ACTIVITIES WE GIVE UP WHEN WE'RE ONLINE

Measuring the Crowd Out Effect of Online Leisure Activity

BACKGROUND

- Since the dot-com revolution, the internet has become an integral part of our life.
- Some activities, like reading the newspaper or watching movies, are now being done through the internet
- Other activities, like surfing the web for personal interest or checking Twitter, Facebook, Instagram, are new activities that have emerged with the dot-com revolution
- **How valuable are the activities we do online as opposed to offline ones?**

LITERATURE
WALLSTEN, SCOTT. *WHAT ARE WE NOT DOING WHEN WE'RE ONLINE*. NO. W19549. NATIONAL BUREAU OF ECONOMIC RESEARCH, 2013.

- "Estimating the value of the Internet is difficult in part not just because many online activities do not require monetary payment, but also because these activities may crowd out other, offline, activities."
- This paper "estimates the opportunity cost of online leisure time. The analysis suggests that the opportunity cost of online leisure is less time spent on a variety of activities, including leisure, sleep, and work."

Opportunity Cost of Online Leisure Time = Time Foregone on Other Activity

CROWD-OUT EFFECT

WALLSTEN'S DATASET: ATUS

The American Time Use Survey interviews respondents about:

- how they spent their time on the previous day
- where they were
- whom they were with

- The goal is to measure how people divide their time among life's activities.
- Individuals are randomly selected from a subset of households previously interviewed in the Current Population Survey (CPS).

WALLSTEN'S METHODOLOGY

- Using ATUS dataset for years 2003-2011, he estimates 18 versions of equations (one for each major activity and one for an unknown activity)
- He uses the coefficient (and t-statistic) on the computer leisure variable from each regression as a measurement of the crowd-out effect of computer leisure on each major category

$$\text{major activity}_i = f\begin{pmatrix} \text{computer leisure}_i, \text{income}_i, \text{education}_i, \text{age}_i, \text{sex}_i, \text{race}_i, \text{married}_i, \\ \text{number children in household}_i, \text{occupation}_i \\ \text{Spanish-speaking only}_i, \text{labor force status}_i, (\text{metro, suburban, rural})_i, \\ \text{year}_t, \text{survey day of week}_i \end{pmatrix}$$

CHALLENGE – 'WHO HAS INTERNET?'

- "While I know the ages of all household members, the data do not indicate whether a household has Internet access."
- However, I can identify some households that have access. In particular, any ATUS respondent who spends any time at home involved in computer leisure, e- mail, or using a computer for volunteer work must have home Internet access. Following Goldfarb and Prince, I estimate the following two simultaneous equations using two-stage least squares:

$$
\begin{aligned}
&(1)\ \text{home internet access}_i = \\
&\quad f\begin{pmatrix} \text{income}_i, \text{education}_i, \text{age}_i, \text{sex}_i, \text{race}_i, \text{married}_i, \text{number children in household}_i, \\ \text{Spanish-speaking only}_i, \text{labor force status}_i, (\text{metro, suburban, rural})_i, \\ \text{leisure excluding computer use}_i, \text{year}_t, \text{survey day of week}_i, \text{teenager in house}_i \end{pmatrix} \\
&(2)\ \text{computer use for leisure}_i = f\left((\boldsymbol{Z}), \widehat{\text{home internet access}}_i\right)
\end{aligned}
$$

‘WHO HAS INTERNET?’ - IMPLICATIONS

- His method implied that only 17 percent of households had access to the internet in 2010, when the US Census estimated that more than 70 percent actually had access.

- **For that reason, we will be taking a different approach for estimating the internet access variable on the ATUS dataset.**

- However, “The fitted propensity to have access increases by about 70 percent while actual home Internet access increased by about 78 percent* during that same time period.”

* As per PEW RESEARCH data

‘WHO HAS INTERNET?’ – MY APPROACH: BUILDING A DECISION TREE CLASSIFIER

- The ATUS dataset is in fact a subsample of a larger dataset: CPS (Current Population Survey).
- Unlike the ATUS dataset, the CPS dataset includes a variable that indicates whether a subject has internet access or not.
- As such, we will be using similar type of variables as the one used in Wallsten's regression, and will look at the common variables found in both the ATUS and CPS datasets in order to construct a decision tree algorithm, classifiying our records into: subject has internet access or subject does not have internet access.

CPS DATASET – COMPUTER AND INTERNET USE SURVEY

- The Current Population Survey (CPS) interviews around 56,000 households monthly, scientifically selected on the basis of area of residence to represent the nation as a whole, individual states, and other specified areas.
- The main purpose of the survey is to collect information on the employment situation, as well as other information on demographic characteristics such as age, sex, race, marital status, educational attainment, family relationship, occupation and industry etc.
- Starting 2011, CPS has included the optional Computer and Internet Use Survey. This was done for the years 2011, 2013, 2015. We will be using the aforestated datasets.

MY METHODOLOGY VS. WALLSTEN'S

	Wallsten	My Methodology
Datasets	ATUS	ATUS and CPS
Internet Prediction	Two Stage Regression	Decision Tree Classifier
Crowd Out Effect Measure	Linear Regression and Coefficient Analysis	Linear Regression and Coefficient Analysis

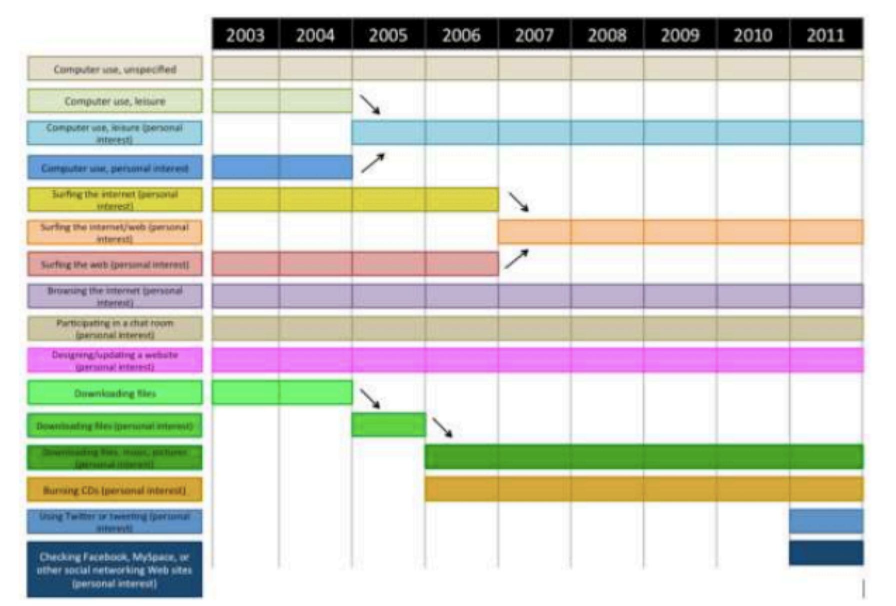
VARIABLES

Variables Related To	ATUS Variables CODE	CPS Variables CODE
Income	Weekly earnings TRERNWA	Weekly earnings PRERNWA
Education	Highest level of school completed PEEDUCA	Highest level of school completed PEEDUCA
Age	Age PEAGE	Age PEAGE
Sex	Sex PESEX	Sex PESEX
Married	Presence of Spouse in the household TRSPRES	Presence of Spouse in the household TRSPRES
Number of Children in the Household and Age of Youngest Child	Number of household children TRCHILDNUM	Number of household children TRCHILDNUM
	Age of Youngest Child TRYHHCHILD	Age of youngest child PRCHLD
Spanish-Speaking	Are you Spanish, Hispanic or Latino? PEHSPNON	Are you Spanish, Hispanic or Latino? PEHSPNON
Labor Force Status	Total hours worked per week TEHRUSLT	Total hours worked per week PEHRUSLT
	Do you have more than one job? TEMJOT	Do you have more than one job? PEMJOT
	Labor force status (Employed, Unemployed etc.) TELFS	Labor Force Status PEMLR
Metropolitan Status (Metro, Suburban, Rural)	Metropolitan Status (2000 and 2010 definitions) GTMETSA	Metropolitan Status (2000 and 2010 definitions) GTMETSA

VARIABLES YEARS 2011 TO 2015

Smaller Sample than Wallsten's sample. Two Reasons:

