Preferred Senior Living Areas in Boston

An Analysis on OpenStreetMap dataset

Bidisha Das

Big Data for Cities Final Project

Table of Contents

Introduction

A safe and stable environment for our Seniors is a very important part of city building and planning. Also, according to the U.S Census Bureau (2010) more people were 65 years and over in 2010 than in any previous census. Between 2000 and 2010, the population 65 years and over increased at a faster rate (15.1 percent) versus the total U.S population (9.7 percent). Hence in a way it makes sense to look for what or where Seniors prefer to live. A very common form of Senior living is Assisted Senior Living Facilities. Assisted living which has been a rapidly growing segment of Senior housing over the past several years is a residential setting that provides with personal services. This includes activities, reading services, health related services, 24-hour emergency services etc. The main purpose is to provide a residential like life to Seniors. Now the main purpose of Assisted living is to give a home to Seniors, but their location is another factor which should be taken into consideration. A National Survey by Senior Living article tells us despite being rapidly growing these Senior livings have substantial differences across states (Hawes et all. 2003, A National Survey of Assisted Living Facilities).[1] Also, the location of the livings matter a in terms of choosing one Assisted living over another. It has always been seen that Assisted Livings which are in prominent locations with closer proximity to amenities are more popular than others which are not so close to amenities. This paper tries to explore the various components which are required by Seniors to be around Assisted/Senior living communities.

Background

Everyone deserves a life full of joy, comfort, and overall well-being and Senior being the most important part of our community and its future deserve a safe and stable environment. Keeping their needs in mind the following visualization was developed.

[1] Hawes et all. 2003, A National Survey of Assisted Living Facilities (43). *The Gerontologist*. https://academic.oup.com/gerontologist/article/43/6/875/863127

