

Team 80 - There Goes the Neighborhood: Predicting Gentrification in the United States

One of the biggest challenges with gentrification is that people are not aware that it is happening in a community until it is too late to address it and native residents have already been displaced. If local governments are aware that their communities are likely to become gentrified, they can allocate resources, gather community input and enact policies to avoid widespread displacement of native residents. In this project, we develop a model to predict which U.S. counties are at risk of becoming gentrified based on several predictors.

```
## Warning: package 'caTools' was built under R version 3.6.3
```

```
## Warning: package 'DMwR' was built under R version 3.6.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 3.6.3
```

```
## Loading required package: grid
```

```
## Registered S3 method overwritten by 'quantmod':  
##   method             from  
##   as.zoo.data.frame zoo
```

```
## Warning: package 'mlbench' was built under R version 3.6.3
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
## Warning: package 'caret' was built under R version 3.6.3
```

```
## Loading required package: ggplot2
```

```
##  
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':  
##  
##   margin
```

```
## Warning: package 'xlsx' was built under R version 3.6.3
```

```
## Warning: package 'psych' was built under R version 3.6.3
```

```
##  
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':  
##  
##    %+%, alpha
```

```
## The following object is masked from 'package:randomForest':  
##  
##    outlier
```

```
## The following object is masked from 'package:DMwR':  
##  
##    crossValidation
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##  
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:randomForest':  
##  
##    combine
```

```
## The following objects are masked from 'package:stats':  
##  
##    filter, lag
```

```
## The following objects are masked from 'package:base':  
##  
##    intersect, setdiff, setequal, union
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages -----  
----- tidyverse 1.3.0 --
```

```
## v tibble 3.0.1      v stringr 1.4.0
## v readr  1.3.1      v forcats 0.5.0
## v purrr   0.3.4
```

```
## Warning: package 'tibble' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.3
```

```
## Warning: package 'forcats' was built under R version 3.6.3
```

```
## -- Conflicts -----
----- tidyverse_conflicts() --
## x psych::%+%( )      masks ggplot2::%+%( )
## x psych::alpha( )    masks ggplot2::alpha( )
## x dplyr::combine( )  masks randomForest::combine( )
## x dplyr::filter( )   masks stats::filter( )
## x dplyr::lag( )      masks stats::lag( )
## x purrr::lift( )     masks caret::lift( )
## x ggplot2::margin( ) masks randomForest::margin( )
```

```
## Warning: package 'ggpubr' was built under R version 3.6.3
```

```
## Warning: package 'rstatix' was built under R version 3.6.3
```

```
##
## Attaching package: 'rstatix'
```

```
## The following object is masked from 'package:stats':
##
##   filter
```

```
## Warning: package 'neuralnet' was built under R version 3.6.3
```

```
##
## Attaching package: 'neuralnet'
```

```
## The following object is masked from 'package:dplyr':
##
##   compute
```

```
## Warning: package 'partykit' was built under R version 3.6.3
```

```
## Loading required package: libcoin
```

```
## Warning: package 'libcoin' was built under R version 3.6.3
```

```
## Loading required package: mvtnorm
```

```
## Warning: package 'mvtnorm' was built under R version 3.6.3
```

Our main dataset is sourced from the “2020 Healthiest Communities rankings”. The 2020 Healthiest Communities rankings were created in collaboration with the Aetna Foundation, an independent affiliate of CVS Health. The University of Missouri Extension Center for Applied Research and Engagement Systems performed data collection and analysis. This data is publicly available through the following website which has a profile for each U.S. county and shows their respective performance across 84 metrics in 10 categories (Population Health, Equity, Education, Economy, Housing, Food & Nutrition, Environment, Public Safety, Community Vitality, and Infrastructure):

<https://www.usnews.com/news/healthiest-communities/district-of-columbia/district-of-columbia>

(<https://www.usnews.com/news/healthiest-communities/district-of-columbia/district-of-columbia>)

First, we load in our data after we have webscraped it and notice that it is tall.

```
df <- read.csv("https://raw.githubusercontent.com/omarp120/Predicting-Gentrification/main/us_new
s_factors.csv")
head(df)
```

```
##   X      County_Name State_Name
## 1 0 los-alamos-county new-mexico
## 2 1 los-alamos-county new-mexico
## 3 2 los-alamos-county new-mexico
## 4 3 los-alamos-county new-mexico
## 5 4 los-alamos-county new-mexico
## 6 5 los-alamos-county new-mexico
##
##                               Variables County    US Peer_Group
## 1                        Hospital Bed Availability /1k    2.8    2.0      2.1
## 2                Population With No Health Insurance    3.0% 10.6%      8.1%
## 3                Primary Care Doctor Availability /100k    2.5    1.1      1.5
## 4        Adults With No Leisure-Time Physical Activity 12.5% 25.9%     22.9%
## 5 Medicare Beneficiaries With Recent Primary Care Visit 31.0% 21.0%     27.0%
## 6                        Smoking Rate    10.0% 17.3%     15.6%
##   State
## 1    3.0
## 2 11.0%
## 3    1.6
## 4 19.3%
## 5 14.6%
## 6 15.5%
```

We also drop any columns that we will not use at this time, like the state and peer group comparisons for each county.

```
df2 <- subset(df, select = c(County_Name, State_Name, Variables, County))
head(df2)
```

```
##      County_Name State_Name
## 1 los-alamos-county new-mexico
## 2 los-alamos-county new-mexico
## 3 los-alamos-county new-mexico
## 4 los-alamos-county new-mexico
## 5 los-alamos-county new-mexico
## 6 los-alamos-county new-mexico
##                                     Variables County
## 1                        Hospital Bed Availability /1k    2.8
## 2                Population With No Health Insurance    3.0%
## 3            Primary Care Doctor Availability /100k    2.5
## 4    Adults With No Leisure-Time Physical Activity    12.5%
## 5 Medicare Beneficiaries With Recent Primary Care Visit    31.0%
## 6                                Smoking Rate    10.0%
```

Here we transform our data from tall to wide so that each column represents a variable. We have a total of 500 counties and 84 different variables that we can use as inputs for our model.

```
df_wide <- df2 %>% spread(Variables, County)
head(df_wide)
```

##	County_Name	State_Name	Accidental Death Rate /100k	
## 1	ada-county	idaho	37.8	
## 2	adair-county	iowa	50.3	
## 3	addison-county	vermont	45.2	
## 4	alameda-county	california	24.7	
## 5	albany-county	wyoming	49.2	
## 6	albemarle-county	virginia	32.6	
##	Adults in Poor or Fair General Health			
## 1			11.6%	
## 2			11.7%	
## 3			11.4%	
## 4			12.7%	
## 5			16.0%	
## 6			12.9%	
##	Adults With No Leisure-Time Physical Activity Affordable Housing Shortfall			
## 1			14.6%	-76.3
## 2			25.2%	-28.0
## 3			19.2%	-60.5
## 4			14.5%	-66.2
## 5			17.6%	-72.9
## 6			16.9%	-80.6
##	Air-Quality Hazard Air Toxics Exposure Disparity Index Score			
## 1		0.68		1.32
## 2		0.24		N/A
## 3		0.25		4.10
## 4		0.53		1.44
## 5		0.22		2.13
## 6		0.46		7.52
##	Airborne Cancer Risk Area With Tree Canopy Average Weekly Wage			
## 1		35.96	1.0%	\$970
## 2		18.68	1.6%	\$706
## 3		21.15	39.5%	\$923
## 4		30.96	2.7%	\$1,466
## 5		17.18	0.7%	\$784
## 6		33.78	55.1%	\$1,049
##	Babies Born With Low Birth Weight Business Growth Rate			
## 1			6.4%	1.2%
## 2			5.7%	8.1%
## 3			6.6%	7.3%
## 4			7.3%	1.1%
## 5			8.8%	10.5%
## 6			6.3%	10.7%
##	Cancer Incidence Rate /100k Census Self-Response Rate Change in Housing Value			
## 1		455.3	72.8%	32.2%
## 2		449.5	75.0%	12.6%
## 3		413.7	70.3%	2.9%
## 4		392.8	61.5%	39.0%
## 5		378.1	71.2%	10.5%
## 6		424.5	72.8%	8.6%
##	Child Care Facilities /100k Children Meeting Standards in Grade 4 ELA			
## 1		1.7		53.0%
## 2		0.0		68.7%
## 3		10.8		54.5%

