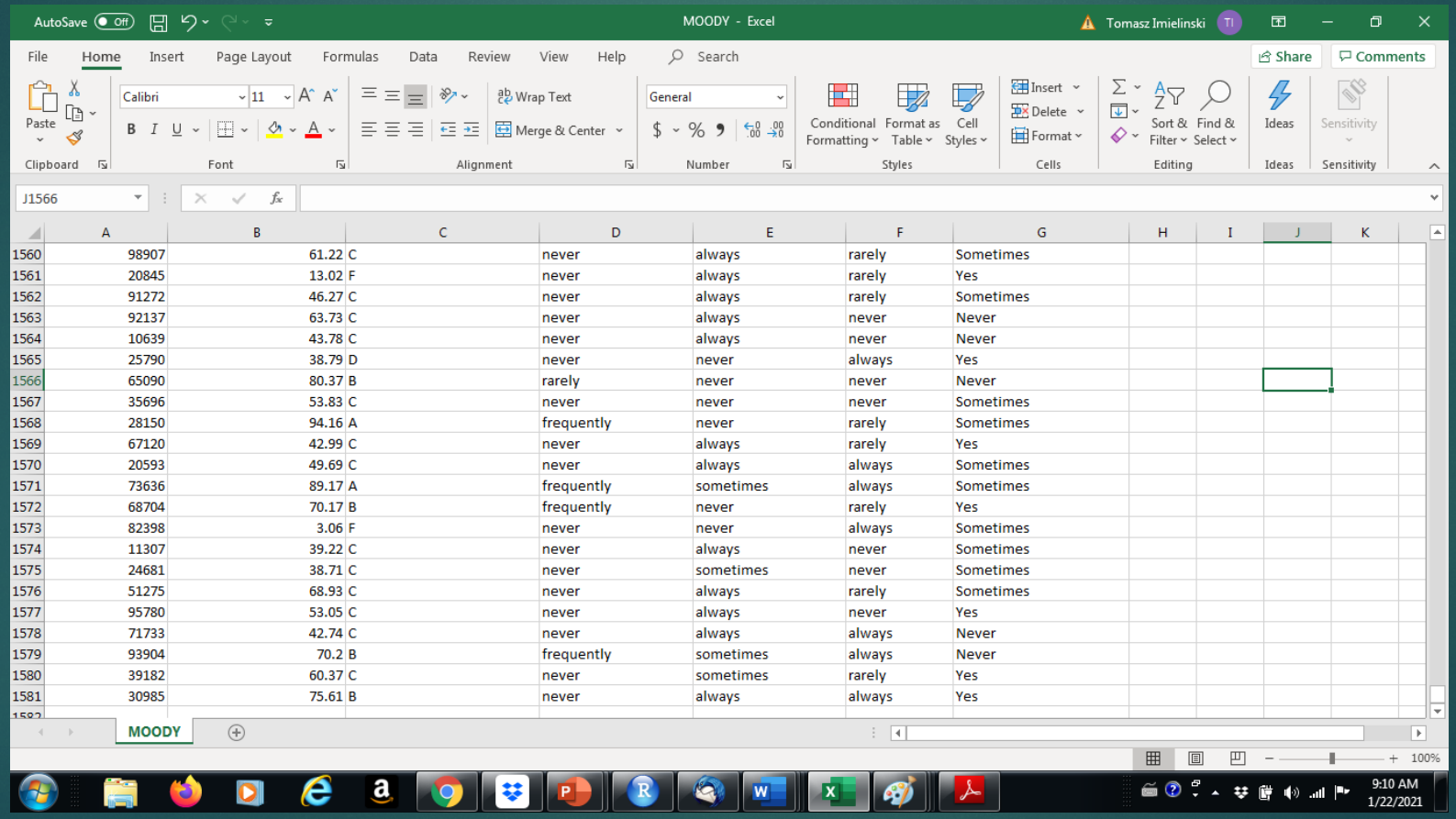
Plots

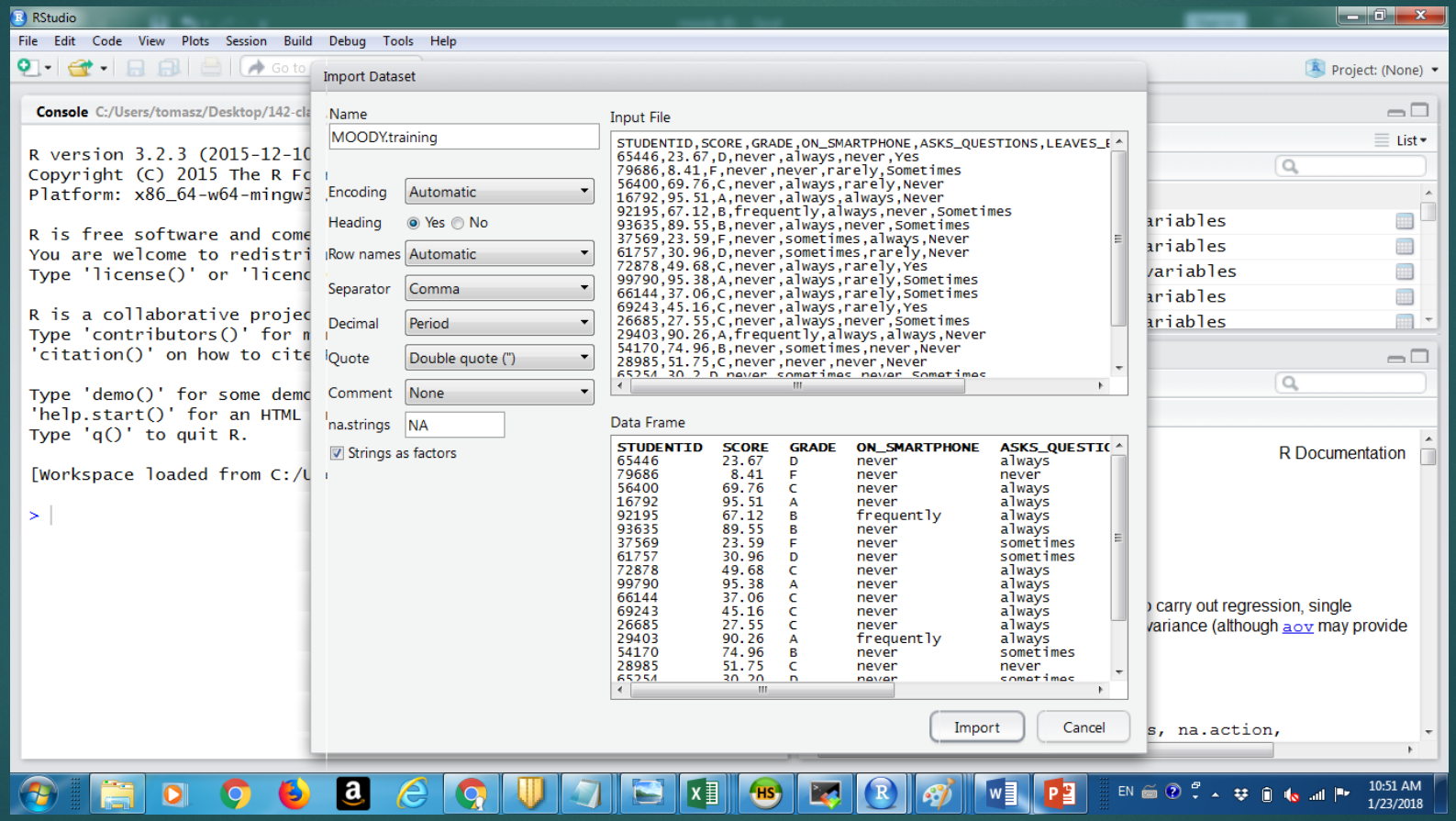
DATA101.CS.RUTGERS.EDU/LABORATORY

Synthetic Data Sets



	A	B	C	D	E	F	G
1560	98907	61.22	C	never	always	rarely	Sometimes
1561	20845	13.02	F	never	always	rarely	Yes
1562	91272	46.27	C	never	always	rarely	Sometimes
1563	92137	63.73	C	never	always	never	Never
1564	10639	43.78	C	never	always	never	Never
1565	25790	38.79	D	never	never	always	Yes
1566	65090	80.37	B	rarely	never	never	Never
1567	35696	53.83	C	never	never	never	Sometimes
1568	28150	94.16	A	frequently	never	rarely	Sometimes
1569	67120	42.99	C	never	always	rarely	Yes
1570	20593	49.69	C	never	always	always	Sometimes
1571	73636	89.17	A	frequently	sometimes	always	Sometimes
1572	68704	70.17	B	frequently	never	rarely	Yes
1573	82398	3.06	F	never	never	always	Sometimes
1574	11307	39.22	C	never	always	never	Sometimes
1575	24681	38.71	C	never	sometimes	never	Sometimes
1576	51275	68.93	C	never	always	rarely	Sometimes
1577	95780	53.05	C	never	always	never	Yes
1578	71733	42.74	C	never	always	always	Never
1579	93904	70.2	B	frequently	sometimes	always	Never
1580	39182	60.37	C	never	sometimes	rarely	Yes
1581	30985	75.61	B	never	always	always	Yes

R studio


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Project: (None)
Console C:/Users/tomasz/Desktop/142-cl
R version 3.2.3 (2015-12-10
Copyright (C) 2015 The R Fo
Platform: x86_64-w64-mingw3
R is free software and come
You are welcome to redistri
Type 'license()' or 'licenc
R is a collaborative projec
Type 'contributors()' for m
'citation()' on how to cite
Type 'demo()' for some demo
'help.start()' for an HTML
Type 'q()' to quit R.
[Workspace loaded from C:/U
>
Import Dataset
Name
MOODY.training
Encoding Automatic
Heading Yes No
Row names Automatic
Separator Comma
Decimal Period
Quote Double quote (")
Comment None