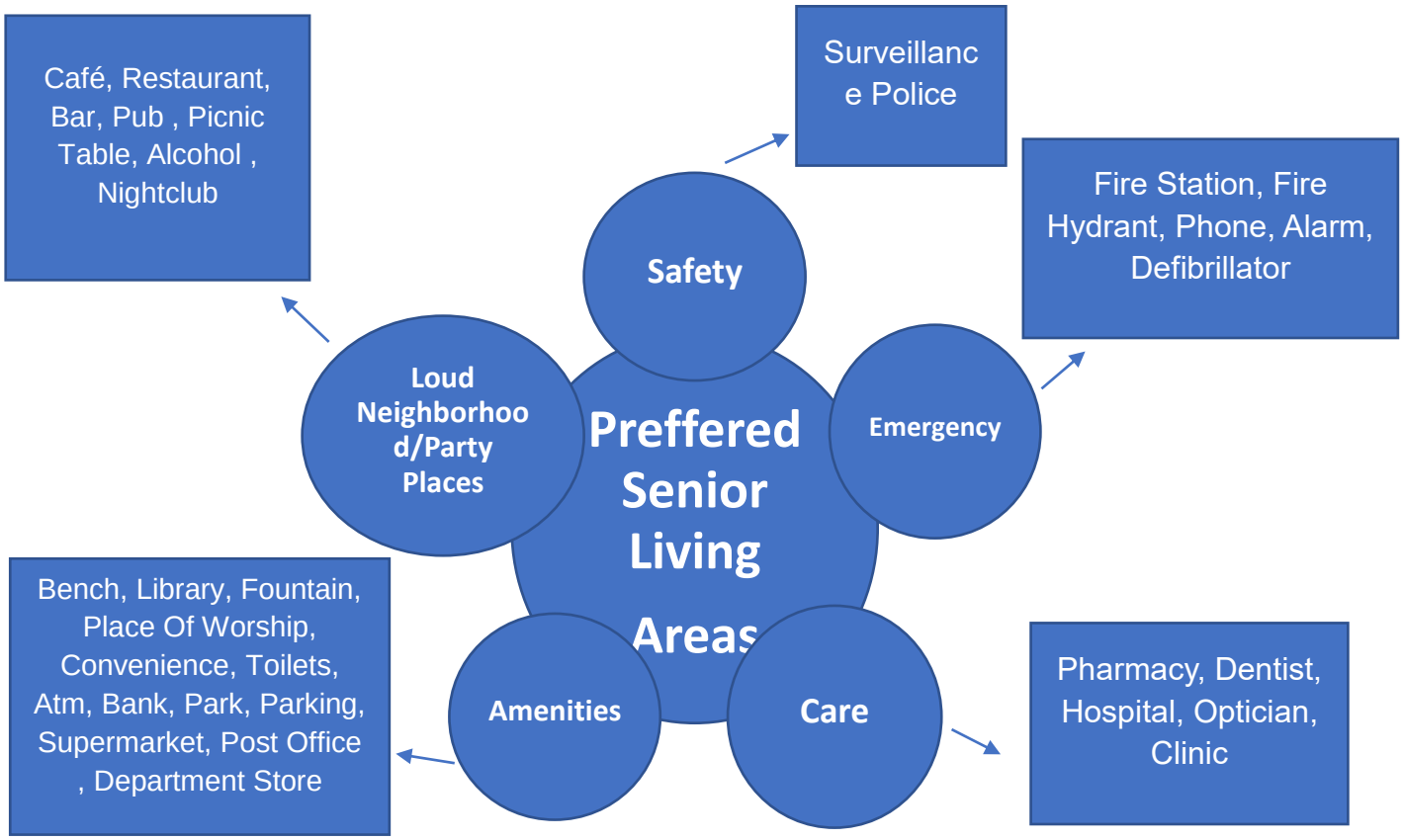


Figure 1- Latent Construct- Preferred Senior Living

The main components for Preferred Senior Living include Emergency, Care, Safety, Amenities and Loud Neighborhood/Party_Places. The component Emergency further includes buildings like Fire Stations, Phone, Alarm etc. The component Care further includes buildings like Hospital, Optician, Dentist etc. The component of Amenities includes the regularly needed Amenities like park, shops, benches, supermarkets etc. The last component Loud Neighborhood includes places like Bar, Pub, Restaurant and Cafes which are presumed to be places from where Seniors would like to stay away as they are loud in nature.

Methods

a) Main Dataset

The base dataset with record level file for this project is the OpenStreetMap dataset. The database contains OpenStreetMap (OSM) point locations from three counties in the Boston Metropolitan Area. This dataset has 19,534 rows and 11 columns. OSM is a collaborative project

that produces a free, open, and editable map of the world. The original dataset contains 11 variables by which these locations can be described; some of these variables are identifying characteristics of the building/places and the others are geographic information about the building/spaces. These two categories of variables can be further broken down. Of the application characteristics variables, ID and Name help to identify a place/thing, while Tag 1/Type 1 and Tag 2/Type 2 help to classify it. For the geographic information variables, the X (longitude) and Y (latitude) variables are original to OSM, while the Place Geography, Place Geography Type, and Block ID variables were appended to the data after they were scraped.

a. Merged datasets

The OpenStreetMap dataset was further merged with American Community Survey data to make more sense of it. The American Community Survey is produced annually by the U.S. Census Bureau in one, three, and five-year estimates. It details basic information on demographics, race and ethnicity, economics, education levels, transportation modes, family and households' characteristics, etc.

New Measures

To identify the Preferred Senior Living from the dataset some modifications were done to create new measures.

a) New measure created at record level

The column BLK_ID_10 was divided to CT_ID_10. CT_ID_10 is the portion of the FIPS code that represents state (first two digits), county (digits 3-5) and census tract (digits 6-11). We maintain it in this format in order to conduct analysis at the tract level without interference by tracts

of the same number in other states and/or counties as well as to preserve the complete structure of the code.

b) New measure created after merging datasets

The column 'Type 1' from OpenStreetMap was further divided into 5 categories according to the component of Preferred Senior Living. The new column was named as Seniortype. The column includes entries from Emergency, Care, Amenities, Safety and Loud Neighborhood/Party Places.

Observation from city walk

To understand my analysis in depth I did few city walks in the City of Boston. The most prominent and relevant city walk was the one where I decided to visit already existing Senior Living areas to test my idea and its components. I visited two Senior Livings in the area of Allston/Brighton. My first stop was the Providence House Assisted Senior Living and my second stop was Care Connection at Sussman Ho. I had various observations and conclusions I could make from my walk.

1) First and the most important observation, I realized though I thought Seniors would not prefer living near a busy or loud neighborhood with cafes and busy places, but my walk tells me a completely different story. The area near the Susman house Senior Living was busy with a lot of cafes and restaurants around but still the Seniors lived there.
2) The second observation was Senior Livings are always in between or close to neighborhoods. Both the Senior Living I visited were in residential areas.
3) My third observation was Seniors like to stay in places surrounded by parks, benches and proper walkways, this I observed in both the places I visited.

