

# DOWNLOADING CENSUS PUMS DATA: USEFUL SOURCES AND AN EXAMPLE

Shulamith Gross

Statistics and Computer Information Systems

Zicklin School of Business

BARUCH COLLEGE

5/15/2007 Baruch Census Workshop

# Some US government Surveys that provide accessible PUMS data

- Please consult your handout number 1
- Not mentioned: Dataferret ---
- A collaboration between The Census Bureau and The Center for Disease Control
- It does not mean the data is particularly health data. The collaboration appears to mostly enhance both data suppliers technically

# Why Bother With FTP files?

- All other methods I have tried, from American FactFinder to The Data Ferret try to force you to select your variables first and do your thinking later. Next to impossible...
- Here, as long as you secure a fast external disk drive, you can download and store all the variables you are likely to want to study, in groups or alone. This I believe is the only way to discover anything really interesting.....

# Handout # 2 / TRANSPORTATION PROGRAM

- In case you wonder---I got plenty of help from Census people in putting together this program
- It makes full use of several states' data from the 5% PUMS CENSUS 2000 datafiles
- This is a fairly sophisticated SAS program to first create the file of transformed variables, derived from both the Person and the Housing records in each state, and then merging those reduced files.
- Then tables of interest are created with correctly estimated counts and percentages

# KNOW YOUR VARIABLES: labels, types, classes, and class labels

- *Take a look at Handout #2* which was provided to me by Census personnel suggested by Len Gaines
- Since the CENSUS 2000 project is rather mature, such FORMAT files exist both at Universities and at the Census Bureau. They very generously share them with new users.

# From IPUM CENSUS 2000 codebook

- Chapter 6.
- Data Dictionary U.S. Census Bureau, Census 2000
- 

<b>Variable name</b>	<b>Character_location</b>	<b>Description</b>
• <b>ACRES</b>	<b>H138</b>	<b>Acreage</b>
• <b>ACRESA</b>	<b>H139</b>	<b>Acreage Allocation Flag</b>
• <b>AGSALES</b>	<b>H140</b>	<b>Sales of Agricultural Products in 1999</b>
• <b>AGSALESA</b> <b>Flag</b>	<b>H141</b>	<b>Sales of Agricultural Products in 1999 Allocation</b>
• <b>AREATYP1</b>	<b>H42-43</b>	<b>Metropolitan Area: Super-PUMA Relationship to MA</b>
• <b>BEDRMS</b>	<b>H124</b>	<b>Number of Bedrooms</b>
• <b>BEDRMSA</b>	<b>H125</b>	<b>Number of Bedrooms Allocation Flag</b>
• <b>BLDGSZ</b>	<b>H115-H116</b>	<b>Size of Building</b>
• <b>BLDGSZA</b>	<b>H117</b>	<b>Size of Building Allocation Flag</b>
• <b>BUSINES</b>	<b>H136</b>	<b>Commercial Business on Property</b>
• <b>BUSINESA</b>	<b>H137</b>	<b>Commercial Business on Property Allocation Flag</b>
• <b>CKITCH</b>	<b>H128</b>	<b>Complete Kitchen Facilities</b>
• <b>CKITCHA</b>	<b>H129</b>	<b>Complete Kitchen Facilities Allocation Flag</b>

# Main points about the program

- Much care taken with variables and their definitions. Heavy use of Formats to create new agglomerated categories from existing categorical variables
- A SAS macro is used to avoid rewriting the same program for 4 different states: NY, NJ, PA, CT. This is of course not absolutely necessary

# FILES CREATED

- For each state to include information about commuters from 3 outside states, NYS, and other boroughs
- Facility with sorting and merging files from the 4 states is required
- The last part of the program that actually uses the PUMS merged file created follows

# The program

- libname census 'D:\2007 CENSUS\';run;
- **proc freq** data=census.pums\_p;
- table incgrp\*trvmns\*home;
- weight pweight;
- where work=1;
- **run;**
- /\* and then use SAS Report facility to create a beautiful table \*/ SEE OUTPUT HANDOUT # 4

# PROJECT #2

- Now that we know who uses Taxis to commute to the City, we ask a simpler question
- Who are the cabbies?
- Create a PUMS file for NYS alone with new variables SEE HANDOUT # 5
- See output in Handout # 6. Output is regular SAS output.

# Much more can be done

- I used only PUMS files from the 5% survey of Census 2000
- Many more questions can be asked, but the data is old
- Other surveys provide microdata even for 2005 and can be used
- You need to devote sometime to surfing the various websites to find what you want
- GOOD LUCK

## HANDOUT # 1

Professor S Gross / Baruch College /  
May 15, 2007/ Census workshop at Baruch College

### POTENTIAL SOURCES OF US MICRODATA

#### American Community Survey (ACS) PUMS:

<http://www.census.gov/acs/www/Products/PUMS/index.htm>

#### Census PUMS:

<http://www.census.gov/main/www/pums.html>

#### Census Survey of Income and Program Participation (SIPP)

*Evaluation of Federal, State and Local Programs*

[http://www.bls.census.gov/sipp\\_ftp.html#sipp](http://www.bls.census.gov/sipp_ftp.html#sipp)

#### Current Population Survey (CPS):

<http://www.census.gov/cps/>

The Current Population Survey (CPS) is a monthly survey of about 50,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics. The survey has been conducted for more than 50 years.