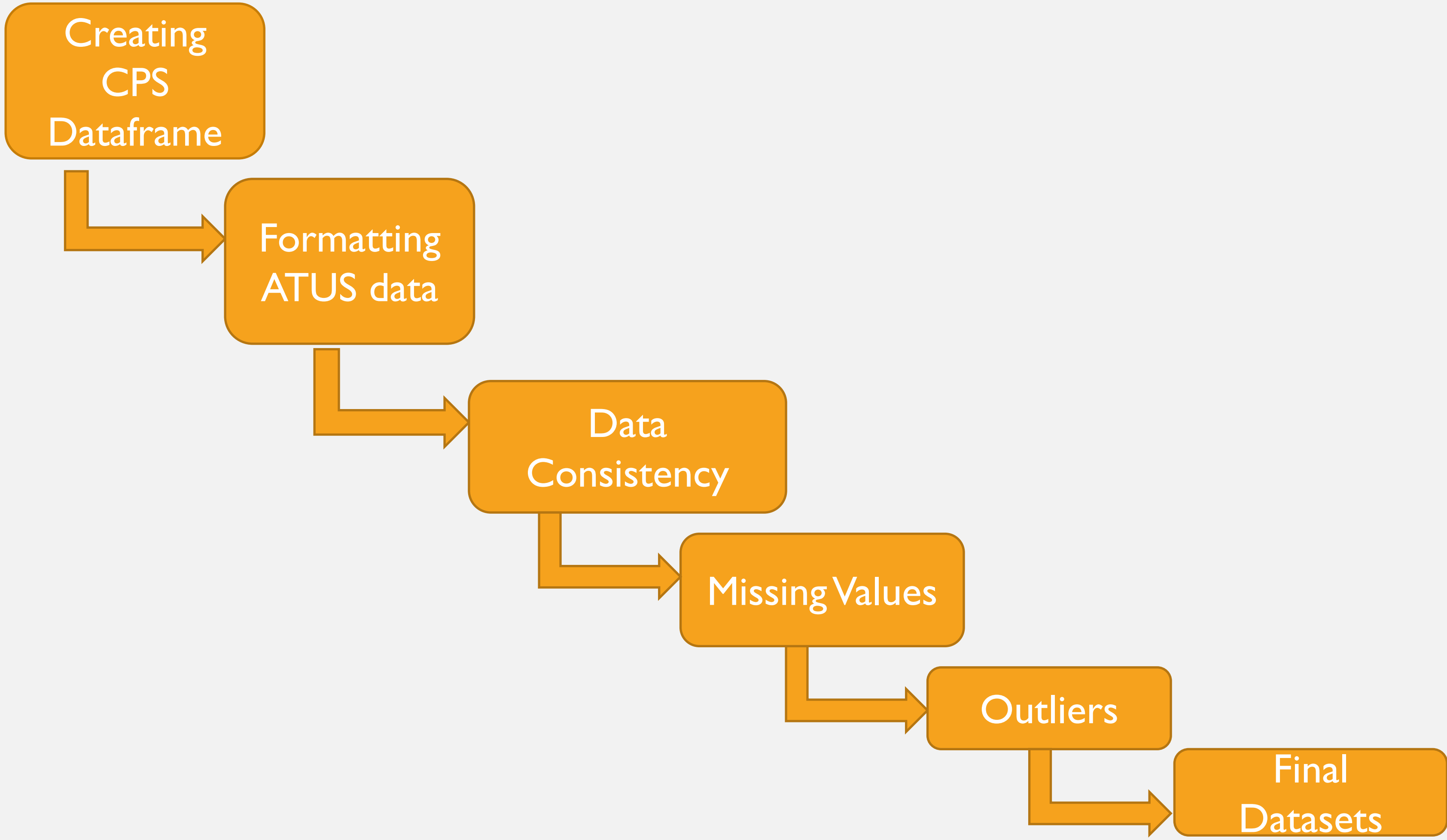
1. Limited Open Source Access to the CPS dataset (data used to build DT)
2. Since 2011 only, the ATUS question for leisure computer time spent has changed to include: Facebook, Instagram, Twitter and other– The most commonly used online platforms. Programming and designing/updating website (personal interest) has also been included.



TECHNICAL REPORT – PHASE 1 DATA FORMATTING AND DATA CLEANING

TABLE OF CONTENT

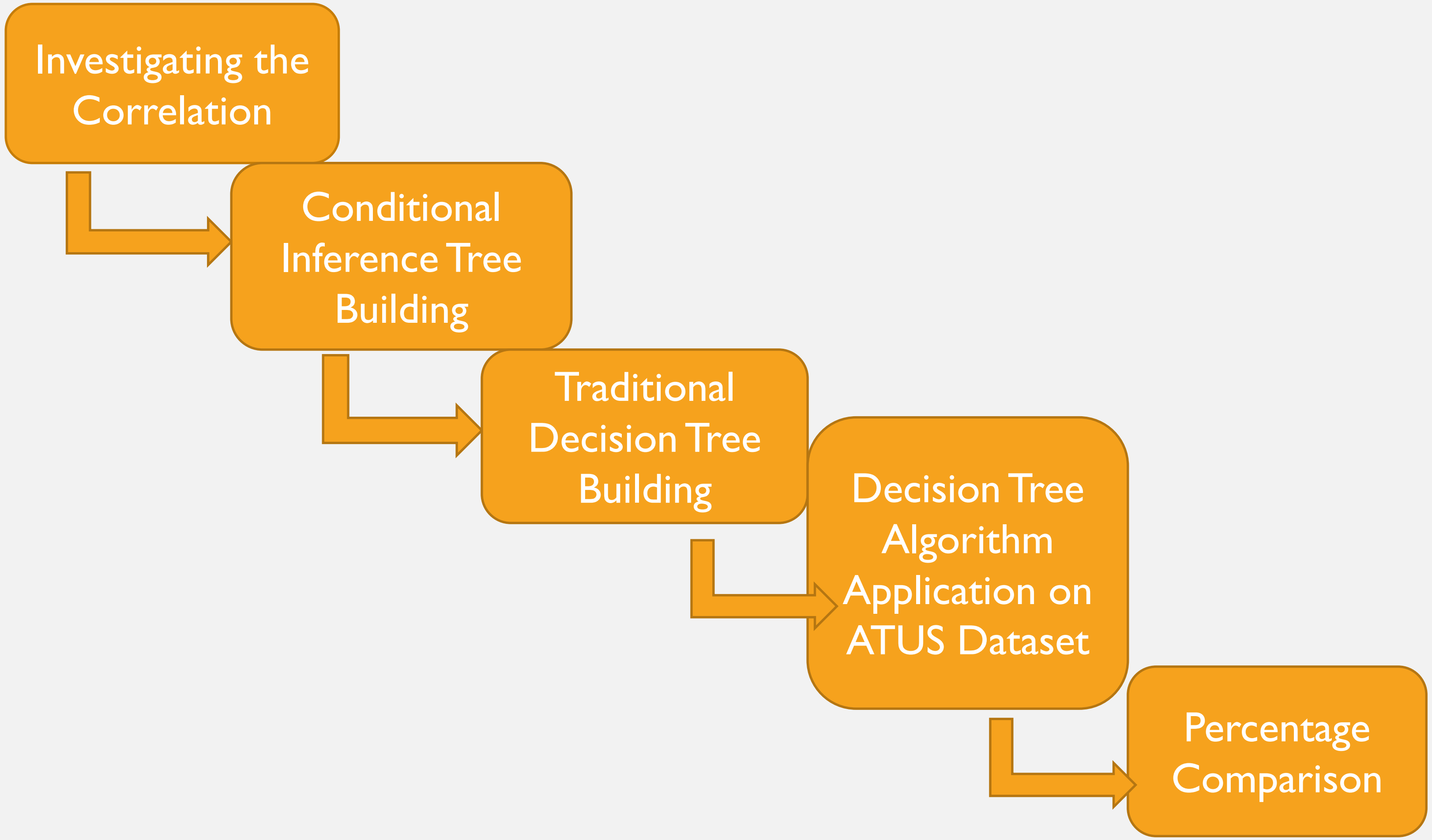
TECHNICAL STEPS – R, SQL (RMYSQL, SQLDF)



