Doing Data Science

Unit 10

Faizan Javed
Data Science @ SMU

Admin notes

Live Session Unit 10 assignment due next week

Case Study 1

Everyone did well

Minor issues with plots/visualization, data presentation order

Github: minor issues with readme, project organization

#Q1: How many breweries are present in each state

```
ct <- data.frame(count(brews, brews$State))
names(ct) <- c("State", "Count")
ct[order(-ct$Count),]
```

	State	Count
6	CO	47
5	CA	39
23	MI	32
38	OR	29

...

#Q2 Print first 6 and last 6 observations

rbind(head(df),tail(df))

	BreweryID	BeerName	BeerID	ABV	IBU	Style	Ounces
1	1	Get Together	2692	0.045	50	American IPA	16
2	1	Maggie's Leap	2691	0.049	26	Milk / Sweet Stout	16
3	1	Wall's End	2690	0.048	19	English Brown Ale	16
4	1	Pumpion	2689	0.060	38	Pumpkin Ale	16
5	1	Stronghold	2688	0.060	25	American Porter	16
6	1	Parapet ESB	2687	0.056	47	Extra Special / Strong Bitter (ESB)	16
2405	556	Pilsner Ukiah	98	0.055	NA	German Pilsener	12
2406	557	Heinnieweisse Weissebier	52	0.049	NA	Hefeweizen	12
2407	557	Snapperhead IPA	51	0.068	NA	American IPA	12
2408	557	Moo Thunder Stout	50	0.049	NA	Milk / Sweet Stout	12
2409	557	Porkslap Pale Ale	49	0.043	NA	American Pale Ale (APA)	12
2410	558	Urban Wilderness Pale Ale	30	0.049	NA	English Pale Ale	12

	Company	City	State
1	NorthGate Brewing	Minneapolis	MN
2	NorthGate Brewing	Minneapolis	MN
3	NorthGate Brewing	Minneapolis	MN

#Q3 Report the number of NAs

```
names(df)<- c("BreweryID","BeerName", "BeerID",
"ABV","IBU","Style","Ounces","Company","City","State")

sapply(X=df, FUN=function(x) sum(is.na(x))) # This is equivalent...

colSums(is.na(df)) # ...to this!
```

