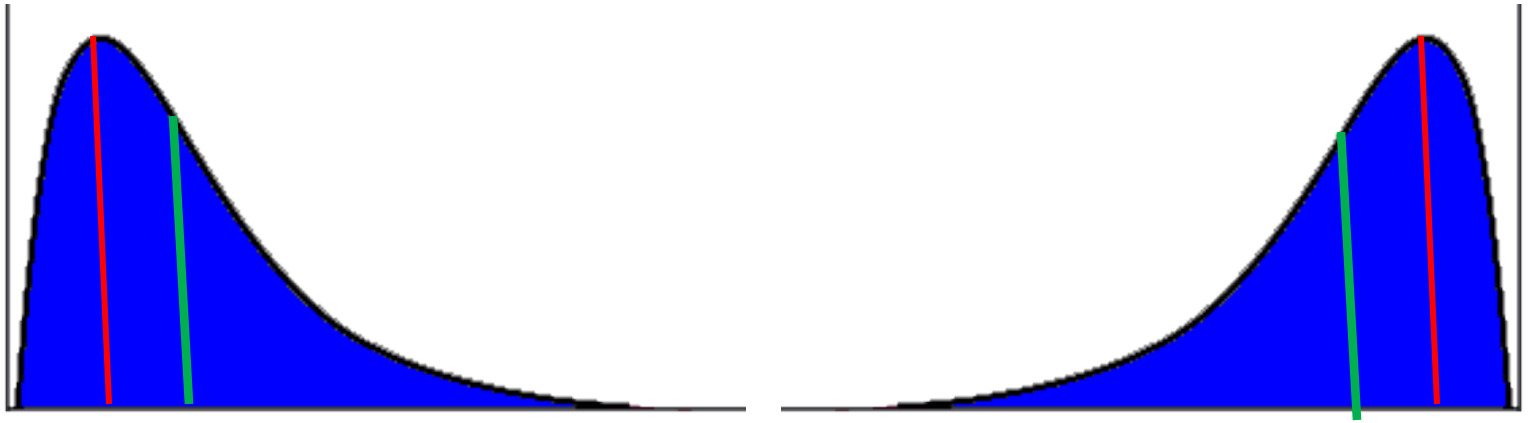
Note: for population we sum over ALL items in the population (big N). For sample we sum over all the units in the sample (we do not have ALL the units in the population)

Other measures (median)

- The Mean is used in computing other statistics (such as the variance) - Not appropriate for skewed distributions.
- **Median** = the midpoint of the data after being ranked (sorted in ascending order).
  - There are as many numbers below the median as above the median.
  - The Median is the center number and is good for skewed distributions because it is resistant to extremes.

Median

- In a positively skewed distribution, the tail is to the right and the mean is larger than the median. In a negatively skewed distribution, the tail is to the left and the mean is smaller than the median.



Symmetric Distribution The data values are evenly distributed on both sides of the mean. In a symmetric distribution, the mean is the median.

http://www.joma.org/images/upload_library/47/schneiter/mm_applet.html

Other measures (mode)

- Mode = the most frequent number
  - Skewed Distribution The majority of the values lie together on one side with very few values (the tail) to the other side.
  - Mode is used to describe the most typical case. It can be used with nominal data.
- The mode may or may not exist and there may be more than one value for the mode.

DISPERSION/SPREAD

Definition

- Population Variance = The average of the squares of the distances from the population mean. It is the sum of the squares of the deviations from the mean divided by the population size.
- The units of the variance are the units of the population mean squared.

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Distance of value x from the mean (center) and then squared

Called deviation from the mean

Reflect: what does it represent?

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n - 1}$$

n-1 needed to give this statistic suitable properties (unbiased – later in class). The units on the variance are the units of the population squared.

Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$

$$s = \sqrt{s^2}$$

Unit of measurement same as the xs

Preferable Statistics

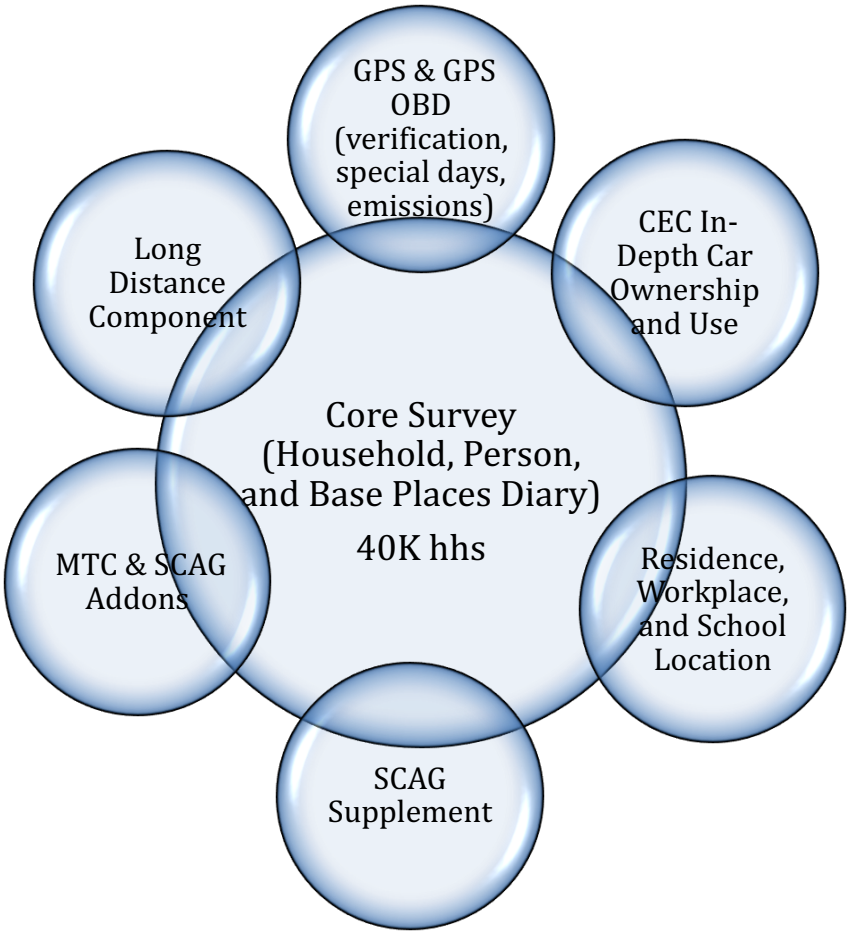
- When they use all the data at hand
- Depending on the situation
- Make sure you do not convey unwanted messages
- See typical statements such as the average American drives 15,000 miles a year (really?)

USING DATA FROM SMALLHHFILE.CSV

Data are from the

CALIFORNIA HOUSEHOLD TRAVEL SURVEY

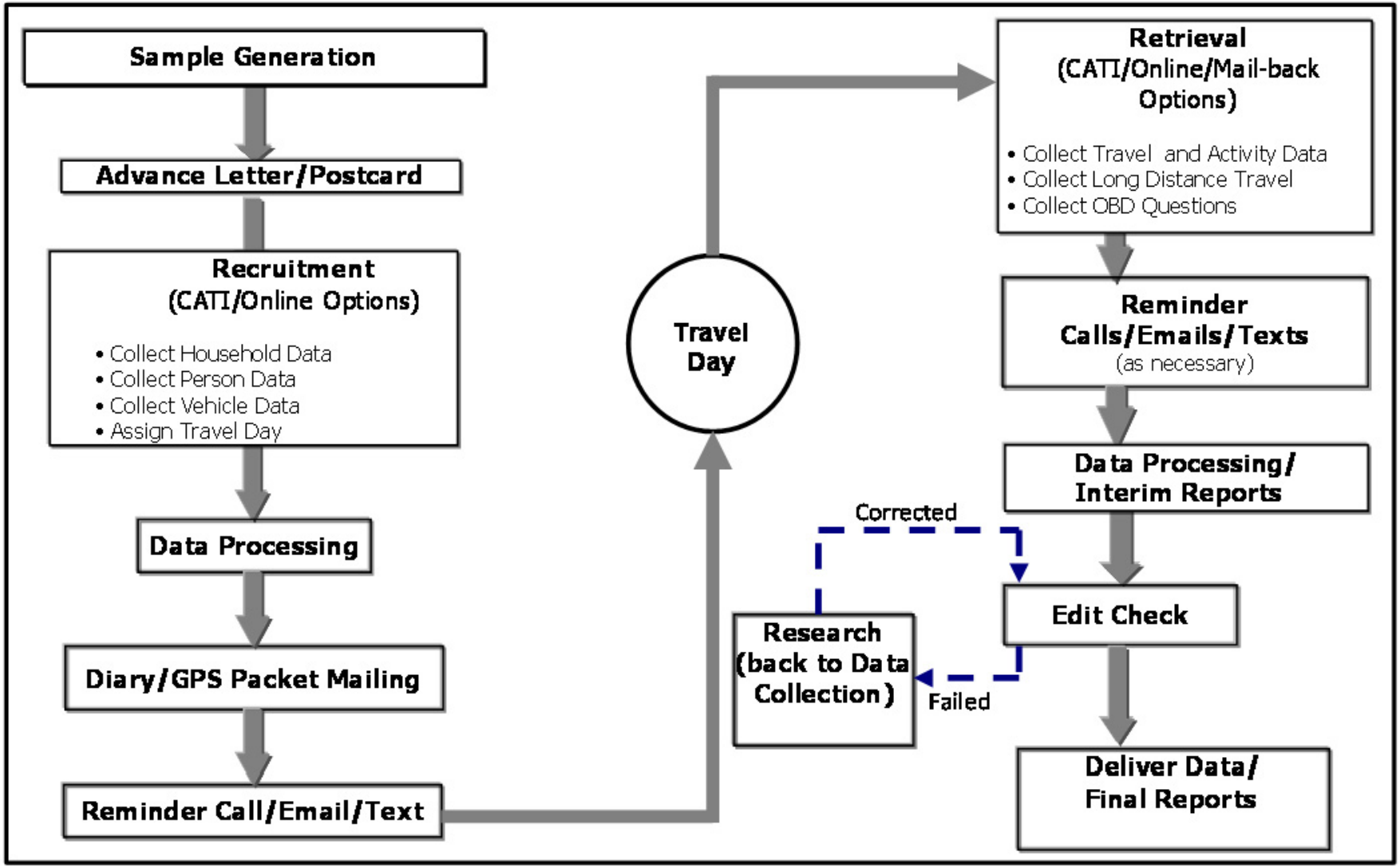
The CHTS Data Collection Overall Scheme Core-Satellite
(One travel day - all days of a year - weekends included)



These are the target numbers for the company NUSTATS (data collection)

- 53,483 California households, with the number of households sampled proportionate to the population in the sampling strata;
- Of these, 48,384 households were to be Non-GPS and 5,099 were to be GPS Households
- Of the GPS households, the desired distribution was:
  - ✓ 400 Wearable devices
  - ✓ 3,099 MTC Wearable devices
  - ✓ 400 Vehicle GPS devices
  - ✓ 800 Vehicle and OBD devices
  - ✓ 400 Energy Commission Vehicle GPS and OBD devices

Figure 4.2.2.1: CHTS Survey Process



Response Mode	START DATES		END DATE
	English	Spanish	
CATI	2/2/2012	4/5/2012	2/14/2013
Online	2/3/2012	4/5/2012	2/14/2013
Mail			2/7/2013

From CHTS/NUSTATS Final Report on GauchoSpace

Sample Type	Sample Used in Main Survey		Recruited Households			Retrieved Households		
	Number (A)	% of Total	Number (B)	% of Total	Recruitment Rate (B)/(A)	Number (C)	% of Total	Retrieval Rate (C')/(B)
ABS Matched/ Listed	1,410,365	66.5%	58658	93.0%	4.2%	38934	91.8%	66.4%
Unmatched	585520	27.6%	2996	4.7%	0.5%	2458	5.8%	82.0%
Energy Commission Samples	121835	5.7%	985	1.6%	0.8%	809	1.9%	82.1%
Kern County Transit Intercept	1353	0.1%	443	0.7%	32.7%	230	0.5%	51.9%
Total	2,119,073	100.0%	63082	100.0%	3.0%	42431	100.0%	67.3%

https://www.tutorialspoint.com/r/index.htm

R AND RSTUDIO

For your laptop or your PC

In the EH 3620 (Descartes Lab) Rstudio is already installed

Before First Lab

1. Install R

Windows:

Basics of R

1.Go to https://cran.r-project.org/bin/windows/base/
2.Click download link and run the `.exe` file to install

Mac:

1.Go to https://cran.r-project.org/bin/macosx/
2.Download appropriate version

1. – OS X 10.9 (Mavericks) or higher: "R-3.3.1.pkg"
2. – OS X 10.6 (Snow Leopard) - 10.8 (Mountain Lion): "R-3.2.1-snowleopard.pkg"

