Plots

CODE IS LIVE

- In our active tutorial (code snippets)
- All examples are there
- PPTs – anything but code (with small exceptions sometimes)

Categorical, Numerical and Ordinal Variables…

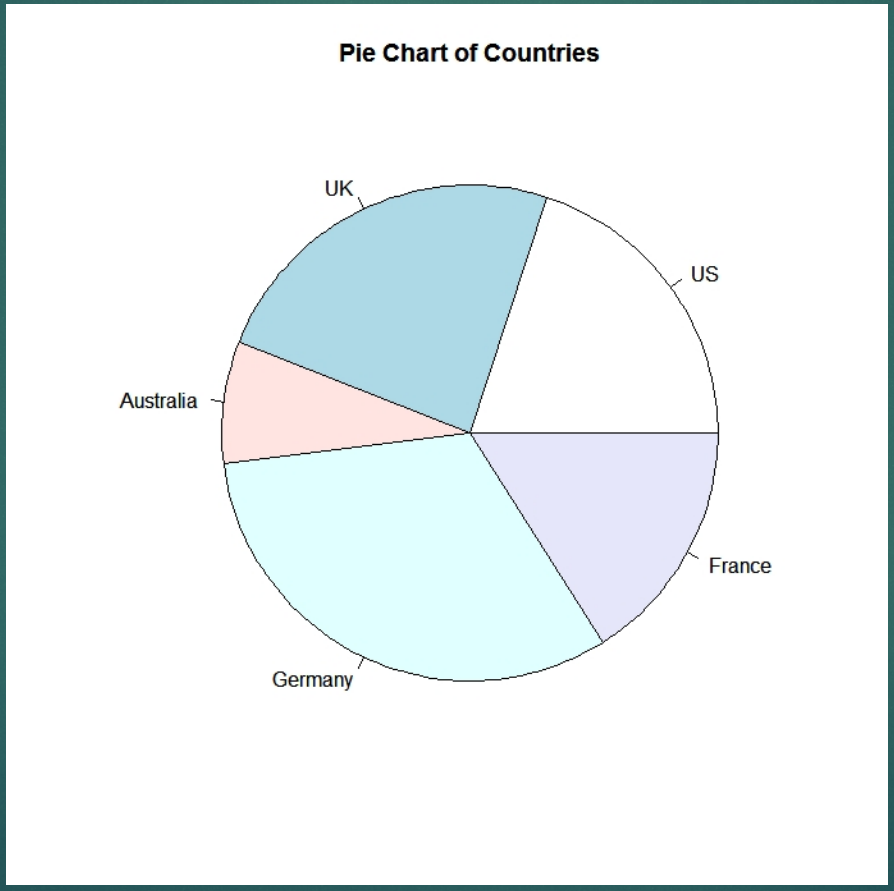
- CAT: Categorical: GRADE like A, B, C, D
- NUM: Numerical: SCORE: like 89.64
- ORD: Ordinal: ordered categorical: D<C<B<A

Which plot to use?

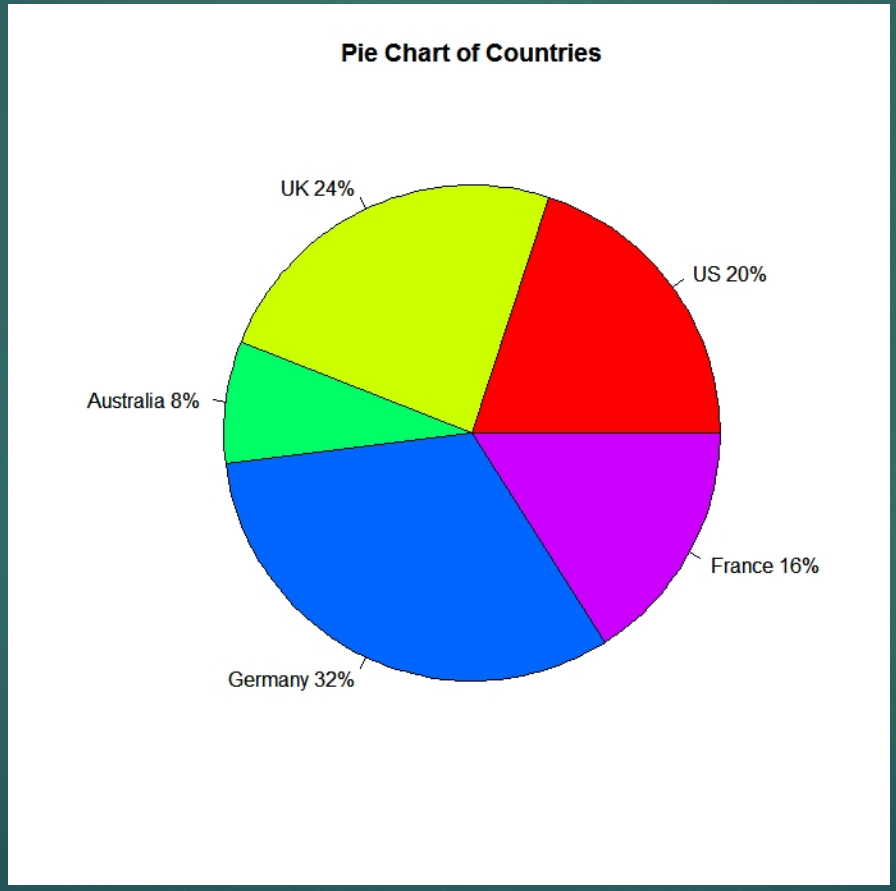
- It all depends on the variables, CAT (categorical), NUM (numerical),