na.strings NA
Strings as factors
Input File
STUDENTID,SCORE,GRADE,ON_SMARTPHONE,ASKS_QUESTIONS,LEAVES_E
65446,23.67,D,never,always,never,Yes
79686,8.41,F,never,never,rarely,Sometimes
56400,69.76,C,never,always,rarely,Never
16792,95.51,A,never,always,always,Never
92195,67.12,B,frequently,always,never,Sometimes
93635,89.55,B,never,always,never,Sometimes
37569,23.59,F,never,sometimes,always,Never
61757,30.96,D,never,sometimes,rarely,Never
72878,49.68,C,never,always,rarely,Yes
99790,95.38,A,never,always,rarely,Sometimes
66144,37.06,C,never,always,rarely,Sometimes
69243,45.16,C,never,always,rarely,Yes
26685,27.55,C,never,always,never,Sometimes
29403,90.26,A,frequently,always,always,Never
54170,74.96,B,never,sometimes,never,Never
28985,51.75,C,never,never,never,Never
Data Frame
STUDENTID SCORE GRADE ON_SMARTPHONE ASKS_QUESTIC
65446 23.67 D never always
79686 8.41 F never never
56400 69.76 C never always
16792 95.51 A never always
92195 67.12 B frequently always
93635 89.55 B never always
37569 23.59 F never sometimes
61757 30.96 D never sometimes
72878 49.68 C never always
99790 95.38 A never always
66144 37.06 C never always
69243 45.16 C never always
26685 27.55 C never always
29403 90.26 A frequently always
54170 74.96 B never sometimes
28985 51.75 C never never
Import
Cancel
R Documentation
carry out regression, single
variance (although aov may provide
s, na.action,
10:51 AM
1/23/2018

After you import the data = PLOT!

We will start from plots and one line commands in R to make them

First step: IMPORT your data into R studio

- Easiest is to have R studio working directory same as directory with your data use setwd(), getwd()
- This way you do not need to type the full path to the file in the statement: read.csv('file name')

Typical commands

- plottype(variable1, variable, tag1= , tag2=,.....file = ...)