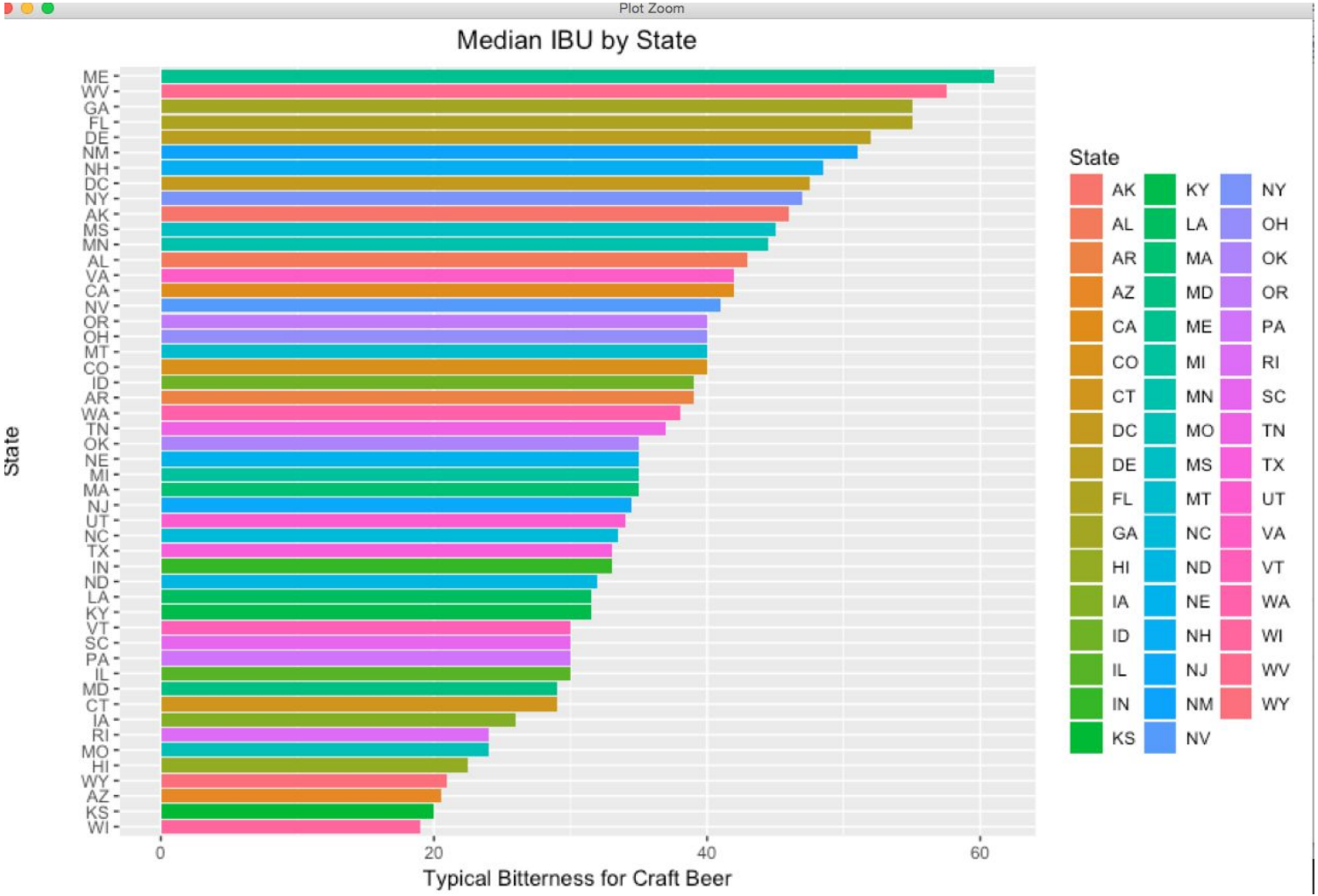
BreweryID	BeerName	BeerID	ABV	IBU	Style	Ounces	Company	City	State
0	0	0	62	1005	0	0	0	0	0

#Q4 Compute the median alcohol content and international bitterness unit for each state

median(df$IBU, na.rm=TRUE)

[1] 35

```
IB<-df %>% # I use dplyr here to do all my steps simultaneously, but you can do this piecemeal in base R, too
    select(State, IBU) %>%
    group_by(State) %>%
    summarize(MedianIBU=round(median(IBU, na.rm=TRUE),4)) %>%
                                        # I chose to round to four digits here
  arrange(desc(MedianIBU))
  # I put them in descending order because my clients are most interested in higher values
data.frame(IB)
```