Results and Analysis

a) Hypothesis Testing – T Test

First step towards analysis of my data was to test what I observed during city walk mathematically. For this purpose, I started off with hypothesis testing and decided to run a t test to understand if Seniors are affected by the number restaurants and cafes or not. My Null and alternate Hypothesis for this analysis is as follows

H0=Mean of variable ageO65 in tracts with more restaurants is equal to the mean of variable ageO65 in tracts with less restaurants

HA= Mean of variable ageO65 in tracts with more restaurants is not equal to the mean of variable ageO65 in tracts with less restaurants

I aggregated the number of restaurants and cafes by tracts. It was observed from the aggregated table each census tract has a minimum of 1 restaurant. Therefore, the census tracts were broken down based on number of restaurants. Thus, if a census tract has more than 5 restaurants then it will be classified as 'many restaurants' otherwise it will be called 'few restaurant'. Similarly, for cafes if a census tract has more than 5 cafes then it will be classified as 'many cafes' otherwise it will be called ‘few cafes’. The T test produced the following result.

Table 1- T Test (Restaurants)

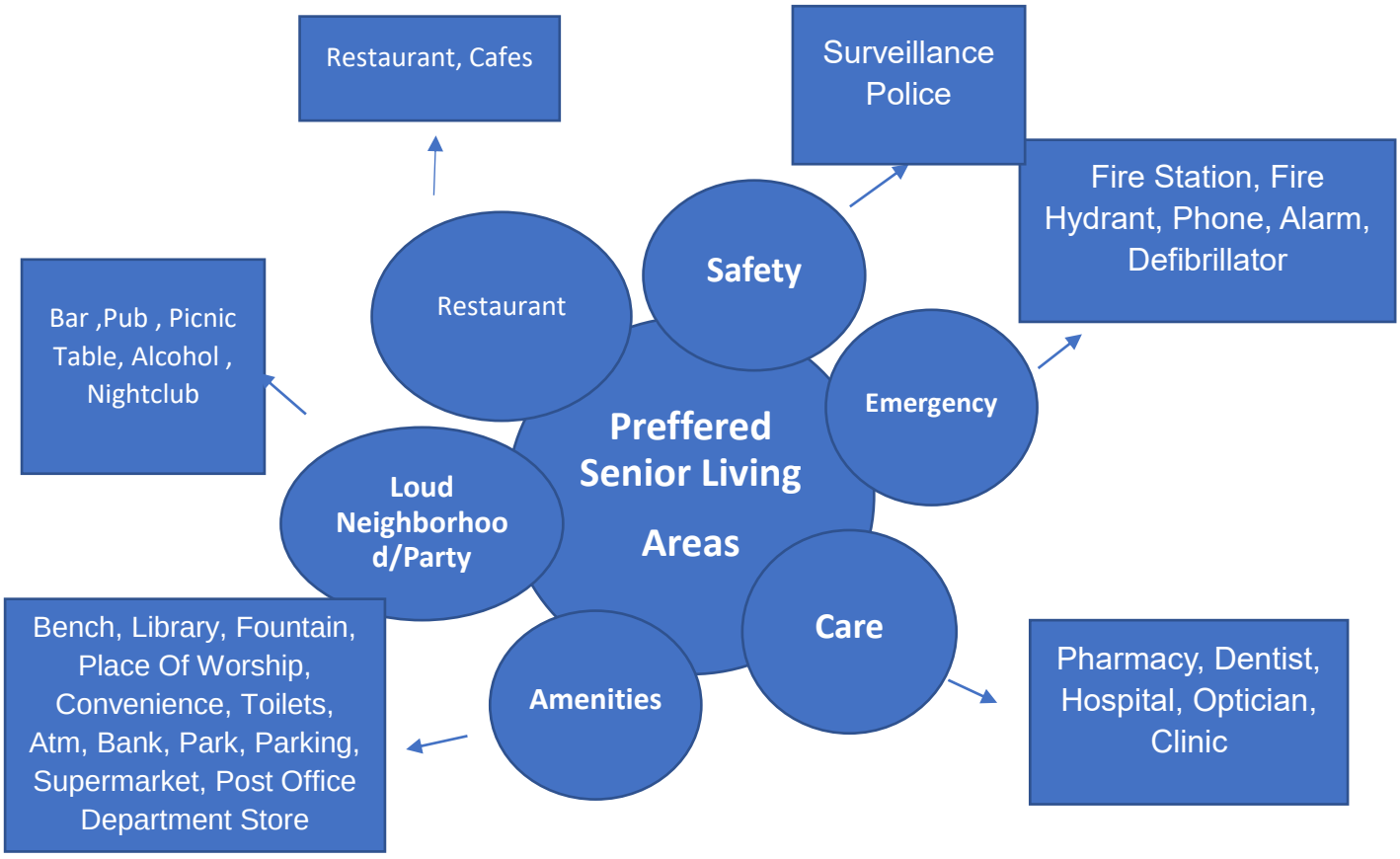
Welch Two Sample t-test
data: final_res$AgeO65 by final_res$res1
t = -0.39426, df = 128.09, p-value = 0.694
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.02864357 0.01912527
sample estimates:
mean in group few restaurants mean in group many restaurants
0.1094743 0.1142334