From basic plots

- Scatter plots
- Boxplots
- Barplots
- Mosaic (HeatMap)
- Pie Charts
- Density Plots
- …

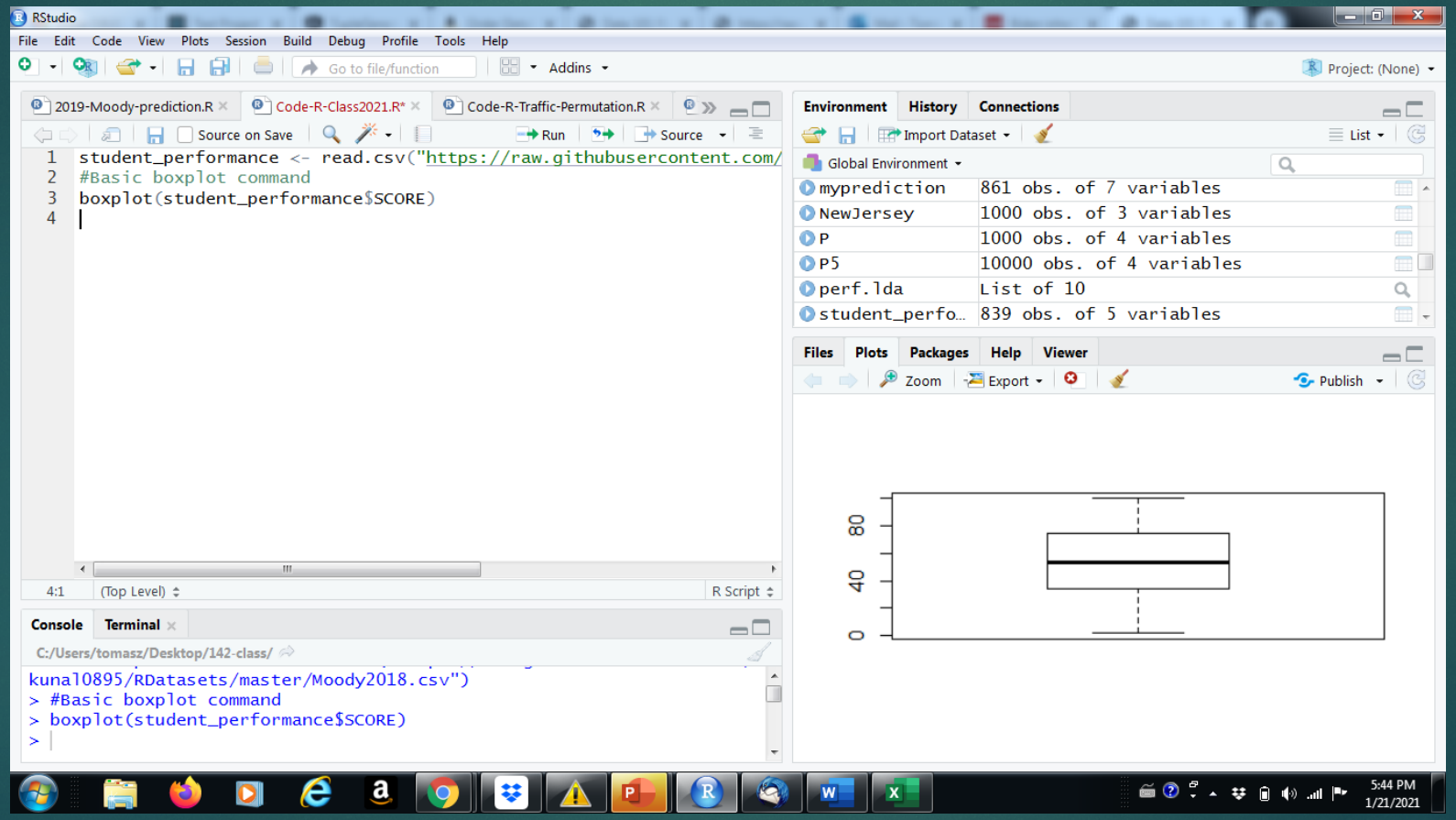
data101.cs.rutgers.edu/laboratory

- DATA TYPES PANEL: DATATYPES/read.csv
- PLOTS/plot, PLOTS/boxplot etc

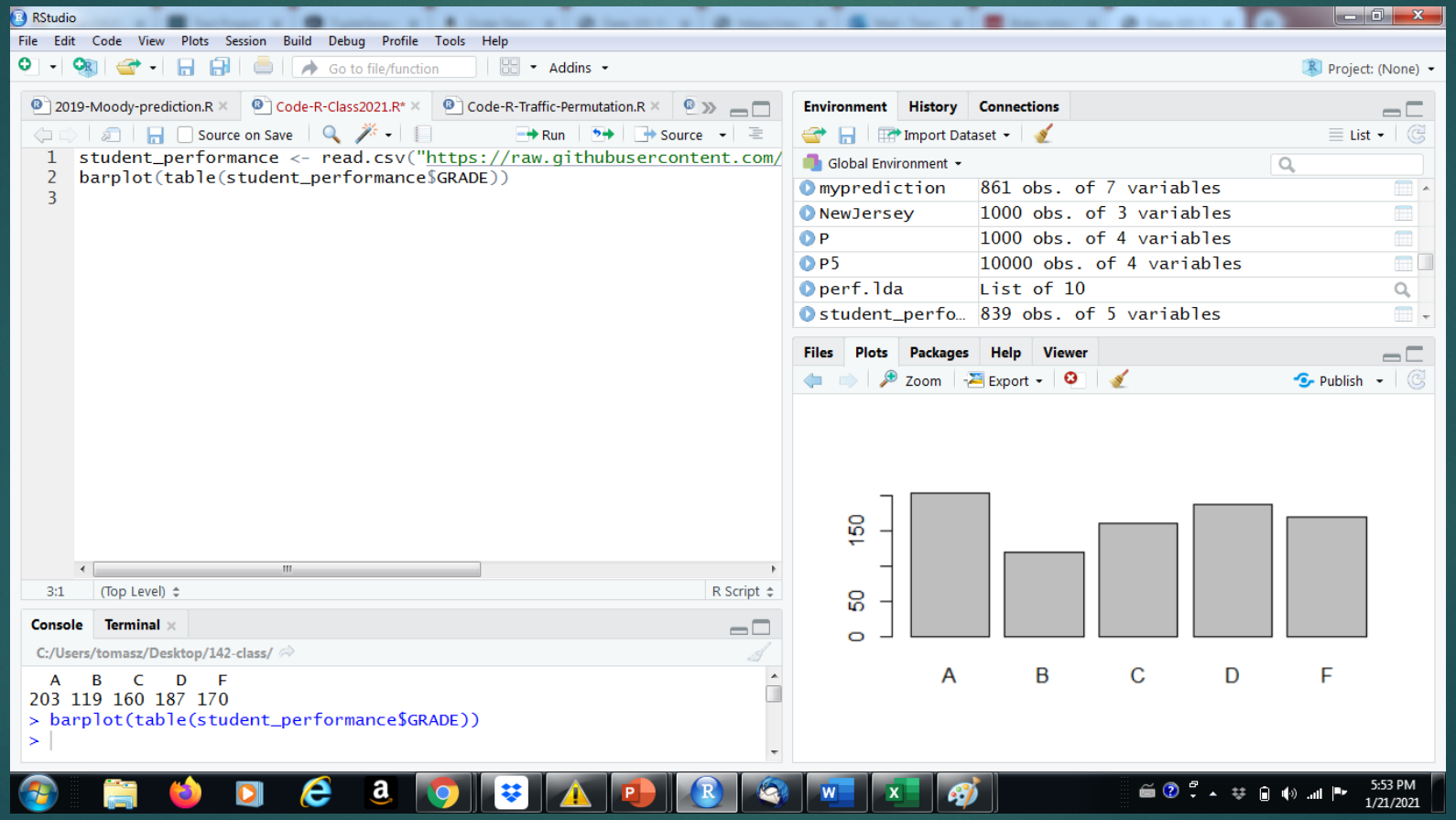
Categorical, Numerical and Ordinal Variables…