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

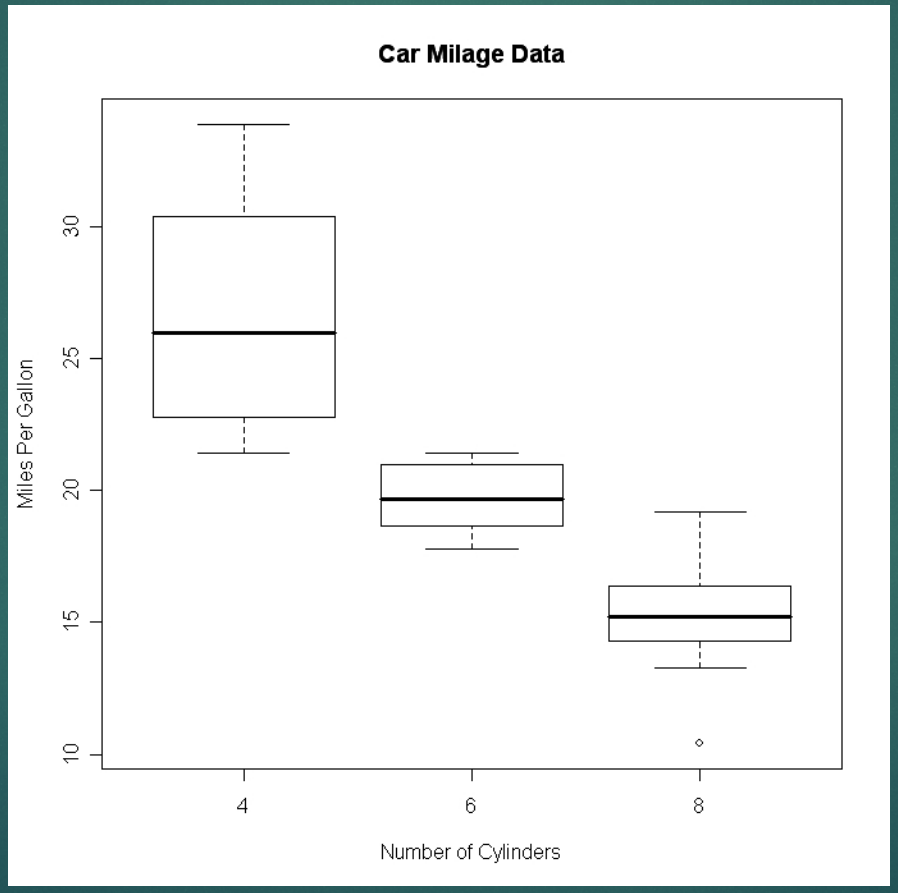
Piecharts TYPE: CAT



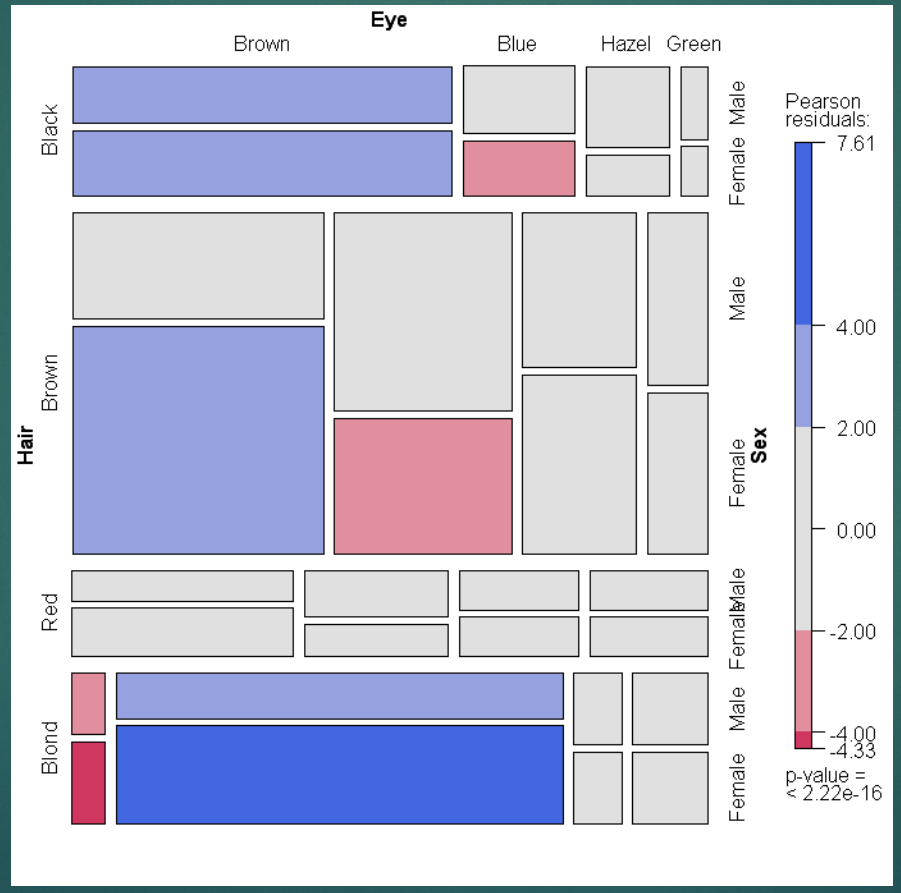
Piecharts


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



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)

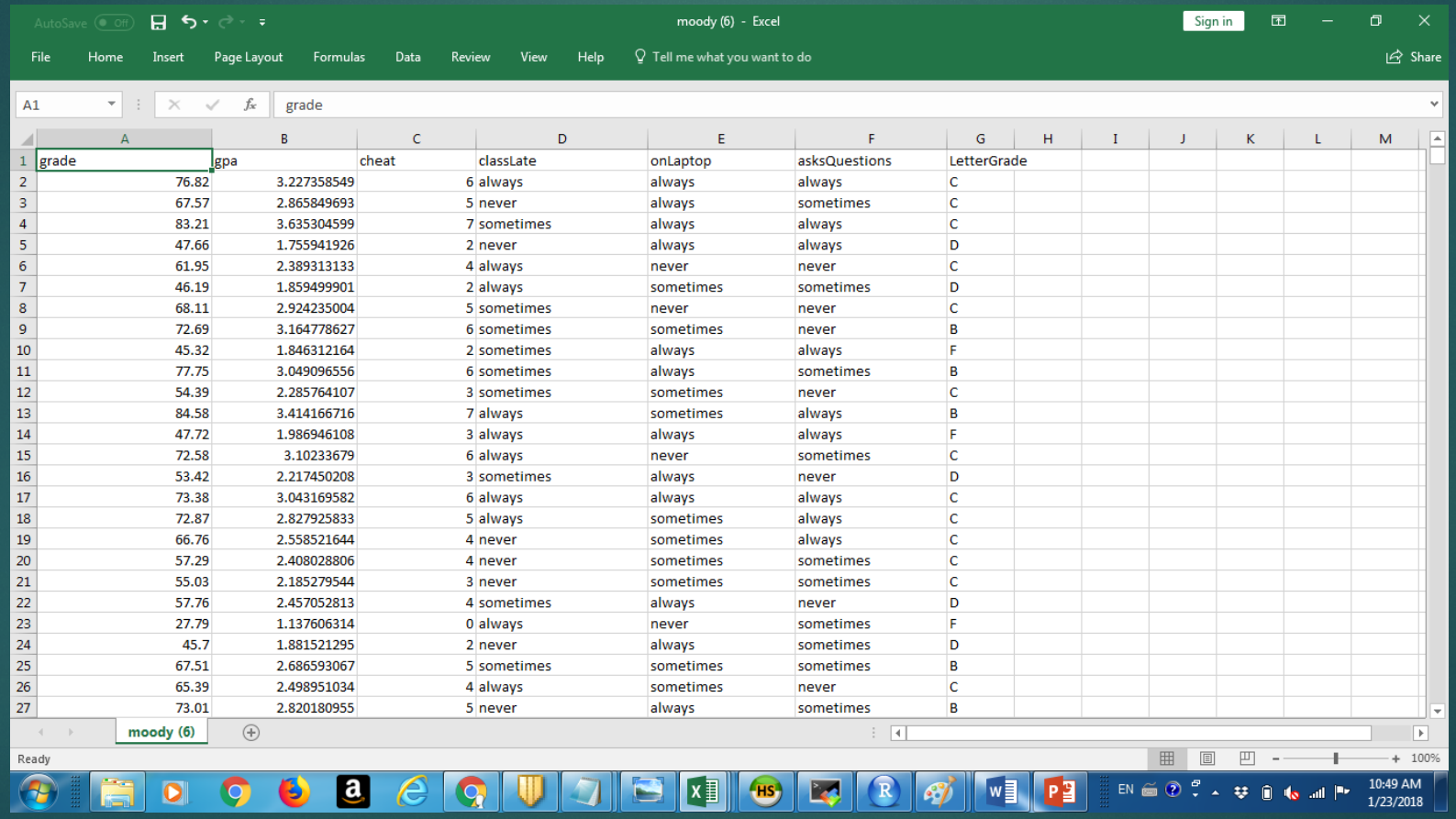
DATA CLEANING – biased data collection, errors, missing data

CURIOSITY (did you know that?)

How much R do I need to know?

- 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)
- **OBJECTIVE: SHORTEST PATH – how to plot/explore/predict/test hypotheses with MINIMAL PROGRAMMING**

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

TRANSFORMATIONS on data frames.

- Subsetting
- Adding new columns
- Aggregating (tapply, table)
- Next class!