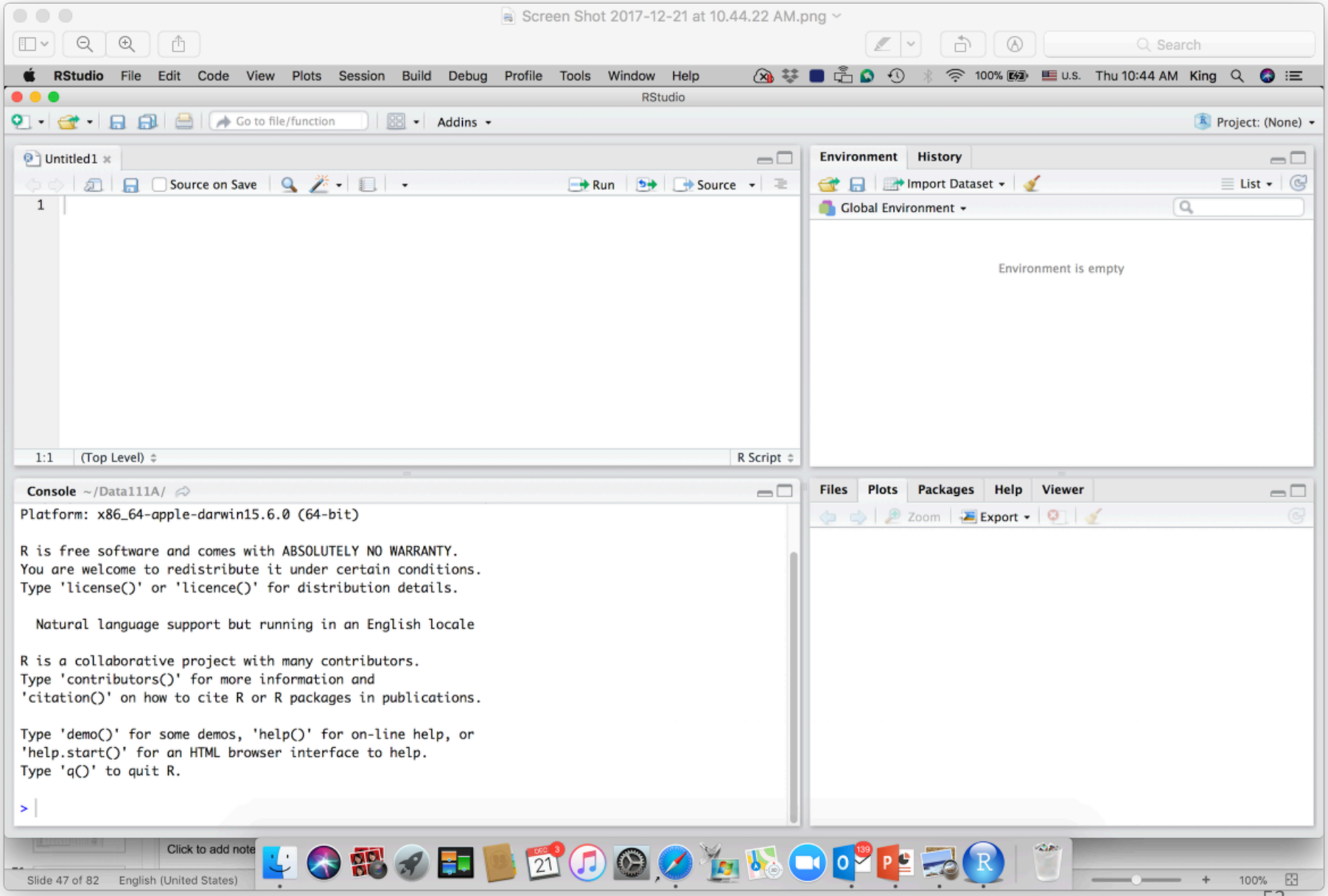
3.Run downloaded `.dmg` file to install

2. Install RStudio

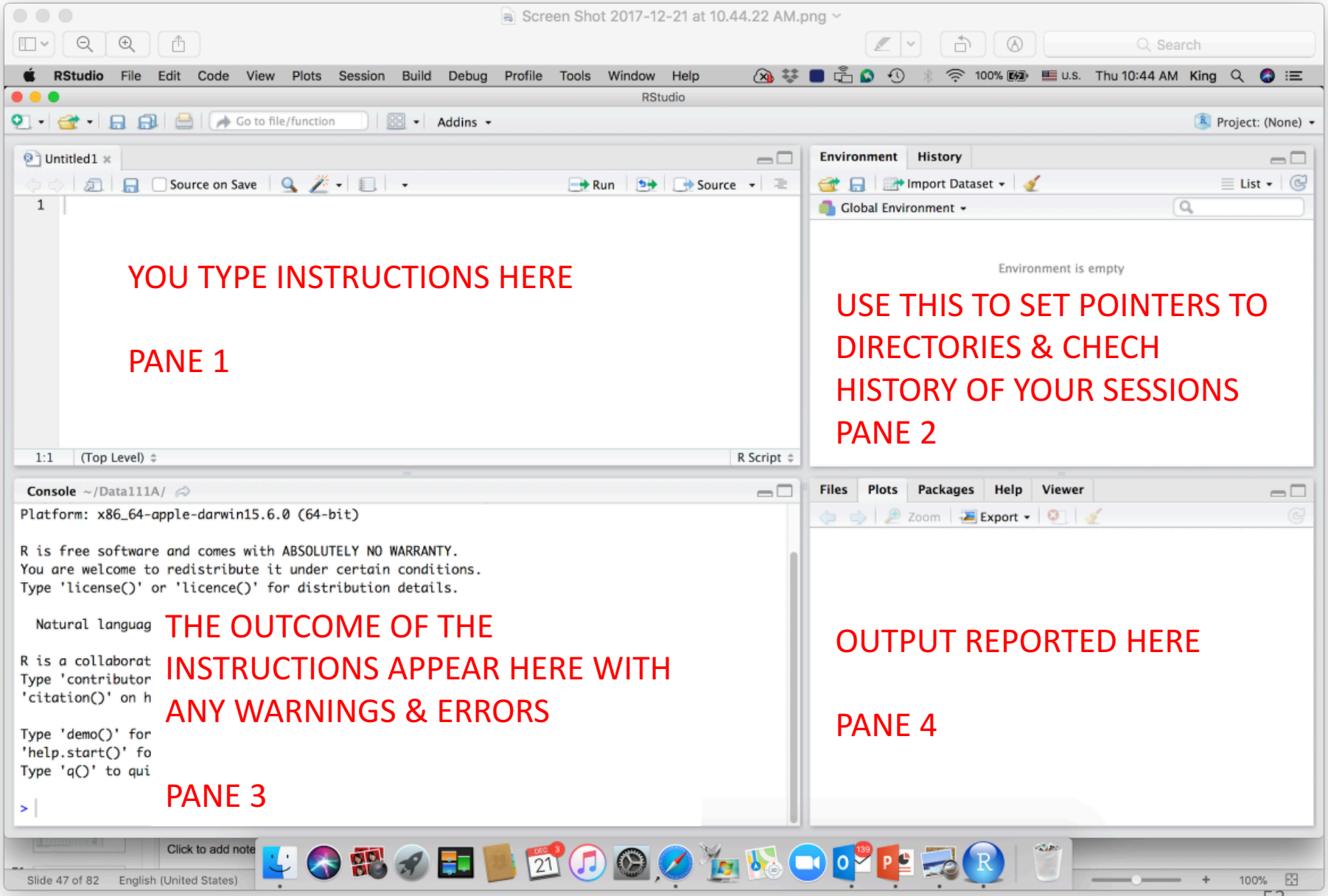
All Systems

1.Go to https://www.rstudio.com/products/rstudio/download3/
2.Download and install appropriate version from "Installers for Supported Platforms"

4 Panes of RStudio



4 Panes of Rstudio (I numbered them to reference later)



Packages are collections of **R** functions, data, and compiled code in a well-defined format. The directory(location) where packages are stored is called the **library**. **R** comes with a standard set of packages.

We will use two sets of instructions:

In Pane 1 we write:

install.packages(”NAME OF PACKAGE")

This will search in the default locations (e.g., https://cran.rstudio.com…) and download and install the package on your local computer

You will the progress of this in pane 3.

Typing in Pane 1
library(NAME OF PACKAGE)

Activates all the functions in this library

In Pane 4 under “Packages” you can find the default packages for your installation and any downloaded packages. Checking the left hand side box does the same as the library instruction above

Preliminaries

- Download files from gauchospace and save in your working directory

- SmallHHfile.csv
- Intro210B.R
- Codebook for SmallHHfile.doc

# Look at the structure of the file
str(SmallHHfile)

The console shows this

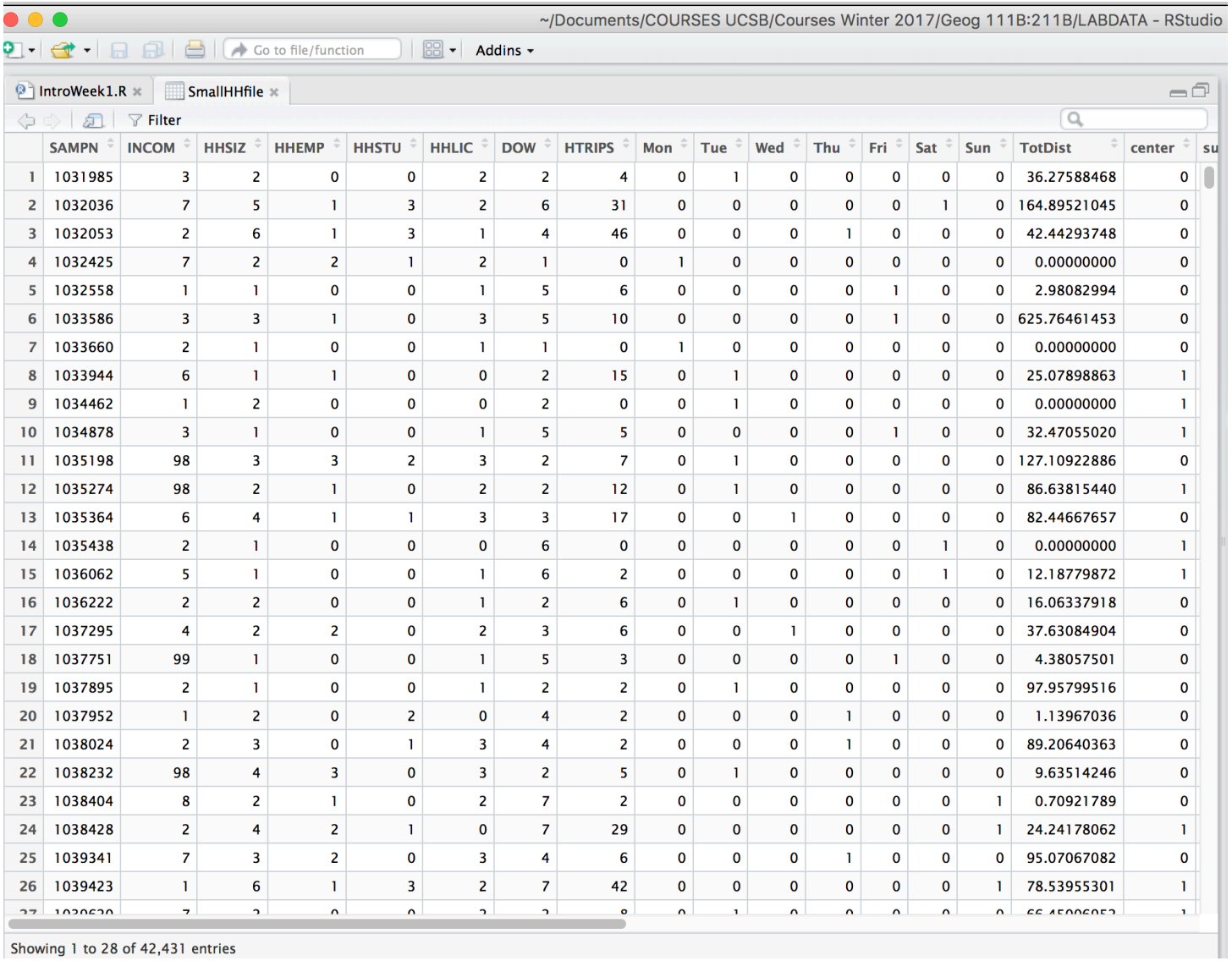
We created in R a data.frame

Each row is an observation and each column is a variable

We have 42431 observations and 31 variables

```
Console ~/Documents/COURSES UCSB/Courses Winter 2017/Geog 111B:211B/LABDATA/
> SmallHHfile <- read.csv("~/Desktop/geog111b/SmallHHfile.csv", header=TRUE)
> str(SmallHHfile)
'data.frame':   42431 obs. of  31 variables:
 $ SAMPN  : int  1031985 1032036 1032053 1032425 1032558 1033586 1033660 1033944 1034462 1034878 ...
 $ INCOM  : int  3 7 2 7 1 3 2 6 1 3 ...
 $ HHSIZ  : int  2 5 6 2 1 3 1 1 2 1 ...
 $ HHEMP  : int  0 1 1 2 0 1 0 1 0 0 ...
 $ HHSTU  : int  0 3 3 1 0 0 0 0 0 0 ...
 $ HHLIC  : int  2 2 1 2 1 3 1 0 0 1 ...
 $ DOW    : int  2 6 4 1 5 5 1 2 2 5 ...
 $ HTRIPS : int  4 31 46 0 6 10 0 15 0 5 ...
 $ Mon    : int  0 0 0 1 0 0 1 0 0 0 ...
 $ Tue    : int  1 0 0 0 0 0 0 1 1 0 ...
 $ Wed    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Thu    : int  0 0 1 0 0 0 0 0 0 0 ...
 $ Fri    : int  0 0 0 0 1 1 0 0 0 1 ...
 $ Sat    : int  0 1 0 0 0 0 0 0 0 0 ...
 $ Sun    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ TotDist: num  36.28 164.9 42.44 0 2.98 ...
 $ center : int  0 0 0 0 0 0 0 1 1 1 ...
 $ suburb : int  0 1 0 0 1 0 0 0 0 0 ...
 $ exurb  : int  1 0 0 1 0 1 1 0 0 0 ...
 $ rural  : int  0 0 1 0 0 0 0 0 0 0 ...
 $ other  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ highinc: int  0 1 0 1 0 0 0 1 0 0 ...
 $ HHVEH  : int  2 1 2 2 0 2 1 0 0 1 ...
 $ HHBIC  : int  2 4 2 3 0 1 1 1 0 2 ...
 $ VEHNEW : int  1 1 2 2 2 2 2 2 2 2 ...
 $ OWN    : int  1 1 2 1 2 2 1 1 2 2 ...
 $ CarBuy : int  1 1 0 0 0 0 0 0 0 0 ...
 $ snglhm : int  1 1 1 1 1 1 1 1 0 0 ...
 $ ownhm  : int  1 1 0 1 0 0 1 1 0 0 ...
 $ MilesPr: num  18.14 32.98 7.07 0 2.98 ...
 $ TrpPrs : num  2 6.2 7.67 0 6 ...
```