## 4	6.1	50.5%
## 5	2.6	68.0%
## 6	4.6	78.0%
## Continuing Education Tax Credits as Share of Total Tax Filings		
## 1	12.3%	
## 2	7.5%	
## 3	8.1%	
## 4	9.8%	
## 5	16.8%	
## 6	9.9%	
## Deaths of Despair /100k Diabetes Prevalence Disability Employment Gap		
## 1	38.0	7.4% 0.71
## 2	43.4	9.4% 0.90
## 3	24.3	7.5% 0.74
## 4	26.8	6.7% 0.64
## 5	44.0	6.2% 0.91
## 6	19.6	6.2% 0.77
## Eviction Rate Extreme Heat Days per Year days Food Environment Index Score		
## 1	0.4%	7.3 9.84
## 2	0.3%	8.3 9.00
## 3	0.1%	8.7 18.97
## 4	0.6%	10.3 9.54
## 5	N/A	3.3 5.43
## 6	2.2%	8.7 9.63
## Gini Index Score Heart Disease Prevalence Among Medicare Beneficiaries		
## 1	0.47	19.0%
## 2	0.43	21.0%
## 3	0.41	22.0%
## 4	0.46	20.0%
## 5	0.47	18.0%
## 6	0.48	20.0%
## High School Graduation Rate Homeownership Rate Hospital Bed Availability /1k		
## 1	84.0%	68.4% 3.4
## 2	93.8%	76.7% 3.5
## 3	N/A	74.1% 0.7
## 4	86.5%	53.3% 3.0
## 5	N/A	49.7% 2.6
## 6	91.0%	62.5% 1.9
## Households in Flood Hazard Zone Households Receiving Public Assistance Income		
## 1	3.1%	2.3%
## 2	0.1%	1.7%
## 3	3.7%	2.2%
## 4	1.9%	3.6%
## 5	2.4%	1.1%
## 6	1.6%	1.0%
## Households Spending at Least 30% of Income on Housing		
## 1	27.0%	
## 2	18.4%	
## 3	31.7%	
## 4	37.6%	
## 5	36.0%	
## 6	27.6%	
## Households With Incomplete Plumbing Facilities		
## 1	0.3%	

## 2	0.0%
## 3	0.6%
## 4	0.4%
## 5	0.5%
## 6	0.3%
## Households With Internet Access Households With No Vehicle	
## 1	98.0% 3.8%
## 2	85.1% 3.9%
## 3	92.5% 4.6%
## 4	100.0% 9.6%
## 5	97.8% 4.8%
## 6	93.0% 4.7%
## Idle Youth (Not Working or Enrolled) Job Diversity Index Score	
## 1	1.4% 0.89
## 2	0.0% 0.60
## 3	0.4% 0.81
## 4	1.9% 0.90
## 5	0.0% 0.74
## 6	0.2% 0.83
## Jobs Within a 45-Minute Commute Labor Force Participation Life Expectancy	
## 1	48,436 65.7% 80.8
## 2	2,000 64.8% 79.4
## 3	3,618 66.8% 80.8
## 4	217,076 66.8% 81.5
## 5	6,848 68.7% 80.5
## 6	20,181 61.5% 81.2
## Local Food Outlets /100k Median Household Income Medical Debt in Collections	
## 1	3.3 \$63,137 15%
## 2	27.3 \$51,859 11%
## 3	27.1 \$65,093 11%
## 4	2.2 \$92,574 7%
## 5	5.3 \$46,865 17%
## 6	8.6 \$75,394 11%
## Medicare Beneficiaries With Depression	
## 1	21.0%
## 2	17.0%
## 3	19.0%
## 4	14.0%
## 5	15.0%
## 6	18.0%
## Medicare Beneficiaries With Recent Primary Care Visit	
## 1	39.0%
## 2	15.0%
## 3	28.0%
## 4	21.0%
## 5	26.0%
## 6	40.0%
## Natural Amenities Index Score	
## 1	1.87
## 2	-1.59
## 3	-0.28
## 4	5.13
## 5	4.91
## 6	-0.02

##	Neighborhood Disparity in Educational Attainment		
## 1		16.01	
## 2		4.81	
## 3		9.98	
## 4		23.41	
## 5		15.11	
## 6		20.53	
##	Neighborhood Disparity in Poverty	Net Migration Rate	Nonprofits /100k
## 1	7.35	8.7%	49.6
## 2	2.14	-5.7%	75.5
## 3	2.62	0.2%	84.2
## 4	9.96	-1.3%	52.9
## 5	15.49	-6.5%	65.8
## 6	19.77	4.9%	65.5
##	Obesity Prevalence	Overcrowded Households	Per-Pupil Expenditures
## 1	25.3%	1.4%	\$7,562
## 2	31.3%	0.5%	\$13,849
## 3	24.8%	0.9%	\$20,325
## 4	19.2%	7.7%	\$13,434
## 5	24.2%	2.4%	\$29,405
## 6	20.3%	1.2%	N/A
##	Per Capita Spending on Health and Emergency Services		
## 1		\$509	
## 2		\$222	
## 3		\$137	
## 4		\$1,363	
## 5		\$383	
## 6		\$749	
##	Poor Mental Health Days per Month		
## 1	3.7		
## 2	3.2		
## 3	3.8		
## 4	3.3		
## 5	3.7		
## 6	3.3		
##	Population Living Close to Emergency Facilities		
## 1		32.5%	
## 2		60.4%	
## 3		24.3%	
## 4		74.9%	
## 5		47.8%	
## 6		20.0%	
##	Population With Advanced Degree	Population With No Health Insurance	
## 1	47.2%	8.7%	
## 2	29.3%	5.5%	
## 3	45.6%	6.2%	
## 4	52.4%	5.8%	
## 5	58.3%	12.5%	
## 6	59.1%	10.0%	
##	Population Within 0.5 Mile of a Park		
## 1	35.2%		
## 2	48.6%		
## 3	13.4%		
## 4	87.0%		

## 5	62.1%	
## 6	17.7%	
##	Population Within 0.5 Mile of Walkable Destinations	
## 1	11.2%	
## 2	0.3%	
## 3	13.5%	
## 4	56.5%	
## 5	38.0%	
## 6	17.6%	
##	Population Without Access to Large Grocery Store Poverty Rate	
## 1	27.5%	11.0%
## 2	6.4%	9.1%
## 3	4.0%	7.7%
## 4	7.6%	10.6%
## 5	36.7%	22.6%
## 6	11.9%	9.8%
##	Premature Death Disparity Index Score Preschool Enrollment	
## 1	0.02	37.6%
## 2	N/A	56.7%
## 3	N/A	68.1%
## 4	0.14	56.4%
## 5	0.01	46.0%
## 6	0.06	58.4%
##	Preventable Hospital Admissions Among Medicare Beneficiaries /100k	
## 1		2,691
## 2		2,208
## 3		2,914
## 4		3,329
## 5		3,326
## 6		3,888
##	Primary Care Doctor Availability /100k Property Crime Rate /100k	
## 1	2.4	1,693.2
## 2	1.0	693.4
## 3	2.0	1,000.5
## 4	2.0	3,602.8
## 5	1.1	1,927.1
## 6	1.9	1,583.0
##	Public Safety Professionals in Population	
## 1	0.8%	
## 2	0.3%	
## 3	0.4%	
## 4	0.5%	
## 5	0.8%	
## 6	0.7%	
##	Racial Disparity in Educational Attainment Racial Disparity in Poverty	
## 1	0.25	0.10
## 2	N/A	N/A
## 3	0.04	0.01
## 4	0.27	0.14
## 5	0.40	0.02
## 6	0.18	0.15
##	Segregation Index Score Share of At-Home Food Expenditures on Fruit/Veg	
## 1	0.20	12.4%
## 2	0.54	12.0%

```
## 3      0.33      11.8%
## 4      0.28      14.5%
## 5      0.24      12.4%
## 6      0.26      12.3%
##      Smoking Rate Teen Birth Rate /1k Toxic Release Index Score Unemployment Rate
## 1      11.8%      13.4      4,847.10      2.5%
## 2      14.0%      14.0      0.01      2.4%
## 3      13.3%      7.8      4.58      2.4%
## 4      10.7%      11.3      115,760.65      3.1%
## 5      16.0%      9.9      4,372.47      3.2%
## 6      12.4%      6.2      3.40      2.7%
##      Unsafe Drinking Water Vacant Houses Vehicle Crash Fatality Rate /100k
## 1      0.1%      1.0%      7.2
## 2      2.3%      1.1%      47.7
## 3      0.0%      0.8%      18.1
## 4      0.0%      0.4%      5.6
## 5      0.0%      3.2%      15.6
## 6      0.1%      0.3%      10.4
##      Violent Crime Rate /100k Voter Participation Rate Walkability Index Score
## 1      230.8      65.5%      8.8
## 2      103.0      65.0%      6.3
## 3      78.6      60.0%      6.2
## 4      628.8      59.8%      13.7
## 5      137.9      54.9%      7.8
## 6      100.2      73.9%      7.7
##      Work Hours Needed to Pay for Affordable Housing
## 1      36.0
## 2      39.0
## 3      44.0
## 4      54.0
## 5      52.0
## 6      53.0
##      Workers Commuting 60 Minutes or More Youth Within 5 Miles of a Public School
## 1      2.5%      99.5%
## 2      9.7%      71.4%
## 3      7.5%      53.9%
## 4      17.3%      99.9%
## 5      2.5%      91.6%
## 6      3.6%      90.8%
```

