Stat 240 week 2

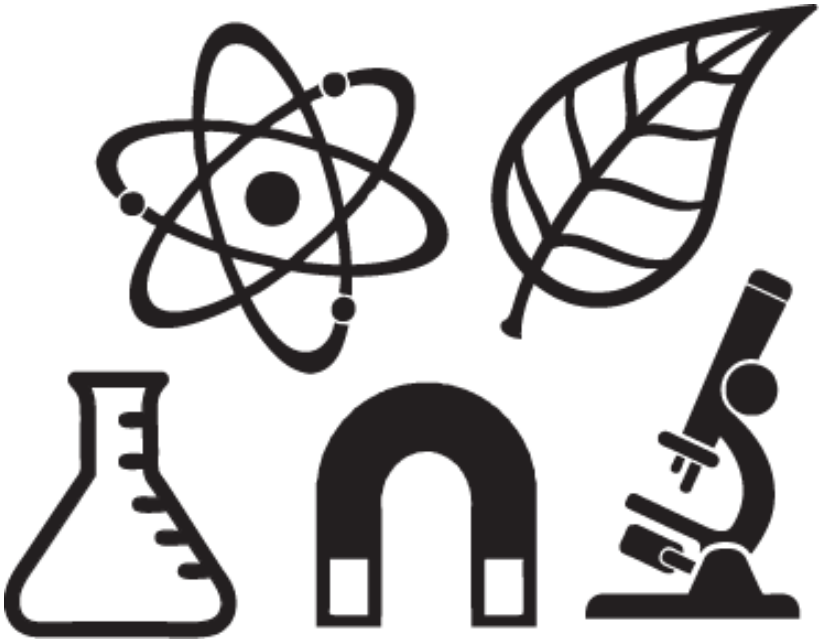
Dr. Dave Campbell
dac5@SFU.ca

SFU CO-OPERATIVE EDUCATION

SCIENCE & ENVIRONMENT CO-OPERATIVE EDUCATION

INFO SESSION
January 17, 2019
11:30 am - 12:20 pm
AQ 3159

EXPLORE CAREER OPTIONS WITH
CO-OP AS PART OF YOUR DEGREE!



This week in Data Science

What is Uber?

What data might they have?

What could it be used for?

Data Science in the city

Uber releases their traffic data:

https://movement.uber.com/use-case/ipa

https://movement.uber.com/use-case/dc

https://movement.uber.com log in with your uber account to see more data

How to prep for computer (or paper) based exams in this course

You need to know how code works and what it does. That comes mainly from experience. As we move along get data sets, look at data sets, …

Create a document / page in your notes with the commands that we use/ learn in class and lab.

Data sets for today:

Jobs from last day

	Title	City	Employer	Experience
1	Satellite Data Analyst	Vancouver, BC	PhotoSat	<NA>
2	Data Scientists and Statisticians	Vancouver, BC	Genome Canada Transplant Consortium	<NA>
3	Data Technician	Surrey, BC	Absolute Results Productions	<NA>
4	Junior Data Engineer [20181130TECH]	Vancouver, BC	Paragon Testing Enterprises Inc	3
5	>Love Data? Apply to become a Research Data Centre Analyst!	Vancouver, BC	Government of Canada	<NA>
6	>STUDENTS: Geoscience Field, GIS Data, Lab/Office and Researc...	Vancouver, BC	Government of Canada	<NA>