\*\*\*CPS RECORD LAYOUT FOR BASIC LABOR FORCE ITEMS available

### HOUSING RECORD

- *Bedrooms*
- *Condominium status*
- *Contract rent (monthly rent)*
- *Cost of utilities and fuels*
- *Family income*
- *Family, subfamily, and household relationships*
- *Farm status and value*
- *Fire, hazard, and flood insurance*
- *Food Stamps reciprocity*
- *Fuels used*
- *Gross rent*
- *House heating fuel*
- *Household income*
- *Household type*
- *Kitchen facilities*
- *Linguistic isolation\**
- *Meals included in rent*
- *Mortgage status and selected monthly owner costs*

- *Plumbing facilities*
- *Presence and age of own children*
- *Presence of subfamilies in household*
- *Property value*
- *Real estate taxes*
- *Residence State*
- *Rooms*
- *Sewage disposal*
- *Source of water*
- *Telephone in housing unit*
- *Tenure*
- *Units in structure*
- *Vacancy status*
- *Vehicles available*
- *Year householder moved into unit*
- *Year structure built*

#### PERSON RECORD

- *Ability to speak English*
- *Age*
- *Ancestry*
- *Citizenship*
- *Class of worker*
- *Disability status*
- *Educational attainment*
- *Fertility*
- *Hispanic origin*
- *Hours worked*
- *Income by type*
- *Industry*
- *Language spoken at home*
- *Last week work status*
- *Marital status*
- *Means of transportation to work*
- *Migration*
- *Military status, periods of active duty military service, veteran period of service*
- *Mobility status*
- *Occupation*
- *Personal care limitation*
- *Place of birth*
- *Place of work*
- *Poverty status*
- *Race*

- *Relationship*
- *School enrollment and type of school*
- *Sex*
- *Time of departure for work*
- *Travel time to work*
- *Vehicle occupancy*
- *Weeks worked*
- *Work status*
- *Work limitation status*
- *Year of entry*

**Social Security Administration Public Use Microdata Files:**

<http://www.ssa.gov/policy/docs/microdata/>

*Benefits and Earning*

*New Beneficiary Data Systems*

*Old Age, Survivors, and Disability Insurance*

*Social Security Insurance (SSI)*

**Center for Disease Control/National Center for Health Statistics**

<http://www.cdc.gov/nchs/dataawh/ftpserv/ftpdata/ftpdata.htm>

*National Health and Nutrition Examination Survey*

*National Health Care Surveys*

*National Vital Statistics System*

*National Health Interview Survey*

*National Immunization Survey*

*Longitudinal Studies of Aging*

**US Department of Labor – Bureau of Labor Statistics**

Consumer Expenditure Survey Microdata (for purchase on CD)

<http://www.bls.gov/cex/csxmico.htm>

American Time Use Survey Micro Data Files:

<http://www.bls.gov/tus/>

**US Department of Education/IES National Center for Education Statistics**

PEQIS

<http://nces.ed.gov/Surveys/peqis/Download/index.asp>

Other Data Sets – Restricted Use (data for researchers)

<http://nces.ed.gov/pubsearch/getpubcats.asp?sid=031#015>

**US Department of Health and Human Services Medical Expenditure Panel Survey (MEPS)**

[http://www.meps.ahrpr.gov/mepsweb/data\\_stats/download\\_data\\_files.jsp](http://www.meps.ahrpr.gov/mepsweb/data_stats/download_data_files.jsp)

**Panel Study of Income Dynamics**

<http://psidonline.isr.umich.edu/data/>

*Survey of 8000 families on economic, health, and social behavior*

**Integrated Public Use Microdata Series**

<http://usa.ipums.org/usa/intro.shtml>

May 15, 2007 SAS program to merge 4 state microdata files from the 5% survey of Census 2000. 1

```
/* HANDOUT #2. Professor Shula Gross, Baruch College. This program is based on code provided to me by Roy Williams of the Bureau of the Census. Any errors are mine alone*/
```

```
/* May 15, 2007.
```

```
SAS program to prepare the Pums file that provide data for the transportation of commuters into NYC example.
```

```
The program uses both the "H" and the "P" records of PUMS 5% Census 2000 data to count the number of New Yorkers and commuters into NYC, by age, education, income, and age groups. Data from four states was initially used, but only four were retained for extracting data about commuters into NYC*/
```

```
/* the program does not contain detailed SAS REPORT code to produce the tables shown in the commuter transportation by income group, means of transportation, and home location tables */
```

```
filename pums_36 'D:\2007 CENSUS\all_New_York\PUMS5_36.TXT';  
/*filename pums_42 'D:\2007 CENSUS\all_Pennsylvania\PUMS5_42.TXT'; */  
filename pums_34 'D:\2007 CENSUS\all_New_Jersey\PUMS5_34.TXT';  
/*filename pums_09 'D:\2007 CENSUS\all_Connecticut\PUMS5_09.TXT'; */
```

```
options obs=max; /* For testing purposes, put 100 instead of max */
```

```
%macro createpums(stid); /* create a macro-var-name for state code here*/
```

```
data pums_p_&stid.; /* a data step with data-name using a global macro-variable name*/
```

```
infile pums_&stid. MISSOVER LRECL=314;  
/* Here are the variables involved in this small exercise */  
Attrib RACE1 length=$1 label='Race Recode 1';  
attrib RACE2 length=$2 label='Race Recode 2';  
attrib RACE3 length=$2 label='Race Recode 3';  
Attrib RACEA length=$1 label='Race Allocation Flag';  
Attrib OCCEN5 length=$3 label='Occupation (Census) for 5% file';  
Attrib OCCENA length=$1 label='Occupation (Census) Allocation  
Flag';  
Attrib OCCSOC5 length=$7 label='Occupation (SOC) for 5% file';  
Attrib RECTYPE length=$1 label='Record Type';  
Attrib SERIALNO length=$7 label='Housing/Group Quarters (GQ) Unit  
Serial Number'; /* this matches the p and the h records */  
Attrib PNUM length=$2 label='Person Sequence Number';  
Attrib PWEIGHT length= 4 label='Person Weight';  
Attrib RELATE length=$2 label='Relationship';  
Attrib SEX length=$1 label='Sex';  
Attrib AGE length= 4 label='Age';  
Attrib HISPAN length=$2 label='Hispanic or Latino Origin';  
Attrib RACE length=$1 label='Race Recode 1';  
Attrib POWST1 length=$3 label='Place of Work State or Foreign  
Country Code for 1% file';  
Attrib POWPUMA5 length=$5 label='Place of Work PUMA';  
Attrib POWPUMA1 length=$5 label='Place of Work SuperPUMA';  
Attrib TRVMNS length=$2 label='Means of Transportation to Work';  
Attrib TRVTIME length= 4 label='Travel Time to Work';  
Attrib INCTOT length= 8 label='Persons Total Income in 1999';
```