In this section, we export our file to Excel so that we can convert our variables to the correct numerical formats and ratios (double or integer), and then we re-import our file into R. This will allow us to properly impute missing values.

```
#write.csv(df_wide, 'all-us-counties-transformed.csv')
df_trans <- read.csv("all-us-counties-transformed.csv")
head(df_trans)
```

##	County_me	State_me	AccidentalDeathRate
## 1	ada-county	idaho	0.000378
## 2	adair-county	iowa	0.000503
## 3	addison-county	vermont	0.000452
## 4	alameda-county	california	0.000247
## 5	albany-county	wyoming	0.000492
## 6	albemarle-county	virginia	0.000326
##	AdultsinPoororFairGeneralHealth	AdultsWithNoLeisure.TimePhysicalActivity	
## 1	0.116	0.146	
## 2	0.117	0.252	
## 3	0.114	0.192	
## 4	0.127	0.145	
## 5	0.160	0.176	
## 6	0.129	0.169	
##	AffordableHousingShortfall	Air.QualityHazard	
## 1	-76.3	0.68	
## 2	-28.0	0.24	
## 3	-60.5	0.25	
## 4	-66.2	0.53	
## 5	-72.9	0.22	
## 6	-80.6	0.46	
##	AirToxicsExposureDisparityIndexScore	AirborneCancerRisk	AreaWithTreeCanopy
## 1	1.32	35.96	0.010
## 2	NA	18.68	0.016
## 3	4.10	21.15	0.395
## 4	1.44	30.96	0.027
## 5	2.13	17.18	0.007
## 6	7.52	33.78	0.551
##	AverageWeeklyWage	BabiesBornWithLowBirthWeight	BusinessGrowthRate
## 1	970	0.064	0.012
## 2	706	0.057	0.081
## 3	923	0.066	0.073
## 4	1466	0.073	0.011
## 5	784	0.088	0.105
## 6	1049	0.063	0.107
##	CancerIncidenceRate	CensusSelf.ResponseRate	ChangeinHousingValue
## 1	0.004553	0.728	0.322
## 2	0.004495	0.750	0.126
## 3	0.004137	0.703	0.029
## 4	0.003928	0.615	0.390
## 5	0.003781	0.712	0.105
## 6	0.004245	0.728	0.086
##	ChildCareFacilitiesRate	ChildrenMeetingStandardsinGrade4ELARate	
## 1	0.000017	0.530	
## 2	0.000000	0.687	
## 3	0.000108	0.545	
## 4	0.000061	0.505	
## 5	0.000026	0.680	
## 6	0.000046	0.780	
##	ContinuingEducationTaxCreditsasShareofTotalTaxFilings	DeathsofDespairRate	
## 1	0.123	0.000380	
## 2	0.075	0.000434	
## 3	0.081	0.000243	

## 4		0.098	0.000268
## 5		0.168	0.000440
## 6		0.099	0.000196
##	DiabetesPrevalence DisabilityEmploymentGap EvictionRate		
## 1	0.074	0.71	0.004
## 2	0.094	0.90	0.003
## 3	0.075	0.74	0.001
## 4	0.067	0.64	0.006
## 5	0.062	0.91	NA
## 6	0.062	0.77	0.022
##	ExtremeHeatDaysperYeardays FoodEnvironmentIndexScore GiniIndexScore		
## 1	7.3	9.84	0.47
## 2	8.3	9.00	0.43
## 3	8.7	18.97	0.41
## 4	10.3	9.54	0.46
## 5	3.3	5.43	0.47
## 6	8.7	9.63	0.48
##	HeartDiseasePrevalenceAmongMedicareBeneficiaries HighSchoolGraduationRate		
## 1		0.19	0.840
## 2		0.21	0.938
## 3		0.22	NA
## 4		0.20	0.865
## 5		0.18	NA
## 6		0.20	0.910
##	HomeownershipRate HospitalBedAvailabilityRate HouseholdsInFloodHazardZone		
## 1	0.684	0.0034	0.031
## 2	0.767	0.0035	0.001
## 3	0.741	0.0007	0.037
## 4	0.533	0.0030	0.019
## 5	0.497	0.0026	0.024
## 6	0.625	0.0019	0.016
##	HouseholdsReceivingPublicAssistanceIncome		
## 1		0.023	
## 2		0.017	
## 3		0.022	
## 4		0.036	
## 5		0.011	
## 6		0.010	
##	HouseholdsSpendingatLeast30.ofIncomeonHousing		
## 1		0.270	
## 2		0.184	
## 3		0.317	
## 4		0.376	
## 5		0.360	
## 6		0.276	
##	HouseholdsWithIncompletePlumbingFacilities HouseholdsWithInternetAccess		
## 1		0.003	0.980
## 2		0.000	0.851
## 3		0.006	0.925
## 4		0.004	1.000
## 5		0.005	0.978
## 6		0.003	0.930
##	HouseholdsWithNoVehicle IdleYouth.NotWorkingorEnrolled.		
## 1	0.038	0.014	

## 2	0.039	0.000
## 3	0.046	0.004
## 4	0.096	0.019
## 5	0.048	0.000
## 6	0.047	0.002
##	JobDiversityIndexScore	JobsWithi45.MinuteCommute
## 1	0.89	48436
## 2	0.60	2000
## 3	0.81	3618
## 4	0.90	217076
## 5	0.74	6848
## 6	0.83	20181
##	LifeExpectancy	LocalFoodOutletsRate
## 1	80.8	0.000033
## 2	79.4	0.000273
## 3	80.8	0.000271
## 4	81.5	0.000022
## 5	80.5	0.000053
## 6	81.2	0.000086
##	MedianHouseholdIncome	
## 1		63137
## 2		51859
## 3		65093
## 4		92574
## 5		46865
## 6		75394
##	MedicalDebtinCollections	MedicareBeneficiariesWithDepression
## 1	0.15	0.21
## 2	0.11	0.17
## 3	0.11	0.19
## 4	0.07	0.14
## 5	0.17	0.15
## 6	0.11	0.18
##	MedicareBeneficiariesWithRecentPrimaryCareVisit	NaturalAmenitiesIndexScore
## 1		0.39
## 2		1.87
## 3		-1.59
## 4		-0.28
## 5		5.13
## 6		4.91
##	NeighborhoodDisparityinEducationalAttainment	NeighborhoodDisparityinPoverty
## 1		0.40
## 2		-0.02
## 3		
## 4		
## 5		
## 6		
##	NetMigrationRate	NonprofitsRate
## 1	0.087	0.000496
## 2	-0.057	0.000755
## 3	0.002	0.000842
## 4	-0.013	0.000529
## 5	-0.065	0.000658
## 6	0.049	0.000655
##	ObesityPrevalence	OvercrowdedHouseholds
## 1	0.253	0.014
## 2	0.313	0.005
## 3	0.248	0.009
## 4	0.192	0.077
## 5	0.242	0.024
## 6	0.203	0.012
##	Per.PupilExpenditures	PerCapitaSpendingonHealthandEmergencyServices
## 1	7562	509
## 2	13849	222
## 3	20325	137
## 4	13434	1363
## 5	29405	383
## 6	NA	749