- CAT: Categorical: GRADE like A, B, C, D
- NUM: Numerical: SCORE: like 89.64
- ORD: Ordinal: ordered categorical: D<C<B<A

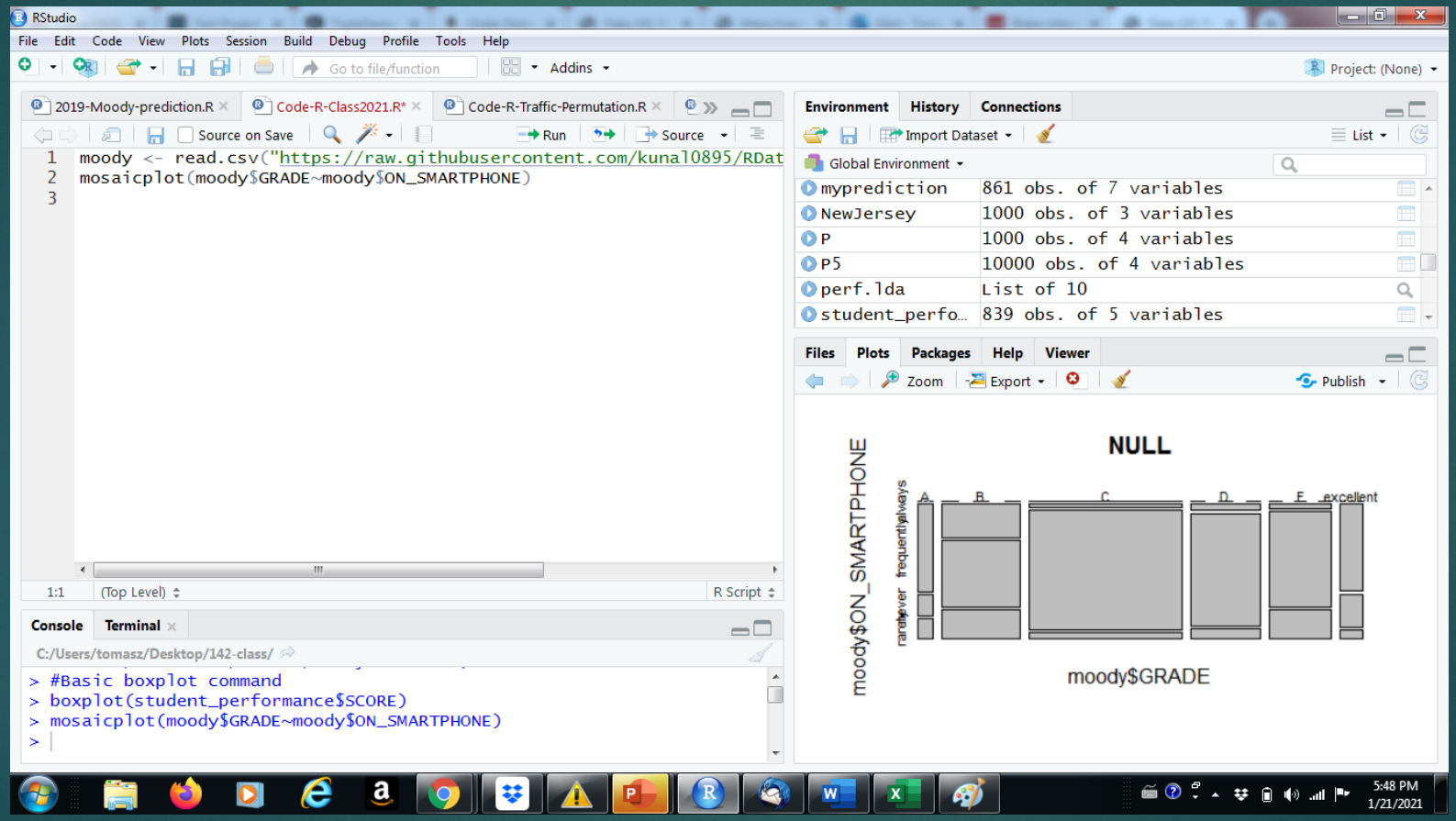
http://data101.cs.rutgers.edu/laboratory/pages/boxplots