May 15, 2007 SAS program to merge 4 state microdata files from the 5% survey of Census 2000. 2

```
/* two files will be created, one extracting variables from the "P"
(person) records, and one extracting variables from the "H" (housing)
records */
input RECTYPE 1 @;
if RECTYPE = "P" then do;
  input SERIALNO 2-8
    PNUM 9-10
    PWEIGHT 13-16
    RELATE 17-18
    SEX 23
    AGE 25-26
    HISPAN 28-29
    RACE 38
    RACE1 38
      RACE2 39-40
      RACE3 41-42
      RACEA 43
    POWST1 157-159
    POWPUMA5 161-165
    POWPUMA1 166-170
    TRVMNS 191-192
    TRVTIME 200-202
      OCCCEN5 223-225
    OCCCENA 226
    OCCSOC5 227-233
    INCTOT 297-303
  ;

  if powst1 ne "000"; /* Include only those who worked */
  output;
end; /* If "P" Do ... */

Data Pums_h_&stid.; /* here we use the global macro variable-name for
state */

infile pums_&stid. MISSOVER LRECL=314;

Attrib RECTYPE length=$1 label='Record Type';
Attrib SERIALNO length=$7 label='Housing/Group Quarters (GQ) Unit
Serial Number';
Attrib STATE length=$2 label='State Code';
Attrib PUMA5 length=$5 label='Public Use Microdata Area Code
(PUMA)';
Attrib PUMA1 length=$5 label='Super Public Use Microdata Area
Code (SuperPUMA)';
Attrib PERSONS length= 3 label='Number of person records following
this housing record';

input RECTYPE 1 @;
if RECTYPE = "H" then do;
  input SERIALNO 2-8
    STATE 10-11
    PUMA5 14-18
    PUMA1 19-23
    PERSONS 106-107
  ;
  if persons gt 0;
```

```
output;
end;

/****SORT ****/
proc sort data=pums_p_&stid.; /* sort P and H files by SERIALNO */
  by serialno;
run;

proc sort data=pums_h_&stid.;
  by serialno;

data pums_&stid.; /* MERGE H file and mini-P file by SERIALNO */
  merge pums_p_&stid. (in=record) pums_h_&stid.;
  by serialno;
  if record; /* record identifies it is a P-record and not an
H-record */

  select (State); /* Define a variable HOME using state of home for
commuters */
  when ("36") do; /* New York State */
    select (substr(puma5,1,3)); /* select-when is a way to
describe mutually exclusive groups. Substring function selects digit 1
to 3 of char var puma5 */

    when ("038") Home = 1; /* Manhattan */
    when ("037") Home = 2; /* Bronx */
    when ("039") Home = 2; /* S.I. */
    when ("040") Home = 2; /* Brooklyn */
    when ("041") Home = 2; /* Queens */
    otherwise Home = 3;
  end; /* Select Substr */
end; /* When NYS */
when ("34") Home = 4; /* NJ */
when ("42") Home = 5; /* PA */
when ("09") Home = 6; /* CT */
otherwise Home = 9;
end; /* Select State */

/*****create variable work which is location, in NYC of
place of work. The grouping is based on Powst1 in
the merged H and P file for NYS*****/
Select (Powst1);
when ("036") do; /* New York State */
select (substr(powpuma5,1,3));
  when ("038") Work = 1; /* Manhattan */
  when ("037") Work = 2; /* Bronx */
  when ("039") Work = 2; /* S.I. */
  when ("040") Work = 2; /* Brooklyn */
  when ("041") Work = 2; /* Queens */
  otherwise Work = 3;
  end; /* Select Substr */
end; /* When NYS */
when ("034") Work = 4; /* NJ */
when ("042") Work = 5; /* PA */
when ("009") Work = 6; /* CT */
otherwise Work = 9;
end; /* Select State */
```

```

/*****create grouped income variable using IncTot *****/
  if (IncTot < 20000) then IncGrp = 1;
  else if (20000 <= IncTot < 40000) then IncGrp = 2;
  else if (40000 <= IncTot < 75000) then IncGrp = 3;
  else if (75000 <= IncTot < 120000) then IncGrp = 4;
  else if (120000 <= IncTot < 180000) then IncGrp = 5;
  else IncGrp = 6;
/*****create grouped age variable using Age *****/
  if (Age < 30) then AgeGrp = 1;
  else if (30 <= Age < 40) then AgeGrp = 2;
  else if (40 <= Age < 50) then AgeGrp = 3;
  else if (50 <= Age < 60) then AgeGrp = 4;
  else if (60 <= Age < 70) then AgeGrp = 5;
  else AgeGrp = 6;

  /****use created formats****/
  format trvmns $trvmns. home HomeWork. work HomeWork.
  sex $sex. AgeGrp agegrp. incgrp incgrp. race $race.; /* here we
use the proc format supplied earlier */
  run;

%mend; /* End of macro definition */

/* Now run the macro for each state in turn */

/* %createpums(09); */
%createpums(34); /* NJ */
%createpums(36); /* NYS */
/* %createpums(42); */
run; /* completed creating pums files for Ny and NJ */

data census.pums_p; /*CONCATENATE files sequentially to create desired
Pums file in library census*/
  set pums_34 pums_36 /*pums_42 pums_09*/;
run;

proc datasets; /* delete all intermediate files after processing */
  delete /*pums_h_09 pums_p_09 pums_09 */
  pums_h_34 pums_p_34 pums_34
  pums_h_36 pums_p_36 pums_36
  /*pums_h_42 pums_p_42 pums_42*/;
run;

/* Now produce first table of commuters by income group, transportation
mode, and home location */

proc freq data=census.pums_p;
  table incgrp*trvmns*home;
  weight pweight;
  where work=1;
run;
```

**/\* HANDOUT # 3. Code received from Roy Williams of the Bureau of the Census.\*/**

\*---This code can be cannibalized to generate specific value-label formats. You will also need to associate the resulting formats with the variable. In most cases the format code name is the same as the name of the variable.