##	PoorMentalHealthDaysperMonth	PopulationLivingClosetoEmergencyFacilities
## 1	3.7	0.325
## 2	3.2	0.604
## 3	3.8	0.243
## 4	3.3	0.749
## 5	3.7	0.478
## 6	3.3	0.200
##	PopulationWithAdvancedDegree	PopulationWithNoHealthInsurance
## 1	0.472	0.087
## 2	0.293	0.055
## 3	0.456	0.062
## 4	0.524	0.058
## 5	0.583	0.125
## 6	0.591	0.100
##	PopulationWithin0.5MileofaPark	PopulationWithin0.5MileofWalkableDestitutions
## 1	0.352	0.112
## 2	0.486	0.003
## 3	0.134	0.135
## 4	0.870	0.565
## 5	0.621	0.380
## 6	0.177	0.176
##	PopulationWithoutAccesstoLargeGroceryStore	PovertyRate
## 1	0.275	0.110
## 2	0.064	0.091
## 3	0.040	0.077
## 4	0.076	0.106
## 5	0.367	0.226
## 6	0.119	0.098
##	PrematureDeathDisparityIndexScore	PreschoolEnrollment
## 1	0.02	0.376
## 2	NA	0.567
## 3	NA	0.681
## 4	0.14	0.564
## 5	0.01	0.460
## 6	0.06	0.584
##	PreventableHospitalAdmissionsAmongMedicareBeneficiariesRate	
## 1	0.0269	
## 2	0.0221	
## 3	0.0291	
## 4	0.0333	
## 5	0.0333	
## 6	0.0389	
##	PrimaryCareDoctorAvailabilityRate	PropertyCrimeRate
## 1	2.4e-05	0.0169
## 2	1.0e-05	0.0069
## 3	2.0e-05	0.0100
## 4	2.0e-05	0.0360
## 5	1.1e-05	0.0193
## 6	1.9e-05	0.0158
##	PublicSafetyProfessiolsinPopulation	RacialDisparityinEducatiolAttainment
## 1	0.008	0.25
## 2	0.003	NA
## 3	0.004	0.04
## 4	0.005	0.27

```
## 5          0.008          0.40
## 6          0.007          0.18
##  RacialDisparityinPoverty SegregationIndexScore
## 1          0.10          0.20
## 2          NA          0.54
## 3          0.01          0.33
## 4          0.14          0.28
## 5          0.02          0.24
## 6          0.15          0.26
##  ShareofAt.HomeFoodExpendituresonFruit.Veg SmokingRate TeenBirthRate
## 1          0.124          0.118          0.0134
## 2          0.120          0.140          0.0140
## 3          0.118          0.133          0.0078
## 4          0.145          0.107          0.0113
## 5          0.124          0.160          0.0099
## 6          0.123          0.124          0.0062
##  ToxicReleaseIndexScore UnemploymentRate UnsafeDrinkingWater VacantHouses
## 1          4847.10          0.025          0.001          0.010
## 2          0.01          0.024          0.023          0.011
## 3          4.58          0.024          0.000          0.008
## 4          115760.65          0.031          0.000          0.004
## 5          4372.47          0.032          0.000          0.032
## 6          3.40          0.027          0.001          0.003
##  VehicleCrashFatalityRate ViolentCrimeRate.100k VoterParticipationRate
## 1          0.000072          0.002308          0.655
## 2          0.000477          0.001030          0.650
## 3          0.000181          0.000786          0.600
## 4          0.000056          0.006288          0.598
## 5          0.000156          0.001379          0.549
## 6          0.000104          0.001002          0.739
##  WalkabilityIndexScore WorkHoursNeededtoPayforAffordableHousing
## 1          8.8          36
## 2          6.3          39
## 3          6.2          44
## 4          13.7          54
## 5          7.8          52
## 6          7.7          53
##  WorkersCommuting60MinutesorMore YouthWithin5Miles of a Public School
## 1          0.025          0.995
## 2          0.097          0.714
## 3          0.075          0.539
## 4          0.173          0.999
## 5          0.025          0.916
## 6          0.036          0.908
```

Next, we use KNN imputation to fill in any missing values, using the nearest 2 neighbors to approximate the values to impute with.

```
df_imputed <- knnImputation(df_trans[,3:86], k=2)
```

Here we calculate some summary statistics for all of our variables.


```
summary(df_imputed)
```

```

## AccidentalDeathRate AdultsInPoororFairGeneralHealth
## Min. :0.0001910 Min. :0.0830
## 1st Qu.:0.0003658 1st Qu.:0.1130
## Median :0.0004395 Median :0.1225
## Mean :0.0004448 Mean :0.1241
## 3rd Qu.:0.0005144 3rd Qu.:0.1330
## Max. :0.0009590 Max. :0.2690
## AdultsWithNoLeisure.TimePhysicalActivity AffordableHousingShortfall
## Min. :0.0940 Min. :-96.00
## 1st Qu.:0.1790 1st Qu.: -67.90
## Median :0.2075 Median : -54.35
## Mean :0.2100 Mean : -50.73
## 3rd Qu.:0.2400 3rd Qu.: -38.10
## Max. :0.3370 Max. : 52.80
## Air.QualityHazard AirToxicsExposureDisparityIndexScore AirborneCancerRisk
## Min. :0.1000 Min. : 0.000 Min. : 9.35
## 1st Qu.:0.2200 1st Qu.: 1.038 1st Qu.: 17.46
## Median :0.2700 Median : 2.272 Median : 21.20
## Mean :0.3173 Mean : 2.780 Mean : 25.07
## 3rd Qu.:0.4000 3rd Qu.: 3.842 3rd Qu.: 28.53
## Max. :1.5000 Max. :21.900 Max. :525.56
## AreaWithTreeCanopy AverageWeeklyWage BabiesBornWithLowBirthWeight
## Min. :0.0000 Min. : 570.0 Min. :0.02600
## 1st Qu.:0.0290 1st Qu.: 744.8 1st Qu.:0.05900
## Median :0.0945 Median : 834.0 Median :0.06500
## Mean :0.1685 Mean : 897.4 Mean :0.06646
## 3rd Qu.:0.2592 3rd Qu.: 970.5 3rd Qu.:0.07300
## Max. :0.7920 Max. :2590.0 Max. :0.12600
## BusinessGrowthRate CancerIncidenceRate CensusSelf.ResponseRate
## Min. :0.00900 Min. :0.002410 Min. :0.3540
## 1st Qu.:0.06600 1st Qu.:0.004123 1st Qu.:0.6980
## Median :0.08500 Median :0.004491 Median :0.7270
## Mean :0.08321 Mean :0.004420 Mean :0.7255
## 3rd Qu.:0.10400 3rd Qu.:0.004745 3rd Qu.:0.7610
## Max. :0.18700 Max. :0.005932 Max. :1.0000
## ChangeInHousingValue ChildCareFacilitiesRate
## Min. : -0.04200 Min. :0.000e+00
## 1st Qu.: 0.06275 1st Qu.:2.400e-05
## Median : 0.11700 Median :5.600e-05
## Mean : 0.13795 Mean :9.309e-05
## 3rd Qu.: 0.17800 3rd Qu.:1.363e-04
## Max. : 0.58900 Max. :1.493e-03
## ChildrenMeetingStandardsinGrade4ELARate
## Min. :0.2190
## 1st Qu.:0.5030
## Median :0.5795
## Mean :0.6033
## 3rd Qu.:0.7200
## Max. :0.9200
## ContinuingEducationTaxCreditsasShareofTotalTaxFilings DeathsofDespairRate
## Min. :0.0410 Min. :0.0001050
## 1st Qu.:0.0900 1st Qu.:0.0002840
## Median :0.1040 Median :0.0003505

```

```

## Mean      :0.1031                               Mean      :0.0003532
## 3rd Qu.:0.1160                               3rd Qu.:0.0004002
## Max.      :0.2330                               Max.      :0.0006800
## DiabetesPrevalence DisabilityEmploymentGap EvictionRate
## Min.      :0.01500   Min.      :0.2200   Min.      :0.00000
## 1st Qu.:0.06200   1st Qu.:0.7100   1st Qu.:0.00300
## Median :0.07300   Median :0.7700   Median :0.00800
## Mean      :0.07478   Mean      :0.7754   Mean      :0.01095
## 3rd Qu.:0.08800   3rd Qu.:0.8300   3rd Qu.:0.01600
## Max.      :0.19300   Max.      :1.2200   Max.      :0.07900
## ExtremeHeatDaysperYeardays FoodEnvironmentIndexScore GiniIndexScore
## Min.      : 2.300           Min.      : 0.000           Min.      :0.3600
## 1st Qu.: 7.000           1st Qu.: 8.935           1st Qu.:0.4100
## Median : 8.700           Median :10.895           Median :0.4300
## Mean      : 8.916           Mean      :12.719           Mean      :0.4299
## 3rd Qu.:10.300           3rd Qu.:15.220           3rd Qu.:0.4500
## Max.      :26.000           Max.      :58.360           Max.      :0.6000
## HeartDiseasePrevalenceAmongMedicareBeneficiaries HighSchoolGraduationRate
## Min.      :0.1400           Min.      :0.6540
## 1st Qu.:0.2000           1st Qu.:0.8700
## Median :0.2300           Median :0.9100
## Mean      :0.2276           Mean      :0.8998
## 3rd Qu.:0.2500           3rd Qu.:0.9350
## Max.      :0.3400           Max.      :0.9750
## HomeownershipRate HospitalBedAvailabilityRate HouseholdsInFloodHazardZone
## Min.      :0.2410   Min.      :0.000000   Min.      :0.00000
## 1st Qu.:0.6953   1st Qu.:0.001100   1st Qu.:0.02100
## Median :0.7445   Median :0.001900   Median :0.03400
## Mean      :0.7340   Mean      :0.002542   Mean      :0.04743
## 3rd Qu.:0.7850   3rd Qu.:0.003000   3rd Qu.:0.05700
## Max.      :0.8970   Max.      :0.042900   Max.      :0.72500
## HouseholdsReceivingPublicAssistanceIncome
## Min.      :0.00000
## 1st Qu.:0.01200
## Median :0.01700
## Mean      :0.01763
## 3rd Qu.:0.02200
## Max.      :0.05600
## HouseholdsSpendingatLeast30.ofIncomeonHousing
## Min.      :0.1070
## 1st Qu.:0.2000
## Median :0.2375
## Mean      :0.2493
## 3rd Qu.:0.2963
## Max.      :0.4440
## HouseholdsWithIncompletePlumbingFacilities HouseholdsWithInternetAccess
## Min.      :0.0000           Min.      :0.2200
## 1st Qu.:0.0010           1st Qu.:0.8480
## Median :0.0030           Median :0.9495
## Mean      :0.0038           Mean      :0.8970
## 3rd Qu.:0.0050           3rd Qu.:0.9842
## Max.      :0.0280           Max.      :1.0000
## HouseholdsWithNoVehicle IdleYouth.NotWorkingorEnrolled. JobDiversityIndexScore
## Min.      :0.00500           Min.      :0.00000           Min.      :0.1200