#Look at the contents of the file
View(SmallHHfile)

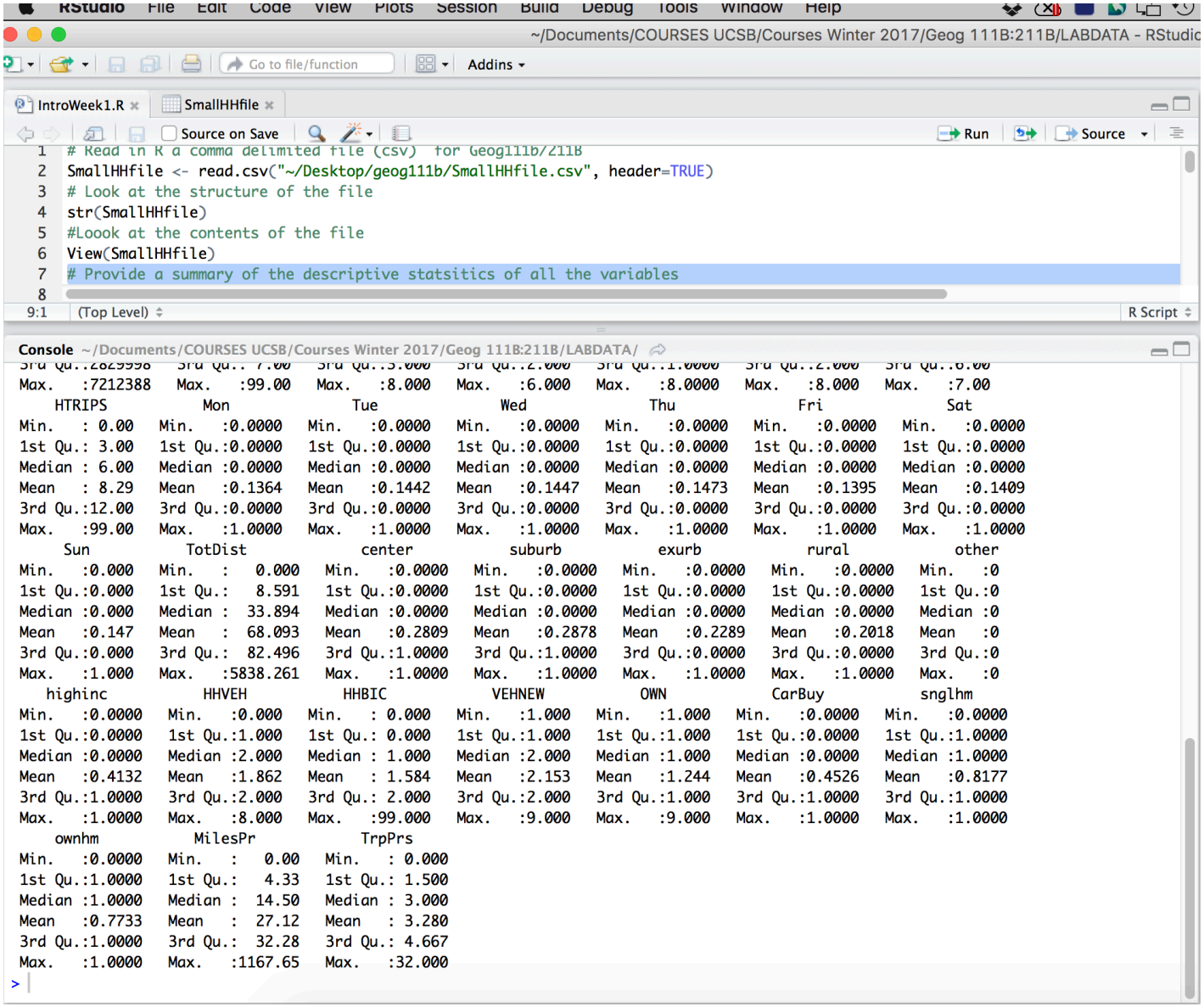


	SAMPN	INCOM	HHSIZ	HHEMP	HHSTU	HHLIC	DOW	HTRIPS	Mon	Tue	Wed	Thu	Fri	Sat	Sun	TotDist	center	su
1	1031985	3	2	0	0	2	2	4	0	1	0	0	0	0	0	36.27588468	0	
2	1032036	7	5	1	3	2	6	31	0	0	0	0	0	1	0	164.89521045	0	
3	1032053	2	6	1	3	1	4	46	0	0	0	1	0	0	0	42.44293748	0	
4	1032425	7	2	2	1	2	1	0	1	0	0	0	0	0	0	0.00000000	0	
5	1032558	1	1	0	0	1	5	6	0	0	0	0	1	0	0	2.98082994	0	
6	1033586	3	3	1	0	3	5	10	0	0	0	0	1	0	0	625.76461453	0	
7	1033660	2	1	0	0	1	1	0	1	0	0	0	0	0	0	0.00000000	0	
8	1033944	6	1	1	0	0	2	15	0	1	0	0	0	0	0	25.07898863	1	
9	1034462	1	2	0	0	0	2	0	0	1	0	0	0	0	0	0.00000000	1	
10	1034878	3	1	0	0	1	5	5	0	0	0	0	1	0	0	32.47055020	1	
11	1035198	98	3	3	2	3	2	7	0	1	0	0	0	0	0	127.10922886	0	
12	1035274	98	2	1	0	2	2	12	0	1	0	0	0	0	0	86.63815440	1	
13	1035364	6	4	1	1	3	3	17	0	0	1	0	0	0	0	82.44667657	0	
14	1035438	2	1	0	0	0	6	0	0	0	0	0	0	1	0	0.00000000	1	
15	1036062	5	1	0	0	1	6	2	0	0	0	0	0	1	0	12.18779872	1	
16	1036222	2	2	0	0	1	2	6	0	1	0	0	0	0	0	16.06337918	0	
17	1037295	4	2	2	0	2	3	6	0	0	1	0	0	0	0	37.63084904	0	
18	1037751	99	1	0	0	1	5	3	0	0	0	0	1	0	0	4.38057501	0	
19	1037895	2	1	0	0	1	2	2	0	1	0	0	0	0	0	97.95799516	0	
20	1037952	1	2	0	2	0	4	2	0	0	0	1	0	0	0	1.13967036	0	
21	1038024	2	3	0	1	3	4	2	0	0	0	1	0	0	0	89.20640363	0	
22	1038232	98	4	3	0	3	2	5	0	1	0	0	0	0	0	9.63514246	0	
23	1038404	8	2	1	0	2	7	2	0	0	0	0	0	0	1	0.70921789	0	
24	1038428	2	4	2	1	0	7	29	0	0	0	0	0	0	1	24.24178062	1	
25	1039341	7	3	2	0	3	4	6	0	0	0	1	0	0	0	95.07067082	0	
26	1039423	1	6	1	3	2	7	42	0	0	0	0	0	0	1	78.53955301	1	

Showing 1 to 28 of 42,431 entries

Pane 1 after we run "view" shows the data.frame we created from the SmallHHfile.csv

# Provide a summary of the descriptive statistics of all the variables
summary(SmallHHfile)



The summary gives you descriptive statistics for all the variables in the data file

Descriptive statistics from the psych package “describe” function

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
SAMPN	1	42431	2588378.63	1641345.14	1971814.00	2195483.36	847148.74	1031985	7212388.00	6180403.00	2.04	3.09	7968.16
INCOM	2	42431	13.18	26.29	5.00	5.51	2.97	1	99.00	98.00	2.92	6.62	0.13
HHSIZ	3	42431	2.57	1.37	2.00	2.40	1.48	1	8.00	7.00	1.03	0.90	0.01
HHEMP	4	42431	1.22	0.88	1.00	1.18	1.48	0	6.00	6.00	0.47	0.33	0.00
HHSTU	5	42431	0.64	1.02	0.00	0.44	0.00	0	8.00	8.00	1.66	2.52	0.00
HHLIC	6	42431	1.86	0.85	2.00	1.81	0.00	0	8.00	8.00	0.60	1.70	0.00
DOW	7	42431	4.02	1.99	4.00	4.02	2.97	1	7.00	6.00	0.00	-1.24	0.01
HTRIPS	8	42431	8.29	7.78	6.00	7.14	5.93	0	99.00	99.00	1.72	4.88	0.04
Mon	9	42431	0.14	0.34	0.00	0.05	0.00	0	1.00	1.00	2.12	2.49	0.00
Tue	10	42431	0.14	0.35	0.00	0.06	0.00	0	1.00	1.00	2.03	2.10	0.00
Wed	11	42431	0.14	0.35	0.00	0.06	0.00	0	1.00	1.00	2.02	2.08	0.00
Thu	12	42431	0.15	0.35	0.00	0.06	0.00	0	1.00	1.00	1.99	1.96	0.00
Fri	13	42431	0.14	0.35	0.00	0.05	0.00	0	1.00	1.00	2.08	2.33	0.00
Sat	14	42431	0.14	0.35	0.00	0.05	0.00	0	1.00	1.00	2.06	2.26	0.00
Sun	15	42431	0.15	0.35	0.00	0.06	0.00	0	1.00	1.00	1.99	1.98	0.00
TotDist	16	42431	68.09	118.52	33.89	45.44	45.13	0	5838.26	5838.26	8.38	196.69	0.58
center	17	42431	0.28	0.45	0.00	0.23	0.00	0	1.00	1.00	0.97	-1.05	0.00
suburb	18	42431	0.29	0.45	0.00	0.23	0.00	0	1.00	1.00	0.94	-1.12	0.00
exurb	19	42431	0.23	0.42	0.00	0.16	0.00	0	1.00	1.00	1.29	-0.34	0.00
rural	20	42431	0.20	0.40	0.00	0.13	0.00	0	1.00	1.00	1.49	0.21	0.00
other	21	42431	0.00	0.00	0.00	0.00	0.00	0	0.00	0.00	NaN	NaN	0.00
highinc	22	42431	0.41	0.49	0.00	0.39	0.00	0	1.00	1.00	0.35	-1.88	0.00
HHVEH	23	42431	1.86	1.00	2.00	1.81	1.48	0	8.00	8.00	0.80	2.26	0.00
HHBIC	24	42431	1.58	3.79	1.00	1.20	1.48	0	99.00	99.00	20.40	513.75	0.02
VEHNEW	25	42431	2.15	2.02	2.00	1.57	1.48	1	9.00	8.00	2.38	4.20	0.01
OWN	26	42431	1.24	0.56	1.00	1.16	0.00	1	9.00	8.00	5.96	67.49	0.00
CarBuy	27	42431	0.45	0.50	0.00	0.44	0.00	0	1.00	1.00	0.19	-1.96	0.00
snglhm	28	42431	0.82	0.39	1.00	0.90	0.00	0	1.00	1.00	-1.65	0.71	0.00
ownhm	29	42431	0.77	0.42	1.00	0.84	0.00	0	1.00	1.00	-1.31	-0.29	0.00
MilesPr	30	42431	27.12	43.46	14.50	18.40	18.19	0	1167.65	1167.65	5.15	47.24	0.21
TrpPrs	31	42431	3.28	2.58	3.00	3.02	2.22	0	32.00	32.00	1.27	3.68	0.01

```
install.packages("psych")
library(psych)
describe(SmallHHfile)
```