Parking tickets in Chicago

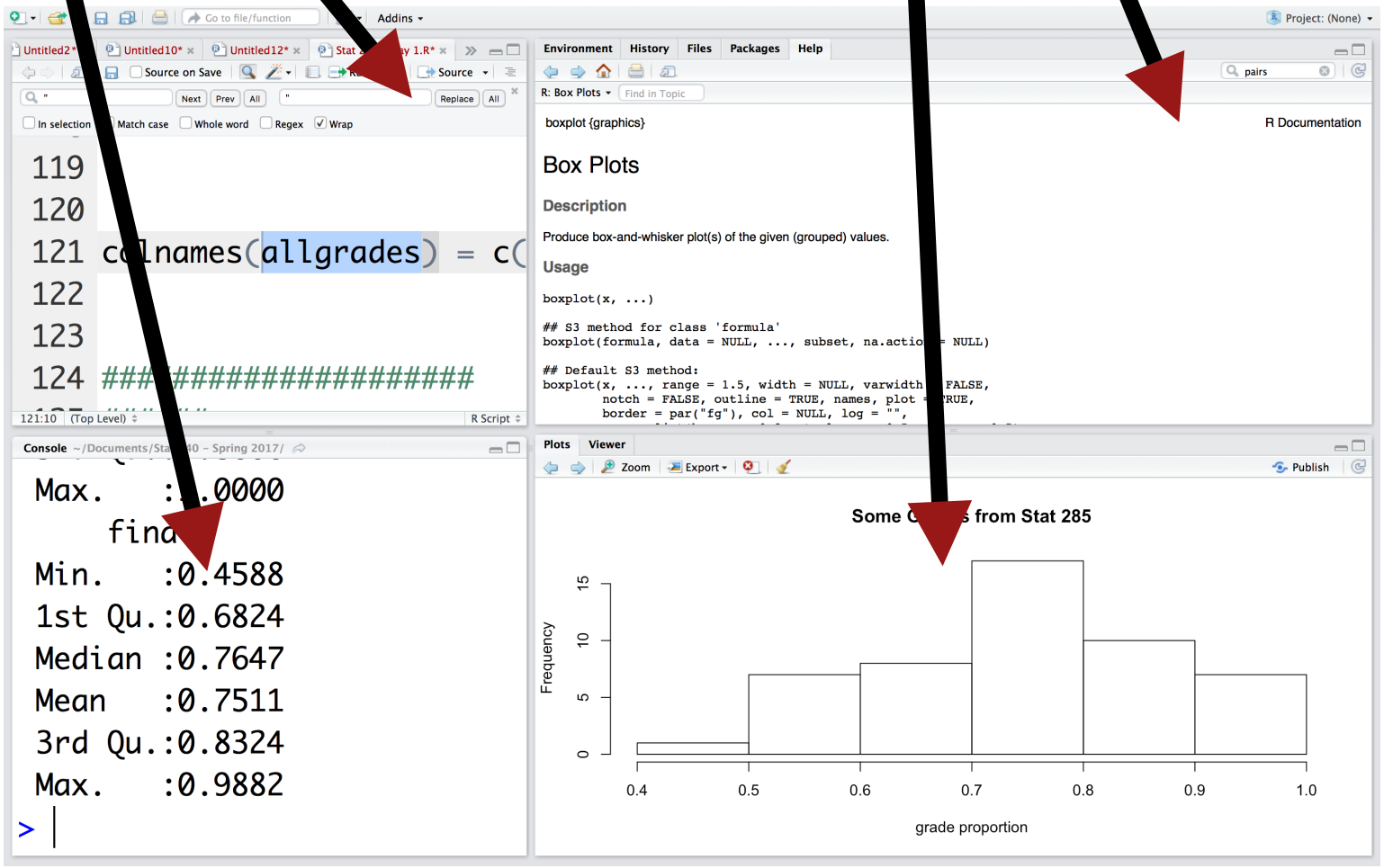
	license_plate_state	license_plate_type	zipcode	violation_code	violation_description	unit	unit_description	vehicle_make
1	IL	PAS	600762418	0964090E	RESIDENTIAL PERMIT PARKING	502	DOF	NISS
2	IL	PAS	604252077	0964080B	NO STANDING/PARKING TIME RESTRICTED	502	DOF	JEEP
3	IL	PAS	94066	0964080B	NO STANDING/PARKING TIME RESTRICTED	502	DOF	HYUN
4	IL	TRK	606255415	0964100A	WITHIN 15' OF FIRE HYDRANT	8	CPD	FORD
5	IL	PAS	606133955	0964090E	RESIDENTIAL PERMIT PARKING	502	DOF	HOND
6	IL	PAS	60623	0964080B	NO STANDING/PARKING TIME RESTRICTED	502	DOF	CHRY

editor: edit code here

console: run code here

help

plots



RStudio: Use the Editor!

ctrl+enter (command+enter) runs a highlighted chunk of text in the editor or a line of code

Where are you?

getwd() #get working directory

setwd(_____) # change the working directory, getwd() first to know the format.