Table 2- T Test (Cafes)

Welch Two Sample t-test
data: final_cafe$AgeO65 by final_cafe$cafe1
t = -0.39426, df = 128.09, p-value = 0.694
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.02864357 0.01912527
sample estimates:
mean in group few cafe mean in group many cafe
0.1094743 0.1142334

We can see from the results of T test that mean difference is not significant and there is not much difference. But the p value is greater than 0.05 in this case we cannot reject the null hypothesis. At a 5% significance and a 95% confidence interval we fail to reject the null hypothesis and fail to accept the alternate hypothesis because the p value is greater than 0.05. Therefore, we can say that Seniors are not really affected by presence of restaurants and cafes in an area, this I also observed in the city walk. This hypothesis testing gives me a new component to be added in my latent

Figure 2- Modified Latent Construct



b) Hypothesis Testing- Anova

I will be running ANOVA between Age65 and LndType. The variable LndType contains 4 levels - Residential, Institution, Downtown and Park. From this ANOVA we will compare means of age 65 among 4 different land types and check if differences are statistically significant.

The Null and alternate hypothesis are as follows

H0: All 4 land types means are equal, i.e. no relationship between land types and age65

HA: Not all 4 land types means are equal, i.e. there is relationship between land types and age65

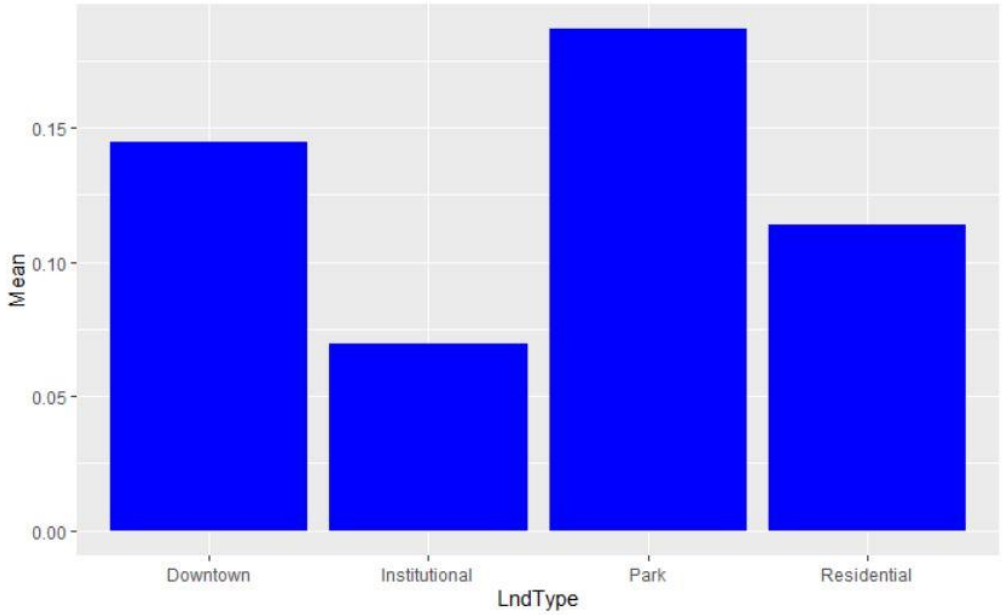
The results indicated that f value is 6.115 and p value is significantly low, in other words the variation of age65 means among different land types is much larger than the variation of age65 within each land type. Hence, we can conclude that for our confidence interval of 95% we reject the null hypothesis and accept the alternative hypothesis that there is significant relationship between land types and age65.