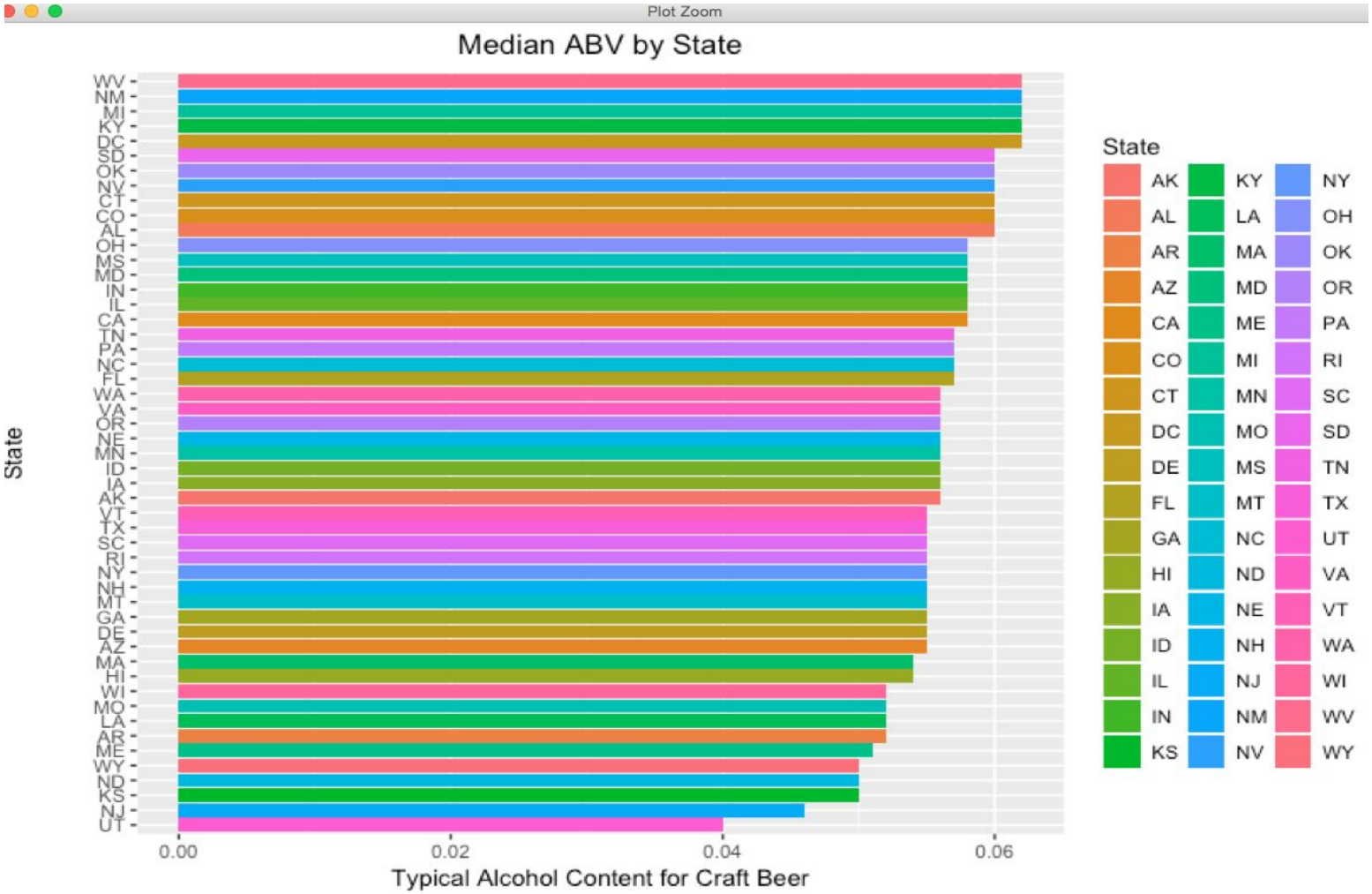

Plot Zoom
Median IBU by State
State
ME
WV
GA
FL
DE
NM
NH
DC
NY
AK
MS
MN
AL
VA
CA
NV
OR
OH
MT
CO
ID
AR
WA
TN
OK
NE
MI
MA
NJ
UT
NC
TX
IN
ND
LA
KY
VT
SC
PA
IL
MD
CT
IA
RI
MO
HI
WY
AZ
KS
WI
0
20
40
60
Typical Bitterness for Craft Beer
State
AK
AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
HI
IA
ID
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
MT
NC
ND
NE
NH
NJ
NM
NV
NY
OH
OK
OR
PA
RI
SC
TN
TX
UT
VA
VT
WA
WI
WV
WY

```
median(df$ABV, na.rm=TRUE)
```

```
[1] 0.056
```

```
AB<-df %>%
  select(State, ABV) %>%
  group_by(State) %>%
  summarize(MedianABV=round(median(ABV, na.rm=TRUE),3)) %>%
  arrange(desc(MedianABV))
data.frame(AB)
```

```
ggplot(AB, aes(reorder(State, MedianABV), MedianABV)) +
# I chose to reorder the bars in descending order, rather than alphabetical
  geom_bar(aes(fill=State), stat="identity") + # the bar colors and what the values mean
  ggtitle("Median ABV by State") + # The Title
  theme(plot.title = element_text(hjust = 0.5)) + # Centers the Title
  xlab("State\n\n") + # Gives an X-axis name
  ylab("Typical Alcohol Content for Craft Beer") + # Gives an informative Y-axis name
  coord_flip() # Flips the coordinates - I do this because it's easier to see States on the Y-Axis
```