Console

Environment Commands

```
ls()

save.image( file="FileContainsEverything.Rdata")

save(ThisThing,
file="FileJustContainsThisThing.Rdata")

rm(NameSomething2Delete)
```

Getting help for the function “load”

help(load)

?load

args(load)

Plotting the data

histogram of chicago data:

```
hist(chicago[,"fine_level1_amount"],main="Fines"
,xlab="cost”,30) # causes problems for quotes

hist(chicago[,"fine_level1_amount"],main="Fines"
,xlab="cost",30) # no problem
```

Plotting the data

Boxplot of ticket prices:

```
boxplot(chicago[,”fine_level1_amount"],main="First price",ylab="Cost")

boxplot(chicago[,”fine_level2_amount"],main="If unpaid”,ylab="Cost")
```

Story telling

Why would it cost more if the ticket is disputed?

Are tickets being paid?

Dummy variables; logicals as binary

paid (1) vs not yet (many possible reasons) (0)

paid = c(0, 1, 0, 0, 0,….

hist(paid)

Putting 2 Vectors Together

```
tickets = cbind(paid,chicago)

#Just a few rows:

tickets[21:23, ]

#Just the Paid:

tickets[,1]
```

Putting 2 Vectors Together

```
tickets = cbind(paid,chicago)

#Just a few rows:

tickets[21:23, ]

#Just the Tutorials:

tickets[ ,1]

tickets[  ,"paid"]

#Some rows for one specific column:

tickets[21:23  ,”paid”]
```

The Second Dimension!

```
plot(TutsGrades)
```

Control your Plots (see lab)

plotting: you can change the colour (col), the line width (lwd), the point type (pch),…

```
plot(price1,due,col=2,pch=3)

plot(price1,due,col=2,pch=1+paid)

plot(price1,due,col=2+ paid,pch=2)

plot(price1,due,col=2+paid,pch=3)
```

Split data into tutorials and no tutorials - composite logicals

more logicals: ==, != and combining logicals

```
No=c(paid==0 &chicago[,"fine_level1_amount"]<100)

Yes =   c(paid    !=0  &
chicago[     ,"fine_level1_amount"]
<100)
```

Hard to read when the spacing isn't consistent

Split data into bonus and no bonus - composite logicals

more logicals: ==, != and combining logicals

No=c(paid==0 &chicago[,"fine_level1_amount"]<100)

Yes=c(paid!=0 &chicago[,"fine_level1_amount"]<100)

Alignment makes code readable

Control your Plots (see lab)

```
par(mfrow=c(1,2))

boxplot(chicago[paid==1,"fine_level1_amount"],
ylim=c(0,1000),main="paid Ticket Cost")

boxplot(chicago[paid==0,"fine_level1_amount"],
ylim=c(0,1000),main="Unpaid Ticket Cost”)

#Plus reading error messages…
```

```
>  boxplot(chicago[paid==0,"fine_level1_amount"],
ylim=c(0,1000),main=“ Unpaid Ticket Cost”)
Error: unexpected symbol in:
"boxplot(chicago[paid==0,"fine_level1_amount"],
ylim=c(0,1000),main="Unpaid Ticket Cost”)
```

