1.

import delimited "E:\MA 2015\STAT506\ps1\_1.csv", clear // import data

duplicates drop id, force //drop duplicate instances

**Number of distinct rows : 1000 - 48 = 952**

sum weight //summary of variable weight

**Mean of the weight variable = 23.97419**

2.

replace state=subinstr(state,"'","",.) //remove all ' from state

replace state=subinstr(state," ","",.) //remove all space from state

replace state=subinstr(state," ","",.) //remove tab characters from state

replace state=subinstr(state,"(","",.) //remove all ( from state

replace state=subinstr(state,")","",.) //remove all ) from state

replace state=subinstr(state,`"""',"",.) //remove all " from state

replace state="AZ" if state=="arizona" // replace with standard state code Arizona as example

replace state="AZ" if state=="Arizona"

replace state="AZ" if state=="Az"

replace state="AZ" if state=="az"

replace state=”” if state==”unknown” //set the missing values in state for unknown other, etc

tabstat weight,statistics(count mean var) by(state) // summary of weight by state

Summary for variables: weight

by categories of: state (State)

state | N mean variance

------------+------------------------------

AK | 65 24.02723 .8887433

AL | 17 24.18806 .643503

AZ | 53 24.15232 .9332646

CA | 42 23.9525 1.100607

CO | 21 24.07752 .9156415

CT | 68 23.97341 .8446218

DE | 24 23.76842 .9865896

FL | 17 24.38982 .5793145

GA | 30 23.91697 .7918015

HI | 113 24.01686 .8595722

IA | 25 23.86236 .6807452

ID | 21 24.19957 .622905

IL | 21 24.11414 1.074431

IN | 36 23.94733 1.084266

KS | 53 23.99196 1.106582

KY | 27 24.06144 1.181412

LA | 30 23.7457 1.277298

MA | 27 24.01222 .7120282

MD | 30 24.05577 .9165989

ME | 26 24.20592 1.126625

MI | 51 23.90506 .785935

OI | 22 23.91991 .9331523

------------+------------------------------

Total | 819 24.00849 .9099547

-------------------------------------------

3.

split score1score2,p(/) //split scores by the slash in the middle

variables created as string:

score1sco~21 score1sco~22

encode(score1score21),gen(score1) //transfer string variable into numeric variable

encode(score1score22),gen(score2)

replace score1=. if score1score21=="NA" // replace with standard missing values

replace score2=. if score1score21=="NA"

sum score1,detail // find the quantiles of score1

score1

-------------------------------------------------------------

Percentiles Smallest

1% 4 1

5% 12 2

10% 25 2 Obs 947

25% 67 3 Sum of Wgt. 947

50% 200 Mean 191.4646

Largest Std. Dev. 123.614

75% 312 370

90% 351 370 Variance 15280.43

95% 361 370 Skewness -.0588099

99% 370 370 Kurtosis 1.464382

sum score2, detail // find the quantiles of score2

score2

-------------------------------------------------------------

Percentiles Smallest

1% 3 1

5% 11 1

10% 24 1 Obs 945

25% 64 1 Sum of Wgt. 945

50% 175 Mean 181.8762

Largest Std. Dev. 121.8425

75% 302 368

90% 341 368 Variance 14845.59

95% 353 368 Skewness .0144776

99% 365 368 Kurtosis 1.452249

4.

**Some are coded as yyyymmdd, some as mmddyyyy, some as ydm or ymd or mdy or dmy not clear to sort.**

gen visit1\_new=date(visit1, "YMD") // generate standard data format

replace visit1\_new=date(visit1,"MDY") if visit1\_new==. // generate data matching other format

format visit1\_new %td datetime // format to datetime

5.

reg weight visit1\_new visit2\_new visit3\_new visit4\_new visit5\_new // regression weight along visits

Source | SS df MS Number of obs = 880

-------------+---------------------------------- F(5, 874) = 149.13

Model | 371.211773 5 74.2423546 Prob > F = 0.0000

Residual | 435.096241 874 .497821786 R-squared = 0.4604

-------------+---------------------------------- Adj R-squared = 0.4573

Total | 806.308014 879 .917301495 Root MSE = .70556

------------------------------------------------------------------------------

weight | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

visit1\_new | -8.66e-06 .0000224 -0.39 0.699 -.0000526 .0000352

visit2\_new | -.000012 .0000232 -0.52 0.605 -.0000574 .0000335

visit3\_new | 2.46e-06 .0000228 0.11 0.914 -.0000422 .0000471

visit4\_new | .0003077 .0000113 27.26 0.000 .0002856 .0003299

visit5\_new | .0000109 .0000229 0.48 0.634 -.0000341 .0000559

\_cons | 19.62679 .7331614 26.77 0.000 18.18783 21.06575

------------------------------------------------------------------------------

**According to the coefficients and the significance, Visit4 seems to be most strongly related to the weight variable.**