R GRAPHICS

ggplot2

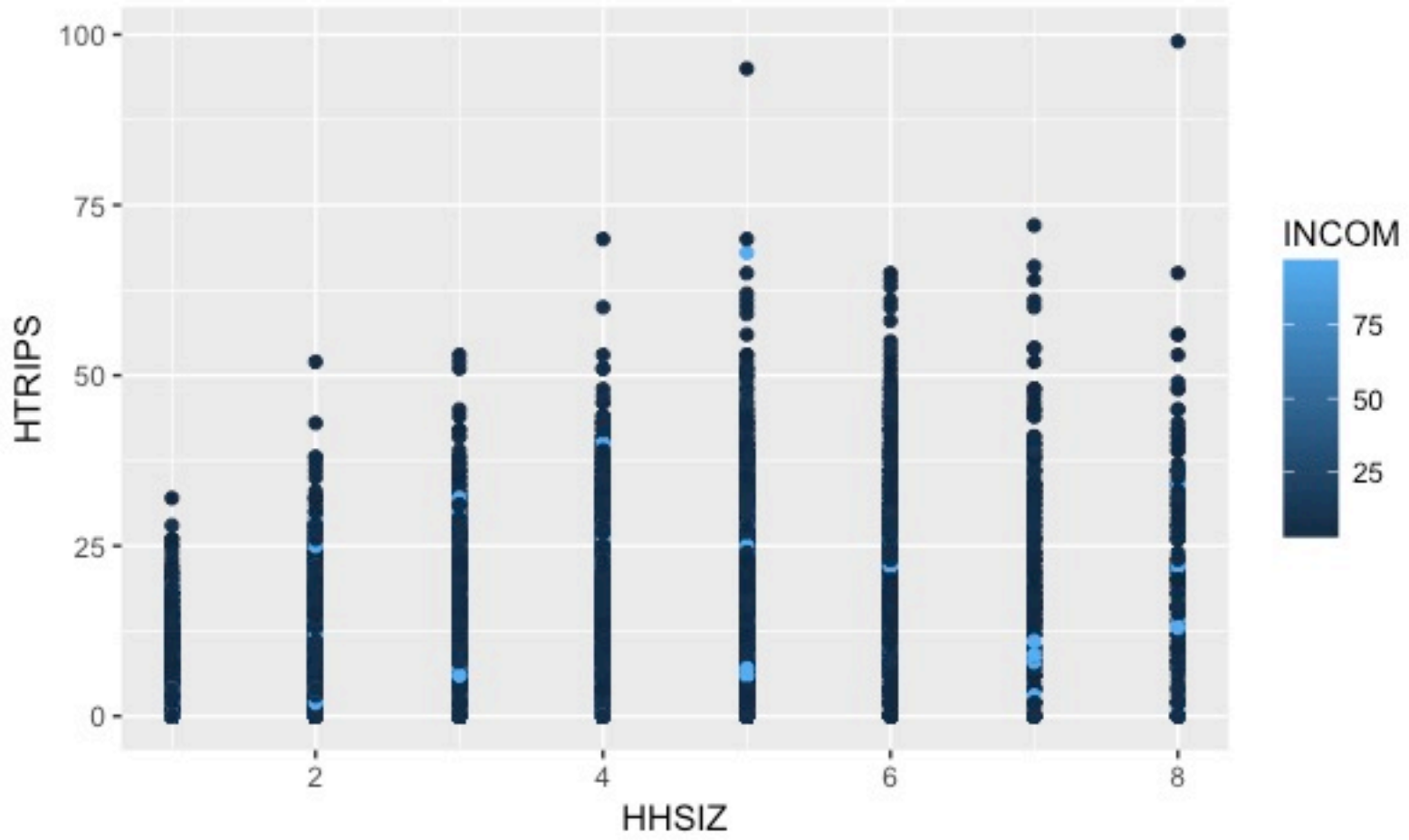
- ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details

- Example follows

Ggplot example

- ggplot(SmallHHfile, aes(HHSIZ, HTRIPS, colour = INCOM)) + geom_point()

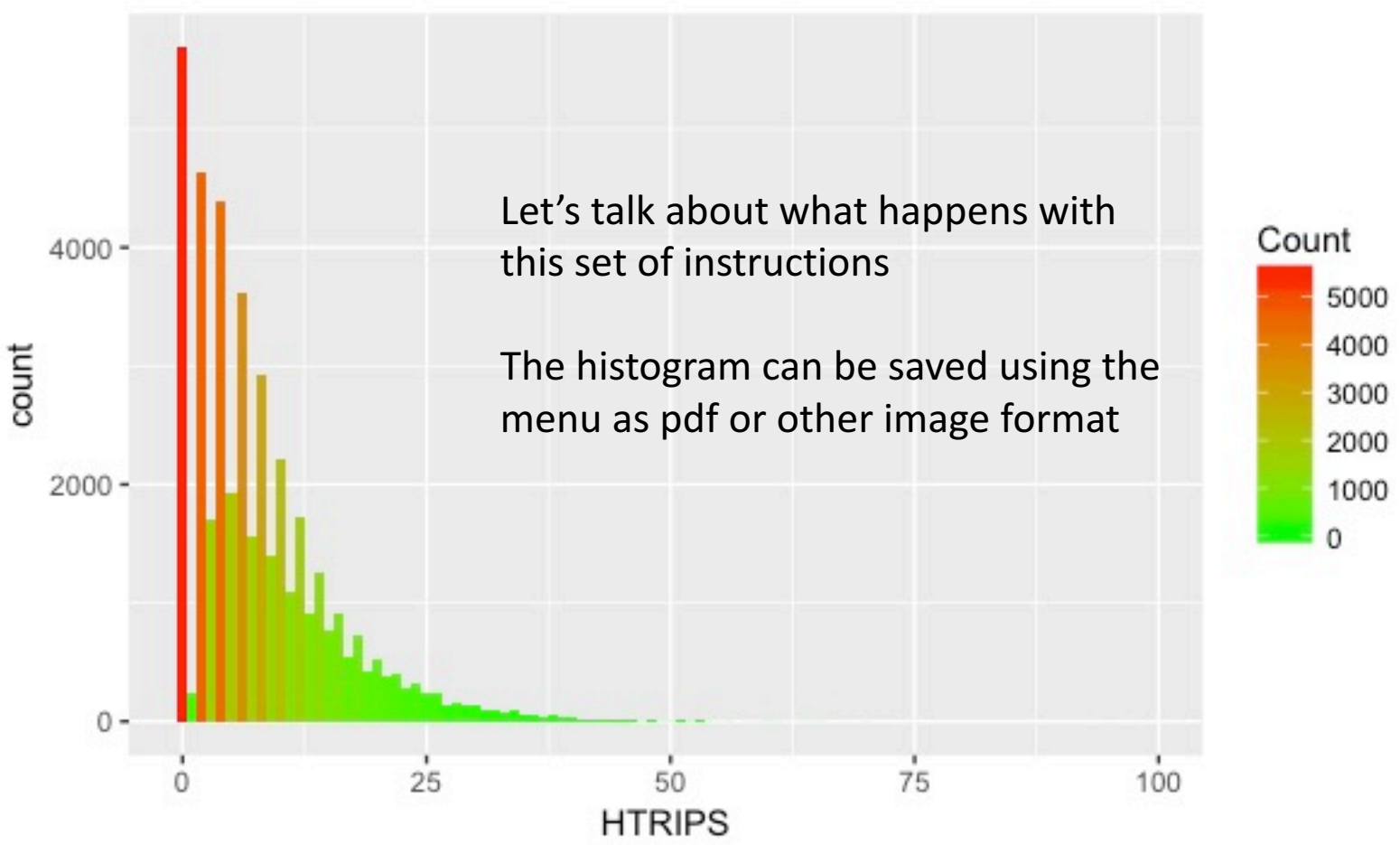
- From the data SmallHHfile
- Use aes (aesthetics) with HHSIZ as x and HTRIPS as y and plot points in a graph that changes colors based on the values of variable INCOM
- After the graph is created I use export in pane 4 and save it to a file that is JPEG



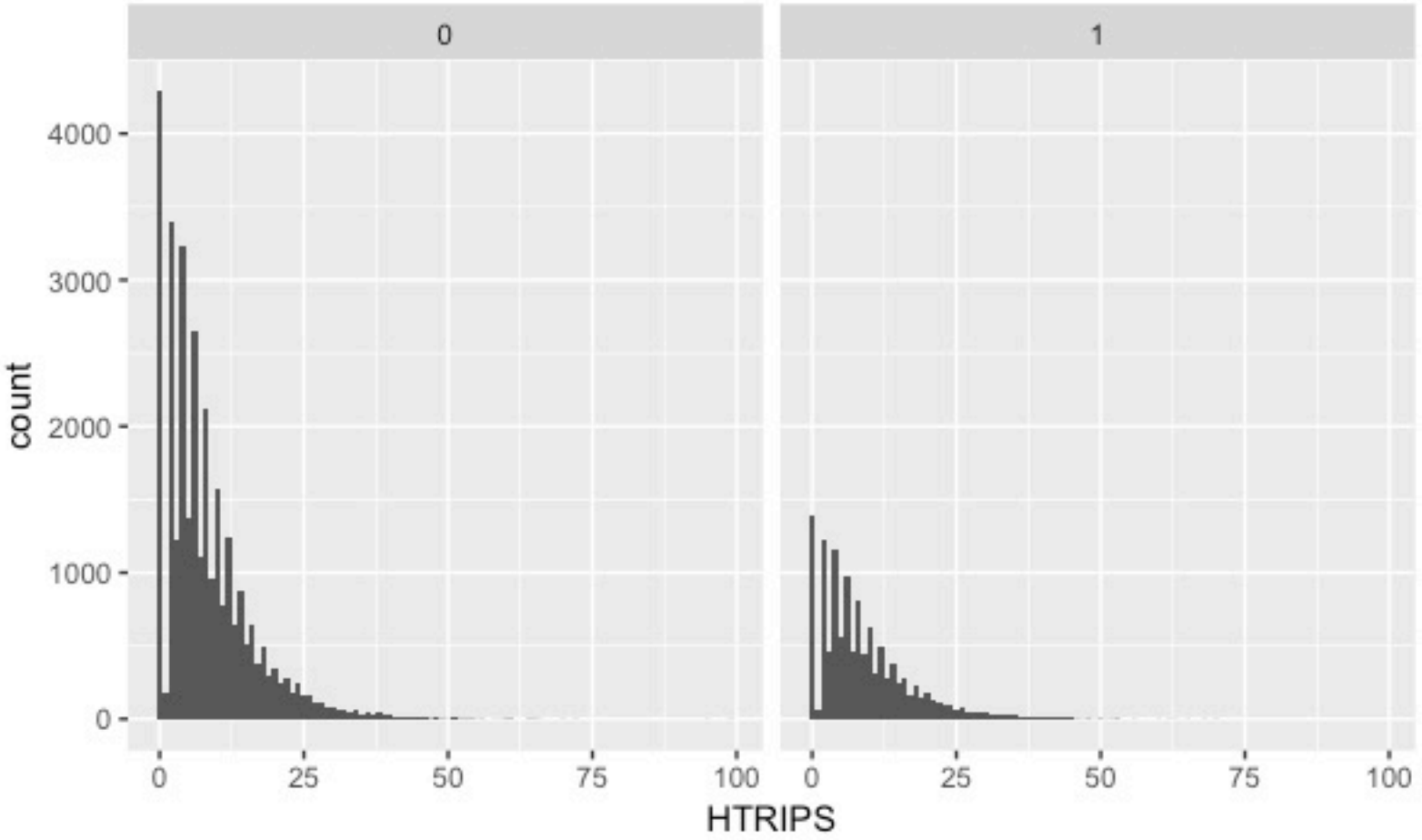
Can you tell the story?

HISTOGRAMS

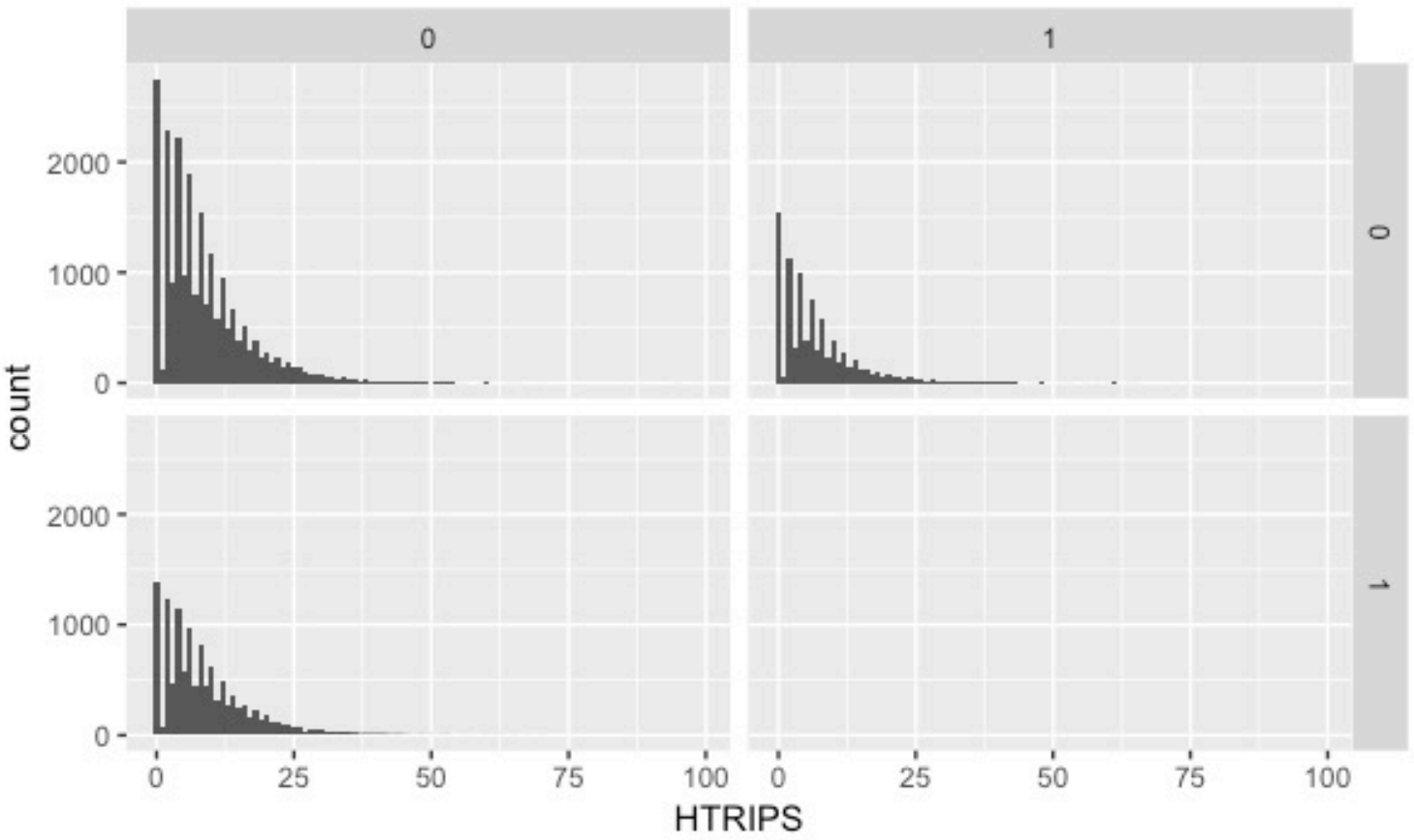
```
m<-ggplot(SmallHHfile, aes(HTRIPS))
m + geom_histogram(binwidth = 1)
m + geom_histogram(aes(fill = ..count..), binwidth = 1) + scale_fill_gradient("Count", low = "green", high = "red")
```



```
m <- m + geom_histogram(binwidth = 1)
m + facet_grid(. ~ center)
```