Course Grades Data

```
allgrades = matrix(ncol=4,byrow=T,c(0, 0.5714286, 0.3269231, 0.4588235 ,
                                    1, 0.3928571, 0.6153846, 0.5176471     ,
                                    0, 0.2857143, 0.4038462, 0.5411765     ,
                                    0, 0.3035714, 0.3653846, 0.5411765     ,
                                    0, 0.3214286, 0.6153846, 0.5529412     ,
                                    0, 0.5535714, 0.5769231, 0.5529412     ,
                                    1, 0.6250000, 0.5576923, 0.5647059     ,
....
```

```
colnames(allgrades)=c("assign","MT","Attend","finalEX","final Score")

round(allgrades,3)
```

	assign	MT	Attend	finalEX	final Score
[1,]	5.786	0.000	0.000	0.000	5.786
[2,]	20.905	4.085	2.667	0.000	27.657
[3,]	24.279	0.679	0.000	7.873	32.831
[4,]	40.446	1.378	3.333	5.343	50.501
[5,]	30.175	4.241	3.333	11.545	49.295
[6,]	36.085	1.428	4.000	10.767	52.280

More data commands

matrix(values, nrows, ncols, byrow=T)

colnames(allgrades)

head(allgrades)

Plotting the data

```
    par(mfrow=c(2,2))
hist(allgrades[,"assign"],main="assign")
hist(allgrades[,"MT"],main="MT")
hist(allgrades[,"finalEX"],main="finalEX")
hist(allgrades[,"Attend"],main="Attend")

    #or make box-plots
boxplot(allgrades[,"assign"],main="assign")
boxplot(allgrades[,"MT"],main="MT")
boxplot(allgrades[,"finalEX"],main="finalEX")
boxplot(allgrades[,"Attend"],main="Attend")
```

Split data into tutorial regulars and non-regulars

```
allgrades[,"final"]

allgrades[,"final"]>0

allgrades[allgrades[,"Attend"]>0,"final"]

allgrades[,"Attend"]==0

allgrades[,"Attend"]<4
```

Split data into tutorial regulars and non-regulars (Equivalently, using data.frames)

```
allgrades.dat.frm = as.data.frame(allgrades)

allgrades.dat.frm$final

allgrades.dat.frm$final>0

allgrades.dat.frm[allgrades.dat.frm$final>0,"final"]

allgrades.dat.frm$Attend==0

allgrades.dat.frm$Attend<4
```

Data Frame vs Matrix

	Data Frame	Matrix
Building	data.frame()	matrix(…)
Contents	numbers, factors, non-numbers	just numbers
Math	Might not let you do what you want to do	lets you do Math 232 or Math 240 operations
Converting	as.data.frame()	as.matrix()

Numbers vs Factors

	numbers	factors
Building	data.frame()	as.factor(…)
Contents	1, 3.14159, 1:10,…	categories (levels) like “did bonus”, “didn’t do bonus”
Math	Can do arithmetic, inner products, max, min, mean, var,…	let’s you count contents, can be sorted alphabetically
Converting	as.numeric() turns factors into numbers (*** may lead to confusion)	as.factor() turns numbers into factor levels

Detour back to Jobs

```
Jobs[1:5,c("Title","Experience")]

as.numeric(Jobs[1:5,"Experience"])

# Not as expected…  time to fix….
```

Split data into tutorial regulars and non-regulars - composite logicals

```
NotAllTuts   = allgrades[allgrades[,"final"]>0 &allgrades[,"Attend"]<4,]

AllTuts      = allgrades[allgrades[,"final"]>0 &allgrades[,"Attend"]==4,]
```

Split data into tutorials and not all tutorials

```
#20 histogram bins

par(mfrow=c(2,1))

hist(NotAllTuts[,"final"],20)

hist(      AllTuts[,"final"],20)
```

Split data into tutorials and not all tutorials

```
#20 histogram bins and common x - axis

par(mfrow=c(2,1))

hist(NotAllTuts[,"final"],20,xlim=c(0,100))

hist(      AllTuts[,"final"],20,xlim=c(0,100))
```

Define a sequence

```
seq(from = 1, to = 10, by = 1/1000)

#No need to define variables if they are in the 'standard' order

seq(1,     10,    1/1000)
```

