```r
# List of packages required for this analysis
pkg <- c("dplyr", "ggplot2", "knitr", "devtools")
new.pkg <- pkg[!(pkg %in% installed.packages())]
if (length(new.pkg))
  install.packages(new.pkg, repos = "http://cran.rstudio.com")

# Load packages
library(dplyr)
library(ggplot2)
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.2.3
```

```r
# fixing NAs/blanks
urgentCareOnly.df[c("VYEAR")][is.na(urgentCareOnly.df[c("VYEAR")])] <- 2011
urgentCareOnly.df[urgentCareOnly.df == "Blank"] <- NA
urgentCareOnly.df[urgentCareOnly.df == "Missing Data"] <- NA
urgentCareOnly.df[urgentCareOnly.df == "Missing data"] <- NA
```

```r
cont.vars <- c("PCTPOVR", "PBAMORER", "URBANRUR", "AGER", "RACER", "SEX", "PAYTYPER")
Cont.model <- urgentCareOnly.df[cont.vars]
Cont.model <- na.omit(Cont.model)
```

```r
probtable.func <- function(variable){
    # get name
    category <- deparse(substitute(variable))
    category <- data.frame(name = category, variable = "", Freq = 100.00, stringsAsFactors = FALSE)
    # get proportions
    table.sums <- table(variable)
    variable.tbl <- round(prop.table(table.sums)*100, 2)
    variable.df <- as.data.frame(variable.tbl, stringsAsFactors = FALSE)
    # add id rows
    variable.df <- cbind(data.frame(name = "", stringsAsFactors = FALSE), variable.df)
```

```
    variable.df <- rbind( variable.df[0,0], category, variable.df[] )


    variable.df
}
```

Methods

In this thesis I am attempting to accomplish three distinct goals: to examine the current health-seeking trends in America, something that has yet to have been done in medical sociology with urgent care in consideration; then, to examine the question of *whom*: by clustering the groups who have chosen to go to urgent care; and finally, to use the quantitative data and sociological theories to hypothesize on the reasons behind the decisions.

Using National Ambulatory Medical Care Survey data for the years 2008-2012, I utilize unsupervised machine learning methods in order to develop a descriptive categories of urgetn care center patients. From the exploratory anlaysis, I then test the resulting variables in a logarithmic regression analysis, devopling a model of urgent care seekers.

The initial exploratory nature of the analysis is motivated by the current lack of statistical analysis and social theory surround urgent care centers, despite their rapid progression as an acceptable replacement for primary care (erhm). The results of the findings will hopefuly allow those intent on locating urgent care centers within the larger contextual framework of the American health care system to draw on the revealed typologies of urgent care seekers in order to better understand the industry's rapid growth.

Data

Empirical exploration of the theories proposed in chapter 1 require data that provides an abundance of variables which may or may not be statistically important but which we cannot initally rule out, as well as a large size since the phenomenon is still comparatively rare when talking about how patients access primary care in the US. I thus chose the National Ambulatory Medical Care Survey (NAMCS), which is a national survey designed to provide researchers in the medical and social science fields "accurate and reliable information about the provision and use of ambulatory medical care services in the United States".

Ambulatory care is defined by the survey as health services or acute care services provided to patients on an outpatinet basis, without an overnight stay, and every year NAMCS surveys visits to non-federal

employed office-based physicians are collected from a representative sample of the United States. These surveys contain information about how the patients utilize physician services and hospital outpatient and emergency department services, the conditions most often treated, and the diagnostic and therapeutic services rendered, including medications prescribed. This data served the purpose of the current study because it is both representative of the larger trends in the United States and includes specific information regarding urgent care centers which can be used to explore that particular American health care trend in particular.

I specifically examined the group of visits were coded as having been at "Urgent Care Centers/Freestanding clinics" by the NAMCS. While the combined years produced a dataset of 123,123 observations, only 3,863 of those occurred at urgent care centers (about 3 percent). Of these visits, I limited the analysis to patients over the age of 18, bringing the sample to 3,224 visits to urgent care centers.