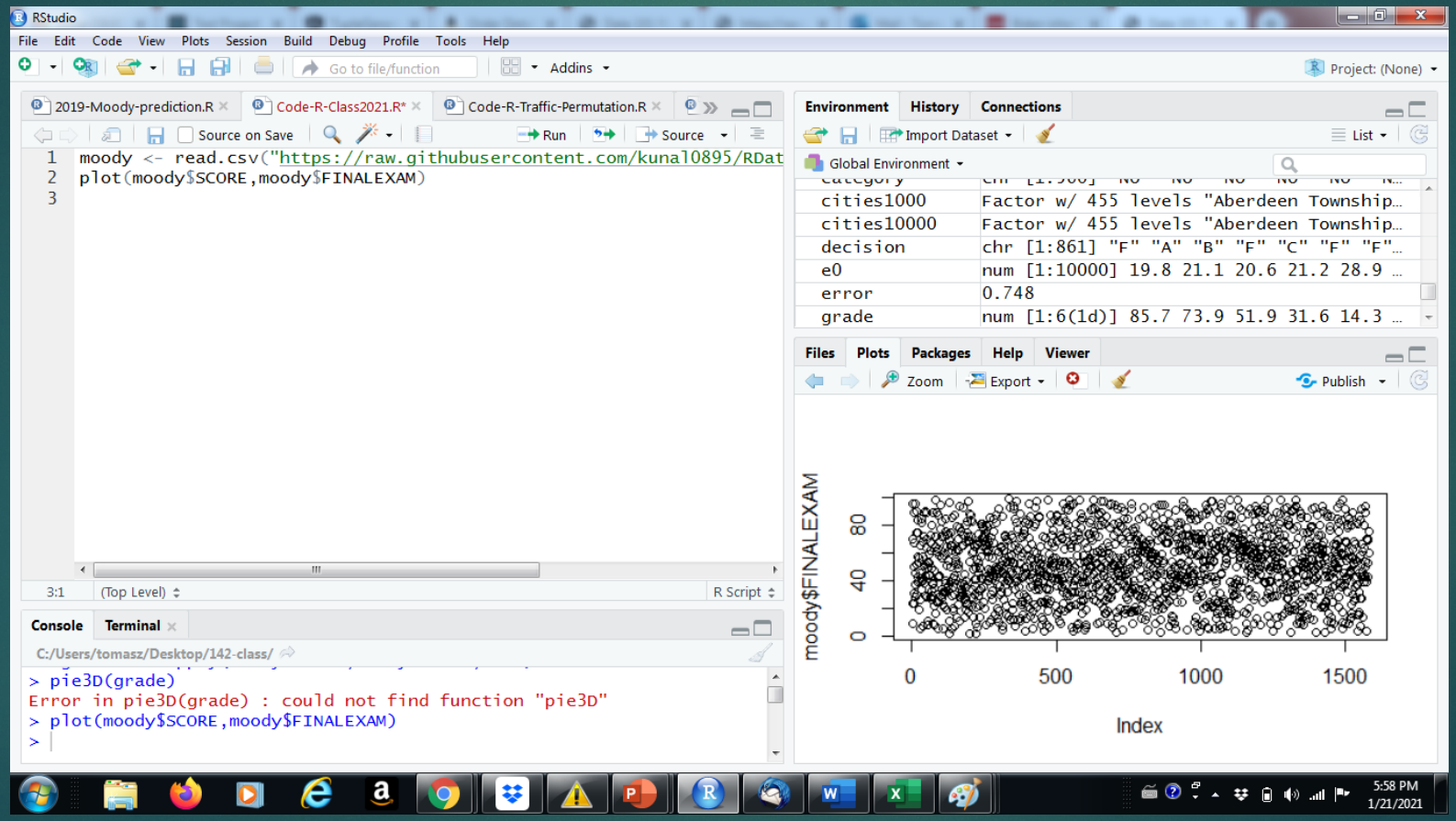
http://data101.cs.rutgers.edu/laboratory/pages/barplot



http://data101.cs.rutgers.edu/laboratory/pages/mosaicplotnew



http://data101.cs.rutgers.edu/laboratory/pages/plot

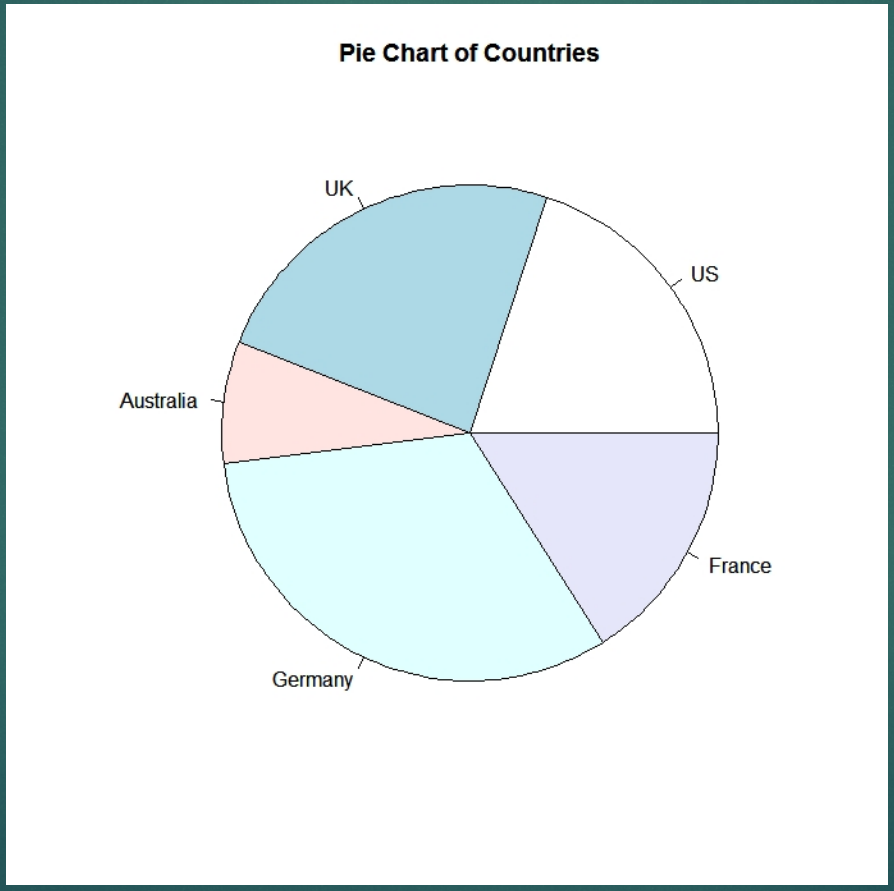


Which plot to use?