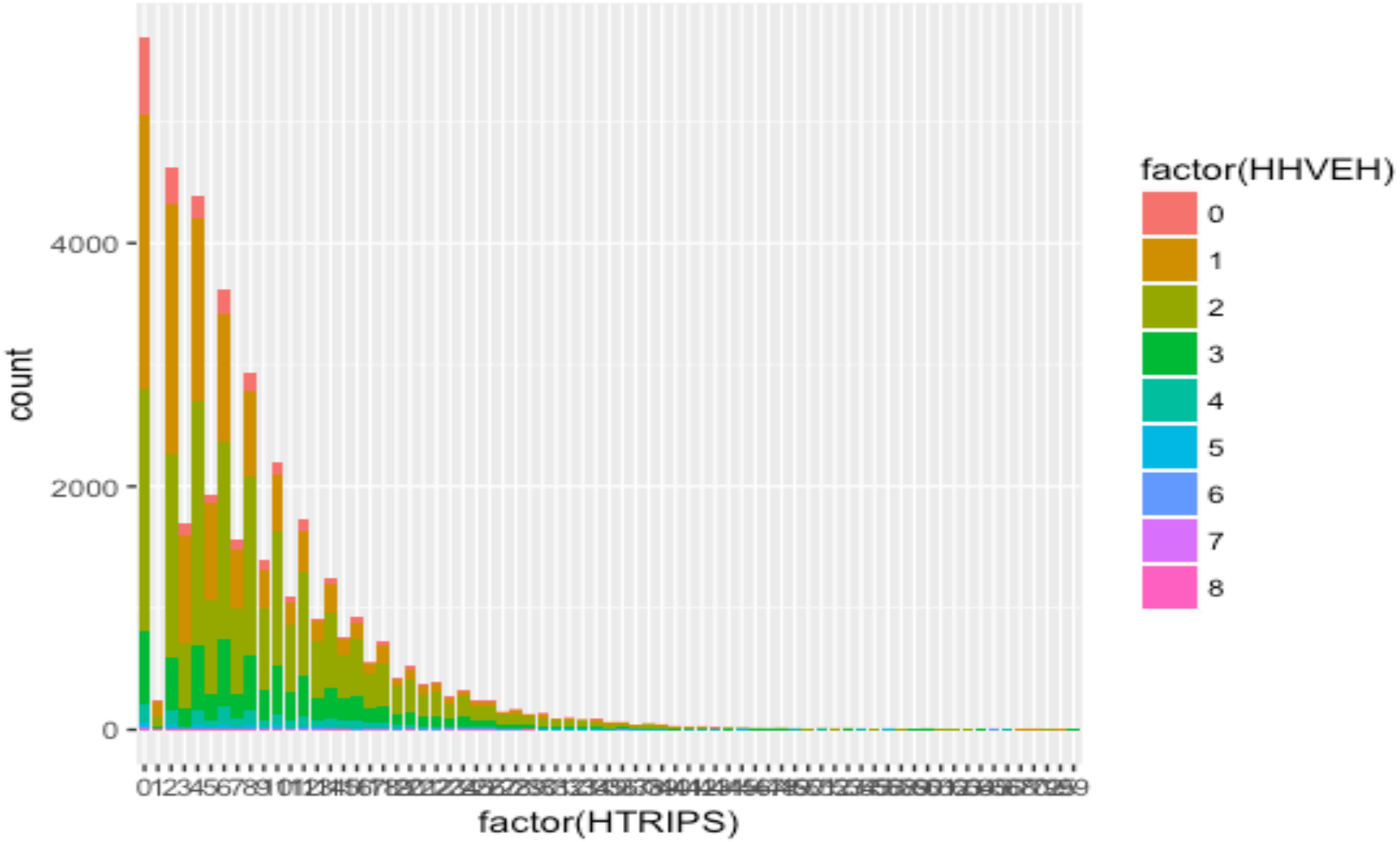
```
# Use facets
m <- m + geom_histogram(binwidth = 1)
m + facet_grid(center ~ rural)
```



Histograms of HTRIPS for combinations of values between center & rural (residence of household)

More options in: http://www.cookbook-r.com/Graphs/Facets_(ggplot2)/

# colored categories
k <- ggplot(SmallHHfile, aes(factor(HTRIPS), fill = factor(HHVEH)))
k + geom_bar()



Box Plot

- A box plot=a basic graphing tool that displays centering, spread, and distribution of a continuous data set (a variable)
- Provides a 5 point summary of the data :
  - The box represents the middle 50% of the data.
  - The median is the point where 50% of the data is above it and 50% below it.
  - The 25th quartile is where, at most, 25% of the data fall below it.
  - The 75th quartile is where, at most, 25% of the data is above it.
  - The whiskers cannot extend any further than 1.5 times the length of the inner quartiles.
  - Data points outside this will show up as outliers.

Example from SPSS

Provides a 5 point summary of the data :

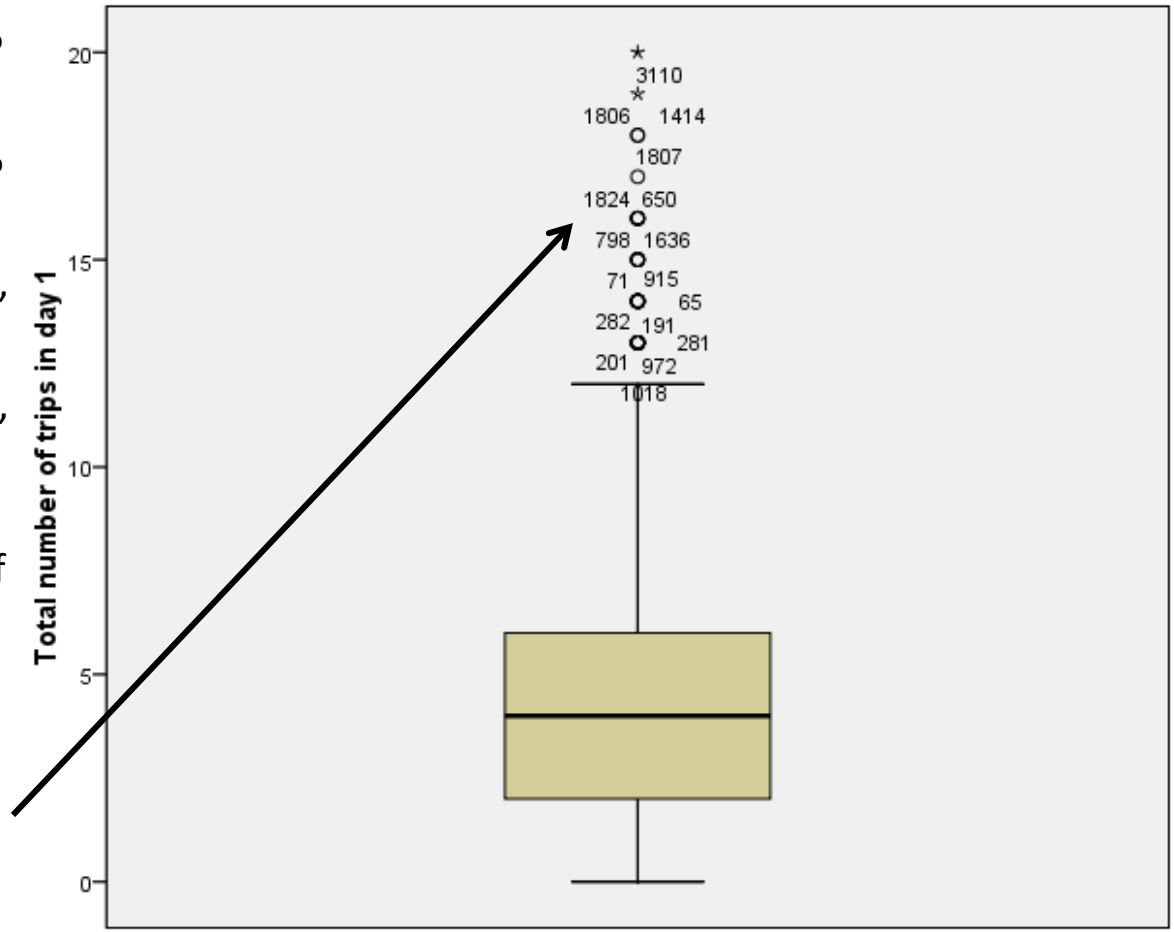
The box represents the middle 50% of the data.

The median is the point where 50% of the data is above it and 50% below it.
The 25th quartile is where, at most, 25% of the data fall below it.

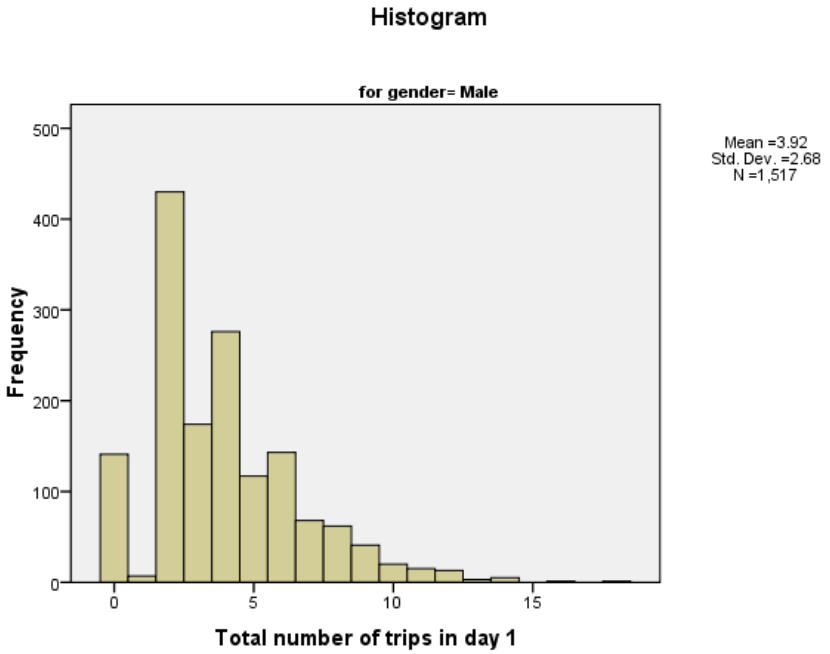
The 75th quartile is where, at most, 25% of the data is above it.

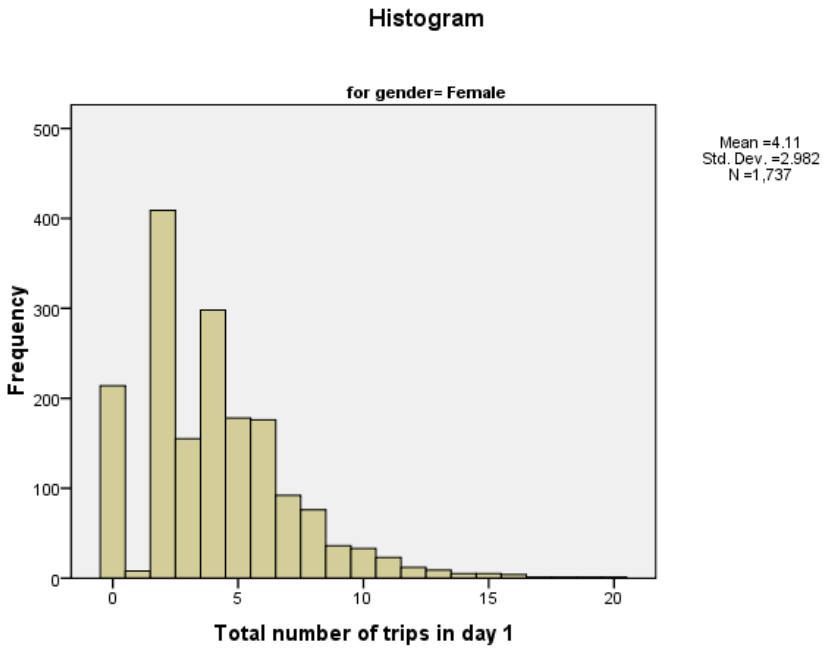
The whiskers cannot extend any further than 1.5 times the length of the inner quartiles.