But format codes can only be 7 characters and cannot end with digits so we have some exceptions, most of which involve using a trailing underscore in the format name to replace the ending digit, e.g. instead of areatyp1 we use the name areatyp\_.

We do not have format for allocation flags nor for many variables with Yes/No values.

Most (but not all) of this code was generated by program gen\_valfmats.sas that read a csv data dictionary file from the Bureau. But *\*very substantial\** post-editing was done on the original output from that program.

Note the library= option on the proc format statement.

Last revised: 6/5/2003 8:10AM.

---\*;

proc format library=pums2000.formats1 length=16; \*<----- library= (optional) used to save these permanently in a format catalog. We plan to store the corresponding format codes for the PUMS 5 pct sample in a catalog named formats5, hence the name formats1 instead of simply formats.

To access these formats you would need to specify : options fmtsearch=(pums2000.formats1 ....)

-----;

\*--The pictca format is used for displaying values that fall within a topcoded range. It uses a comma9. format

followed by the string "(tca)" indicating the number is a (state) topcode average. We reference it using an

other=[pictca.]

spec in several of the value statements that follow. (See condfee, for example)--;

picture pictca

50-high='0,000,099 (tca)';

\*\*\*\*Begin value formats for variables on the H file\*\*\*\*;

Value \$ ACRES

' '='Not in universe (vacant or GQ; occupied and BLDGSZ>3) '  
'1'='Less than 1 acre '  
'2'='1.0 to 9.9 acres '  
'3'='10 acres or more '  
;

Value \$ AGSALES

' '='Not in universe (vacant or GQ; occupied and ACRES=1 or BDLGSZ>3 '

'0'='None '  
'1'='\$1 to \$999 '  
'2'='\$1,000 to \$2,499 '  
'3'='\$2,500 to \$4,999 '  
'4'='\$5,000 to \$9,999 '  
'5'='\$10,000 or more '

```

;
Value $ AREATYP_
'11'='Contains only metropolitan territory inside central city (MSA
part of fully-identified MSA) '
'12'='Contains only metropolitan territory outside central city
(MSA part of fully-identified MSA) '
'13'='Contains only metropolitan territory both inside and outside
central city (MSA part of fully ide '
'14'='Contains an entire MSA (and no other territory) '
'21'='Contains only metropolitan territory inside central city (MSA
part of partially-identified MSA) '
'22'='Contains only metropolitan territory outside central city
(MSA part of partially-identified MSA) '
'23'='Contains only metropolitan territory both inside and outside
central city (MSA part of partially '
'31'='Contains only metropolitan territory inside central city
(PMSA part of fully-identified PMSA and '
'32'='Contains only metropolitan territory outside central city
(PMSA part of fully-identified PMSA an '
'33'='Contains only metropolitan territory both inside and outside
central city (PMSA part of fully-id '
'34'='Contains an entire PMSA (and no other territory) (PMSA
belongs to a fully-identified CMSA) '
'41'='Contains only metropolitan territory inside central city
(PMSA part of fully-identified PMSA and '
'42'='Contains only metropolitan territory outside central city
(PMSA part of fully-identified PMSA an '
'43'='Contains only metropolitan territory both inside and outside
central city (PMSA part of fully-id '
'44'='Contains an entire PMSA (and no other territory) (PMSA
belongs to a partially-identified CMSA) '
'51'='Contains only metropolitan territory inside central city
(PMSA part of partially-identified PMSA '
'52'='Contains only metropolitan territory outside central city
(PMSA part of partially-identified PMS '
'53'='Contains only metropolitan territory both inside and outside
central city (PMSA part of partiall '
'70'='Contains both metropolitan and nonmetropolitan territory '
'80'='Contains only metropolitan territory in two or more partial
and/or entire MSAs/PMSAs/CMSAs '
'90'='Contains only nonmetropolitan territory '

```

```

;
Value $ BEDRMS
. ='Not in universe (GQ) '
0 ='No bedrooms '
5 ='5 or more bedrooms '

```

```

;
Value $ BLDGSZ
' '='Not in universe (GQ) '
'01'='A mobile home '
'02'='A one-family house detached from any other house '
'03'='A one-family house attached to one or more houses '
'04'='A building with 2 apartments '
'05'='A building with 3 or 4 apartments '
'06'='A building with 5 to 9 apartments '
'07'='A building with 10 to 19 apartments '
'08'='A building with 20 to 49 apartments '

```

'09'='A building with 50 or more apartments '  
'10'='Boat, RV, van, etc

WORKERS IN MANHATTAN

Percent of each income group by mode of transportation and location of residence

Transportation	Income	NYC Man	NYC Other	NYS Other	NJ
Car, truck, or van	Under \$40K	6%	15%	34%	27%
	\$40K to \$75K	5%	22%	36%	28%
	\$75K to \$120K	5%	26%	32%	30%
	\$120K to \$250K	5%	25%	28%	30%
	Over \$250K	8%	30%	29%	39%
Bus or trolley bus	Under \$40K	11%	9%	8%	36%
	\$40K to \$75K	11%	12%	5%	34%
	\$75K to \$120K	9%	11%	3%	26%
	\$120K to \$250K	10%	13%	1%	23%
	Over \$250K	9%	8%	2%	16%
Subway	Under \$40K	45%	71%	7%	17%
	\$40K to \$75K	49%	59%	5%	13%
	\$75K to \$120K	48%	56%	3%	10%
	\$120K to \$250K	41%	52%	3%	9%
	Over \$250K	38%	53%	2%	4%
Railroad	Under \$40K	1%	2%	49%	18%
	\$40K to \$75K	0%	3%	54%	23%
	\$75K to \$120K	0%	4%	61%	32%
	\$120K to \$250K	0%	5%	68%	35%
	Over \$250K	0%	4%	65%	37%
Taxicab	Under \$40K	3%	0%	0%	0%
	\$40K to \$75K	4%	0%	0%	0%
	\$75K to \$120K	7%	0%	0%	0%
	\$120K to \$250K	10%	1%	0%	.
	Over \$250K	16%	3%	1%	0%
Walked	Under \$40K	26%	1%	1%	0%
	\$40K to \$75K	24%	1%	0%	0%
	\$75K to \$120K	23%	0%	0%	0%
	\$120K to \$250K	23%	1%	0%	0%
	Over \$250K	21%	1%	0%	.
Other method	Under \$40K	10%	1%	1%	2%
	\$40K to \$75K	8%	2%	0%	2%
	\$75K to \$120K	8%	3%	0%	2%
	\$120K to \$250K	10%	4%	0%	3%
	Over \$250K	7%	1%	1%	3%