Table 3-Anova of age 65 and land type

[1] "aov" "lm"				
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dataanova$LndType	3	0.1214	0.04045	6.115	0.000581 ***
Residuals	159	1.0519	0.00662		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Further in order to differentiate which land type is different from other the TukeyHSD test was run along with Anova. The results are depicted by the graph below



Graph 1- Mean of Age 65 people according to land types

I can see from the graph that Parks and downtown are above average, followed by Residential and Intuition has the least mean. Thus, we can say that Age 65 people/Seniors prefer Parks and downtown areas the most for living and Institutional areas are least preferred areas.

c) Aggregation according to tracts

In this section the components of Preferred Senior Living namely Emergency, Care, Safety, Loud Neighborhood/Party Places, Amenities and Restaurants along with number of people belonging to age 65 years and above were aggregated according to tracts. This was done to get an overview of counts. The head of the table looks like follows.

Table 4- Head of Aggregation by tract

Tract	Emergency	Safety	Party Places/ Loud Nbhd	Care	Amenities	Restaurant	N65
000100	1	10	2	2	41	7	417
000201	0	0	0	0	9	1	396
000202	0	0	0	2	6	5	459
000301	1	0	0	0	5	3	297
000302	0	0	1	0	4	2	424
000401	1	0	0	0	0	2	1083

Also, component wise maps of neighborhoods were created in order to get an overall idea for the mean of the component variables by each neighborhood.