TECHNICAL REPORT – PHASE 2 CORRELATION, DECISION TREE BUILDING AND INTERNET ACCESS PERCENTAGE COMPARISON

TABLE OF CONTENT

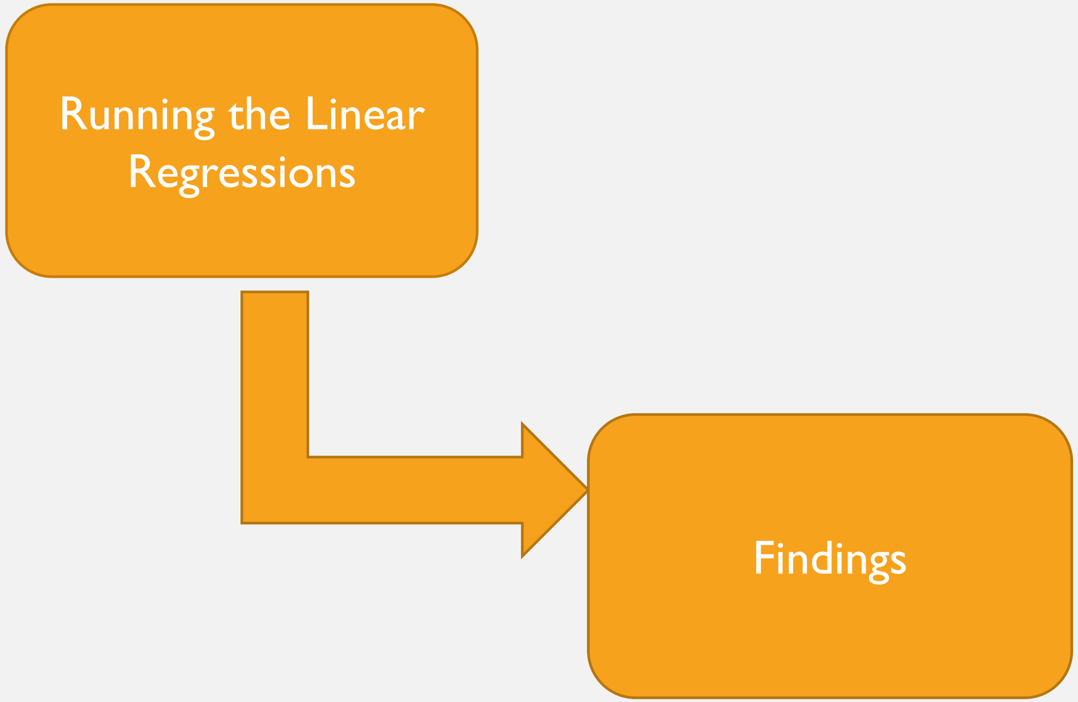
TECHNICAL STEPS – R (RPART AND CTREE)



TECHNICAL REPORT – PHASE 3 LINEAR REGRESSION AND COEFFICIENT ANALYSIS

TABLE OF CONTENT

TECHNICAL STEPS – R AND TABLEAU



PHASE 1

PHASE 1 - STEP 1 CPS DATAFRAME

Dat file read in a single line

```
138009000100198 72011 220100 1 1 1-1 115-1-1-1  38893409 1 2 1 2 2 0 289001 2
2-1-1-1-1 36301 115000154102000      2 1-1540 1 2 1 2   244 1-1 9   1   1 1 1 2
1 2 57 57 57 1 0 0 2 1 1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1 2-124-1-1 24 2-1 2-1 2-1
11640-1 40-1      2 5 4 3 2 3 5 2-1-1-1-1-1 -1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-
1-1-1-1-1-1-1-1-1-1-1-1 -1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1 1-118 1 1 2
2-1-118 -1  -1 1 9-1-1-1-1-1-1-1 4 2      -1-1      -1-1-1 1 2 4-1 1 6-132-1 2-1
1 8-1 1-1 1 1 1 1 0-1-1-1-1  -1  -1  -1  -10-1      -10          -1       -10  -
1-1-1-1-1-1-1-1-1-1-1-1  38893409  52811218         0  37458457  36305764 0 0  -
1-1 0 0 0 1 0-1 050 0 0 0 0 1 0 0 0-1-1-1   0 0-1 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1
1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1 0 0 0-1-1-1-1-1-1-1-1    -1-1-1-1-
1-1-1-1-1-1-1-1    1 1-1-1-1-1-1 2 1 1 1 1 1 0  371544046870 850  -1  -111-1-1-
1-1-1 0-1-1-1-1-1 1 1 1 1 1 2 2 2 2 2 2 2 0 0 0 0 0 0 0              -1-1-1-
1-1-1-1-1-1-1-1 22010 1-1 2 2 1 2 2 2 2-1-1-1-1-1-1 1 1 1 1 2150-1 1 2 2-1-1-1-1
1 1 2 1-1-1 2-1 1 1 2 1 2 2-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-
1-1-1-1-1-1-1-1-1-1-1-1-1-1-1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0         0
138009000100198 72011 220100 1 1 1-1 115-1-1-1  38893409 1 2 1 2 2 0 289001 2
2-1-1-1-1 36301 115000154102000      2 3-1550 1 1 2 2   245 1-1 9   2   1 2 1 2
1 2 57 57 57 1 0 0 1 1 1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1 2-124-1-1 24 2-1 2-1 2-1
2-124-1 24-1      2 5 4 3 2 3 5 2-1-1-1-1-1 -1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-
1-1-1-1-1-1-1-1-1-1-1-1 -1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1 1-115 1 1 2
2-11118 -1  -1 1 7-1-1-1-1-1-1-1 5 4      -1-1      -1-1-1 1 2 4-1 1 6-140-1 1-1
110-1 1-1 1 1 1 1 0-1-1-1-1  -1  -1  -1  -10-1      -10          -1       -10  -
1-1-1-1-1-1-1-1-1-1-1-1  38893409  54722054         0  38893409  38349642 0 0  -
1-1 0 0 0 1 0-1 050 0 0 0 0 1 0 0 0-1-1-1   0 0-1 1 1 0 1 0 1 1 0 0 1 0 0 1 0 1
1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1 0 0 0-1-1-1-1-1-1-1-1    -1-1-1-1-
1-1-1-1-1-1-1-1    1 2-1-1-1-1-1-1 1 1 1 1 1 1  384235797890 430  -1  -115-1-1-
1-1-1 0-1-1-1-1-1 1 1 1 1 1 2 2 2 2 2 2 2 0 0 0 0 0 0 0               1 1 1 1
2 2 2 2 2 2 1 22010 1-1 2 2 1 2 2 2 2-1-1-1-1-1-1 1 1 1 1 2150-1 1 2 2-1-1-1-1 1
1 2 1-1-1 2-1 1 1 2 2 2 2 1 3 2 2 1 2 2 1 2 2 2 2 2 2 1 2 2 2 2 2 1-1-1-1-1-1-1-
1 1 1 2 2 2 2 1-1-1-1-1-1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  61565527
```

PHASE 1 - STEP 1 CPS DATAFRAME

```
CPS_2011 <- read.delim("july2011cps.dat", header = FALSE, sep="\t")
CPS_2013 <- read.delim("july2013cps.dat", header = FALSE, sep="\t")
CPS_2015 <- read.delim("july2015cps.dat", header = FALSE, sep="\t")
```

```
CPS_2015 <- as.data.frame(sapply(CPS_2015[,1], as.character))
CPS_2013 <- as.data.frame(sapply(CPS_2013[,1], as.character))
CPS_2011 <- as.data.frame(sapply(CPS_2011[,1], as.character))

colnames(CPS_2011) <- "RAW_DATA"
colnames(CPS_2013) <- "RAW_DATA"
colnames(CPS_2015) <- "RAW_DATA"
```

	RAW_DATA
1	138009000100198 72011 220100 1 1 1-1 115-1-1...
2	138009000100198 72011 220100 1 1 1-1 115-1-1...
3	400109960881499 72011-122700-1 1-1-1 0-1-1 3...

PHASE 1 - STEP 1 CPS DATAFRAME

```
first_row_2011 <- CPS_2011[1,1]
first_row_2011 <- as.character(first_row_2011)
nchar(first_row_2011)

## [1] 1259

first_row_2013 <- CPS_2013[1,1]
first_row_2013 <- as.character(first_row_2013)
nchar(first_row_2013)

## [1] 1173

first_row_2015 <- CPS_2015[1,1]
first_row_2015 <- as.character(first_row_2015)
nchar(first_row_2015)

## [1] 1174
```

July 2011 Computer and Internet Use Supplement Data. The July supplement data are in locations 951-1259. (See Attachment 7)

July 2013 Computer and Internet Use Supplement Data. The July supplement data are in locations 951-1173. (See Attachment 7)

July 2015 Computer and Internet Use Supplement Data. The July supplement data are in locations 951-1174. (See Attachment 7)

PHASE 1 - STEP 1 CPS DATAFRAME

- Each row contains has a length 1259. These represent the answers of the subjects on more than 100 questions. To know which answer corresponds to which question, we need to check the CPS codebook.
- The location of the answers are highlighted in the codebook.
- Example shown in the picture.