```

```

## 1st Qu.:0.03300      1st Qu.:0.00700      1st Qu.:0.5700
## Median :0.04400      Median :0.01400      Median :0.7200
## Mean :0.04721      Mean :0.01822      Mean :0.6802
## 3rd Qu.:0.05500      3rd Qu.:0.02400      3rd Qu.:0.8300
## Max. :0.77000      Max. :0.15900      Max. :0.9200
## JobsWithi45.MinuteCommute LaborForceParticipation LifeExpectancy
## Min. : 289      Min. :0.3840      Min. :77.30
## 1st Qu.: 2630      1st Qu.:0.6310      1st Qu.:79.70
## Median : 8052      Median :0.6600      Median :80.40
## Mean : 33923      Mean :0.6565      Mean :80.48
## 3rd Qu.: 34160      3rd Qu.:0.6870      3rd Qu.:81.20
## Max. :1059922      Max. :0.7970      Max. :86.80
## LocalFoodOutletsRate MedianHouseholdIncome MedicalDebtinCollections
## Min. :0.000e+00      Min. : 31468      Min. :0.0000
## 1st Qu.:3.175e-05      1st Qu.: 55878      1st Qu.:0.0700
## Median :7.500e-05      Median : 62977      Median :0.1200
## Mean :1.040e-04      Mean : 67930      Mean :0.1156
## 3rd Qu.:1.433e-04      3rd Qu.: 77788      3rd Qu.:0.1500
## Max. :5.810e-04      Max. :136268      Max. :0.3100
## MedicareBeneficiariesWithDepression
## Min. :0.1000
## 1st Qu.:0.1600
## Median :0.1700
## Mean :0.1727
## 3rd Qu.:0.1900
## Max. :0.2400
## MedicareBeneficiariesWithRecentPrimaryCareVisit NaturalAmenitiesIndexScore
## Min. :0.0200      Min. : -1.9700
## 1st Qu.:0.1800      1st Qu.: -1.3100
## Median :0.2600      Median : -0.3748
## Mean :0.2579      Mean : 0.5583
## 3rd Qu.:0.3225      3rd Qu.: 1.6275
## Max. :0.5600      Max. :11.1700
## NeighborhoodDisparityinEducationalAttainment NeighborhoodDisparityinPoverty
## Min. : 2.690      Min. : 0.590
## 1st Qu.: 7.835      1st Qu.: 2.993
## Median :11.570      Median : 4.220
## Mean :11.877      Mean : 5.060
## 3rd Qu.:15.418      3rd Qu.: 5.968
## Max. :27.940      Max. :26.340
## NetMigrationRate NonprofitsRate ObesityPrevalence OvercrowdedHouseholds
## Min. : -0.12500      Min. :0.0001570      Min. :0.1230      Min. :0.00000
## 1st Qu.: -0.02000      1st Qu.:0.0004337      1st Qu.:0.2490      1st Qu.:0.00900
## Median : 0.00650      Median :0.0005485      Median :0.2930      Median :0.01350
## Mean : 0.02802      Mean :0.0006040      Mean :0.2882      Mean :0.01733
## 3rd Qu.: 0.05125      3rd Qu.:0.0006985      3rd Qu.:0.3270      3rd Qu.:0.02000
## Max. : 0.76400      Max. :0.0020290      Max. :0.4680      Max. :0.10500
## Per.PupilExpenditures PerCapitaSpendingonHealthandEmergencyServices
## Min. : 6198      Min. : 0.0
## 1st Qu.:11986      1st Qu.: 330.0
## Median :13717      Median : 445.0
## Mean :14937      Mean : 549.4
## 3rd Qu.:17218      3rd Qu.: 628.8
## Max. :45688      Max. :8862.0

```

```

## PoorMentalHealthDaysperMonth PopulationLivingCloseToEmergencyFacilities
## Min. :2.500 Min. :0.0390
## 1st Qu.:3.100 1st Qu.:0.3040
## Median :3.300 Median :0.4160
## Mean :3.327 Mean :0.4176
## 3rd Qu.:3.600 3rd Qu.:0.5350
## Max. :4.500 Max. :1.0000
## PopulationWithAdvancedDegree PopulationWithNoHealthInsurance
## Min. :0.1390 Min. :0.02300
## 1st Qu.:0.3460 1st Qu.:0.05300
## Median :0.4105 Median :0.06900
## Mean :0.4252 Mean :0.07785
## 3rd Qu.:0.4858 3rd Qu.:0.09600
## Max. :0.8230 Max. :0.23800
## PopulationWithin0.5MileofaPark PopulationWithin0.5MileofWalkableDestinations
## Min. :0.0000 Min. :0.0000
## 1st Qu.:0.1638 1st Qu.:0.0640
## Median :0.2960 Median :0.1305
## Mean :0.3283 Mean :0.1806
## 3rd Qu.:0.4600 3rd Qu.:0.2532
## Max. :1.0000 Max. :0.9680
## PopulationWithoutAccessToLargeGroceryStore PovertyRate
## Min. :0.0000 Min. :0.02700
## 1st Qu.:0.1258 1st Qu.:0.07100
## Median :0.2140 Median :0.09000
## Mean :0.2287 Mean :0.09257
## 3rd Qu.:0.3160 3rd Qu.:0.10900
## Max. :0.7550 Max. :0.24700
## PrematureDeathDisparityIndexScore PreschoolEnrollment
## Min. :0.00000 Min. :0.1430
## 1st Qu.:0.02000 1st Qu.:0.4050
## Median :0.03000 Median :0.4880
## Mean :0.03260 Mean :0.4882
## 3rd Qu.:0.03668 3rd Qu.:0.5645
## Max. :0.17000 Max. :0.9640
## PreventableHospitalAdmissionsAmongMedicareBeneficiariesRate
## Min. :0.00860
## 1st Qu.:0.03008
## Median :0.03685
## Mean :0.03825
## 3rd Qu.:0.04442
## Max. :0.10020
## PrimaryCareDoctorAvailabilityRate PropertyCrimeRate
## Min. :0.000e+00 Min. :0.00050
## 1st Qu.:1.100e-05 1st Qu.:0.00880
## Median :1.400e-05 Median :0.01310
## Mean :1.502e-05 Mean :0.01381
## 3rd Qu.:1.900e-05 3rd Qu.:0.01730
## Max. :1.000e-04 Max. :0.05480
## PublicSafetyProfessionalsinPopulation RacialDisparityinEducationalAttainment
## Min. :0.000000 Min. :0.0100
## 1st Qu.:0.004000 1st Qu.:0.1372
## Median :0.006500 Median :0.1771
## Mean :0.007142 Mean :0.2100