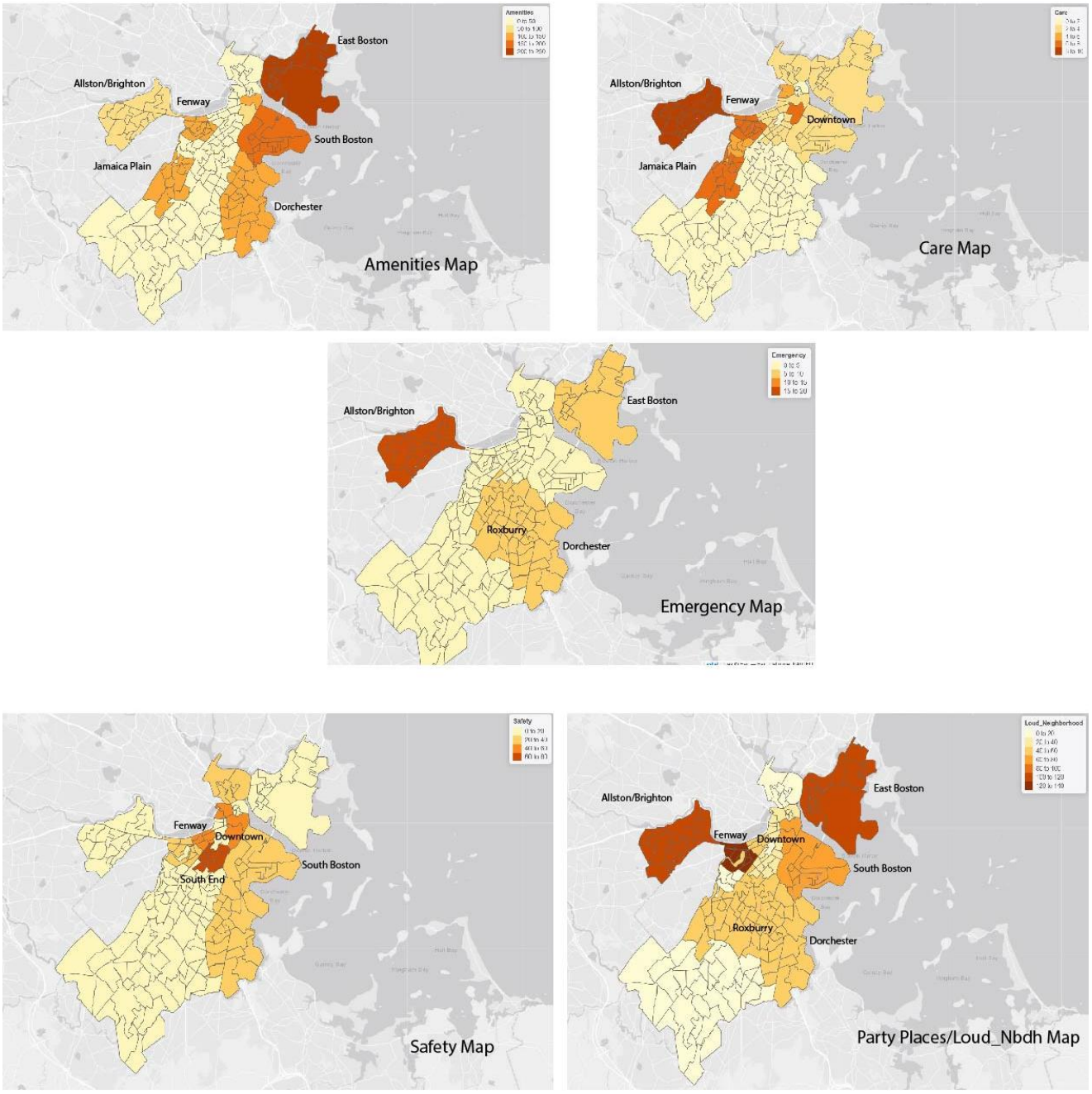


Figure 3- Maps of components according to neighborhoods

The darker the color more is the number of building/places in each category of the component variables.

d) Correlation and Regression

Now that I had the count by tracts for the components and the number of people who were age 65 and above, I decided to run a correlation to look for relationship between the variables. The correlation matrix shows N65 has relatively high correlations with Care and Safety, relatively moderate correlation with Restaurant and Party places, low correlation with Amenities and negative correlation with Emergency.

Table 5- Correlation

N65	Correlation Coef.
Emergency	-0.006358332
Care	0.190058299
Party Places	0.036339896
Safety	0.155153995
Amenities	0.016327074
Restaurant	0.085747479
N65	1.000000000

Overall, the correlation coefficients are low. However relatively, Care and Safety have the maximum correlation which suggests that tracts with high number of Seniors have a high number of Care and Safety places. On the other hand, there is no significant correlation of Emergency, Party Places, Restaurants and Amenities which means that the tracts with high number of Seniors may or may not have high number of these places. Thus, we can say that Seniors prefer living in tracts with more Care and Safety places and are not much affected by other places.

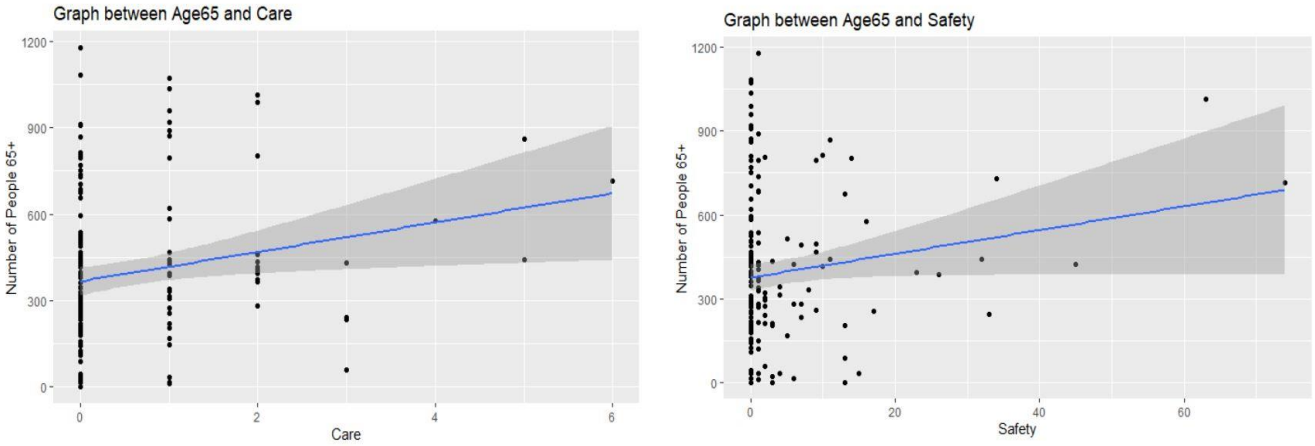
Next step was to run a multivariate Regression to check for the line of best fit and find out the effect of component variables on Senior population. I ran the Regression Analysis with N65 as the target variable and Emergency, Care, Safety, Amenities, Restaurant and Party Places as the independent variables. The Regression Analysis tells me that Care is the most significant variable with a p value less than 0.05 followed by Safety. This analysis makes sense as Seniors would prefer to stay close to care places like hospitals, nursing homes given their health and accessibility issues.

```
Residuals:
    Min      1Q  Median      3Q     Max
-471.54 -177.66  -50.38  133.10  806.42

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          367.3165    28.3147  12.973   <2e-16 ***
DataProject$Emergency  2.8408    29.8465   0.095   0.9243
DataProject$Care      50.3015    24.7822   2.030   0.0442 *
DataProject$PP       -18.6930    18.4168  -1.015   0.3117
DataProject$Restaurant  1.4073     5.5495   0.254   0.8002
DataProject$Safety      4.2584     2.9362   1.450   0.1491
DataProject$Amenities  -0.6055     1.2669  -0.478   0.6334
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3- Regression Analysis

Further visualizations were created to make more sense of the inferential statistic calculation.



Graph 2 & 3- Regression graph of Age 65 with Care and Safety

The two graphs represent positive correlation between Number of people above age 65 and Care/Safety. The blue line represent correlation between the variables and the grey area represents the standard error, the range in which the regression predict values for Number of people above age 65. The graph depicts the same results are discussed above. The areas which Seniors preffer are areas with more Care buildings nearby like Hospitals, Clinics and Safety buildings like Police stations, surveillance etc.

Discussion and Conclusion

Final Score Calculation

For the purpose of calculating the Preferred Senior Living, I have used the technique of weighted sum. Noting from the previous observations I have assigned a weight of 1 to Amenities, and Emergency, a weight of 3 to Care as this is a highly correlated variable, a weight of 2 to Restaurants and Safety as they were next on correlation. But Party Places/Loud Neighborhood has been assigned a weight of -1. Before calculating the score, my next step was to normalize the values in order to do bring all the subcomponents on a common scale without distorting differences in the range of values.

Final score = 3XCare+2XRestaurant+2XSafety+1Emergency+1Amenities -1PP

The final score depicts the following table, the top 3 entries have been listed here.

Table 6- Final Score Table

tract	Emergency	Care	Party Places	Restaurant	Amenities	Safety	N65	Score
070101	0.28	1	1	1	0.4	1	714	4.70
020301	0	0.8	0.05	0.05	0.12	0.4	441	3.41
120400	0.14	0.83	0.05	0.16	0.41	0	862	2.99

070101 and 020301 tracts belong to Downtown and West End areas while 120400 belongs to areas around Jamaica Plain. From these scores I found that most preferred Senior living tracts are the ones in and around Downtown Boston. Which also was depicted in the Anova test that Seniors prefer Park and residential tracts around Downtown for living. Preferably because this area has

high numbers of Care and Safety Buildings along with amenities like lots of Parks, Benches and walkways.

This concludes my study with some major learning from this study:

1. First, seniors are not affected by restaurants and cafes nearby
2. Second, seniors preffer living in areas which have care and safety units nearby
3. Third, Seniors preffer living in Downtown Boston and nearby areas.

Reference

Hawes et all. 2003, A National Survey of Assisted Living Facilities (43). *The Gerontologist*. https://academic.oup.com/gerontologist/article/43/6/875/863127

Appendix

title: "R Notebook"

output: html_notebook

#loading libraries