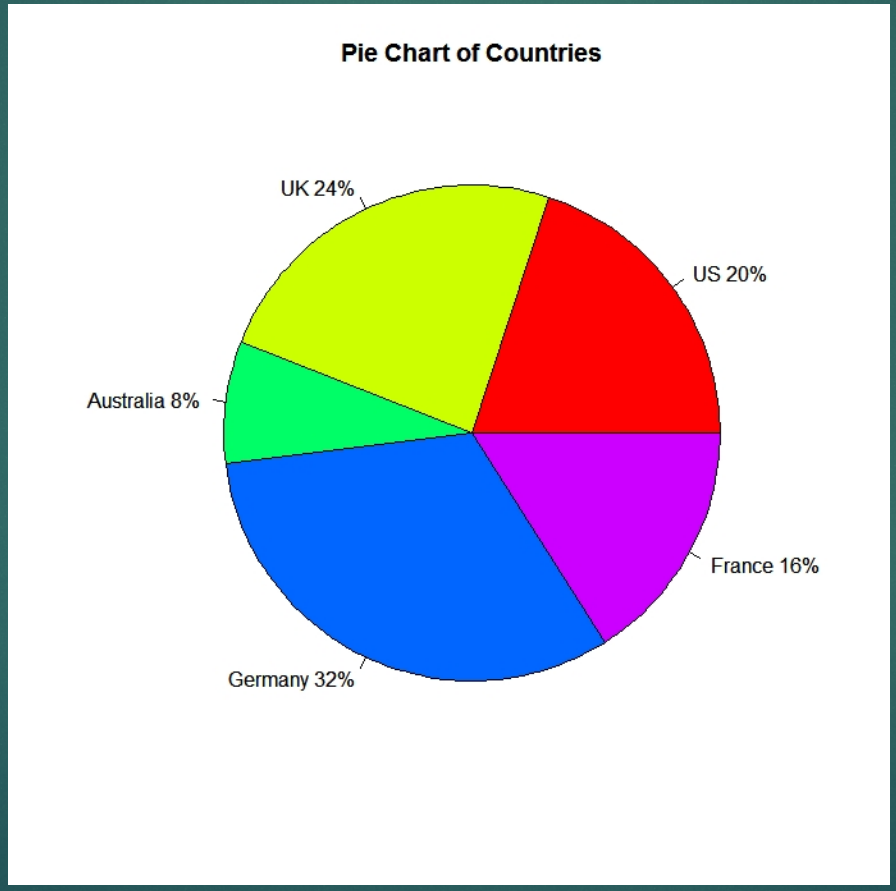
- It all depends on the variables, CAT (categorical), NUM (numerical),

- NUM x NUM scatter plot
- CAT x CAT mosaic plot
- CAT x NUM box plot
- NUM box plot, histogram
- CAT bargraph…..