WORKERS IN MANHATTAN

Percent of each income group by mode of transportation and location of residence

Income	Transportation	NYC Man	NYC Other	NYS Other	NJ
Under \$40K	Car, truck, or van	6%	15%	34%	27%
	Bus or trolley bus	11%	9%	8%	36%
	Subway	45%	71%	7%	17%
	Railroad	1%	2%	49%	18%
	Taxicab	3%	0%	0%	0%
	Walked	26%	1%	1%	0%
	Other method	10%	1%	1%	2%
\$40K to \$75K	Car, truck, or van	5%	22%	36%	28%
	Bus or trolley bus	11%	12%	5%	34%
	Subway	49%	59%	5%	13%
	Railroad	0%	3%	54%	23%
	Taxicab	4%	0%	0%	0%
	Walked	24%	1%	0%	0%
	Other method	8%	2%	0%	2%
\$75K to \$120K	Car, truck, or van	5%	26%	32%	30%
	Bus or trolley bus	9%	11%	3%	26%
	Subway	48%	56%	3%	10%
	Railroad	0%	4%	61%	32%
	Taxicab	7%	0%	0%	0%
	Walked	23%	0%	0%	0%
	Other method	8%	3%	0%	2%
\$120K to \$250K	Car, truck, or van	5%	25%	28%	30%
	Bus or trolley bus	10%	13%	1%	23%
	Subway	41%	52%	3%	9%
	Railroad	0%	5%	68%	35%
	Taxicab	10%	1%	0%	.
	Walked	23%	1%	0%	0%
	Other method	10%	4%	0%	3%
Over \$250K	Car, truck, or van	8%	30%	29%	39%
	Bus or trolley bus	9%	8%	2%	16%
	Subway	38%	53%	2%	4%
	Railroad	0%	4%	65%	37%
	Taxicab	16%	3%	1%	0%
	Walked	21%	1%	0%	.
	Other method	7%	1%	1%	3%

WORKERS IN MANHATTAN

Count of each income group by mode of transportation and location of residence

Transportation	Income	NYC Man	NYC Other	NYS Other	NJ
Car, truck, or van	Under \$40K	18,586	86,584	24,396	21,496
	\$40K to \$75K	8,225	52,606	31,965	23,964
	\$75K to \$120K	4,097	14,270	15,403	14,108
	\$120K to \$250K	2,243	3,629	6,422	5,394
	Over \$250K	4,050	2,500	7,585	8,104
Bus or trolley bus	Under \$40K	32,069	53,605	5,632	29,125
	\$40K to \$75K	17,380	28,218	4,113	29,117
	\$75K to \$120K	6,930	6,358	1,599	11,834
	\$120K to \$250K	4,351	1,850	325	4,096
	Over \$250K	4,525	658	508	3,423
Subway	Under \$40K	135,278	416,217	5,207	13,815
	\$40K to \$75K	79,699	140,555	4,058	10,828
	\$75K to \$120K	36,309	31,307	1,575	4,417
	\$120K to \$250K	17,812	7,365	581	1,582
	Over \$250K	18,455	4,394	576	820
Railroad	Under \$40K	1,584	13,198	35,222	14,535
	\$40K to \$75K	512	8,191	47,721	19,737
	\$75K to \$120K	201	2,247	29,144	14,741
	\$120K to \$250K	59	730	15,740	6,138
	Over \$250K	155	342	16,762	7,693
Taxicab	Under \$40K	8,475	2,718	98	145
	\$40K to \$75K	6,774	713	68	107
	\$75K to \$120K	5,432	14	72	31
	\$120K to \$250K	4,451	100	18	.
	Over \$250K	7,794	210	246	25
Walked	Under \$40K	77,469	5,154	571	338
	\$40K to \$75K	38,858	1,326	291	226
	\$75K to \$120K	17,419	129	134	93
	\$120K to \$250K	10,009	86	101	46
	Over \$250K	10,330	67	97	.
Other method	Under \$40K	29,251	8,435	468	1,402
	\$40K to \$75K	12,523	4,938	336	1,665
	\$75K to \$120K	5,971	1,542	74	1,144
	\$120K to \$250K	4,441	504	53	517
	Over \$250K	3,405	55	147	705

WORKERS IN MANHATTAN

Count of each income group by mode of transportation and location of residence

Income	Transportation	NYC Man	NYC Other	NYS Other	NJ
Under \$40K	Car, truck, or van	18,586	86,584	24,396	21,496
	Bus or trolley bus	32,069	53,605	5,632	29,125
	Subway	135,278	416,217	5,207	13,815
	Railroad	1,584	13,198	35,222	14,535
	Taxicab	8,475	2,718	98	145
	Walked	77,469	5,154	571	338
	Other method	29,251	8,435	468	1,402
\$40K to \$75K	Car, truck, or van	8,225	52,606	31,965	23,964
	Bus or trolley bus	17,380	28,218	4,113	29,117
	Subway	79,699	140,555	4,058	10,828
	Railroad	512	8,191	47,721	19,737
	Taxicab	6,774	713	68	107
	Walked	38,858	1,326	291	226
	Other method	12,523	4,938	336	1,665
\$75K to \$120K	Car, truck, or van	4,097	14,270	15,403	14,108
	Bus or trolley bus	6,930	6,358	1,599	11,834
	Subway	36,309	31,307	1,575	4,417
	Railroad	201	2,247	29,144	14,741
	Taxicab	5,432	14	72	31
	Walked	17,419	129	134	93
	Other method	5,971	1,542	74	1,144
\$120K to \$250K	Car, truck, or van	2,243	3,629	6,422	5,394
	Bus or trolley bus	4,351	1,850	325	4,096
	Subway	17,812	7,365	581	1,582
	Railroad	59	730	15,740	6,138
	Taxicab	4,451	100	18	.
	Walked	10,009	86	101	46
	Other method	4,441	504	53	517
Over \$250K	Car, truck, or van	4,050	2,500	7,585	8,104
	Bus or trolley bus	4,525	658	508	3,423
	Subway	18,455	4,394	576	820
	Railroad	155	342	16,762	7,693
	Taxicab	7,794	210	246	25
	Walked	10,330	67	97	.
	Other method	3,405	55	147	705