Some limitations to the data should be noted. There may be related errors given that as the popularity of urgetn care centers have risen, so too have the number that participated in the NAMCS. In 2008, there were 842 visits surveyed compared to 1168 surveyed in 2010.

Variable Selection and Summary Statistics

The variables I chose to include in my analysis were chosen with both data availaility and sociological theory in mind. Initially I am interested in looking at demographics surrounding age, race, gender, and socio economic status. Inital clusters included dummy variables for self-pay, private insurance, race (white, black, hispanic, asian, 2+), sex, rural, and wealthy. To identify the subsets, I took random samples from the cases which were recording as having been at urgent care centers using the dplyr package sample_n, which allows for a random campling of rows from a table. from the cases which were recording as having been at urgent care centers. For the three years in question a total of 2459 visits to urgent care were surveyed. In order to get workable training data for the unsupervised clustering, I set aside a random sample of 250 visits for test data, and proceeded to randomly sample from the 2209 cases available, running the analysis on 250 observations at a time.

```
#pretty names
PercentPoverty <- Cont.model$PCTPOVR
PercentBachelors <- Cont.model$PBAMORER
UrbanCategory <- Cont.model$URBANRUR
AgeGroup <- Cont.model$AGER
Race <- Cont.model$RACER
```

```r
Sex <- Cont.model$SEX

PaymentType <- Cont.model$PAYTYPER


t1 <- probtable.func(Sex)

t2 <- probtable.func(AgeGroup)

t3 <- probtable.func(Race)

t4 <- probtable.func(PaymentType)

t5 <- probtable.func(UrbanCategory)

t6 <- probtable.func(PercentPoverty)

t7 <- probtable.func(PercentBachelors)


demographics.tbl <- rbind(t1, t2, t3, t4, t5, t6)
```

Table 1: Summary Statistics 1

Variable	Category	Percentage
Sex		100.00
	Female	57.10
	Male	42.90
AgeGroup		100.00
	15-24 years	8.68
	25-44 years	22.89
	45-64 years	29.63
	65-74 years	13.93
	75 years and over	12.58
	Under 15 years	12.30
Race		100.00
	Black	9.15
	Other	3.87
	White	86.98
PaymentType		100.00
	All sources of payment are blank	0.44
	Medicaid	10.67
	Medicare	26.37
	No charge	0.58
	Other	3.07
	Private insurance	46.90
	Self-pay	4.86
	Unknown	1.93
	Worker's compensation	5.17
UrbanCategory		100.00
	Large central metro	24.60
	Large fringe metro	14.62
	Medium metro	34.55
	Micropolitan/noncore (nonmetro)	15.98
	Missing data	0.00
	Small metro	10.25
PercentPoverty		100.00
	Missing data	0.00
	Quartile 1 (Less than 5.00 percent)	18.16
	Quartile 2 (5.00-9.99 percent)	30.24
	Quartile 3 (10.00-19.99 percent)	39.22
	Quartile 4 (20.00 percent or more)	12.38

Below are the summary statistics for an example sample:

```
summarize_each(sample1, funs(mean))
```

insert table here

Within just this one random sample, we can observe that urgent care visitor are mostly white, urban and not wealthy. There is an almost even mix of those who have private insurance and those who dont, only about 7% pay out of pocket, and there seems to be an even mix of sexes.

Four other methodological decisions were made by following either statistical convention or similar previous anlayses. All variables were created as dummy variables, and the distances between these were standardized on a scale from 0 to 1, so as to prevent any skewdness which might result. Second, I chose to use the measure of distance known as the Jaccard method, which is specifically created to measure the distance between 0 to 1 scaled variables. It also has the unique feature of not including as significant pairs which both have 0 for a parameter. This is substantively important since though two visits may both have 0's for Private Insurance for example, the fact that they both don't have private insurance is not enough to consider them theoretically simliar by negation: one may be on medicare while the other may be uninsured. Third, for the actual grouping themselves, I have chosen the standard Ward's method, which attempts to minimize the variance within groups and thus maximizes the homogenaity within groups. Fourth, in keeping with similar exploratory anlyses, I have limited the clusters to a theoretically interesting number while keeping a managable representation of reality.