IAME	SIZE	DESCRIPTION	LOCATION
		Additional valid entries for unedited items: -1 (blank), -2 (don't know), -3 (refused).	
IRHHID	15	HOUSEHOLD IDENTIFIER (Part 1) EDITED UNIVERSE: ALL HHLD's IN SAMPLE Part 1. See Characters 71-75 for Part 2 of the Household Identifier. Use Part 1 only for matching backward in time and use in combination with Part 2 for matching forward in time.	1 - 15
IRMONTH	2	MONTH OF INTERVIEW EDITED UNIVERSE: ALL HHLDs IN SAMPLE VALID ENTRIES 01 MIN VALUE 12 MAX VALUE	16 - 17
IRYEAR4	4	YEAR OF INTERVIEW EDITED UNIVERSE:	18 - 21

PHASE 1 - STEP 1 CPS DATAFRAME

STRUCTURING 2011, 2013, 2015 DATASET INTO DATAFRAME

```r
#2013 Dataset:
CPS_2013$PERSON_TYPE <- substr(CPS_2013$RAW_DATA, 161,162)
CPS_2013$LABOUR_FORCE_STATUS <- substr(CPS_2013$RAW_DATA, 180,181)
CPS_2013$AGE <- substr(CPS_2013$RAW_DATA, 122,123)
CPS_2013$HISPANIC <- substr(CPS_2013$RAW_DATA, 157,158)
CPS_2013$SEX <- substr(CPS_2013$RAW_DATA, 129,130)
CPS_2013$MORE_THAN_1_JOB <- substr(CPS_2013$RAW_DATA, 214,215)
CPS_2013$HOURS_PER_WEEK <- substr(CPS_2013$RAW_DATA, 224,226)
CPS_2013$FULL_TIME_PART_TIME <- substr(CPS_2013$RAW_DATA, 2,3)
CPS_2013$WEEKLY_EARNINGS <- substr(CPS_2013$RAW_DATA, 527,534)
CPS_2013$EDUCATION <- substr(CPS_2013$RAW_DATA, 137,138)
CPS_2013$SPOUSE <- substr(CPS_2013$RAW_DATA, 125,126)
CPS_2013$CHILDREN <- substr(CPS_2013$RAW_DATA, 635,636)
CPS_2013$AGE_YOUNGEST_CHILD <- substr(CPS_2013$RAW_DATA, 633,634)
CPS_2013$METROPOLITAN_STATUS <- substr(CPS_2013$RAW_DATA, 105,105)
CPS_2013$HOME_INTERNET_ACCESS <- substr(CPS_2013$RAW_DATA, 977,978)
```

```r
CPS_2015 <- as.data.frame(sapply(CPS_2015[,1:16], as.numeric))
CPS_2013 <- as.data.frame(sapply(CPS_2013[,1:16], as.numeric))
CPS_2011 <- as.data.frame(sapply(CPS_2011[,1:16], as.numeric))
CPS_prep <- rbind(CPS_2011, CPS_2013, CPS_2015)
```

VIEW END RESULT

	LABOUR_FORCE_STATUS	AGE	RACE	HISPANIC	SEX	MORE_THAN_1_JOB	HOURS_PER_WEEK	FULL_TIME_PART_TIME	WEEKLY_EARNINGS	EDUCATION	SPOUSE	CHILDREN	AGE_YOUNGEST_CHILD	METROPOLITAN_STATUS
1	1	54	1	2	1	2	24	2	NA	44	1	0	NA	1
2	1	55	1	2	2	2	24	2	NA	45	1	0	NA	1
3	1	57	1	2	1	2	50	1	NA	39	1	0	NA	1
4	1	54	1	2	2	2	20	2	NA	39	1	0	NA	1
5	3	54	1	2	1	NA	NA	NA	NA	34	1	0	NA	1
6	5	54	1	2	2	NA	NA	NA	NA	34	1	0	NA	1

PHASE 1 - STEP 2 ATUS DATAFRAME

SELECTING DEMOGRAPHIC VARIABLES

```r
ATUS_years_select <- da36268.0001[which(da36268.0001$TUYEAR>=2011),]

Variables <- c(9, 8, 6, 13, 10, 24, 16, 17,5,20,15,21, 4)

ATUS_prep <- ATUS_years_select[, Variables]

names(ATUS_prep)[names(ATUS_prep)=="TELFS"] <- "LABOUR_FORCE_STATUS"
names(ATUS_prep)[names(ATUS_prep)=="TEAGE"] <- "AGE"
names(ATUS_prep)[names(ATUS_prep)=="PEHSPNON"] <- "HISPANIC"
names(ATUS_prep)[names(ATUS_prep)=="TESEX"] <- "SEX"
names(ATUS_prep)[names(ATUS_prep)=="TEMJOT"] <- "MORE_THAN_1_JOB"
names(ATUS_prep)[names(ATUS_prep)=="TEHRUSLT"] <- "HOURS_PER_WEEK"
names(ATUS_prep)[names(ATUS_prep)=="TRDPFTPT"] <- "FULL_TIME_PART_TIME"
names(ATUS_prep)[names(ATUS_prep)=="TRERNWA"] <- "WEEKLY_EARNINGS"
names(ATUS_prep)[names(ATUS_prep)=="PEEDUCA"] <- "EDUCATION"
names(ATUS_prep)[names(ATUS_prep)=="TRSPPRES"] <- "SPOUSE"
names(ATUS_prep)[names(ATUS_prep)=="TRCHILDNUM"] <- "CHILDREN"
names(ATUS_prep)[names(ATUS_prep)=="TRYHHCHILD"] <- "AGE_YOUNGEST_CHILD"
names(ATUS_prep)[names(ATUS_prep)=="GTMETSTA"] <- "METROPOLITAN_STATUS"
```

SELECTING TIME VARIABLES (AN EXAMPLE)

```r
#Phone calls
Variable_Phone <- (c(which( colnames(ATUS_years_select)=="T160101"), which( c
olnames(ATUS_years_select)=="T169989")))
Variable_Phone

## [1] 388 396

Pre_Phone <- ATUS_years_select[,388:396]
ncol(Pre_Phone)

## [1] 9

Pre_Phone$Phone_calls <- rowSums(Pre_Phone[1:9])
Phone_calls <- data.frame(Pre_Phone$Phone_calls)
```

PHASE 1 - STEP 2 ATUS DATAFRAME

VIEW DEMOGRAPHIC VARIABLES

```
##           LABOUR_FORCE_STATUS AGE         HISPANIC        SEX
## 112039 (5) Not in labor force  62 (2) Non-Hispanic (2) Female
## 112040 (1) Employed - at work  22 (2) Non-Hispanic (2) Female
## 112041 (1) Employed - at work  33 (2) Non-Hispanic   (1) Male
##        MORE_THAN_1_JOB HOURS_PER_WEEK FULL_TIME_PART_TIME WEEKLY_EARNINGS
## 112039            <NA>             NA                <NA>              NA
## 112040          (2) No             40       (1) Full time             150
## 112041          (2) No             42       (1) Full time             350
##                                                        EDUCATION
## 112039                                           (37) 11th grade
## 112040 (39) High school graduate - diploma or equivalent [GED]
## 112041                                           (36) 10th grade
##                                          SPOUSE CHILDREN
## 112039 (3) No spouse or unmarried partner present        1
## 112040 (3) No spouse or unmarried partner present        0
## 112041                        (1) Spouse present        1
##        AGE_YOUNGEST_CHILD  METROPOLITAN_STATUS
## 112039                  9     (1) Metropolitan
## 112040                 NA     (1) Metropolitan
## 112041                 15 (2) Non-metropolitan
```

VIEW TIME VARIABLES (AN EXAMPLE)

	CASEID	Phone_calls	Consumer_Purchases	Gov_and_Civic_Obligations	HH_Services	Professional_Care	Volunteer	Religion	Helping_NONHH	Sports	Education
1	1	0	0	0	0	0	0	0	0	200	0
2	2	0	0	0	0	0	0	0	0	0	0
3	3	60	60	0	0	0	0	0	0	0	0
4	4	0	0	0	0	0	0	0	0	0	0
5	5	0	0	0	0	15	0	0	0	60	0
6	6	0	0	0	0	0	0	0	0	0	0

PHASE 1 – STEP 3 DATA CONSISTENCY CPS VS. ATUS DEMOGRAPHICS

CPS

```
CPS_prep[1:3,]

##   PERSON_TYPE LABOUR_FORCE_STATUS AGE HISPANIC SEX MORE_THAN_1_JOB
## 1           2                   1  54        2   1               2
## 2           2                   1  55        2   2               2
## 3          -1                  -1  -1       -1  -1              -1
##   HOURS_PER_WEEK FULL_TIME_PART_TIME WEEKLY_EARNINGS EDUCATION SPOUSE
## 1             24                  38              -1        44      1
## 2             24                  38              -1        45      1
## 3             -1                   0              -1        -1     -1
##   CHILDREN AGE_YOUNGEST_CHILD METROPOLITAN_STATUS HOME_INTERNET_ACCESS
## 1        0                  0                   1                    1
## 2        0                  0                   1                    1
## 3       -1                 -1                   1                   -1
```