**/\* HANDOUT # 5 / Professor Shula gross, Baruch College / May 15, 2007 /**

**Program to create the file needed to analyze the income, gender, education, and age groups of Manhattan taxi Drivers. Only NYS Census 2000 PUMS data is used. The output is PROC FREQ output\*/**

**proc format;**

**Value \$ EDUC**

'00'='Not in universe (Under 3 years) '  
'01'='No schooling completed '  
'02'='Nursery school to 4th grade '  
'03'='5th grade or 6th grade '  
'04'='7th grade or 8th grade '  
'05'='9th grade '  
'06'='10th grade '  
'07'='11th grade '  
'08'='12th grade, no diploma '  
'09'='High school graduate '  
'10'='Some college, but less than 1 year '  
'11'='One or more years of college, no degree '  
'12'='Associate degree '  
'13'='Bachelor s degree '  
'14'='Master s degree '  
'15'='Professional degree '  
'16'='Doctorate degree '

**;**

**Value \$ SEX**

'1'='Male '  
'2'='Female '

**;**

**Value \$ HISPAN**

'01'='Not Hispanic or Latino '  
'02'='Mexican '  
'03'='Puerto Rican '  
'04'='Cuban '  
'05'='Dominican '  
'06'='Costa Rican '  
'07'='Guatemalan '  
'08'='Honduran '  
'09'='Nicaraguan '  
'10'='Panamanian '  
'11'='Salvadoran '  
'12'='Other Central American '  
'13'='Argentinean '

'14'='Bolivian '  
'15'='Chilean '  
'16'='Colombian '  
'17'='Ecuadoran '  
'18'='Paraguayan '  
'19'='Peruvian '  
'20'='Uruguayan '  
'21'='Venezuelan '  
'22'='Other South American '  
'23'='Spaniard '  
'24'='Other Spanish or Latino '

;

**Value \$ RACE**

'1'='White alone '  
'2'='Black or African American alone '  
'3'='American Indian alone '  
'4'='Alaska Native alone '  
'5'='American Indian and Alaska Native tribes specified, and American Indian or  
Alaska Native, not sp '  
'6'='Asian alone '  
'7'='Native Hawaiian and Other Pacific Islander alone '  
'8'='Some other race alone '  
'9'='Two or more major race groups '

;

**Value \$ TRVMNS**

'00'='Not in universe '  
'01'='Car, truck, or van '  
'02'='Bus or trolley bus '  
'03'='Streetcar '  
'04'='Subway '  
'05'='Railroad '  
'06'='Ferryboat '  
'07'='Taxicab '  
'08'='Motorcycle '  
'09'='Bicycle '  
'10'='Walked '  
'11'='Worked at home '  
'12'='Other method '

;

**Value HomeWork**

1 = 'NYC Manhattan'  
2 = 'NYC Other Boro'  
3 = 'NY Suburb'

4 = 'NJ'  
5 = 'PA'  
6 = 'CT'  
9 = 'Other'  
;

Value AgeGrp  
1 = 'Under 30'  
2 = "30's"  
3 = "40's"  
4 = "50's"  
5 = "60's"  
6 = "Over 70"  
;

Value IncGrp  
1 = 'Under \$20K'  
2 = '\$20K to \$40K'  
3 = '\$40K to \$75K'  
4 = '\$75K to \$120K'  
5 = '\$120K to \$180K'  
6 = 'Over \$180K'  
;

run;

data cabbies;

set Census.pums\_p;