```

```

## 3rd Qu.:0.009000          3rd Qu.:0.2903
## Max. :0.050000          Max. :0.6500
## RacialDisparityinPoverty SegregationIndexScore
## Min. :0.0100          Min. :0.1200
## 1st Qu.:0.0900          1st Qu.:0.2700
## Median :0.1138          Median :0.3400
## Mean :0.1265          Mean :0.3639
## 3rd Qu.:0.1600          3rd Qu.:0.4425
## Max. :0.3300          Max. :0.7500
## ShareofAt.HomeFoodExpendituresonFruit.Veg SmokingRate TeenBirthRate
## Min. :0.1120          Min. :0.0670 Min. :0.00290
## 1st Qu.:0.1180          1st Qu.:0.1340 1st Qu.:0.01030
## Median :0.1210          Median :0.1440 Median :0.01340
## Mean :0.1231          Mean :0.1434 Mean :0.01438
## 3rd Qu.:0.1290          3rd Qu.:0.1550 3rd Qu.:0.01773
## Max. :0.1500          Max. :0.2010 Max. :0.04470
## ToxicReleaseIndexScore UnemploymentRate UnsafeDrinkingWater VacantHouses
## Min. : 0          Min. :0.01600 Min. :0.00000 Min. :0.00000
## 1st Qu.: 0          1st Qu.:0.02600 1st Qu.:0.00000 1st Qu.:0.00700
## Median : 301          Median :0.03000 Median :0.00000 Median :0.01400
## Mean : 115337          Mean :0.03097 Mean :0.03708 Mean :0.02229
## 3rd Qu.: 8423          3rd Qu.:0.03500 3rd Qu.:0.01425 3rd Qu.:0.02925
## Max. :16326582          Max. :0.06900 Max. :1.00000 Max. :0.16600
## VehicleCrashFatalityRate ViolentCrimeRate.100k VoterParticipationRate
## Min. :0.000e+00          Min. :0.000000 Min. :0.4270
## 1st Qu.:8.175e-05          1st Qu.:0.000837 1st Qu.:0.6280
## Median :1.210e-04          Median :0.001264 Median :0.6715
## Mean :1.523e-04          Mean :0.001531 Mean :0.6717
## 3rd Qu.:1.960e-04          3rd Qu.:0.001911 3rd Qu.:0.7200
## Max. :6.540e-04          Max. :0.007605 Max. :0.8790
## WalkabilityIndexScore WorkHoursNeededtoPayforAffordableHousing
## Min. : 3.200          Min. : 18.00
## 1st Qu.: 6.200          1st Qu.: 37.00
## Median : 6.700          Median : 43.00
## Mean : 7.247          Mean : 45.44
## 3rd Qu.: 7.700          3rd Qu.: 51.00
## Max. :16.200          Max. :112.00
## WorkersCommuting60MinutesorMore YouthWithin5MilesOfaPublicSchool
## Min. :0.00700          Min. :0.3670
## 1st Qu.:0.03700          1st Qu.:0.7268
## Median :0.05150          Median :0.8475
## Mean :0.07044          Mean :0.8297
## 3rd Qu.:0.08425          3rd Qu.:0.9650
## Max. :0.34800          Max. :1.0000

```

There are too many different correlation pairs to attempt observing as we have 84 different variables, so we instead look at which variables have the strongest relationships in our dataset.

Below are the top 20 strongest correlation pairs in our dataset. It appears that the two variables with the strongest correlation are NeighborhoodDisparityinEducationAttainment and PopulationWithAdvancedDegree.

```
#devtools::install_github("Laresbernardo/Lares")
library(lares)

corr_cross(df_imputed, # name of dataset
  max_pvalue = 0.05, # display only significant correlations (at 5% level)
  top = 20
)
```

```
## Returning only the top 20. You may override with the 'top' argument
```

```
## Warning in theme_lares2(legend = "top"): Font Arial Narrow is not installed, has
## other name, or can't be found
```

Ranked Cross-Correlations

20 most relevant



Correlations with p-value < 0.05

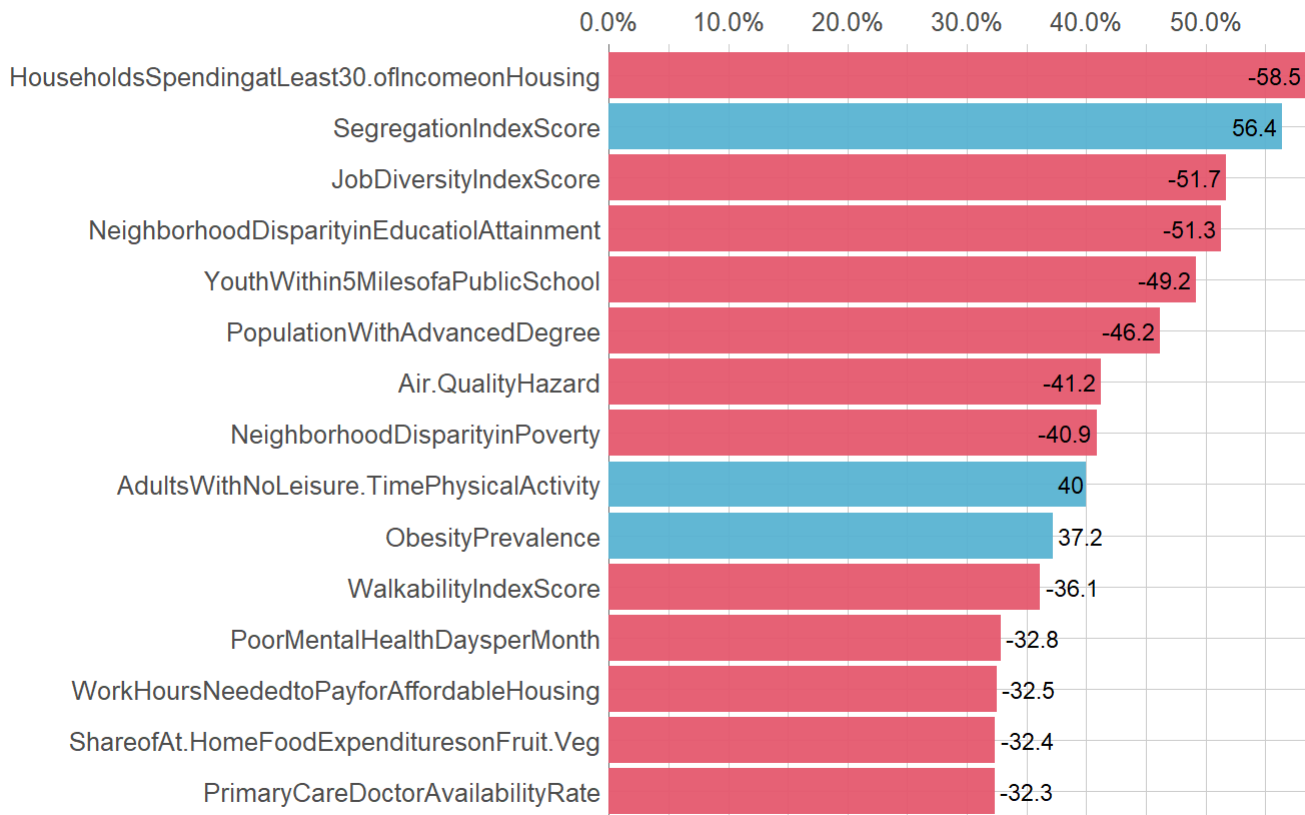
Next, we look at which variables are most highly correlated with our 'AffordableHousingShortfall' variable which is our proxy variable for predicting gentrification. The data source defines this variable as the "Availability of affordable housing relative to a community's low-income population. Negative numbers indicate a shortfall."

It appears that "Households Spending at least 30% of income on Housing" is the strongest predictor of an affordable housing shortfall, or gentrification.

```
corr_var(df_imputed, # name of dataset
  AffordableHousingShortfall, # name of variable to focus on
  top = 15 # display top 5 correlations
)
```

Correlations of AffordableHousingShortfall [%]

Top 15 out of 83 variables (original & dummy)



Next, we work on developing 7 different models for predicting gentrification using our proxy AffordableHousingShortfall variable. We will select the “best” model by evaluating them in terms of their RMSE.

We start by splitting our data into a training and test set, using 80% of our data to train our models and holding out 20% to test them.

```
#data splitting
set.seed(101)

sample = sample.split(df_imputed$AffordableHousingShortfall, SplitRatio = .8)

train = subset(df_imputed, sample == TRUE)
test = subset(df_imputed, sample == FALSE)

train_X = subset(train, select = -AffordableHousingShortfall)
train_y = train[, 'AffordableHousingShortfall']

test_X = subset(test, select = -AffordableHousingShortfall)
test_y = test[, 'AffordableHousingShortfall']
```

Linear Regression Model

Linear regression is an attractive model because representation is simply done. The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). In this instance, we will be predicting AffordableHousingShortfall values through three different

linear regression techniques.