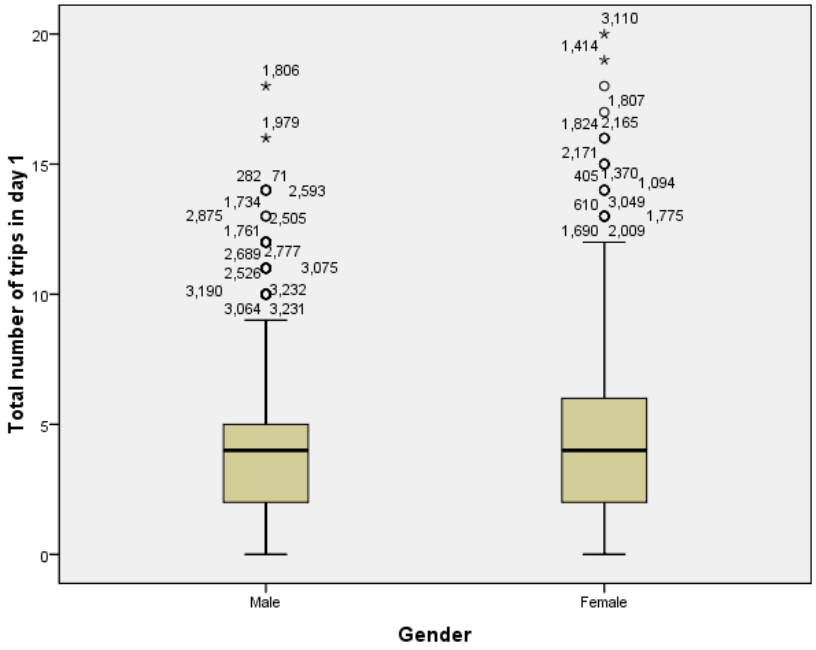
Data points outside this range of values will show up as outliers.



"extreme outliers are * at 3 times the height of the darker box

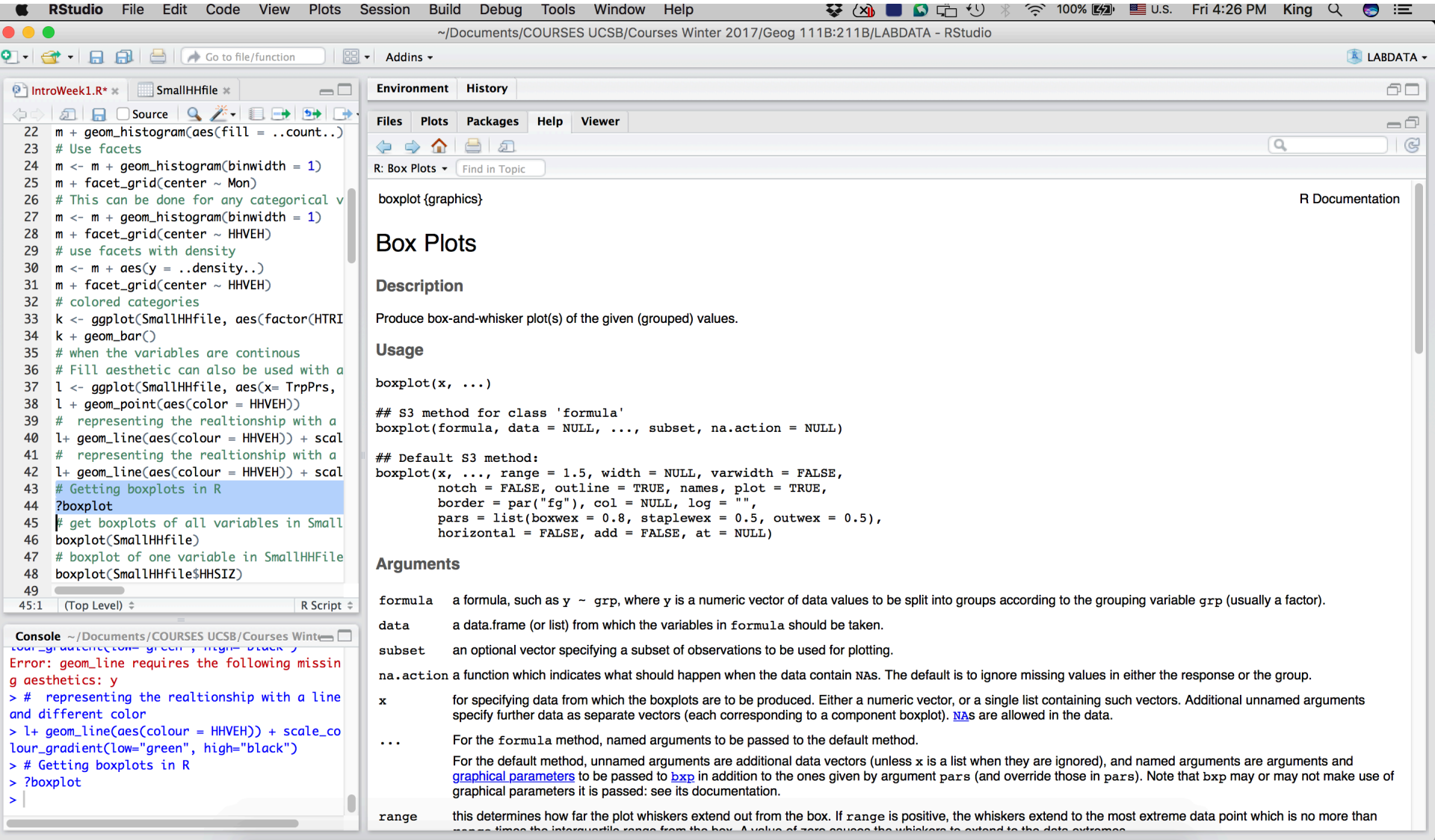

Histogram
for gender= Male
Mean =3.92
Std. Dev. =2.68
N =1,517
Frequency
Total number of trips in day 1


Histogram
for gender= Female
Mean =4.11
Std. Dev. =2.982
N =1,737
Frequency
Total number of trips in day 1


Total number of trips in day 1
Male
Female
Gender

Getting boxplots in R
?boxplot

This is one way of getting help about an instruction/function

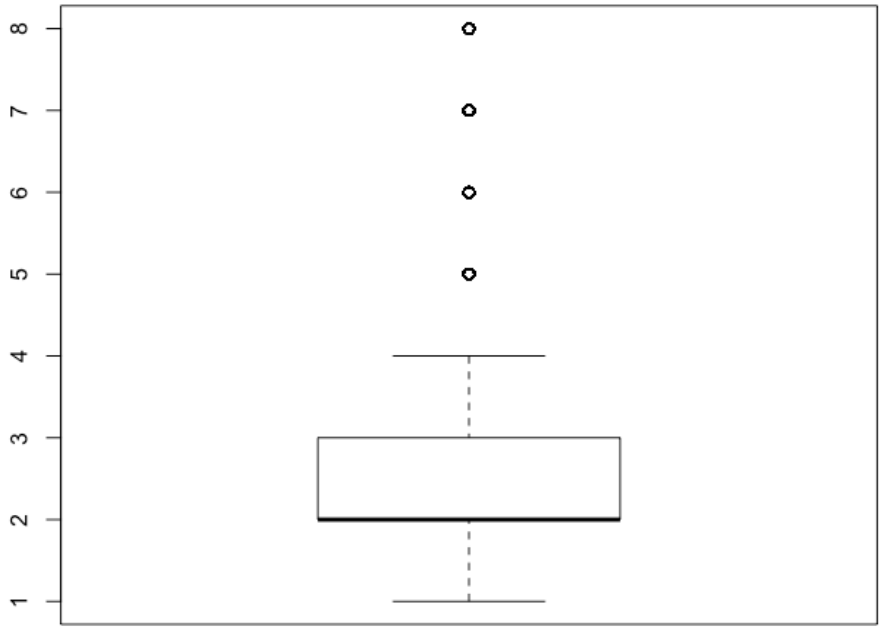


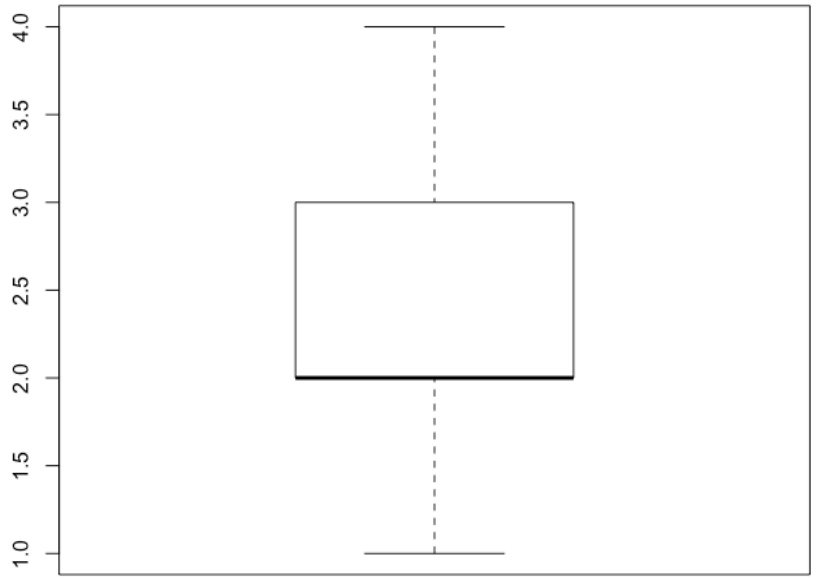
```
# get boxplots of all variables in SmallHHfile
boxplot(SmallHHfile)
```

We can do this for an entire database but we get something too big to visualize

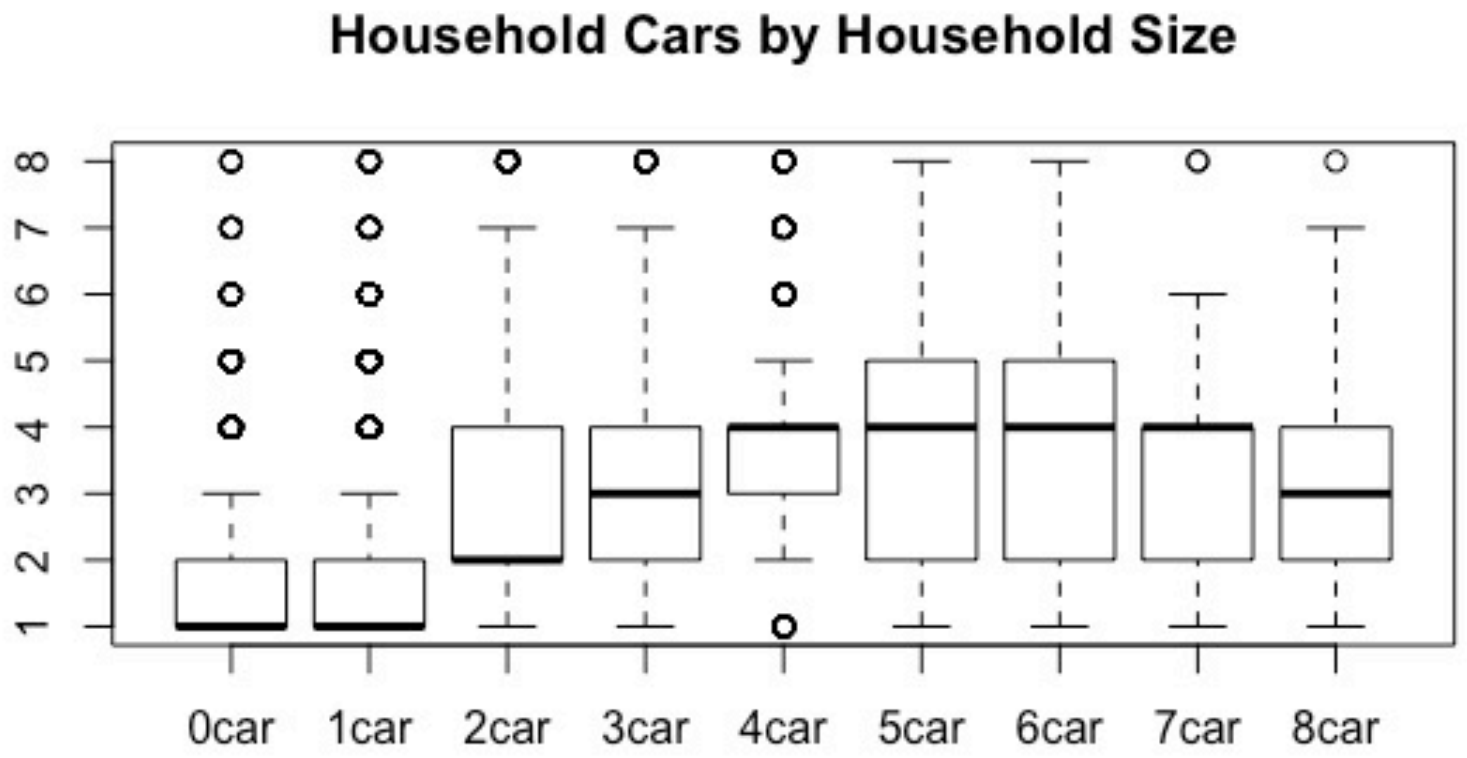
```
# get boxplots of all variables in SmallHHfile
boxplot(SmallHHfile)
# boxplot of one variable in SmallHHFile NOTE the $
boxplot(SmallHHfile$HHSIZ)
# switching off the drawing of outliers
boxplot(SmallHHfile$HHSIZ, outline = FALSE)
```