Plot Zoom
Median ABV by State
State
WV
NM
MI
KY
DC
SD
OK
NV
CT
CO
AL
OH
MS
MD
IN
IL
CA
TN
PA
NC
FL
WA
VA
OR
NE
MN
ID
IA
AK
VT
TX
SC
RI
NY
NH
MT
GA
DE
AZ
MA
HI
WI
MO
LA
AR
ME
WY
ND
KS
NJ
UT
0.00
0.02
0.04
0.06
Typical Alcohol Content for Craft Beer
State
AK
AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
HI
IA
ID
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
MT
NC
ND
NE
NH
NJ
NM
NV
NY
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY

#Q5 which state has the maximum ABV beer? Which state has the most bitter beer?

Which state has the maximum alcoholic beer?

```
df[which.max(df$ABV),c("State","BeerName","ABV")]
```

	State	BeerName	ABV
375	CO	Lee Hill Series Vol. 5 - Belgian Style Quadrupel Ale	0.128

Which state has the most bitter beer?

```
df[which.max(df$IBU),c("State","BeerName","IBU")]
```

	State	BeerName	IBU
1857	OR	Bitter Bitch Imperial IPA	138

Q6: Summary statistics for the ABV variable

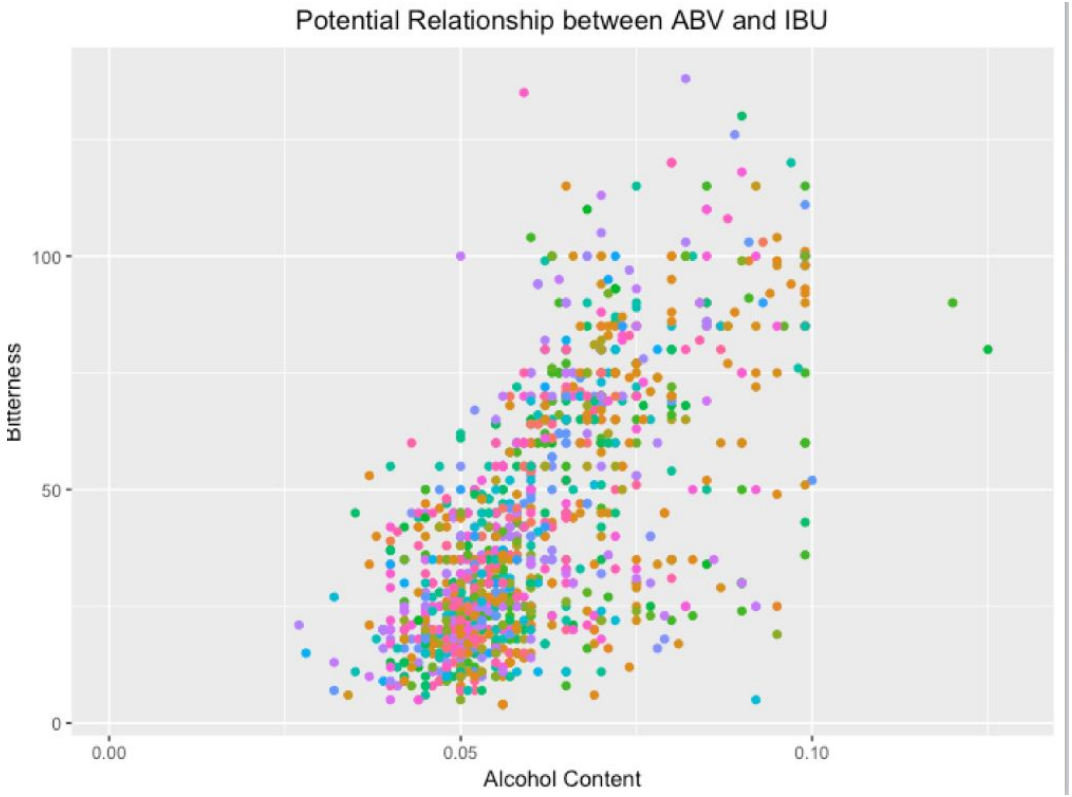
```
summABV<-data.frame(cbind(summary(df$ABV)))

names(summABV)<-"Summary Statistics (ABV)"

round(summABV,3)
```