GLM

The generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

```
control = trainControl(method = 'cv', number = 5,
  verboseIter = FALSE, savePredictions = TRUE, allowParallel = T)
```

```
set.seed(17)
GLM_train = train(AffordableHousingShortfall ~ ., data = train, metric = 'RMSE', method = 'glm',
  preProcess = c('center', 'scale'), trControl = control)
GLM_reg_pred <- predict(GLM_train, test_X)

GLM_train
```

```
## Generalized Linear Model
##
## 400 samples
## 83 predictor
##
## Pre-processing: centered (83), scaled (83)
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 320, 320, 320, 320, 320
## Resampling results:
##
##      RMSE      Rsquared   MAE
## 19.19956  0.4346039  13.68569
```

Using RMSE as a benchmark for linear regression modelling, we determined our GLM model performed quite well at 19.2. We will attempt two other iterations with different linear techniques to determine if our baseline can be improved.

glmnet

glmnet is an extremely efficient procedure for fitting the entire lasso or elastic-net regularization path for linear regression, logistic and multinomial regression models, Poisson regression and the Cox model. Two recent additions are the multiple-response Gaussian, and the grouped multinomial regression. The algorithm uses cyclical coordinate descent in a path-wise fashion to determine the best linear fit.

```
set.seed(17)
glmnet_train = train(AffordableHousingShortfall ~ ., data = train, metric = 'RMSE', method = 'glmnet',
  preProcess = c('center', 'scale'), trControl = control)
glmnet_reg_pred <- predict(glmnet_train, test_X)

glmnet_train
```

```
## glmnet
##
## 400 samples
## 83 predictor
##
## Pre-processing: centered (83), scaled (83)
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 320, 320, 320, 320, 320
## Resampling results across tuning parameters:
##
##   alpha  lambda      RMSE      Rsquared    MAE
##   0.10   0.02846147  19.12225  0.4370419  13.63454
##   0.10   0.28461473  18.73773  0.4488048  13.37479
##   0.10   2.84614733  17.73569  0.4756701  12.57925
##   0.55   0.02846147  19.04562  0.4395182  13.57765
##   0.55   0.28461473  18.30699  0.4588184  13.05158
##   0.55   2.84614733  17.60362  0.4802538  12.55867
##   1.00   0.02846147  18.96178  0.4421959  13.51636
##   1.00   0.28461473  18.03679  0.4643275  12.84684
##   1.00   2.84614733  18.11644  0.4569902  13.00166
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were alpha = 0.55 and lambda = 2.846147.
```

The final values used for the GLM Net model were alpha = 0.55 and lambda = 2.85, which produced an RMSE of 17.6. This is a slight improvement over the GLM model. The close results can be explained by low lambda value, where a zero lambda is in effect a standard glm model.

Partial Least Squares

Partial least squares (PLS) is a method for constructing predictive models when the factors are many and highly collinear. Partial least squares is a popular method for soft modelling in industrial applications. We believed that it would be an effective model to demonstrate.

```
set.seed(17)
pls_train = train(AffordableHousingShortfall ~ ., data = train , metric = 'RMSE', method = 'pls'
,preProcess = c('center', 'scale'), trControl = control)
pls_reg_pred <- predict(pls_train, test_X)

pls_train
```

```
## Partial Least Squares
##
## 400 samples
## 83 predictor
##
## Pre-processing: centered (83), scaled (83)
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 320, 320, 320, 320, 320
## Resampling results across tuning parameters:
##
##   ncomp  RMSE      Rsquared  MAE
##   1      18.99295  0.3939885  13.54142
##   2      18.11396  0.4505461  12.81498
##   3      18.22649  0.4573401  12.98211
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was ncomp = 2.
```

The final value used for the model was $ncomp = 2$, and it produced an RMSE of 18.11. The GLM Net model outperformed this with a RMSE of 17.6.

Non-linear Regression Models

In this section, we are going to fit a simple neural network using the `neuralnet` package.

Here we confirm that there are no more empty data:

```
#describe(bev)
apply(df_imputed,2,function(x) sum(is.na(x)))
```

```

##                               AccidentalDeathRate
##                               0
##                               AdultsInPoororFairGeneralHealth
##                               0
##                               AdultsWithNoLeisure.TimePhysicalActivity
##                               0
##                               AffordableHousingShortfall
##                               0
##                               Air.QualityHazard
##                               0
##                               AirToxicsExposureDisparityIndexScore
##                               0
##                               AirborneCancerRisk
##                               0
##                               AreaWithTreeCanopy
##                               0
##                               AverageWeeklyWage
##                               0
##                               BabiesBornWithLowBirthWeight
##                               0
##                               BusinessGrowthRate
##                               0
##                               CancerIncidenceRate
##                               0
##                               CensusSelf.ResponseRate
##                               0
##                               ChangeinHousingValue
##                               0
##                               ChildCareFacilitiesRate
##                               0
##                               ChildrenMeetingStandardsinGrade4ELARate
##                               0
##                               ContinuingEducationTaxCreditsasShareofTotalTaxFilings
##                               0
##                               DeathsofDespairRate
##                               0
##                               DiabetesPrevalence
##                               0
##                               DisabilityEmploymentGap
##                               0
##                               EvictionRate
##                               0
##                               ExtremeHeatDaysperYeardays
##                               0
##                               FoodEnvironmentIndexScore
##                               0
##                               GiniIndexScore
##                               0
##                               HeartDiseasePrevalenceAmongMedicareBeneficiaries
##                               0
##                               HighSchoolGraduationRate
##                               0
##                               HomeownershipRate

```

##	0
##	HospitalBedAvailabilityRate
##	0
##	HouseholdsinFloodHazardZone
##	0
##	HouseholdsReceivingPublicAssistanceIncome
##	0
##	HouseholdsSpendingatLeast30.ofIncomeonHousing
##	0
##	HouseholdsWithIncompletePlumbingFacilities
##	0
##	HouseholdsWithInternetAccess
##	0
##	HouseholdsWithNoVehicle
##	0
##	IdleYouth.NotWorkingorEnrolled.
##	0
##	JobDiversityIndexScore
##	0
##	JobsWithi45.MinuteCommute
##	0
##	LaborForceParticipation
##	0
##	LifeExpectancy
##	0
##	LocalFoodOutletsRate
##	0
##	MedianHouseholdIncome
##	0
##	MedicalDebtinCollections
##	0
##	MedicareBeneficiariesWithDepression
##	0
##	MedicareBeneficiariesWithRecentPrimaryCareVisit
##	0
##	NaturalAmenitiesIndexScore
##	0
##	NeighborhoodDisparityinEducationalAttainment
##	0
##	NeighborhoodDisparityinPoverty
##	0
##	NetMigrationRate
##	0
##	NonprofitsRate
##	0
##	ObesityPrevalence
##	0
##	OvercrowdedHouseholds
##	0
##	Per.PupilExpenditures
##	0
##	PerCapitaSpendingonHealthandEmergencyServices
##	0
##	PoorMentalHealthDaysperMonth

```

## 0
## PopulationLivingCloseToEmergencyFacilities
## 0
## PopulationWithAdvancedDegree
## 0
## PopulationWithNoHealthInsurance
## 0
## PopulationWithin0.5MileOfaPark
## 0
## PopulationWithin0.5MileOfWalkableDestinations
## 0
## PopulationWithoutAccessToLargeGroceryStore
## 0
## PovertyRate
## 0
## PrematureDeathDisparityIndexScore
## 0
## PreschoolEnrollment
## 0
## PreventableHospitalAdmissionsAmongMedicareBeneficiariesRate
## 0
## PrimaryCareDoctorAvailabilityRate
## 0
## PropertyCrimeRate
## 0
## PublicSafetyProfessionalsInPopulation
## 0
## RacialDisparityInEducationAttainment
## 0
## RacialDisparityInPoverty
## 0
## SegregationIndexScore
## 0
## ShareOfAtHomeFoodExpendituresOnFruit.Veg
## 0
## SmokingRate
## 0
## TeenBirthRate
## 0
## ToxicReleaseIndexScore
## 0
## UnemploymentRate
## 0
## UnsafeDrinkingWater
## 0
## VacantHouses
## 0
## VehicleCrashFatalityRate
## 0
## ViolentCrimeRate.100k
## 0
## VoterParticipationRate
## 0
## WalkabilityIndexScore

```

```
##                                0
##      WorkHoursNeededtoPayforAffordableHousing
##                                0
##      WorkersCommuting60MinutesorMore
##                                0
##      YouthWithin5MilesOfaPublicSchool
##                                0
```

Preparing to fit the neural network

Before fitting a neural network, some preparation needs to be done.

As a first step, we are going to address data preprocessing. We will be normalizing the data before training a neural network. We chose to use the min-max method and scale the data in the interval [0,1]. Usually scaling in the intervals [0,1] or [-1,1] tends to give better results.