```{r}

library(tidyverse)

library(ggplot2)

#reading csv

osm=read.csv("osm2.2+Nbhds.csv")

#breaking tract from CT_ID_10

osm=osm %>% mutate(

  tract = substr(as.character(CT_ID_10),6,11))

#filtering data according to components of latent construct (naming the column 'seniortype')

```{r}

osm$seniortype=ifelse(osm$Type1 %in% c('phone', 'alarm', 'defibrillator', 'fire_station'), "Emergency", ifelse(osm$Type1 %in% c('pharmacy','dentist', 'hospital','optician'), "Care", ifelse(osm$Type1 %in% c( 'bar', 'pub' ,'alcohol','nightclub'), "Party Places", ifelse(osm$Type1 %in% c('surveillance','police'),"Safety",ifelse(osm$Type1 %in% c('restaurant','cafe'),"Restaurant", ifelse(osm$Type1 %in% c('bench','library','fountain','place of worship', 'convenience','atm', 'bank', 'parking','supermarket','haridresser', 'post office', 'department store','park'),"Amenities", "other"))))))

#converting seniortype to factor

osm$seniortype=as.factor(osm$seniortype)

#aggreating variables of senior type according to tracts

#Filtering Emergency out of seniortype

Etype=filter(osm, osm$seniortype=='Emergency')

```
#Aggregrating Emergency with neighborhood
Emergency=aggregate(Etype$seniortype~Etype$tract, data = Etype, FUN = length)
colnames(Emergency)=c("tract","Emergency")

#Filtering Safety out of seniortype
Stype=filter(osm, osm$seniortype=='Safety')

#Aggregrating Safety with neighborhood
Safety=aggregate(Stype$seniortype~Stype$tract, data = Stype, FUN = length)
colnames(Safety)=c("tract","Safety")