Attrib RACE1 length=\$1 label='Race Recode 1';  
Attrib OCCCEN5 length=\$3 label='Occupation (Census) for 5% file';  
Attrib OCCCENA length=\$1 label='Occupation (Census) Allocation Flag';  
Attrib OCCSOC5 length=\$7 label='Occupation (SOC) for 5% file';  
Attrib RECTYPE length=\$1 label='Record Type';  
Attrib SERIALNO length=\$7 label='Housing/Group Quarters (GQ) Unit Serial  
Number'; /\* this matches the p and the h records \*/  
Attrib PNUM length=\$2 label='Person Sequence Number';  
Attrib PWEIGHT length= 4 label='Person Weight';  
Attrib RELATE length=\$2 label='Relationship';  
Attrib SEX length=\$1 label='Sex';  
Attrib AGE length= 4 label='Age';  
Attrib HISPAN length=\$2 label='Hispanic or Latino Origin';  
Attrib RACE1 length=\$1 label='Race Recode 1';  
Attrib POWST1 length=\$3 label='Place of Work State or Foreign Country Code  
for 1% file';  
Attrib POWPUMA5 length=\$5 label='Place of Work PUMA';  
Attrib POWPUMA1 length=\$5 label='Place of Work SuperPUMA';  
Attrib TRVMNS length=\$2 label='Means of Transportation to Work';

```
Attrib TRVTIME length= 4 label='Travel Time to Work';
Attrib INCTOT length= 8 label='Persons Total Income in 1999';
```

```
if (IncTot < 20000) then IncGrp = 1;
else if (20000 <= IncTot < 40000) then IncGrp = 2;
else if (40000 <= IncTot < 75000) then IncGrp = 3;
else if (75000 <= IncTot < 120000) then IncGrp = 4;
else if (120000 <= IncTot < 180000) then IncGrp = 5;
else IncGrp = 6;
```

```
If (educ='00' or educ= '01' or educ= '02' or educ= '03' or educ= '04' or educ= '05' or
educ= '06' or educ= '07' or educ= '08' or educ= '09') then edu1='HS grad or less';
Else if (educ='00' or educ= '11' or educ= '12') then edu1=
'up to associate degree';
Else if (educ='13') then edu1='college graduate';
Else edu1='advanced degree';
```

```
if (Age < 30) then AgeGrp = 1;
else if (30 <= Age < 40) then AgeGrp = 2;
else if (40 <= Age < 50) then AgeGrp = 3;
else if (50 <= Age < 60) then AgeGrp = 4;
else if (60 <= Age < 70) then AgeGrp = 5;
else AgeGrp = 6;
```

```
format trvmns $trvmns.
```

```
sex $sex. AgeGrp agegrp. incgrp incgrp. Race1 $race.; /* here we use the proc
format supplied earlier */
```

```
if occen5 = "914" and substr(powpuma5,1,3)="038";
run;
```

```
proc means data=cabbies;
var hours trvtime inctot;
weight pweight;
run;
```

```
proc sort data=cabbies out=sort_cabbies;
by sex;
run;
```

```
proc means data=cabbies;
by sex;
var hours trvtime inctot;
weight pweight;
run;
```

```
proc freq data=sort_cabbies;  
title 'CABBIES IN MAHATTEN 1999';  
table incgrp*agegrp*edu1;  
by sex;  
weight pweight;  
run;
```

```
proc freq data=sort_cabbies;  
title 'CABBIES IN MAHATTEN 1999';  
table incgrp*agegrp*edu1;  
weight pweight;  
run;
```

----- Sex=Male The FREQ Procedure

Table 1 of AgeGrp by edu1 Controlling for IncGrp=Under \$20K

AgeGrp edu1

Frequency					
Percent					
Row Pct					
Col Pct	HS grad or less	advanced degree	college graduat	up to as sociate	Total
Under 30	884	48	74	56	1062
	14.07	0.76	1.18	0.89	16.91
	83.24	4.52	6.97	5.27	
	21.86	8.09	12.35	5.35	
30's	1199	238	346	363	2146
	19.09	3.79	5.51	5.78	34.16
	55.87	11.09	16.12	16.92	
	29.65	40.13	57.76	34.70	
40's	1160	121	111	486	1878
	18.47	1.93	1.77	7.74	29.89
	61.77	6.44	5.91	25.88	
	28.68	20.40	18.53	46.46	
50's	686	110	68	115	979
	10.92	1.75	1.08	1.83	15.58
	70.07	11.24	6.95	11.75	
	16.96	18.55	11.35	10.99	
60's	106	76	0	26	208
	1.69	1.21	0.00	0.41	3.31
	50.96	36.54	0.00	12.50	
	2.62	12.82	0.00	2.49	
Over 70	9	0	0	0	9
	0.14	0.00	0.00	0.00	0.14
	100.00	0.00	0.00	0.00	
	0.22	0.00	0.00	0.00	
Total	4044	593	599	1046	6282
	64.37	9.44	9.54	16.65	100.00

Sex=Male -----The FREQ Procedure

Table 2 of AgeGrp by edu1 Controlling for IncGrp=\$20K to \$40K

AgeGrp	edu1				Total
Frequency	HS grad or less	advanced degree	college graduat	up to as sociate	
Percent					
Row Pct					
Col Pct					
Under 30	266	0	32	52	350
	5.23	0.00	0.63	1.02	6.88
	76.00	0.00	9.14	14.86	
	7.55	0.00	6.06	8.89	
30's	1086	207	210	259	1762
	21.34	4.07	4.13	5.09	34.63
	61.63	11.75	11.92	14.70	
	30.83	45.70	39.77	44.27	
40's	1346	153	246	161	1906
	26.45	3.01	4.83	3.16	37.46
	70.62	8.03	12.91	8.45	
	38.22	33.77	46.59	27.52	
50's	582	75	40	113	810
	11.44	1.47	0.79	2.22	15.92
	71.85	9.26	4.94	13.95	
	16.52	16.56	7.58	19.32	
60's	196	0	0	0	196
	3.85	0.00	0.00	0.00	3.85
	100.00	0.00	0.00	0.00	
	5.57	0.00	0.00	0.00	
Over 70	46	18	0	0	64
	0.90	0.35	0.00	0.00	1.26
	71.88	28.13	0.00	0.00	
	1.31	3.97	0.00	0.00	
Total	3522	453	528	585	5088
	69.22	8.90	10.38	11.50	100.00

----- The FREQ Procedure ----- Sex=Male -----

Table 3 of AgeGrp by edu1 Controlling for IncGrp=\$40K to \$75K

AgeGrp	edu1				Total
Frequency	HS grad or less	advanced degree	college graduat	up to as sociate	
Percent					
Row Pct					
Col Pct					
Under 30	33 1.73 43.42 3.38	17 0.89 22.37 6.20	26 1.36 34.21 9.63	0 0.00 0.00 0.00	76 3.98
30's	159 8.33 39.36 16.29	110 5.77 27.23 40.15	15 0.79 3.71 5.56	120 6.29 29.70 30.93	404 21.17
40's	452 23.69 57.14 46.31	85 4.45 10.75 31.02	77 4.04 9.73 28.52	177 9.28 22.38 45.62	791 41.46
50's	251 13.16 46.65 25.72	62 3.25 11.52 22.63	134 7.02 24.91 49.63	91 4.77 16.91 23.45	538 28.20
60's	81 4.25 81.82 8.30	0 0.00 0.00 0.00	18 0.94 18.18 6.67	0 0.00 0.00 0.00	99 5.19
Total	976 51.15	274 14.36	270 14.15	388 20.34	1908 100.00

Sex=Male

The FREQ Procedure Table 4 of AgeGrp by edu1  
 Controlling for IncGrp=\$75K to \$120K