ATUS

```
##           LABOUR_FORCE_STATUS AGE         HISPANIC        SEX
## 112039 (5) Not in labor force  62 (2) Non-Hispanic (2) Female
## 112040 (1) Employed - at work  22 (2) Non-Hispanic (2) Female
## 112041 (1) Employed - at work  33 (2) Non-Hispanic   (1) Male
##        MORE_THAN_1_JOB HOURS_PER_WEEK FULL_TIME_PART_TIME WEEKLY_EARNINGS
## 112039            <NA>             NA                <NA>              NA
## 112040          (2) No             40       (1) Full time             150
## 112041          (2) No             42       (1) Full time             350
##                                                      EDUCATION
## 112039                                         (37) 11th grade
## 112040 (39) High school graduate - diploma or equivalent [GED]
## 112041                                         (36) 10th grade
##                                           SPOUSE CHILDREN
## 112039 (3) No spouse or unmarried partner present        1
## 112040 (3) No spouse or unmarried partner present        0
## 112041                        (1) Spouse present        1
##        AGE_YOUNGEST_CHILD  METROPOLITAN_STATUS
## 112039                  9     (1) Metropolitan
## 112040                 NA     (1) Metropolitan
## 112041                 15 (2) Non-metropolitan
```

PHASE 1 – STEP 3 DATA CONSISTENCY CPS VS. ATUS DEMOGRAPHICS

ATUS UPDATE EXAMPLE

```
#FULL_TIME_PART_TIME
ATUS_prep <- sqldf(c("UPDATE ATUS_prep SET FULL_TIME_PART_TIME=1 where FULL_T
IME_PART_TIME LIKE '%(1)%'", "SELECT * from ATUS_prep"))

ATUS_prep <- sqldf(c("UPDATE ATUS_prep SET FULL_TIME_PART_TIME=2 where FULL_T
IME_PART_TIME LIKE '%(2)%'", "SELECT * from ATUS_prep"))
```

ATUS RESULT

```
ATUS_prep[1:3,]

##   LABOUR_FORCE_STATUS AGE HISPANIC SEX MORE_THAN_1_JOB HOURS_PER_WEEK
## 1                   5  62        2   2            <NA>             NA
## 2                   1  22        2   2               2             40
## 3                   1  33        2   1               2             42
##   FULL_TIME_PART_TIME WEEKLY_EARNINGS EDUCATION SPOUSE CHILDREN
## 1                <NA>              NA        37      3        1
## 2                   1             150        39      3        0
## 3                   1             350        36      1        1
##   AGE_YOUNGEST_CHILD METROPOLITAN_STATUS
## 1                  9                   1
## 2                 NA                   1
## 3                 15                   2
```

PHASE 1 – STEP 3 DATA CONSISTENCY CPS VS. ATUS DEMOGRAPHICS

TELFS: EDITED: LABOR FORCE STATUS

Notes: Edited: labor force status

Value	Label	Unweighted Frequency	%
1	Employed - at work	102040	59.7 %
2	Employed - absent	4582	2.7 %
3	Unemployed - on layoff	898	0.5 %
4	Unemployed - looking	7477	4.4 %
5	Not in labor force	55845	32.7 %
	Total	**170,842**	**100%**

Based upon 170,842 valid cases out of 170,842 total cases.

PEMLR 2 MONTHLY LABOR FORCE RECODE 180 - 181

EDITED UNIVERSE:
PRPERTYP = 2

VALID ENTRIES

1 EMPLOYED-AT WORK
2 EMPLOYED-ABSENT
3 UNEMPLOYED-ON LAYOFF
4 UNEMPLOYED-LOOKING
5 NOT IN LABOR FORCE-RETIRED
6 NOT IN LABOR FORCE-DISABLED
7 NOT IN LABOR FORCE-OTHER

```
CPS_prep <- sqldf(c("UPDATE CPS_prep SET LABOUR_FORCE_STATUS = 5 WHERE LABOUR_FORCE_STATUS = 6 OR LABOUR_FORCE_STATUS = 7", "SELECT * FROM CPS_prep"))

CPS_prep$LABOUR_FORCE_STATUS <- as.factor(CPS_prep$LABOUR_FORCE_STATUS)
ATUS_prep$LABOUR_FORCE_STATUS <- as.factor(ATUS_prep$LABOUR_FORCE_STATUS)
```

PHASE 1 – STEP 4 DEALING WITH MISSING VALUES

NAS IN ATUS

```
na_count2 <-sapply(ATUS_prep, function(ATUS_prep) sum(length(which(is.na(ATUS_prep)))))/nrow(ATUS_prep)
na_count2

## LABOUR_FORCE_STATUS                 AGE            HISPANIC
##           0.0000000           0.0000000           0.0000000
##                 SEX      MORE_THAN_1_JOB      HOURS_PER_WEEK
##           0.0000000           0.3971669           0.4328107
## FULL_TIME_PART_TIME     WEEKLY_EARNINGS           EDUCATION
##           0.3971669           0.4635909           0.0000000
##              SPOUSE            CHILDREN  AGE_YOUNGEST_CHILD
##           0.0000000           0.0000000           0.6018978
## METROPOLITAN_STATUS
##           0.0000000
```

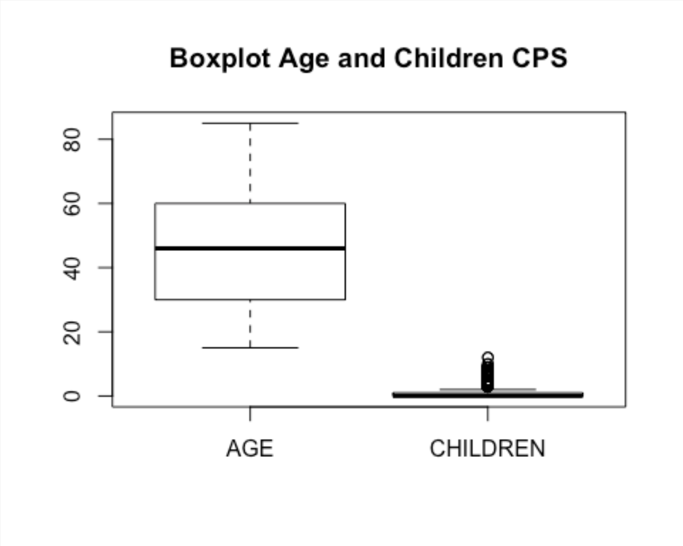
NAS IN CPS

```
na_count <-sapply(CPS_prep, function(CPS_prep) sum(length(which(is.na(CPS_prep)))))/nrow(CPS_prep)
na_count

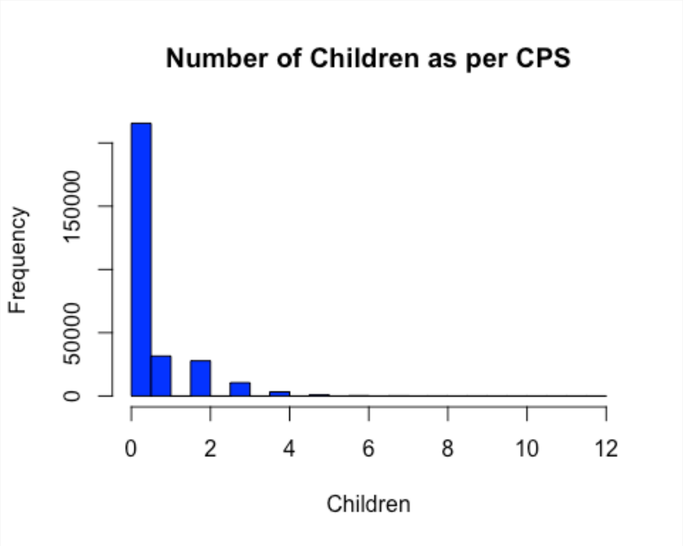
##  LABOUR_FORCE_STATUS                 AGE            HISPANIC
##            0.0000000           0.0000000           0.0000000
##                  SEX     MORE_THAN_1_JOB      HOURS_PER_WEEK
##            0.0000000           0.4139343           0.4139343
##  FULL_TIME_PART_TIME     WEEKLY_EARNINGS           EDUCATION
##            0.4139343           0.8577559           0.0000000
##               SPOUSE            CHILDREN  AGE_YOUNGEST_CHILD
##            0.0000000           0.0000000           0.7436812
##  METROPOLITAN_STATUS HOME_INTERNET_ACCESS
##            0.0000000           0.0000000
```