	Summary Statistics (ABV)
Min.	0.001
1st Qu.	0.050
Median	0.056
Mean	0.060
3rd Qu.	0.067
Max.	0.128
NA's	62.000

#Q7: # Is there an apparent relationship between the IBU and ABV?



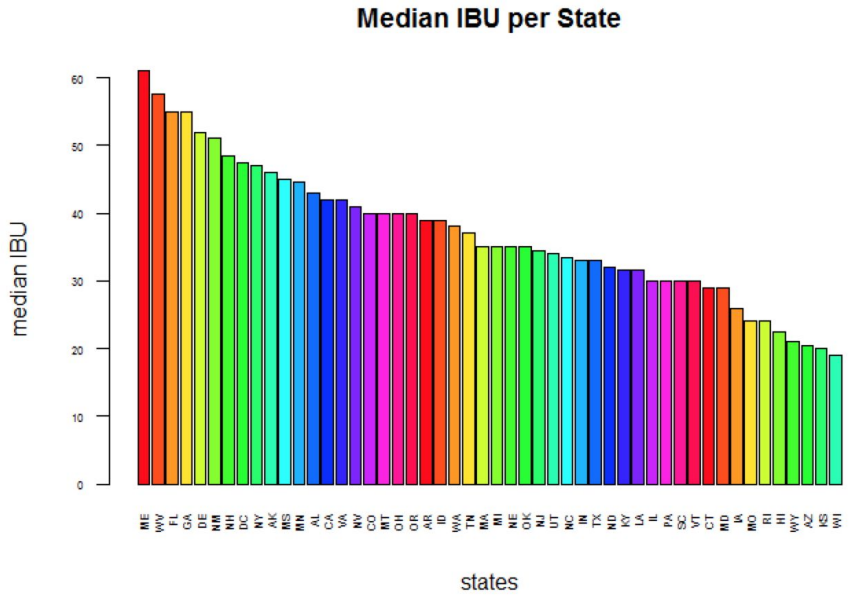
Team Machine Churning

Making bar plot for median IBU and AVB

Creating bar chart for better visual understanding of per state IBU and ABV median.

```
knitr::opts_chunk$set(echo = TRUE)

barplot(medianIBUperstate$IBU, main="Median IBU per State", xlab="states",ylab="median IBU",las=2, col=rainbow(20
), names.arg=medianIBUperstate$State, horiz=FALSE,cex.axis=0.5, cex.names=0.5)
```