This can be thought of as us having a set of

$$X_1, X_2, X_3, ..., X_n$$

observations.

First, we have standardized all the variables we used on a scale from 0 to 1, to prevent the sort of skewed analysis that might result if some variables with a broad range of absolute values were allowed to dominate the data analysis. Second, we have chosen the classic measure of distance known as 'squared Euclidean' to evaluate the similarities between cases, as it gives more importance to greater distances, and thus makes it possible to bring out the differences between countries whose profiles still show high degrees of similarity.12 Third, for the actual groupings themselves, we have adopted the usual Ward's method, which minimizes the variance within groups and thus maximizes their hom- ogeneity. Fourth, in keeping with normal practices for exploratory analyses

In the supervised learning world, the data in question has an obvious response variable, which is tested against a null hypothesis based on theory. For this case, we do not know exactly what the outcome of interest is for those patients who went to urgent care, rather we are interested in who is choosing to go there. For such a data set, clustering methods allow us to examine the data in an *unsupervised* manner — mainly in that we let the statistical software find the patterns rather than test for patterns at the start.

Hierarchal Cluster Analysis

To examine such data, I have chosen to use hierarchical cluster anlaysis. This method allows for grouping patients that have similar characteristics across a set of variables by dividing a set of cases into ever more numerous and specific subsets, thus leading to homogenous empirical types (Rapkin and Luke, 1993). One of the most powerful exploratory aspects of cluster analysis is that you do not need to have a response variable in order to better understand your data. For this project, this is extremely useful since we initially only know who is going to urgent care and who is not, but would like to understand them as a group better before drawing comparisions between patients who visited a traidtional primary care clinic. Also a plus for cluster analysis, such inductive methodologies are based only on quantitative similarities among cases, only two factors may be responsible for trends in the data: the actual structure of the observed phenomenon and the

methodological decisions I made concerning choosing the cases and variables (inclduing the statistical method used to identify subsets).

Because I am interested in two somewhat distinct aspects of the patients of Urgent Care– both their demographics and their patterns of use – The clustering was performed in two batches of parameters which aimed to help us understand the two different side to a visit to urgent care. Variables for *Age, Sex, Race, Urban Type, % Neighborhood Poverty, % Neighborhood college degree attainment, and Pay Type*, what I will refer to as the *demographic variables* from this point forward were first analyzed for subgroups. Secondly, some of the same variables were again analyzed with the behavior parameters of *Injury related, Primary Caregiver, Seen Before?, Past Visits, Major Reason, and the day of the week*, what I will refer to as the *behavior parameters.*

While clustering methods are incresingly being used to generate scientific hypotheses, this analysis rather aims to apply clustering as a method of retroduction on a phenomenon that is currently vastly under studied. The methodological decisions of this study are divided into three areas: (1) the determination of the number of clusters, (2) the examination of how the clusters differ in terms of multiple parameters, and (3) examining the clusters in light of the theoretical hypotheses posed in the introduction.

Determining the number of Clusters

soon to be in the appendix

Because urgent care centers have been greatly ignored by sociologists studying medical practices, there was little theoretical guidance in selecting a likely number of subgroups for the analysis. Similarly, because I am interested in understanding how an unsupervised analysis of patient data will reveal trends, it was particularly important to the analysis that the number of clusters were both methematically acheivable and substantivly small enough for analysis.

To accomplish this, I began with hierarchal clustering of the random samples chosen from the data. Using a mixed-methods tool to calculate the distance matrix between the various observations, I placed each point into an althgorithm which subsequently minimized the variance between clusters. Figure 1 shows the banner for the initial agglomerative cluster methods for the behavioral variables. .

The white lines extending to the right represent clusters which differ from each other. After running the hierarchal, bottom up method for a number of trials, the agglomerative coefficient was always between .73 and .8, indicating that as the height for which the clusters should stop combining. Again, Figure 1 demonstrates that at that height, 8 clusters are have clearly seperated.