```
*clear
clear all
*install a library for later
ssc install catplot


*using it directly on the command field
*Stata as a calculator* seeing how it runs commands, input and output

display 2+2
display (3+5) *2

*take two minutes to play on it yourself



** explaining do files* this is what it looks like in do-file in comparison to the command line
display 2+2
display (3+5) *2


**importing data


*importing data* Excel
*first you can use drop down* show drop down



*using code --
*knowing where your files are - file paths*
pwd

import excel "/Users/simhana99/Desktop/Students.xlsx",firstrow
clear



*saving it as a Stata file* change the dta

save "/Users/simhana99/Desktop/Students.dta"
```

opening a stata file drop down or code

```
use "/Users/simhana99/Desktop/Students.dta"
```

getting to know your dataset
data browser/editor seeing the types of variables

code to examine your dataset

```
describe
codebook
codebook Gender
summarize
```

summarizing variables let's look at Gender and SAT

```
sum Gender
tab Gender
sum SAT
```

*Note : you can only find means, standard deviations, etc. with NUMERIC variables

```
tab SAT
mean SAT
```

summarize variables by splitting into groups

```
tab SAT if Gender=="Female"
tab SAT if Age>25
```

telling it specifically what you want -- more complex

```
tabstat SAT, stat(mean sd max min)
tabstat SAT, by(Gender) stat(mean sd max min)
tabstat SAT Age, stat(mean sd max min)
```

*and if and or not commands

```
tab SAT if Gender=="Female"
tab SAT if Gender!="Male"
tab SAT if Gender=="Female" & Age>20
sum SAT if Major=="Econ" | Major=="Politics"
```

comparing two variables - crosstabs

```
tab Gender Major
tab Gender Major, row column
```

take 15 minutes to get to know the dataset here

new variables

renaming variables

```
rename Major major
label variable major "Student's major"
```

creating new variables

```
gen score2= Averagescoregrade/100
```

```
*more complex*

generate age1=.
replace age1=1 if Age>0 & Age<=25
replace age1=2 if Age>25 & Age<=39
tab age1
label define age1 1 "25 or younger" 2 "older than 25"
label values age1 age1

**why is  age1 now a numeric variable and not a string?
codebook age1
tab age1 major

*we want to make another variable numeric instead of a string
encode major, gen(major1)
encode Gender, gen(gender1)
tab major1
numlabel _all, add
tab gender1
tab major1

*why is this helpful??*
*lets make a variable where we split females into poli majors, econ, math

generate female_major=.
replace female_major=1 if major1==1 & gender1==1
replace female_major=2 if major1==2 & gender1==1
replace female_major=3 if major1==3 & gender1==1
label define female_major 1 "female econ" 2 "female math" 3 "female political"
label values female_major female_major
tab female_major
codebook female_major
```

```
*creating dummy variables*

tab female_major, generate(fmajor)


*sorting*
sort SAT

*drop variables*
drop Major

*drop cases*
drop if SAT<1900

*keep cases*
keep if SAT>1900


*visualizing a variable*
histogram Age, frequency
histogram SAT, percent


*graph continuous data*
twoway scatter SAT Age
*line of best fit
twoway scatter SAT Age, || lfit SAT Age


*graph categorical data*
catplot major1 gender1
catplot major1 gender1, percent(major1)


*analysis - chi2 and ttests
tab major1 gender1, chi2
ttest SAT, by(Gender)
```

log files

saving your data replace original data

```
save "/Users/simhana99/Desktop/Students.dta", replace
```

*usually suggest making a new data file

```
save "/Users/simhana99/Desktop/Students_update.dta"
```

merging files
first using drop down

```
merge 1:1 ID using
"/Users/simhana99/Desktop/Students_update.dta"
```

help Stata can always help you with command
*stackexchange

```
help tabstat
```