Split data into tutorials and not all tutorials

#20 histogram bins with common boundaries and common x - axis

```
par(mfrow=c(2,1))

hist(NotAllTuts[,"final"],seq(0,100,by=5),xlim=c(0,100))

hist(      AllTuts[,"final"],seq(0,100,by=5),xlim=c(0,100))
```

Did people who were succeeding go to tutorials or did people who went to tutorials succeed?

```
# simple plot first:

pairs(allgrades)

# complex but more insightful

pairs(allgrades,col=(allgrades[,"Attend"]==4)+1)
```

Plot data by splitting on a variable (clunky version)

```
par(mfrow=c(3,2))

boxplot(allgrades[allgrades[,”Attend”]==4,”assign"],
main="Assign");
boxplot(allgrades[allgrades[,”Attend”]!=4,"assign"],
main="Assign");

boxplot(allgrades[allgrades[,"Attend"]==4,"MT"], main="MT");
boxplot(allgrades[allgrades[,"Attend"]!=4,"MT"], main="MT");

boxplot(allgrades[allgrades[,”Attend”]==4,"final"], main=“final");
boxplot(allgrades[allgrades[,”Attend"]!=4,"final"],   main=“final");
```

Plot data by splitting on a variable (simpler version)

```
par(mfrow=c(1,3))

boxplot(allgrades[,"assign"]~allgrades[,"Attend"]==4,
main="assign")

boxplot(allgrades[,"MT"]~allgrades[,"Attend"]==4,
main="MT")

boxplot(allgrades[,"finalEX"]~allgrades[,"Attend"]==4,
main="finalEX")
```

boxplots but add data points

```
par(mfrow=c(3,1))

boxplot(allgrades[,"final"]~allgrades[,"Attend"]!=4, main="final")

points(factor(allgrades[,"Attend"]!=4),allgrades[,"final"],col=4)

boxplot(allgrades[,"MT"]~allgrades[,"Attend"]!=4, main="MT")

points(factor(allgrades[,"Attend"]!=4),allgrades[,"MT"],col=4)

boxplot(allgrades[,"assign"]~allgrades[,"Attend"]!=4,
main="assign")

points(factor(allgrades[,"Attend"]!=4),allgrades[,"assign"],col=4)
```

Did people who were succeeding go to tutorials or did people who went to tutorials succeed?

```
# simple plot first:

pairs(allgrades)

# complex but more insightful

pairs(allgrades,col=(allgrades[,"Attend"]==4)+1)
```

Go on Exchange!

Apply by Jan 25: https://www.sfu.ca/students/studyabroad/exchanges.html

View spaces available: https://www.sfu.ca/content/dam/sfu/students/studyabroad/exchange/2018_19%20Exchange%20Terms%20Available%20-%20Jan%2025%20Deadline.pdf

Make this table https://www.sfu.ca/content/dam/sfu/students/studyabroad/pdf/2019_20%20Exchange%20Terms%20Available%20-%20Jan%2025%20Deadline.pdf into a plot

```
exchanges = read.csv("exchange2019.csv")

exchanges2 = read.csv("exchange2019.2.csv")

attach(exchanges2)

plot(table(Country),las=2,xlab="",ylab="Availability",main="Count of Schools & Terms Available per Country”)

detach(exchanges2)

attach(exchanges)

datause = aggregate(apply(exchanges[,4:6],1,sum), by=list(exchanges$Country),FUN=sum,na.rm=T)

colnames(datause) = c("country","Count of # Schools and Terms Availble for Exchange \n Apply by Jan 25 Application Deadline")

library(rworldmap)

spdf = joinCountryData2Map(datause, joinCode="NAME", nameJoinColumn="country")

here = mapCountryData(spdf, nameColumnToPlot="Count of # Schools and Terms Availble for Exchange \n Apply by Jan 25 Application Deadline", catMethod="fixedWidth",numCats = 53,lwd=1)

do.call(addMapLegend, c(here,sigFigs=2,legendLabels="all",legendIntervals="page"))
```