#Filtering Party Places out of seniortype

Ptype=filter(osm, osm$seniortype=='Party Places')

#Aggregrating Loud Neighborhood/party places with neighborhood
Party_Places=aggregate(Ptype$seniortype~Ptype$tract, data = Ptype, FUN = length)
colnames(Party_Places)=c("tract","PP")

#Filtering Amenities out of Aggtype
Atype=filter(osm, osm$seniortype=='Amenities')
Amenities=aggregate(Atype$seniortype~Atype$tract, data=Atype, FUN = length)
colnames(Amenities)=c("tract","Amenities")

#Filtering Care out of seniortype

Ctype=filter(osm, osm$seniortype=='Care')

#Aggregrating Care with neighborhood
Care=aggregate(Ctype$seniortype~Ctype$tract, data = Ctype, FUN = length)
```

```
colnames(Care)=c("tract","Care")

#filtering restaurant out of seniortype
Rtype=filter(osm, osm$seniortype== 'Restaurant')
Restaurant=aggregate(Rtype$seniortype~Rtype$tract, data=Rtype, FUN = length)
colnames(Restaurant)=c("tract","Restaurant")

#merging full outer join of Emergency and care
Data1=merge(Emergency, Care, by='tract',all = TRUE)

#merging full outer join of Party_Places and Safety
Data2=merge(Party_Places, Safety, by='tract', all = TRUE)

#merging full outer join of Data2 and Amenities
Data3=merge(Data2,Amenities, by='tract',all = TRUE)

#Merging full outer join of Data 1 and Data 3

Datafinal=merge(Data1, Data3, by='tract',all = TRUE)

Datafinal=merge(Datafinal,Restaurant, by='tract', all = TRUE)

Datafinal[is.na(Datafinal)]=0

Datafinal
#ACS merged data reading
data=read.csv("osm2.2+ACS+Nbhds2.csv")

#age65 and tract
data1=as.data.frame(data[c(2,33,39)])
```

#dividing tract

data1=data1 %>% mutate(
  tract = substr(as.character(CT_ID_10),6,11))

#unique 65
DT=as.data.frame(data1[c(4,3,2)])
DT=unique(DT)

DT$N65=DT$TotalPop*DT$AgeO65

data3=as.data.frame(DT[c(1,4)])

#merging with Datafinal

DataProject=merge(x=Datafinal,y=data3, by='tract', all.x=TRUE)

DataProject[is.na(DataProject)]=0

write.csv(DataProject,"C:/Users/1992h/Documents/Big Data/new osm/table.csv")
```
```{r}
#correlation
summary(DataProject)

DataProject1=as.data.frame(DataProject[c(2,3,4,5,6,7,8)])

cor(DataProject1)

```
#regression

reg=lm(DataProject$N65~DataProject$Emergency+DataProject$Care+DataProject$PP+DataProject$Restaurant+DataProject$Safety+DataProject$Amenities)

summary(reg)

#reg graph

library(ggplot2)

graph1=ggplot(data=DataProject, aes(x=DataProject$Care, y=DataProject$N65)) + geom_point() + xlab("Care") + ylab("Number of People 65+") + geom_smooth(method=lm)+ ggtitle("Graph between Age65 and Care")

graph1


graph2=ggplot(data=DataProject, aes(x=DataProject$Safety, y=DataProject$N65)) + geom_point() + xlab("Safety") + ylab("Number of People 65+") + geom_smooth(method=lm)+ ggtitle("Graph between Age65 and Safety")

graph2


#calculating beta values

install.packages("QuantPsyc")

library(QuantPsyc)


lm.beta(reg)

lm.beta(reg2)


#normalizing


#normalizing data

Normalize_data=DataProject

FUN=function(x) (x-min(x))/diff(range(x))

Normalize_data$Emergency=FUN(Normalize_data$Emergency)

Normalize_data$Safety=FUN(Normalize_data$Safety)

Normalize_data$Amenities=FUN(Normalize_data$Amenities)
```

```
Normalize_data$Care=FUN(Normalize_data$Care)

Normalize_data$PP=FUN(Normalize_data$PP)

Normalize_data$Restaurant=FUN(Normalize_data$Restaurant)

Normalize_data


#Final Score

#Final score = 3XCare+2XRestaurant+2XSafety+1Emergency+1Amenities -
1XLoud_Neighborhood


Normalize_data$Scores=3*Normalize_data$Care+2*Normalize_data$Safety+1*Normalize_data
$Emergency+1*Normalize_data$Amenities-1*Normalize_data$PP
```