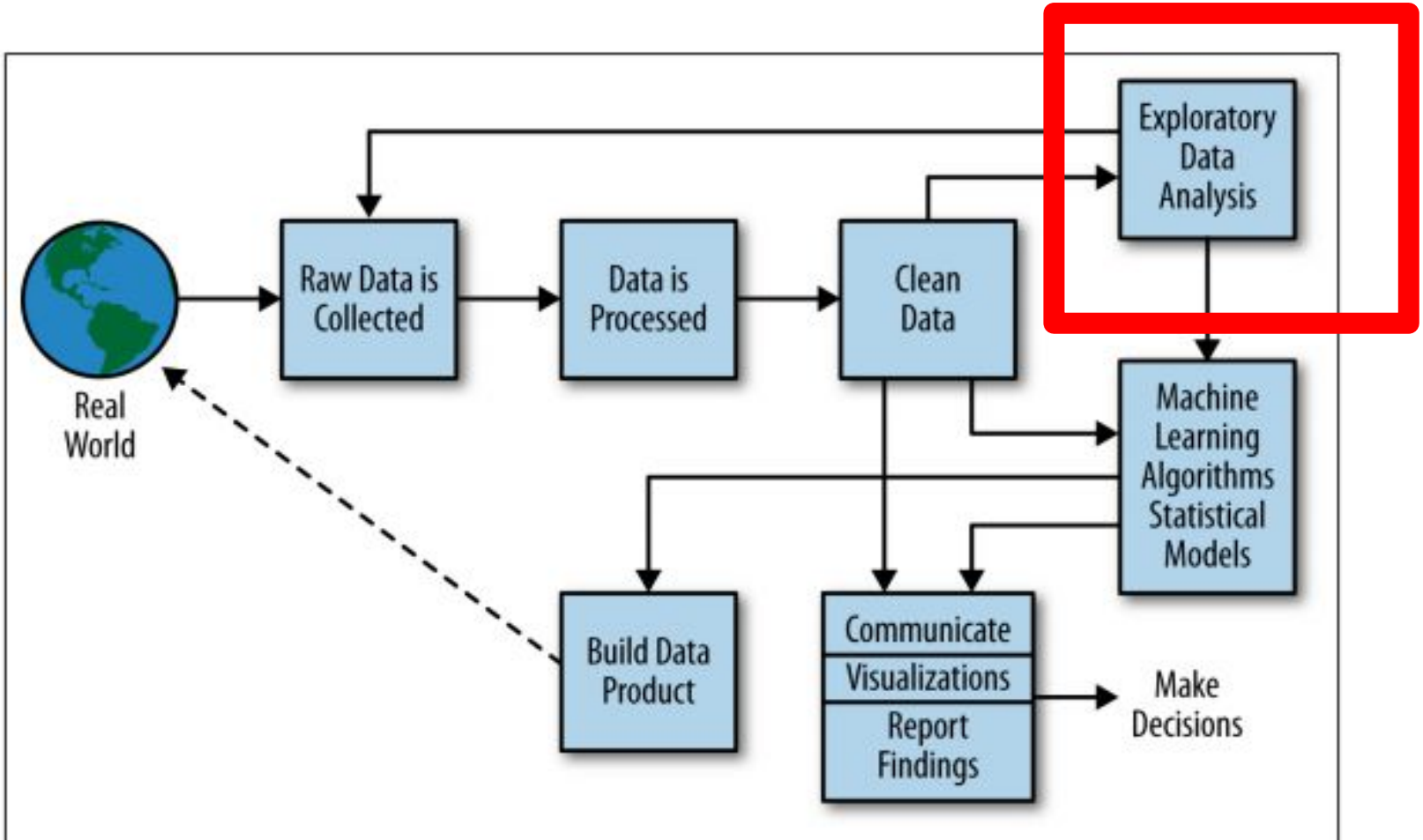


Main topics

Sampling

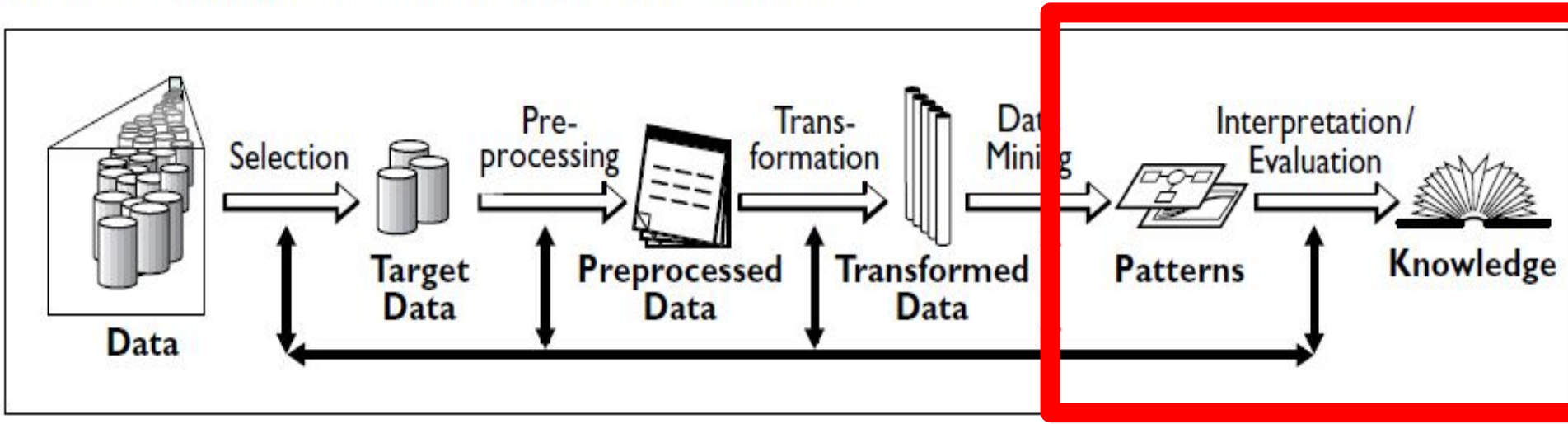
Exploratory Data Analysis (EDA)

The Data Science Process



The Data Science Process

Figure 1. Overview of the steps constituting the KDD process



Exploratory Data Analysis (EDA)

Understanding and exploring clean/tidy data

Plotting distributions of all variables

Time series of data / Modeling

Transforming variables, pairwise relationships between variables

Generating summary statistics of data

(mean, minimum, maximum, upper and lower quartiles, outliers)

Sampling

Population, universe of size N

Observation, characteristics of an object in a population
(may translate to features in machine learning)

Sample, subset of size n

Sampling with replacement

Sampling without replacement

Why sample in the age of Big Data?

Sampling for Big Data, KDD-2014 tutorial: http://nickduffield.net/download/papers/Tutorial_KDD_2014.pdf

Probability distributions

Real world phenomena take a certain mathematical shape

Normal/Gaussian distribution: bell shape → distributions of sums of things.

Take values following the same distribution and sum them - the distribution of their sum follows the normal distribution. Aka the **Central Limit Theorem: implies that statistical methods that work for normal distributions can also be applied to other distributions.**