PHASE 1 – STEP 5 DEALING OUTLIERS

CPS DEMOGRAPHICS

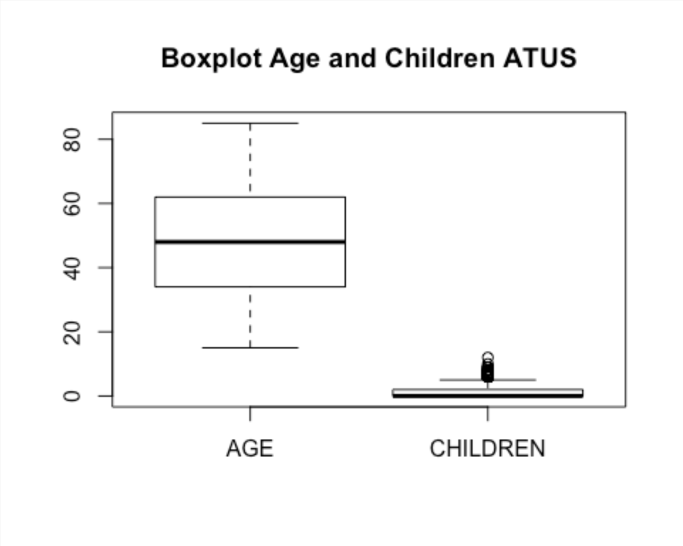


CHILDREN

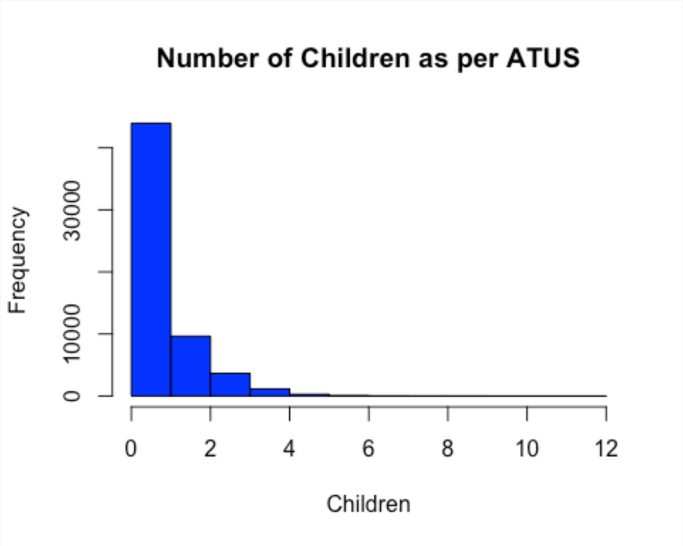


PHASE 1 – STEP 5 DEALING OUTLIERS

ATUS DEMOGRAPHICS

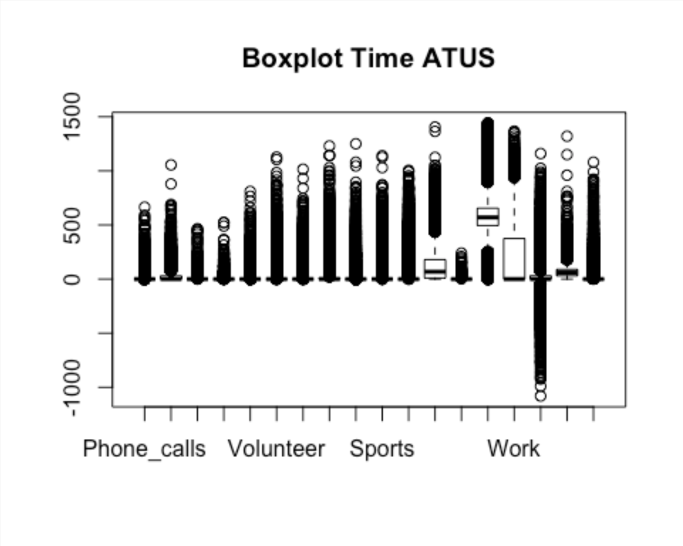


CHILDREN

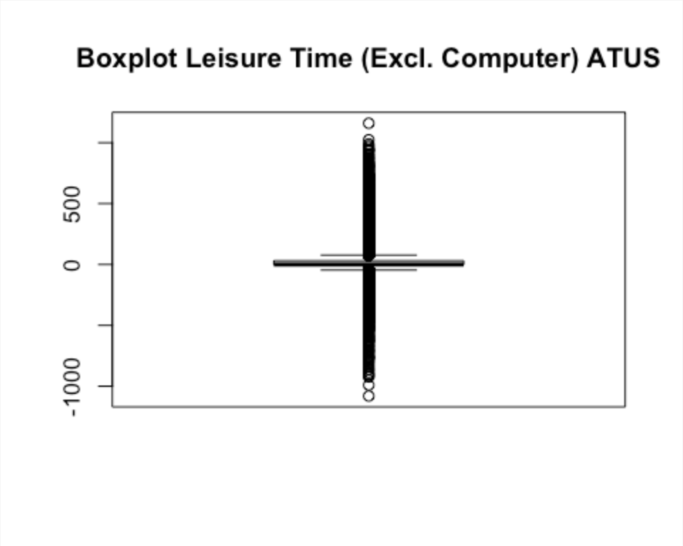


PHASE 1 – STEP 5 DEALING OUTLIERS

ATUS TIME VARIABLES



LEISURE TIME



PHASE 1 – STEP 5 DEALING OUTLIERS

```
summary(ATUS_norm$Leisure_Excl_Computer)

##      Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1080.00     0.00     0.00    29.49    31.00  1160.00
```

There are 5769 records with a negative leisure time. Let's remove those rows.

```
neg <- ATUS_norm$Leisure_Excl_Computer < 0
sum(neg==T)

## [1] 5769
```

PHASE 1 – STEP 6 FINAL DATASETS

```
ATUS[1:3,]

##   LABOUR_FORCE_STATUS AGE HISPANIC SEX EDUCATION SPOUSE CHILDREN
## 1                   5  62        2   2        37      2        1
## 2                   1  22        2   2        39      2        0
## 3                   1  33        2   1        36      1        1
##   METROPOLITAN_STATUS Phone_calls Consumer_Purchases
## 1                   1           0                 20
## 2                   1           0                  0
## 3                   2           0                  0
##   Gov_and_Civic_Obligations HH_Services Professional_Care Volunteer
## 1                         0           0                 0         0
## 2                         0           0                 0         0
## 3                         0           0                 0         0
##   Religion Helping_HH Helping_NONHH Sports Education HH_Activities Travel
## 1        0          0            15      0         0           155      0
## 2        0          0             0      0         0           120      0
## 3        0          0             0      0         0           300      0
##   Personal_Care_Sleep Work Leisure_Excl_Computer Eat_Drink
## 1                 540    0                     0        10
## 2                 600  600                     0        95
## 3                 400    0                     0        25
##   Computer_leisure
## 1                0
## 2                0
## 3                0
```

```
CPS[1:3,]

##   LABOUR_FORCE_STATUS AGE HISPANIC SEX EDUCATION SPOUSE CHILDREN
## 1                   1  54        2   1        44      1        0
## 2                   1  55        2   2        45      1        0
## 3                   1  57        2   1        39      1        0
##   METROPOLITAN_STATUS HOME_INTERNET_ACCESS
## 1                   1                    1
## 2                   1                    1
## 3                   1                    1
```