Outline=false: means do not display outliers





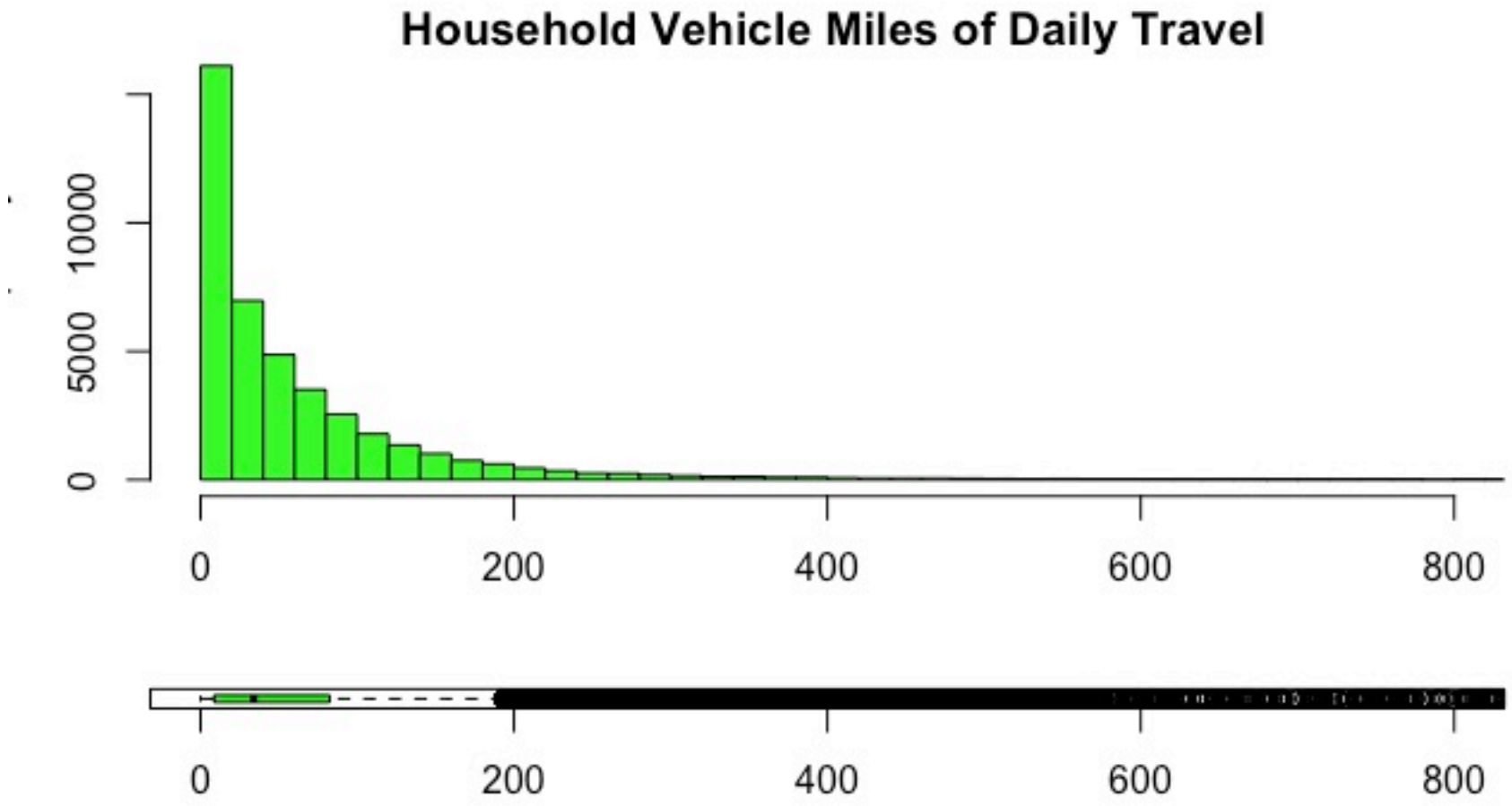
```
boxplot(SmallHHfile$HHSIZ ~ SmallHHfile$HHVEH, outline = TRUE,
main="Household Cars by Household Size", names =
c("0car","1car","2car","3car","4car","5car","6car","7car","8car"))
```

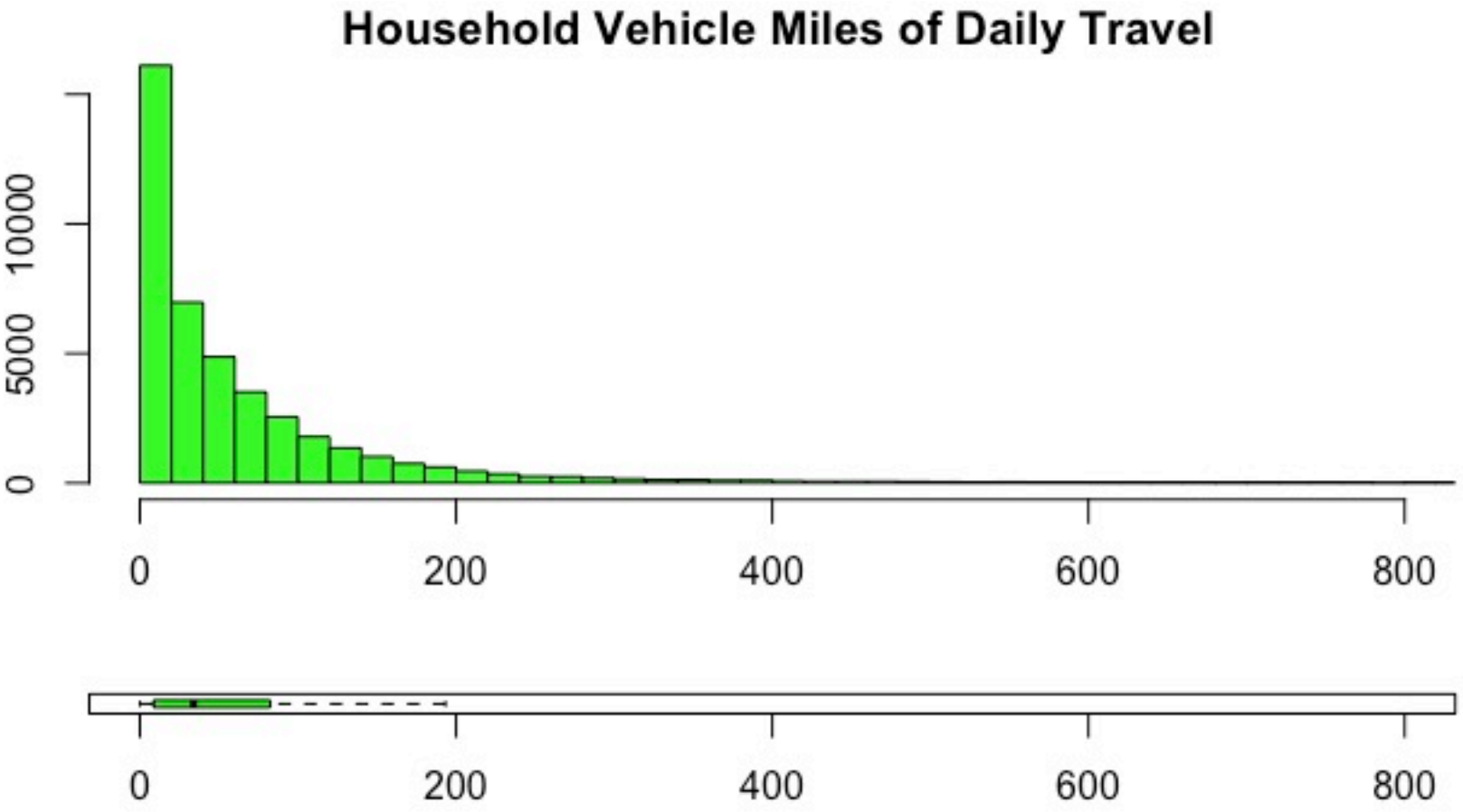


What is the story?

COMBINE HISTOGRAM & BOXPLOT IN SAME GRAPH

```
nf <- layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height =
c(3,1))

par(mar=c(3.1, 3.1, 1.1, 2.1))

hist(SmallHHfile$TotDist, col=rgb(0,1,0,0.9),breaks=400,
xlim=c(0,800),      ylim=c(0,16000), xlab="Number of Travel
Miles", main=" Household Vehicle Miles of Daily Travel")

boxplot(SmallHHfile$TotDist, horizontal=TRUE,  outline=TRUE,
ylim=c(0,800), frame=F, col = "green1")

box()
```


Household Vehicle Miles of Daily Travel
0
5000
10000
0
200
400
600
800
0
200
400
600
800



I just turned off outliers in boxplot
With outline=FALSE

Correlations & Covariances

What happens when we consider two variables at the same time and their relationship?

Measures of Association

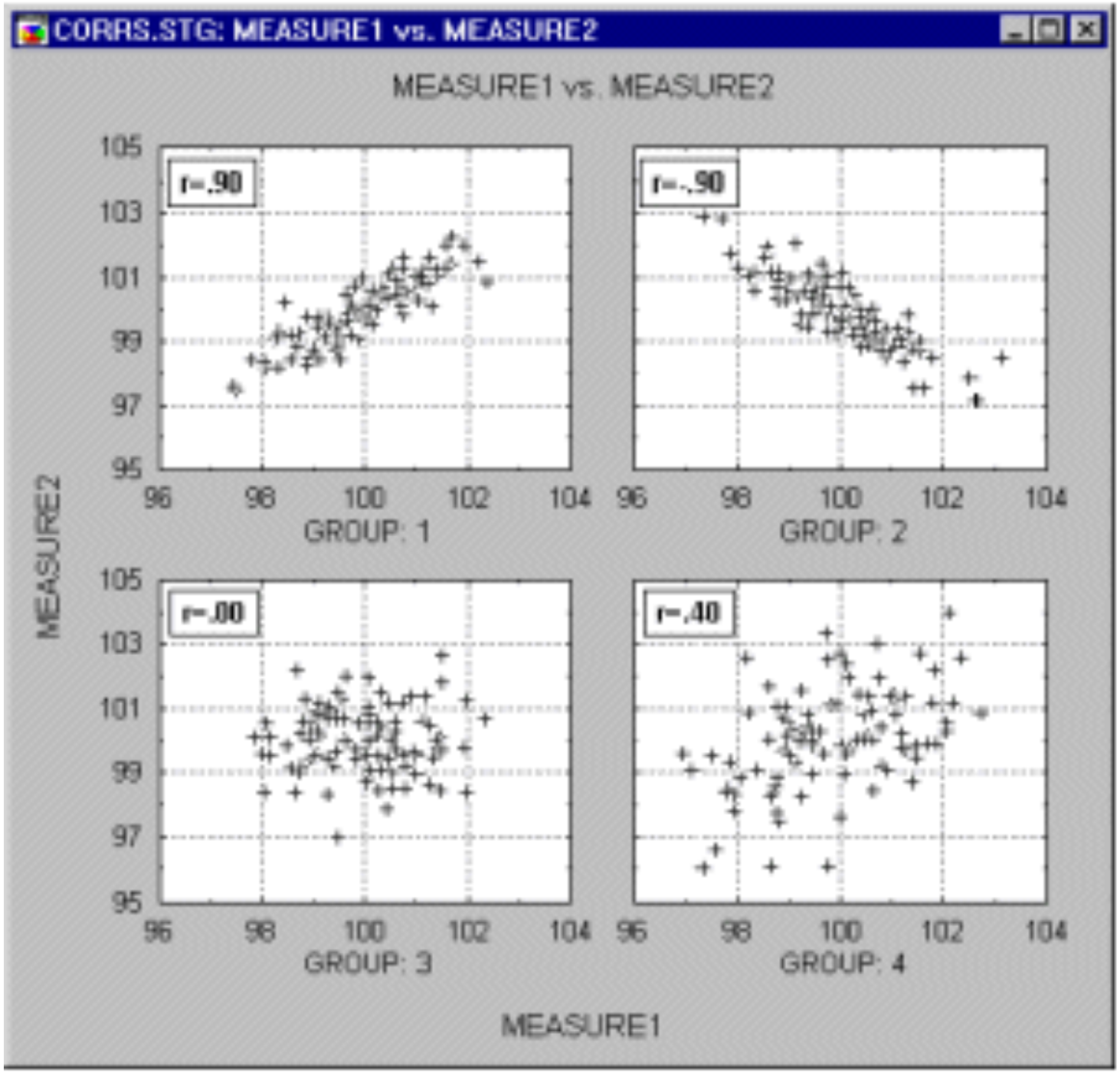
Covariance

$$s_{xy} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Measures of Association

Standardized Covariance = Correlation (Pearson)

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$


CORRS.STG: MEASURE1 vs. MEASURE2
MEASURE1 vs. MEASURE2
r=.90
r=-.90
r=.00
r=.40
105
103
101
99
97
95
96
98
100
102
104
GROUP: 1
GROUP: 2
GROUP: 3
GROUP: 4
MEASURE2
MEASURE1

corstats<-cor(SmallHHfile)
round(corstats, 2)