AgeGrp	edu1				Total
	HS grad or less	advanced degree	college graduat	up to as sociate	
Under 30	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
30's	48 11.68 39.67 18.97	0 0.00 0.00 0.00	0 0.00 0.00 0.00	73 17.76 60.33 64.60	121 29.44
40's	129 31.39 60.28 50.99	24 5.84 11.21 100.00	21 5.11 9.81 100.00	40 9.73 18.69 35.40	214 52.07
50's	56 13.63 100.00 22.13	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	56 13.63
Over 70	20 4.87 100.00 7.91	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	20 4.87
Total	253 61.56	24 5.84	21 5.11	113 27.49	411 100.00

----- Sex=Male -----

The FREQ Procedure

Table 5 of AgeGrp by edu1  
 Controlling for IncGrp=\$120K to \$180K

AgeGrp	edu1				Total
Frequency	HS grad	advanced	college	up to as	
Percent	or less	degree	graduat	sociate	
Row Pct					
Col Pct					
Under 30	0 0.00 . .	0 0.00 . .	0 0.00 . .	0 0.00 0.00	0 0.00
30's	0 0.00 . .	0 0.00 . .	0 0.00 . .	0 0.00 0.00	0 0.00
40's	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
50's	0 0.00 0.00 . .	0 0.00 0.00 . .	0 0.00 0.00 . .	18 100.00 100.00 100.0	18 100.00
Total	0 0.00	0 0.00	0 0.00	18 100.00	18 100.00

----- Sex=Female -----

The FREQ Procedure  
 Table 1 of AgeGrp by edu1  
 Controlling for IncGrp=Under \$20K

AgeGrp	edu1				Total
Frequency	HS grad or less	advanced degree	college graduat	up to as sociate	
Percent					
Row Pct					
Col Pct					
30's	32	0	0	0	32
	26.23	0.00	0.00	0.00	26.23
	100.00	0.00	0.00	0.00	
	29.91	.	.	0.00	
40's	30	0	0	15	45
	24.59	0.00	0.00	12.30	36.89
	66.67	0.00	0.00	33.33	
	28.04	.	.	100.00	
50's	24	0	0	0	24
	19.67	0.00	0.00	0.00	19.67
	100.00	0.00	0.00	0.00	
	22.43	.	.	0.00	
60's	21	0	0	0	21
	17.21	0.00	0.00	0.00	17.21
	100.00	0.00	0.00	0.00	
	19.63	.	.	0.00	
Total	107	0	0	15	122
	87.70	0.00	0.00	12.30	100.00

----- Sex=Female -----

## The FREQ Procedure

Table 2 of AgeGrp by edu1

Controlling for IncGrp=\$20K to \$40K

AgeGrp	edu1				Total
Frequency	HS grad or less	advanced degree	college graduat	up to as sociate	
Percent					
Row Pct					
Col Pct					
30's	0 0.00 0.00 .	26 35.14 44.07 100.00	21 28.38 35.59 58.33	12 16.22 20.34 100.00	59 79.73
40's	0 0.00 0.00 .	0 0.00 0.00 0.00	15 20.27 100.00 41.67	0 0.00 0.00 0.00	15 20.27
50's	0 0.00 . .	0 0.00 . 0.00	0 0.00 . 0.00	0 0.00 . 0.00	0 0.00
60's	0 0.00 . .	0 0.00 . 0.00	0 0.00 . 0.00	0 0.00 . 0.00	0 0.00
Total	0 0.00	26 35.14	36 48.65	12 16.22	74 100.00

----- Sex=Female -----

The FREQ Procedure

Table 3 of AgeGrp by edu1

Controlling for IncGrp=\$40K to \$75K

AgeGrp	edu1				Total
	HS grad or less	advanced degree	college graduat	up to as sociate	
30's	0 0.00 0.00 0.00	0 0.00 0.00 .	0 0.00 0.00 .	17 58.62 100.00 100.00	17 58.62
40's	0 0.00 . 0.00	0 0.00 . .	0 0.00 . .	0 0.00 . 0.00	0 0.00
50's	12 41.38 100.00 100.00	0 0.00 0.00 .	0 0.00 0.00 .	0 0.00 0.00 0.00	12 41.38
60's	0 0.00 . 0.00	0 0.00 . .	0 0.00 . .	0 0.00 . 0.00	0 0.00
Total	12 41.38	0 0.00	0 0.00	17 58.62	29 100.00

The MEANS Procedure

Variable Minimum	Label	N	Mean	Std Dev
HOURS 0	Hours per Week in 1999	551	43.7496411	99.4566580
TRVTIME 0	Travel Time to Work	551	38.9124318	140.0670694
INCTOT -9490.00	Persons Total Income in 1999	551	25027.06	100108.42

Variable	Label	Maximum
HOURS	Hours per Week in 1999	99.0000000
TRVTIME	Travel Time to Work	155.0000000
INCTOT	Persons Total Income in 1999	167000.00

CABBIES IN MAHATTEN 1999

----- Sex=Male -----

The MEANS Procedure

Variable Minimum	Label	N	Mean	Std Dev
HOURS	Hours per Week in 1999	540	43.8756840	99.7102311
TRVTIME	Travel Time to Work	540	38.9761436	141.0706293
INCTOT -9490.00	Persons Total Income in 1999	540	25069.53	100548.15

Variable	Label	Maximum
HOURS	Hours per Week in 1999	99.0000000
TRVTIME	Travel Time to Work	155.0000000
INCTOT	Persons Total Income in 1999	167000.00

----- Sex=Female -----

Variable	Label	N	Mean	Std Dev
HOURS	Hours per Week in 1999	11	36.07111111	82.5244583
TRVTIME	Travel Time to Work	11	35.03111111	77.6458513
INCTOT	Persons Total Income in 1999	11	22439.56	78198.12

Variable	Label	Maximum
HOURS	Hours per Week in 1999	60.0000000
TRVTIME	Travel Time to Work	60.0000000
INCTOT	Persons Total Income in 1999	67500.00