PHASE 2

PHASE 2 – STEP I CORRELATION

```
Correlations <- rcorr(as.matrix(CPS), type="spearman")
flattenCorrMatrix(Correlations$r, Correlations$P)

##                   row               column          cor            p
## 1  LABOUR_FORCE_STATUS                  AGE  0.257411629 0.0000000000
## 2  LABOUR_FORCE_STATUS             HISPANIC  0.006978087 0.0001718765
## 3                  AGE             HISPANIC  0.140601471 0.0000000000
## 4  LABOUR_FORCE_STATUS                  SEX  0.123676710 0.0000000000
## 5                  AGE                  SEX  0.033636000 0.0000000000
## 6             HISPANIC                  SEX  0.003711284 0.0456946389
## 7  LABOUR_FORCE_STATUS            EDUCATION -0.243129492 0.0000000000
## 8                  AGE            EDUCATION  0.090587601 0.0000000000
## 9             HISPANIC            EDUCATION  0.237461835 0.0000000000
## 10                 SEX            EDUCATION  0.021360371 0.0000000000
## 11 LABOUR_FORCE_STATUS               SPOUSE  0.086016573 0.0000000000
## 12                 AGE               SPOUSE -0.275922388 0.0000000000
## 13            HISPANIC               SPOUSE -0.039414484 0.0000000000
## 14                 SEX               SPOUSE  0.040946145 0.0000000000
## 15           EDUCATION               SPOUSE -0.191730753 0.0000000000
## 16 LABOUR_FORCE_STATUS             CHILDREN -0.163317204 0.0000000000
## 17                 AGE             CHILDREN -0.199073061 0.0000000000
## 18            HISPANIC             CHILDREN -0.095136084 0.0000000000
## 19                 SEX             CHILDREN  0.037577201 0.0000000000
## 20           EDUCATION             CHILDREN  0.074135557 0.0000000000
## 21              SPOUSE             CHILDREN -0.284729689 0.0000000000
## 22 LABOUR_FORCE_STATUS  METROPOLITAN_STATUS  0.025444033 0.0000000000
## 23                 AGE  METROPOLITAN_STATUS  0.053411063 0.0000000000
## 24            HISPANIC  METROPOLITAN_STATUS  0.104046255 0.0000000000
## 25                 SEX  METROPOLITAN_STATUS -0.005345389 0.0040015769
## 26           EDUCATION  METROPOLITAN_STATUS -0.075496979 0.0000000000
## 27              SPOUSE  METROPOLITAN_STATUS -0.037985239 0.0000000000
## 28            CHILDREN  METROPOLITAN_STATUS -0.006100686 0.0010208831
## 29 LABOUR_FORCE_STATUS HOME_INTERNET_ACCESS  0.192443743 0.0000000000
## 30                 AGE HOME_INTERNET_ACCESS  0.175807983 0.0000000000
## 31            HISPANIC HOME_INTERNET_ACCESS -0.093050882 0.0000000000
## 32                 SEX HOME_INTERNET_ACCESS  0.016564004 0.0000000000
## 33           EDUCATION HOME_INTERNET_ACCESS -0.260926008 0.0000000000
## 34              SPOUSE HOME_INTERNET_ACCESS  0.133594140 0.0000000000
## 35            CHILDREN HOME_INTERNET_ACCESS -0.072740041 0.0000000000
## 36 METROPOLITAN_STATUS HOME_INTERNET_ACCESS  0.073360600 0.0000000000
```

PHASE 2 – STEP 2 CONDITIONAL INFERENCE TREE BUILDING

We divide our dataset into training and testing (70-30%).

```
train_index <- sample(1:nrow(CPS), 0.7 * nrow(CPS))
train.set <- CPS[train_index,]
test.set <- CPS[-train_index,]
```

Running the model on the training set.

```
internet_ctree_model <- ctree(HOME_INTERNET_ACCESS ~ LABOUR_FORCE_STATUS + AG
E + HISPANIC + SEX + EDUCATION + SPOUSE + CHILDREN  + METROPOLITAN_STATUS, da
ta=train.set)
```

Now let's make our prediction on the test set.

```
internet_ctree_prediction <- predict(internet_ctree_model, test.set)
head(internet_ctree_prediction)

## [1] 1 1 1 2 2 2
## Levels: 1 2

table(internet_ctree_prediction, test.set$HOME_INTERNET_ACCESS)


## internet_ctree_prediction        1     2
##                          1    63886 14778
##                          2     3260  5043
```

```
Mesuring accuracy, precision, recall and F score.
```{r}
library(caret)
confusionMatrix(internet_ctree_prediction, test.set$HOME_INTERNET_ACCESS, mode="everything")
```
```

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1 63968 15043
         2  3041  4910

               Accuracy : 0.792
                 95% CI : (0.7893, 0.7947)
    No Information Rate : 0.7706
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2544
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9546
            Specificity : 0.2461
         Pos Pred Value : 0.8096
         Neg Pred Value : 0.6175
              Precision : 0.8096
                 Recall : 0.9546
                     F1 : 0.8762
             Prevalence : 0.7706
         Detection Rate : 0.7356
   Detection Prevalence : 0.9086
      Balanced Accuracy : 0.6003

       'Positive' Class : 1
```

PHASE 2 – STEP 3 TRADITIONAL DECISION TREE BUILDING

We divide our dataset into training and testing again (70-30%).

```
train_index <- sample(1:nrow(CPS), 0.7 * nrow(CPS))
train.set2 <- CPS[train_index,]
test.set2 <- CPS[-train_index,]
```

Running the model on the training set.

```
internet_rpart_model <- ctree(HOME_INTERNET_ACCESS ~ LABOUR_FORCE_STATUS + AG
E + HISPANIC + SEX + EDUCATION + SPOUSE + CHILDREN  + METROPOLITAN_STATUS, da
ta=train.set2)
```

Now let's make our prediction on the test set.

```
internet_rpart_prediction <- predict(internet_rpart_model, test.set2)
table(internet_rpart_prediction, test.set2$HOME_INTERNET_ACCESS)

##
## internet_rpart_prediction       1     2
##                         1   63908 14776
##                         2    3207  5076
```

Mesuring accuracy, precision, recall and F score.

```
```{r}
confusionMatrix(internet_rpart_prediction, test.set2$HOME_INTERNET_ACCESS, mode="everything")
```
```

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1 64028 15025
         2  2981  4928

               Accuracy : 0.7929
                 95% CI : (0.7902, 0.7956)
    No Information Rate : 0.7706
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.257
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9555
            Specificity : 0.2470
         Pos Pred Value : 0.8099
         Neg Pred Value : 0.6231
              Precision : 0.8099
                 Recall : 0.9555
                     F1 : 0.8767
             Prevalence : 0.7706
         Detection Rate : 0.7363
   Detection Prevalence : 0.9091
      Balanced Accuracy : 0.6012

       'Positive' Class : 1
```

PHASE 2 – STEP 4 AND STEP 5 DECISION TREE APPLICATION ON THE ATUS DATASET 'WHO HAS INTERNET'

Let's see what percentage of the ATUS sample will our model predict as having internet access.

```
ATUS_internet_pred <- predict(internet_rpart_model, ATUS)
summary(ATUS_internet_pred)

##       1     2
##   46386  6649

sum(ATUS_internet_pred == 1)/length(ATUS_internet_pred)

## [1] 0.87463
```

- Our model estimates 87%
  - Note: we only took years 2011, 2013 and 2015 and averaged them
- According to PEW Research Center: the averaged percentage of population using the Internet from 2011 to 2015 is 84%.

PHASE 3

PHASE 3 – STEP 1 LINEAR REGRESSION ANALYSIS

We select rows where ATUS_internet_pred == 1 (TRUE) and we run our 17 version of regression for analysis.

```
ATUS_internet_pred_df <- as.data.frame(ATUS_internet_pred)
ATUS_phase4_prep <- cbind(ATUS, ATUS_internet_pred_df)
ATUS_phase4 <- subset(ATUS_phase4_prep, ATUS_internet_pred ==1)
str(ATUS_phase4)
```

```
## 'data.frame':    46386 obs. of  27 variables:
##  $ LABOUR_FORCE_STATUS      : Factor w/ 5 levels "1","2","3","4",..: 1 1 1
 1 4 1 1 1 1 1 ...
##  $ AGE                      : num  22 33 45 24 29 29 31 35 33 61 ...
##  $ HISPANIC                 : Factor w/ 2 levels "1","2": 2 2 1 2 2 1 2 2
2 2 ...
##  $ SEX                      : Factor w/ 2 levels "1","2": 2 1 1 2 2 1 2 1
2 2 ...
##  $ EDUCATION                : Factor w/ 16 levels "31","32","33",..: 9 6 9
 9 9 10 9 10 10 10 ...
##  $ SPOUSE                   : Factor w/ 2 levels "1","2": 2 1 2 1 2 2 1 2
1 2 ...
##  $ CHILDREN                 : num  0 1 0 2 2 1 0 1 3 0 ...
##  $ METROPOLITAN_STATUS      : Factor w/ 3 levels "1","2","3": 1 2 1 1 2 1
2 2 1 2 ...
##  $ Phone_calls              : num  0 0 0 0 0 0 0 0 0 105 ...
##  $ Consumer_Purchases       : num  0 0 0 0 5 0 0 0 0 25 ...
##  $ Gov_and_Civic_Obligations: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ HH_Services              : num  0 0 0 0 0 0 0 0 15 0 ...
##  $ Professional_Care        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Volunteer                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Religion                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Helping_HH               : num  0 0 0 60 120 320 0 0 230 0 ...
##  $ Helping_NONHH            : num  0 0 0 0 0 0 20 0 0 0 ...
##  $ Sports                   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Education                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ HH_Activities            : num  120 300 2 0 75 70 35 300 445 355 ...
##  $ Travel                   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Personal_Care_Sleep      : num  600 400 540 600 705 670 480 540 575 600
 ...
##  $ Work                     : num  600 0 0 575 0 0 610 0 0 0 ...
##  $ Leisure_Excl_Computer    : num  0 0 0 0 0 315 0 270 0 15 ...
##  $ Eat_Drink                : num  95 25 30 95 25 30 40 60 90 0 ...
##  $ Computer_leisure         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ATUS_internet_pred       : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2
2 2 ...
```

PHASE 3 – STEP I LINEAR REGRESSION ANALYSIS

Now we run the 17 linear regression versions.