	SAMPN	INCOM	HHSIZ	HHEMP	HHSTU	HHLIC	DOW	HTRIPS	Mon	Tue	Wed	Thu	Fri	Sat	Sun	TotDist	center	suburb	exurb	
SAMPN	1.00	0.04	0.08	0.05	0.06	0.05	0.07	0.04	-0.01	-0.03	-0.07	-0.03	0.02	0.08	0.03	0.00	0.07	0.02	-0.05	
INCOM	0.04	1.00	-0.02	0.00	-0.05	0.03	0.01	-0.03	-0.01	0.00	-0.01	0.00	0.01	0.00	0.01	-0.01	0.01	0.02	-0.01	
HHSIZ	0.08	-0.02	1.00	0.46	0.79	0.58	0.00	0.54	0.00	0.00	-0.01	-0.01	0.01	0.00	0.00	0.25	-0.06	0.04	0.03	
HHEMP	0.05	0.00	0.46	1.00	0.29	0.55	0.00	0.31	-0.01	0.00	0.01	0.01	-0.01	0.00	0.00	0.20	0.01	0.03	0.01	
HHSTU	0.06	-0.05	0.79	0.29	1.00	0.28	0.00	0.50	0.01	0.00	-0.01	0.00	0.00	0.00	-0.01	0.20	-0.02	0.03	0.01	
HHLIC	0.05	0.03	0.58	0.55	0.28	1.00	0.00	0.28	0.00	0.00	0.01	-0.01	0.00	0.00	0.00	0.22	-0.11	0.05	0.05	
DOW	0.07	0.01	0.00	0.00	0.00	0.00	1.00	-0.07	-0.60	-0.42	-0.21	0.00	0.20	0.40	0.62	0.02	-0.01	-0.01	-0.01	
HTRIPS	0.04	-0.03	0.54	0.31	0.50	0.28	-0.07	1.00	-0.02	0.05	0.04	0.04	0.03	-0.04	-0.09	0.33	0.04	0.04	0.00	
Mon	-0.01	-0.01	0.00	-0.01	0.01	0.00	-0.60	-0.02	1.00	-0.16	-0.16	-0.17	-0.16	-0.16	-0.16	-0.02	-0.02	-0.01	0.01	
Tue	-0.03	0.00	0.00	0.00	0.00	0.00	-0.42	0.05	-0.16	1.00	-0.17	-0.17	-0.17	-0.17	-0.17	-0.01	0.02	0.01	0.00	
Wed	-0.07	-0.01	-0.01	0.01	-0.01	0.01	-0.21	0.04	-0.16	-0.17	1.00	-0.17	-0.17	-0.17	-0.17	0.00	0.02	0.00	0.00	
Thu	-0.03	0.00	-0.01	0.01	0.00	-0.01	0.00	0.04	-0.17	-0.17	-0.17	1.00	-0.17	-0.17	-0.17	-0.01	0.02	0.01	0.00	
Fri	0.02	0.01	0.01	-0.01	0.00	0.00	0.20	0.03	-0.16	-0.17	-0.17	-0.17	1.00	-0.16	-0.17	0.02	-0.01	0.00	0.00	
Sat	0.08	0.00	0.00	0.00	0.00	0.00	0.40	-0.04	-0.16	-0.17	-0.17	-0.17	-0.16	1.00	-0.17	0.02	-0.01	-0.01	-0.01	
Sun	0.03	0.01	0.00	0.00	-0.01	0.00	0.62	-0.09	-0.16	-0.17	-0.17	-0.17	-0.17	-0.17	1.00	0.00	-0.01	0.00	0.01	
TotDist	0.00	-0.01	0.25	0.20	0.20	0.22	0.02	0.33	-0.02	-0.01	0.00	-0.01	0.02	0.02	0.00	1.00	-0.08	0.00	0.04	
center	0.07	0.01	-0.06	0.01	-0.02	-0.11	-0.01	0.04	-0.02	0.02	0.02	0.02	-0.01	-0.01	-0.01	-0.08	1.00	-0.40	-0.34	
suburb	0.02	0.02	0.04	0.03	0.03	0.05	-0.01	0.04	-0.01	0.01	0.00	0.01	0.00	-0.01	0.00	0.00	-0.40	1.00	-0.35	
exurb	-0.05	-0.01	0.03	0.01	0.01	0.05	-0.01	0.00	0.01	0.00	0.00	0.00	0.00	-0.01	0.01	0.04	-0.34	-0.35	1.00	
rural	-0.04	-0.03	-0.01	-0.05	-0.03	0.02	0.02	-0.08	0.02	-0.03	-0.02	-0.03	0.02	0.03	0.01	0.04	-0.31	-0.32	-0.27	
other	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	

Values do not depend on the scale of variables

DIAGONAL CONTAINS THE CORRELATION OF A VARIABLE WITH ITSELF = 1
THIS IS A SYMMETRIC MATRIX (check)

# pairwise covariances among all variables
covstats<-cov(SmallHHfile)
round(covstats, 3)

Values depend on the scale of the variables

Diagonal is the variance of a variable

We want to study the influence of many variables on one variable = regression model

	HHBIC	VEHNEW	OWN	CarBuy	snglhm	ownhm	MilesPr	TrpPrs
SAMPN	79247.178	13878.628	-17408.611	7209.552	31566.921	18469.356	-1715936.131	-46839.887
INCOM	0.848	2.384	-0.248	-0.294	0.395	0.616	6.138	-1.331
HHSIZ	0.987	-0.130	-0.005	0.085	0.089	0.008	-1.652	-0.143
HHEMP	0.425	-0.101	-0.036	0.089	0.049	0.033	3.046	0.151
HHSTU	0.718	-0.099	0.033	0.058	0.026	-0.030	-0.949	0.109
HHLIC	0.388	-0.064	-0.103	0.072	0.100	0.101	1.974	-0.125
DOW	-0.011	-0.011	-0.005	-0.006	0.012	0.007	1.523	-0.433
HTRIPS	4.227	-0.545	0.045	0.519	0.201	-0.022	58.369	14.097
Mon	-0.003	0.003	0.001	-0.001	0.001	-0.001	-0.242	-0.019
Tue	0.005	0.001	0.000	0.001	-0.003	-0.001	-0.117	0.049
Wed	-0.002	-0.002	-0.001	0.001	-0.001	0.001	-0.045	0.041
Thu	0.000	-0.002	0.001	0.002	-0.001	-0.001	-0.123	0.044
Fri	0.002	0.002	0.000	-0.001	0.002	0.000	0.356	0.025
Sat	0.005	-0.004	-0.001	0.001	0.002	0.001	0.349	-0.041
Sun	-0.008	0.001	-0.001	-0.002	0.001	0.000	-0.179	-0.099
TotDist	33.215	-7.496	-3.844	6.045	4.608	3.764	4363.446	59.895
center	-0.039	0.017	0.041	-0.005	-0.044	-0.040	-1.328	0.133
suburb	0.018	-0.007	-0.010	0.008	0.010	0.010	-0.162	0.016
exurb	0.023	-0.011	-0.015	0.005	0.017	0.014	0.686	-0.031
rural	-0.001	0.001	-0.015	-0.009	0.017	0.016	0.809	-0.117
other	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
highinc	0.197	-0.074	-0.059	0.056	0.040	0.053	2.319	0.119
HHVEH	0.479	-0.073	-0.143	0.076	0.132	0.141	4.036	-0.144
HHBIC	14.332	-0.117	-0.095	0.107	0.130	0.087	3.575	0.353
VEHNEW	-0.117	4.066	0.029	-0.522	-0.021	-0.019	-1.798	0.002
OWN	-0.095	0.029	0.308	-0.017	-0.083	-0.189	-1.419	0.058
CarBuy	0.107	-0.522	-0.017	0.248	0.017	0.015	1.694	0.090
snglhm	0.130	-0.021	-0.083	0.017	0.149	0.081	1.058	-0.065
ownhm	0.087	-0.019	-0.189	0.015	0.081	0.175	1.346	-0.056
MilesPr	3.575	-1.798	-1.419	1.694	1.058	1.346	1888.746	28.655
TrpPrs	0.353	0.002	0.058	0.090	-0.065	-0.056	28.655	6.652

>

Let's look at a small subset using the library dplyr

```
library(dplyr)  # install the library to allow easy subsetting of variables
names(SmallHHfile) # check the names of the columns
NEWFRAME <-select(SmallHHfile, INCOM:HTRIPS) # select the columns from INCOM to HT
corstats<-cor(NEWFRAME) # compute correlations of these few variables
round(corstats, 2) # print correlation with 2 decimal places
# pairwise covariances among all variables
COVMAT = cov(SmallHHfile)
round(COVMAT, 3)
```

```
> round(corstats, 2) # print correlation with 2 decimal
        INCOM HHSIZ HHEMP HHSTU HHLIC   DOW HTRIPS
INCOM    1.00 -0.02  0.00 -0.05  0.03  0.01  -0.03
HHSIZ   -0.02  1.00  0.46  0.79  0.58  0.00   0.54
HHEMP    0.00  0.46  1.00  0.29  0.55  0.00   0.31
HHSTU   -0.05  0.79  0.29  1.00  0.28  0.00   0.50
HHLIC    0.03  0.58  0.55  0.28  1.00  0.00   0.28
DOW      0.01  0.00  0.00  0.00  0.00  1.00  -0.07
HTRIPS  -0.03  0.54  0.31  0.50  0.28 -0.07   1.00
```

```
> round(COVMAT, 3)
         INCOM  HHSIZ  HHEMP  HHSTU  HHLIC    DOW HTRIPS
INCOM  690.918 -0.814 -0.040 -1.354  0.560  0.593 -5.790
HHSIZ   -0.814  1.887  0.557  1.107  0.671  0.010  5.786
HHEMP   -0.040  0.557  0.780  0.262  0.410  0.006  2.126
HHSTU   -1.354  1.107  0.262  1.046  0.239 -0.010  4.002
HHLIC    0.560  0.671  0.410  0.239  0.719 -0.003  1.867
DOW      0.593  0.010  0.006 -0.010 -0.003  3.975 -1.080
HTRIPS  -5.790  5.786  2.126  4.002  1.867 -1.080 60.467
>
```

```
> round(corstats, 2) # print correlation with 2 decim
       INCOM HHSIZ HHEMP HHSTU HHLIC   DOW HTRIPS
INCOM   1.00 -0.02  0.00 -0.05  0.03  0.01  -0.03
HHSIZ  -0.02  1.00  0.46  0.79  0.58  0.00   0.54
HHEMP   0.00  0.46  1.00  0.29  0.55  0.00   0.31
HHSTU  -0.05  0.79  0.29  1.00  0.28  0.00   0.50
HHLIC   0.03  0.58  0.55  0.28  1.00  0.00   0.28
DOW     0.01  0.00  0.00  0.00  0.00  1.00  -0.07
HTRIPS -0.03  0.54  0.31  0.50  0.28 -0.07   1.00
```

This code selects the data from respondents on Mondays only

```
MONDAY <-filter(SmallHHfile, Mon == 1)
NEWMON <-select(MONDAY, INCOM:HTRIPS)
corstats<-cor(NEWMON)
round(corstats, 2)
covstats<-cov(NEWMON)
round(covstats, 2)
```

```
> round(corstats, 2)
        INCOM HHSIZ HHEMP HHSTU HHLIC DOW HTRIPS
INCOM    1.00 -0.01  0.02 -0.05  0.05  NA  -0.03
HHSIZ   -0.01  1.00  0.45  0.79  0.56  NA   0.55
HHEMP    0.02  0.45  1.00  0.28  0.53  NA   0.29
HHSTU   -0.05  0.79  0.28  1.00  0.27  NA   0.51
HHLIC    0.05  0.56  0.53  0.27  1.00  NA   0.28
DOW        NA    NA    NA    NA    NA   1     NA
HTRIPS  -0.03  0.55  0.29  0.51  0.28  NA   1.00
>
```

```
> round(corstats, 2) # print correlation with 2 decimal
        INCOM HHSIZ HHEMP HHSTU HHLIC   DOW HTRIPS
INCOM    1.00 -0.02  0.00 -0.05  0.03  0.01  -0.03
HHSIZ   -0.02  1.00  0.46  0.79  0.58  0.00   0.54
HHEMP    0.00  0.46  1.00  0.29  0.55  0.00   0.31
HHSTU   -0.05  0.79  0.29  1.00  0.28  0.00   0.50
HHLIC    0.03  0.58  0.55  0.28  1.00  0.00   0.28
DOW      0.01  0.00  0.00  0.00  0.00  1.00  -0.07
HTRIPS  -0.03  0.54  0.31  0.50  0.28 -0.07   1.00
> round(corstats, 2)
        INCOM HHSIZ HHEMP HHSTU HHLIC DOW HTRIPS
INCOM    1.00 -0.01  0.02 -0.05  0.05  NA  -0.03
HHSIZ   -0.01  1.00  0.45  0.79  0.56  NA   0.55
HHEMP    0.02  0.45  1.00  0.28  0.53  NA   0.29
HHSTU   -0.05  0.79  0.28  1.00  0.27  NA   0.51
HHLIC    0.05  0.56  0.53  0.27  1.00  NA   0.28
DOW        NA    NA    NA    NA    NA   1     NA
HTRIPS  -0.03  0.55  0.29  0.51  0.28  NA   1.00
>
```