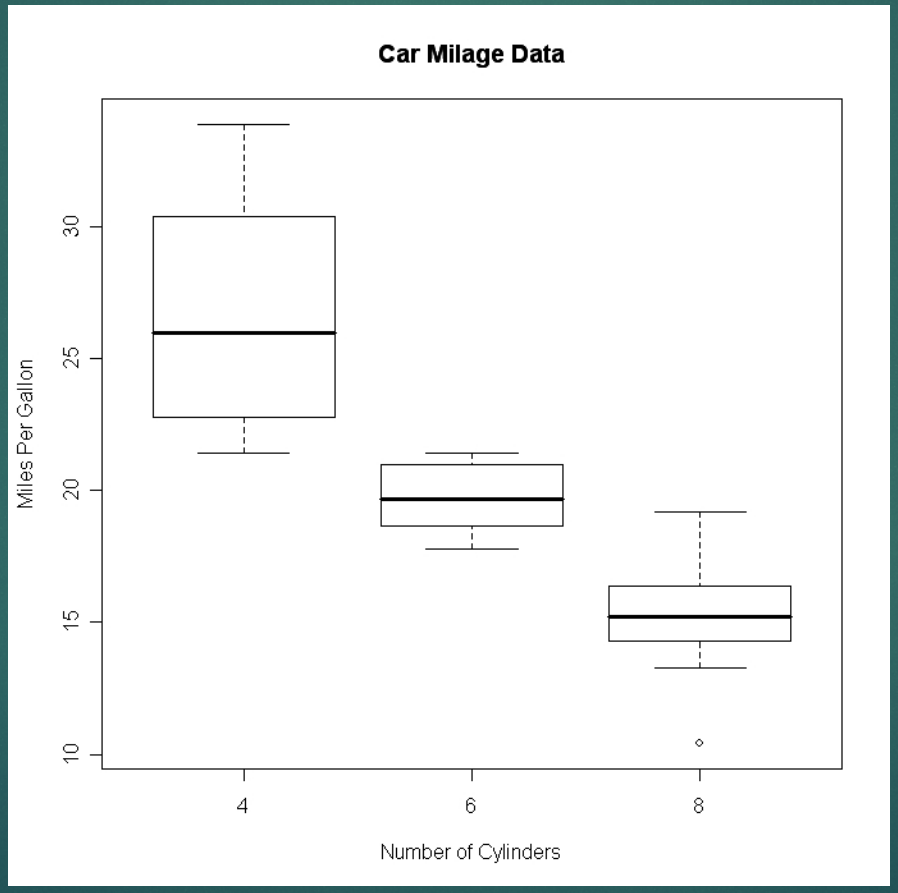
Piecharts TYPE: CAT



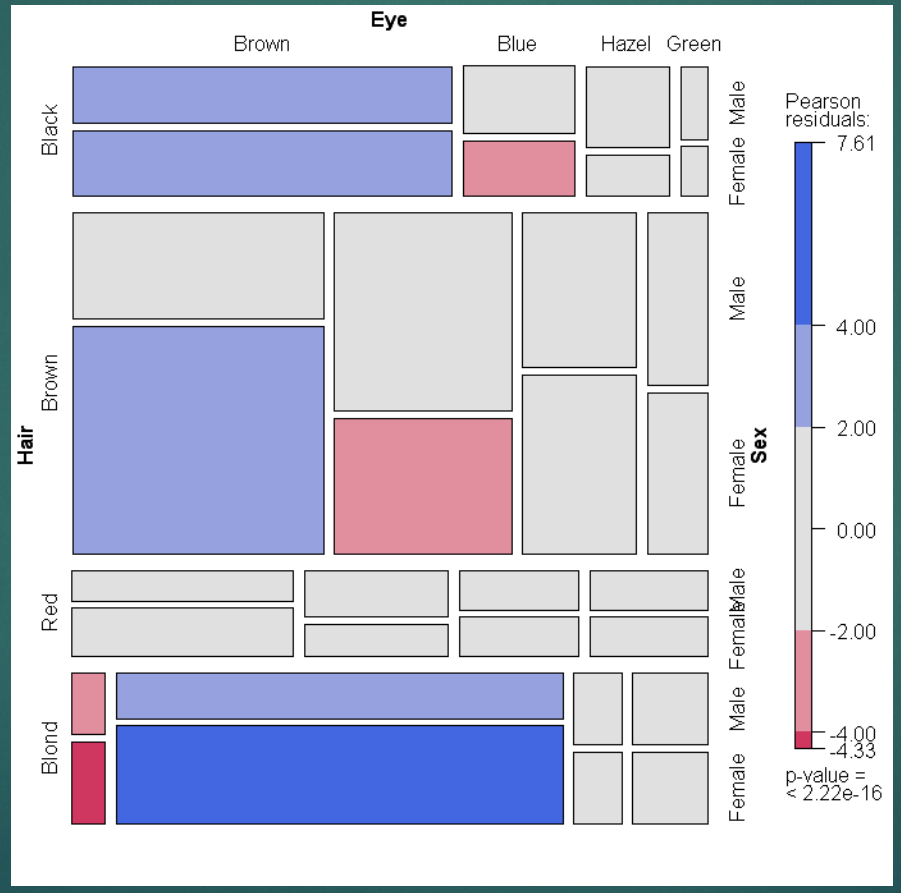
Piecharts


Pie Chart of Countries
UK 24%
US 20%
Australia 8%
France 16%
Germany 32%

Boxplots TYPE: NUM x CAT



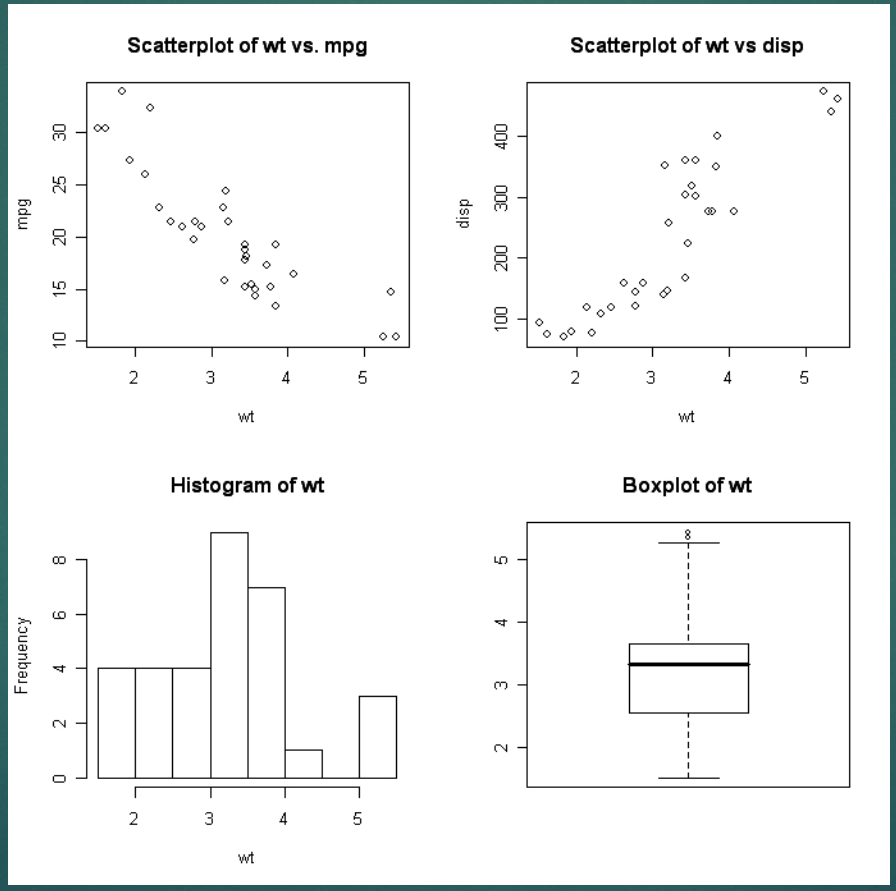
Mosaic Plot TYPE: CAT x CAT X CAT



Combining plots

- Plots/par

Combining Plots



What's interesting?

- Contradictory to our expectations? So called "Bayesian Prior"
- Outliers
- High Correlation
- What are TOP K, Bottom K values

Do you know what I found? – can't wait to show you…..

- Salaries do not depend on education?
- Salaries clearly are positively correlated with education
- IF groom and bride are born under the same sign THEN marriage has much higher chance to survive

Interesting vs actionable

- Wines from Montenegro are much more expensive than French wines
- Californian wines are rated the highest
- Sweden has the highest cost of living
- Greatest basketball players are more than 6’ 7’’ tall

Interesting and/or actionable?

Honda has the best repair record

Vegetarians live 3 years longer

Lincoln tunnel traffic is higher than Holland tunnel traffic on weekends

Out of top 10 richest people in US, 7 of them are under 45

Why look for patterns, trends?