1) Phone calls.

```
attach(ATUS_phase4)

## The following objects are masked _by_ .GlobalEnv:
##
##     ATUS_internet_pred, Computer_leisure, Consumer_Purchases,
##     Eat_Drink, Education, Gov_and_Civic_Obligations, Helping_HH,
##     Helping_NONHH, HH_Activities, HH_Services,
##     Leisure_Excl_Computer, Personal_Care_Sleep, Phone_calls,
##     Professional_Care, Religion, Sports, Travel, Volunteer, Work

phonecalls_regr <- lm(Phone_calls~Computer_leisure+LABOUR_FORCE_STATUS+AGE+HI
SPANIC+SEX+EDUCATION+SPOUSE+CHILDREN+METROPOLITAN_STATUS,data=ATUS_phase4)
summary(phonecalls_regr)

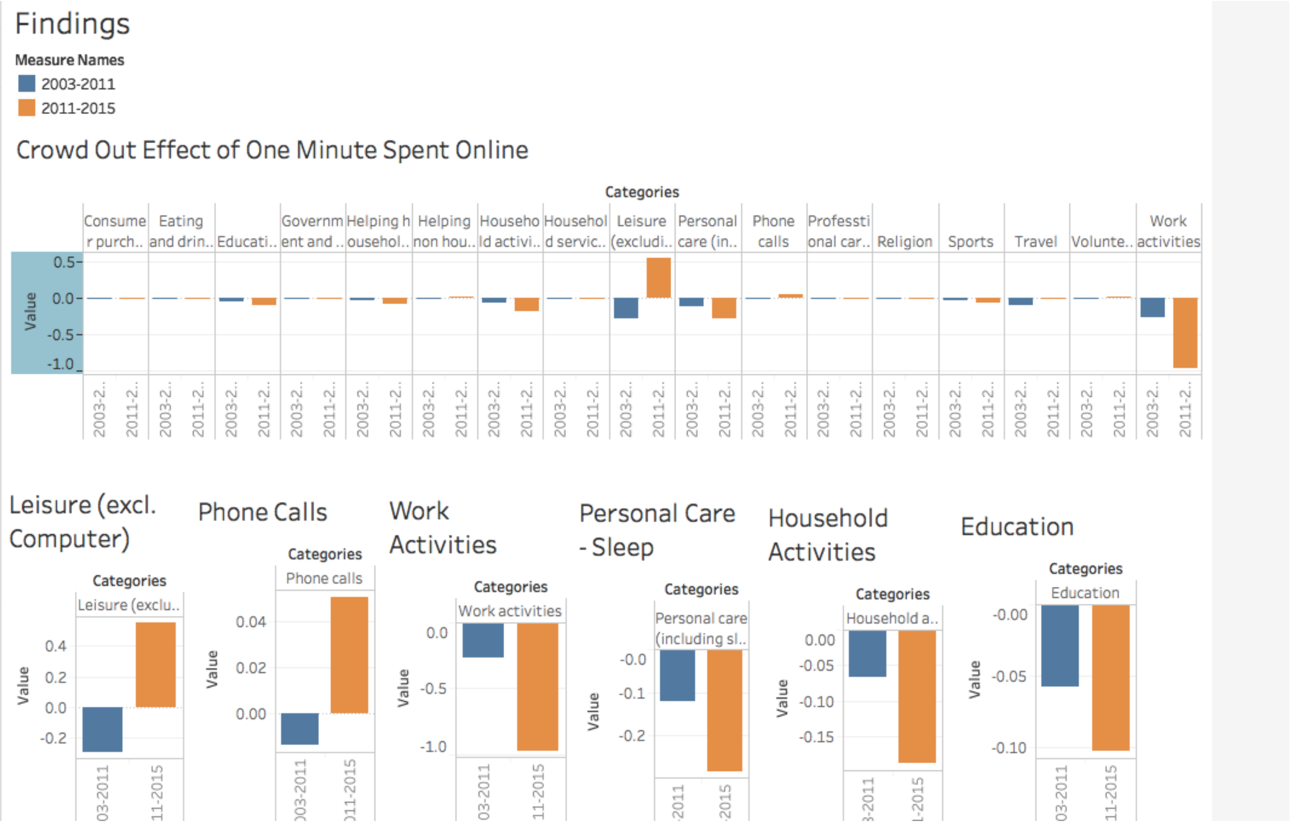
##
## Call:
## lm(formula = Phone_calls ~ Computer_leisure + LABOUR_FORCE_STATUS +
##     AGE + HISPANIC + SEX + EDUCATION + SPOUSE + CHILDREN + METROPOLITAN_ST
ATUS,
##     data = ATUS_phase4)
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -28.42  -8.12  -4.85   -1.19 651.43
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -3.544586   4.568388  -0.776  0.43782
## Computer_leisure       0.050564   0.010148   4.983 6.29e-07 ***
## LABOUR_FORCE_STATUS2   0.692247   0.680748   1.017  0.30921
## LABOUR_FORCE_STATUS3   6.148049   1.614617   3.808  0.00014 ***
## LABOUR_FORCE_STATUS4   4.447254   0.555839   8.001 1.26e-15 ***
## LABOUR_FORCE_STATUS5   2.984066   0.277074  10.770  < 2e-16 ***
## AGE                    0.040972   0.008585   4.773 1.82e-06 ***
## HISPANIC2              1.340558   0.336338   3.986 6.74e-05 ***
## SEX2                   3.695078   0.229197  16.122  < 2e-16 ***
## EDUCATION32            0.226806   5.144774   0.044  0.96484
## EDUCATION33            2.565143   4.816400   0.533  0.59432
## EDUCATION34            3.510445   4.667749   0.752  0.45202
## EDUCATION35            3.692260   4.609933   0.801  0.42317
```

PHASE 3 – FINDINGS AND CONCLUSIONS

Estimated Crowdout Effects of Computer Leisure on Major Categories

Category	2003-2011 coefficients as per Wallsten's findings	2011-2015 coefficients
Leisure (excluding computer)	-0.293***	0.55118***
	(22.34)	(14.737)
Work activities	-0.268***	-0.97216***
	(19.38)	(10.644)
Personal care (including sleep)	-0.121***	-0.28394***
	(12.36)	(4.872)
Travel	-0.0969***	-0.0010751
	(17.36)	(0.918)
Household activities	-0.0667***	-0.18831***
	(7.149)	(3.316)
Education	-0.0574***	-0.10285*
	(8.560)	(3.253)
Sports	-0.0397***	-0.06359*
	(9.17)	(2.499)
Helping household members	-0.0368***	-0.09164**
	(7.589)	(2.977)
Eating and drinking	-0.0254***	-0.004003
	(6.991)	(0.182)
Helping non-household members	-0.0232***	0.01364
	(6.763)	(0.712)
Religion	-0.0146***	-0.01347
	(5.758)	(-0.702)
Volunteer	-0.0120***	0.005886
	(3.503)	0.273
Professional care and services	-0.00360*	-0.016877
	(1.896)	(1.599)
Household services	-0.00129	-0.001102
	(1.583)	(0.283)
Government and civic obligations	-0.000177	-0.0030381
	(0.303)	(0.868)
Consumer purchases	0.00368	-0.008901
	(1.025)	(0.412)
Phone calls	0.0134***	0.050564***
	(7.433)	(4.983)
Absolute t-statistics in parentheses		
***p<0.01, **p<0.05, *p<0.1		

FINDINGS VISUALIZATION USING TABLEAU



CHALLENGES FACED AND LESSONS LEARNT

RESEARCH RELATED

- Qualitative Selection
- Age Of Youngest Child and Marital Status

TECHNICAL RELATED

- Data Cleaning and Formatting
- ‘Too good to be true’ Tree
  - Low F-Score at first
  - Issues:
    - Removing NAs rows and replacing those left with mean and mode
  - Lesson learnt:
    - It’s not enough to count NAs, it is important to visualize them

CONTINUITY

RESEARCH

- Research literature related to my findings and compare the accuracy of my findings to the literatures

TECHNICAL

- Add the 2016 data
- Add more graphics
  - Initial data exploration graphics
  - More findings such as: crowd-out effect by age and others
- Use Wallsten's Regression Methodology (Internet Prediction) and compare with DT

PACKAGES AND TOOLS

```
library("ggplot2")
library("lattice")
library("Formula")
library("survival")
library("Hmisc")
library(grid)
library(mvtnorm)
library(modeltools)
library(stats4)
library(strucchange)
library(zoo)
library(party)
library(sandwich)
library(caret)
library(rpart)
library(randomForest)
library(caTools)
library(sqldf)
library(RMySQL)
```