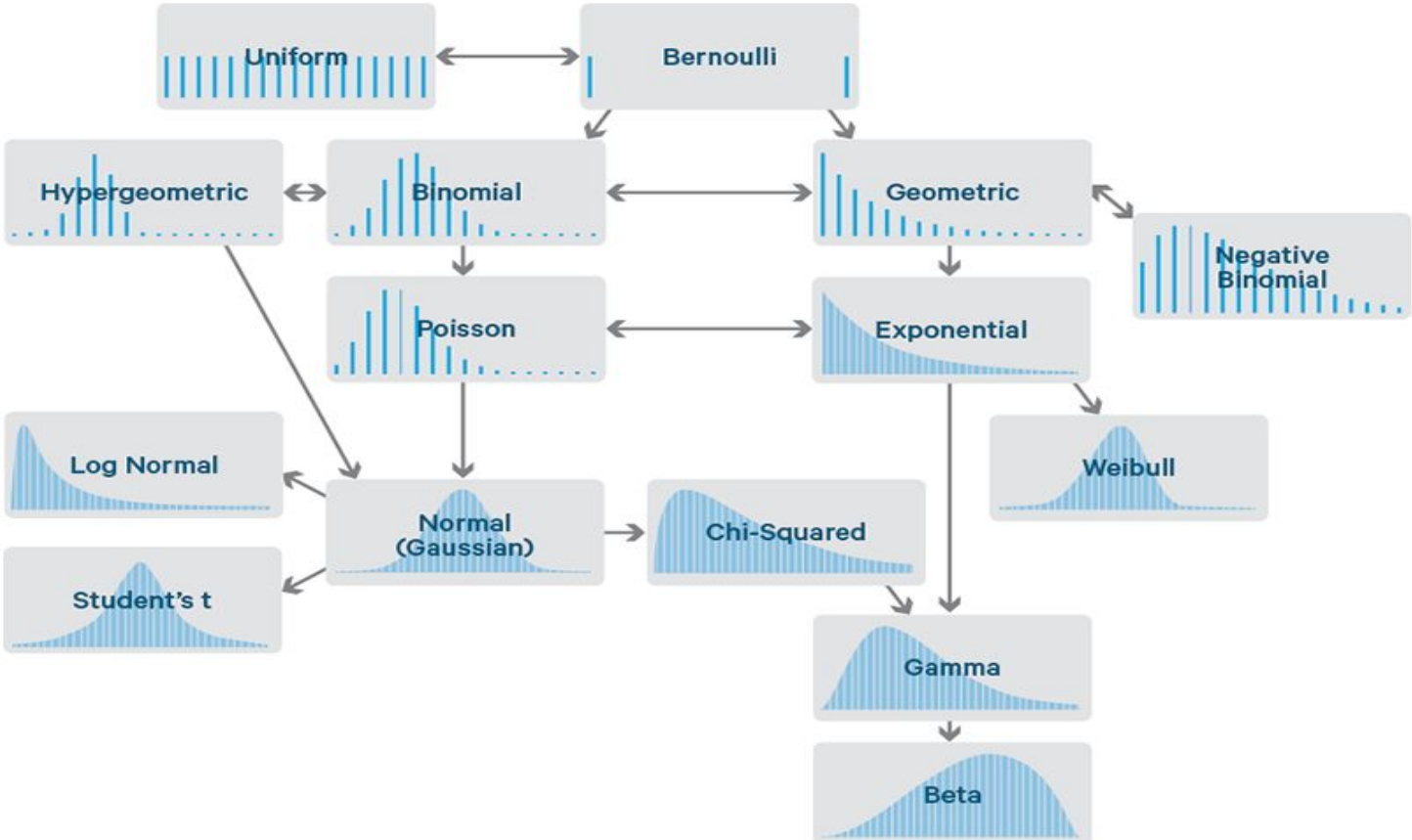
Poisson distribution: number of events occurring in a fixed time interval

e.g., customers calling a support hotline each minute

Relationships between distributions

Data Science crib sheet: https://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/

E.g. Uniform and Bernoulli?



Survey Design

https://zapier.com/learn/forms-surveys/design-analyze-survey/

Common purposes: soliciting feedback, monitoring performance, establishing business metrics

Question/Answer styles:

Categorical/Nominal: specific names/labels (e.g., of products)

E.g., Which feature of our product do you like most? Speed, Ease of Use, etc.

Can't answer questions rating/quantification questions (e.g., "How much..?")

Ordinal: can help answer "how much..?" type questions

E.g., Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree

Interval: useful for collecting segmentation data. Needs to be ordered data, distance between values needs to be meaningful (ideally equally-sized)

E.g., 1-50 employees, 51-100 employees, 100-150 employees

Ratio: richest form of data, and represents precise measurements.

Supports summary statistics such as averages, variance, standard deviation et al

E.g., income level responses, $25,000, $50,000, $105,000

How to select survey responses?

Statistical Inference: Making general statements about populations from samples or the "*process of deducing properties of an underlying probability distribution by analysis of data.*"

Avoid **sampling bias,** e.g., sending surveys via email when you know your users use email, website, and phone equally. Distribute the survey through a variety of channels (unless focused only a particular segment and mode)

How to estimate sample size?

Margin of error: how much your surveys reflect the view of the overall population

Is your sample size statistically significant?

Confidence level: E.g., For a 95% confidence level, if the survey was repeated multiple times, the results would match the results from the actual population 95% of the time.

EDA example:
NYT data (see file Unit-10-NYT-data.R)