ACTIONABLE (we can do something based on the analysis which will benefit someone)

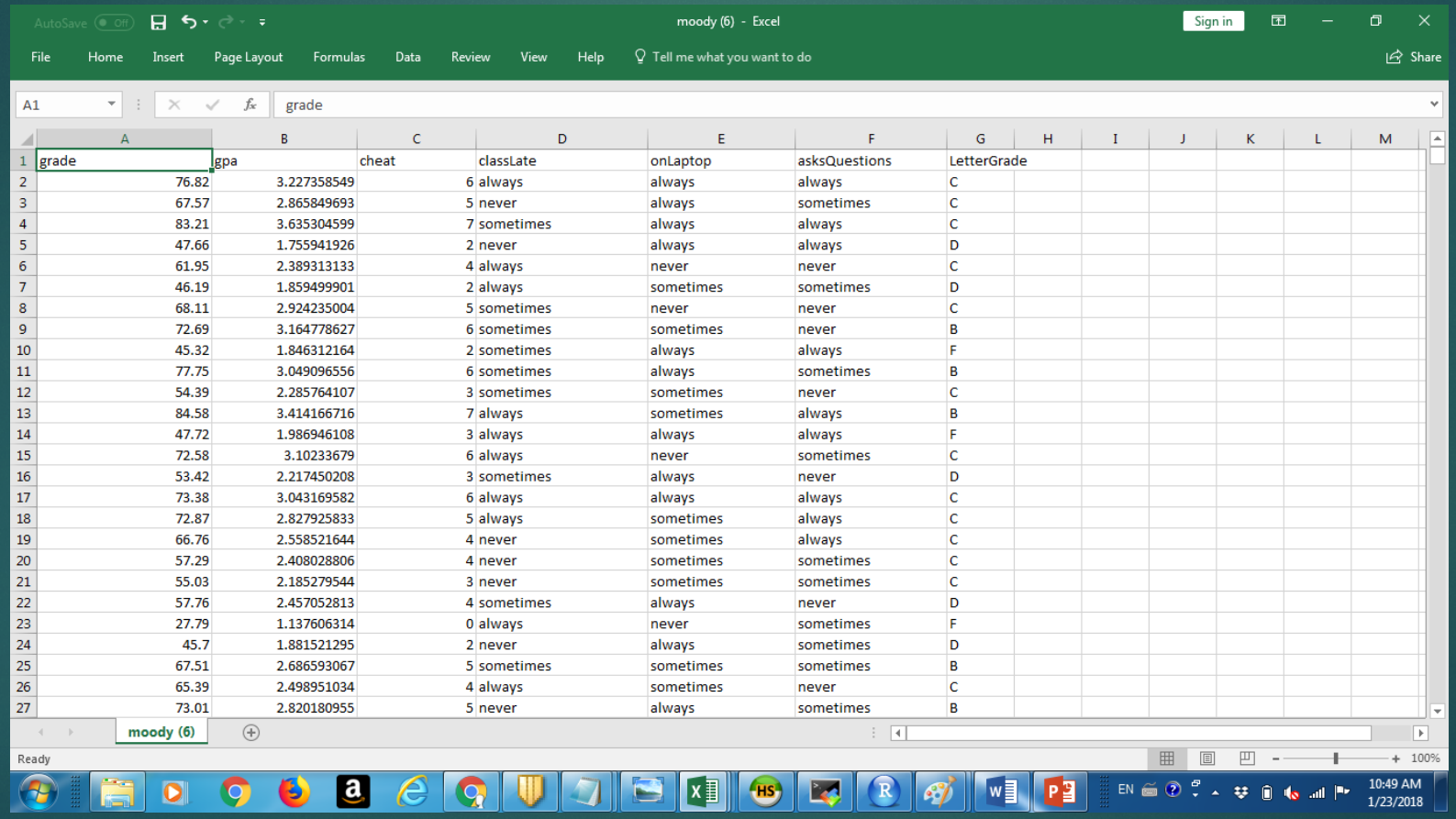
DATA CLEANING – biased data collection, errors, missing data

CURIOSITY (did you know that?)

How much R do I need to know?

- LaboRatory - data101@cs.rutgers.edu/laboratory
- ONE LINERS
- student_performance <- read.csv("MOODY.csv")
- boxplot(student_performance$SCORE, main='My first Boxplot')
- mosaicplot(moody$GRADE~moody$ON_SMARTPHONE)
  gradeTable <- table(student_performance$GRADE)

Professor Moody data set



moody (6) - Excel

	A	B	C	D	E	F	G
1	grade	gpa	cheat	classLate	onLaptop	asksQuestions	LetterGrade
2	76.82	3.227358549	6	always	always	always	C
3	67.57	2.865849693	5	never	always	sometimes	C
4	83.21	3.635304599	7	sometimes	always	always	C
5	47.66	1.755941926	2	never	always	always	D
6	61.95	2.389313133	4	always	never	never	C
7	46.19	1.859499901	2	always	sometimes	sometimes	D
8	68.11	2.924235004	5	sometimes	never	never	C
9	72.69	3.164778627	6	sometimes	sometimes	never	B
10	45.32	1.846312164	2	sometimes	always	always	F
11	77.75	3.049096556	6	sometimes	always	sometimes	B
12	54.39	2.285764107	3	sometimes	sometimes	never	C
13	84.58	3.414166716	7	always	sometimes	always	B
14	47.72	1.986946108	3	always	always	always	F
15	72.58	3.10233679	6	always	never	sometimes	C
16	53.42	2.217450208	3	sometimes	always	never	D
17	73.38	3.043169582	6	always	always	always	C
18	72.87	2.827925833	5	always	sometimes	always	C
19	66.76	2.558521644	4	never	sometimes	always	C
20	57.29	2.408028806	4	never	sometimes	sometimes	C
21	55.03	2.185279544	3	never	sometimes	sometimes	C
22	57.76	2.457052813	4	sometimes	always	never	D
23	27.79	1.137606314	0	always	never	sometimes	F
24	45.7	1.881521295	2	never	always	sometimes	D
25	67.51	2.686593067	5	sometimes	sometimes	sometimes	B
26	65.39	2.498951034	4	always	sometimes	never	C
27	73.01	2.820180955	5	never	always	sometimes	B

Simple and more complex

- DATA -> PLOT
- DATA -> TRANSFORMATION ->PLOT

More complex

- DATA -> TRANSFORMATION ->PLOT