We therefore scale and split the data before moving on:

```
maxs <- apply(df_imputed, 2, max)
mins <- apply(df_imputed, 2, min)
```

Scaled returns a matrix that needs to be coerced into a data.frame.

```
scaled <- as.data.frame(scale(df_imputed, center = mins, scale = maxs - mins))
#scaled
index <- sample(1:nrow(df_imputed), round(0.75*nrow(df_imputed)))
train_ <- scaled[index,]
test_ <- scaled[-index,]
```

Parameters

In this dataset, we are going to use 2 hidden layers with this configuration: 83:5:3:1. The input layer has 83 inputs, the two hidden layers have 5 and 3 neurons and the output layer has, of course, a single output since we are doing regression.

Let's fit the net: Setting the linear.output = True does regression instead of classification.

```
n <- names(train_)
f <- as.formula(paste("AffordableHousingShortfall ~", paste(n[!n %in% "AffordableHousingShortfall"], collapse = " + ")))
nn <- neuralnet(f, data=train_, hidden=c(5,3), linear.output=T)
```

Plot the Neural Network

```
plot(nn)
```

The black lines show the connections between each layer and the weights on each connection while the blue lines show the bias term added in each step. The bias can be thought as the intercept of a linear model. The net is essentially a black box so we cannot say that much about the fitting, the weights and the model. Suffice to say that the training algorithm has converged and therefore the model is ready to be used.

Predicting AffordableHousingShortfall using the neural network

Now we can try to predict the values for the test set and calculate the RMSE. The net will output a normalized prediction, so we need to scale it back in order to make a meaningful comparison (or just a simple prediction).

```
pr.nn <- compute(nn,test_[,1:84])
pr.nn_ <- pr.nn$net.result*(max(df_imputed$AffordableHousingShortfall)-min(df_imputed$Affordable
HousingShortfall))+min(df_imputed$AffordableHousingShortfall)
test.r <- (test_$AffordableHousingShortfall)*(max(df_imputed$AffordableHousingShortfall)-min(df_
imputed$AffordableHousingShortfall))+min(df_imputed$AffordableHousingShortfall)
MSE.nn <- sum((test.r - pr.nn_)^2)/nrow(test_)
```

The RMSE for our Neural Network is 31.55 which does not perform as well as our GLM net model.

```
nn.rmse <- sqrt(MSE.nn)
print(nn.rmse)
```

```
## [1] 31.54765
```

Tree Models

In this next section, we consider various tree models to predict the AffordableHousingShortfall, or gentrification, of a U.S. county given the 83 other indicators.

Basic Regression Tree

Classification and regression trees can be generated through rpart to create simple tree models. Tree-based models consist of one or more nested if-then statements for the predictors that partition the data. A model is used to predict the outcome within these partitions.

```
treeb <- train(x = train_X, y = train_y, method = "rpart", preProcess = c('center', 'scale'))
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo, :
## There were missing values in resampled performance measures.
```

```
treeb
```



```
## CART
##
## 400 samples
## 83 predictor
##
## Pre-processing: centered (83), scaled (83)
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 400, 400, 400, 400, 400, 400, ...
## Resampling results across tuning parameters:
##
##   cp          RMSE      Rsquared    MAE
## 0.06071757 21.18997 0.2847308 15.21984
## 0.07429463 21.34260 0.2719466 15.34983
## 0.32829427 21.59173 0.2794404 15.95130
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.06071757.
```

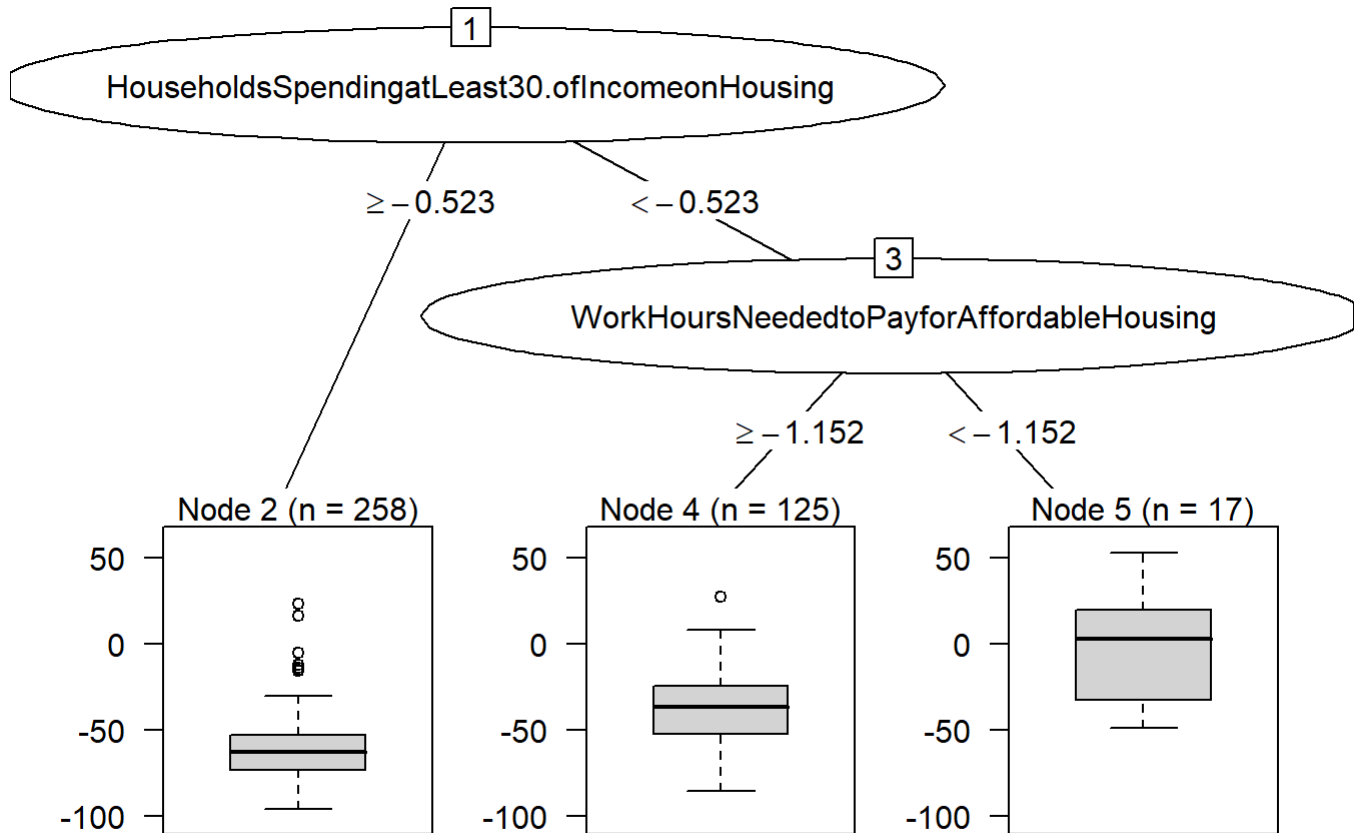
```
treebPred <- predict(treeb, newdata = test_X)
treeb.results <- postResample(pred = treebPred, obs = test_y)
treeb.results
```

```
##          RMSE    Rsquared      MAE
## 24.0467330 0.1727561 17.6478419
```

The RMSE for this basic regression tree model is 24.05 with an RM^2 of 0.17.

We also plot this specific tree below:

```
plot(as.party(treeb$finalModel))
```



Next, we will try out a Random Forest model to see if we can improve upon this model.

Random Forest

Random Forest is an ensemble model where each tree splits out a class prediction and the class with the most contributions becomes the model's prediction value. Random Forest creates as many trees on the subset of the data and combines the output of all the trees. This thus reduces problems in overfitting and reduces the variance.

```
rf <- train(x = train_X, y = train_y, method = "rf", preProcess = c('center', 'scale'))
rf
```

```
## Random Forest
##
## 400 samples
## 83 predictor
##
## Pre-processing: centered (83), scaled (83)
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 400, 400, 400, 400, 400, 400, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared  MAE
##    2    18.49052  0.4460399 12.84827
##   42    18.01823  0.4521409 12.43415
##   83    18.27414  0.4352865 12.62710
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 42.
```

```
rfPred <- predict(rf, newdata = test_X)
rf.results <- postResample(pred = rfPred, obs = test_y)
rf.results
```

```
##      RMSE  Rsquared      MAE
## 18.8211486  0.4164222 13.3131592
```

The RMSE for this model is 18.82 with a RM^2 of 0.416. This Random Forest model performs better than our Basic Regression Tree model. We will also consider an XGBoost model to see if we can find any improvements.

XGBoost

XGBoost is another ensemble model, this time using the gradient boosting framework which is a special case of boosting where errors are minimized by gradient descent algorithm. XGBoost only manages numeric vectors, and luckily all of our variables are numeric.

```
xgb <- train(x = train_X, y = train_y, method = "xgbTree")
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
precated in favor of reg:squarederror.  
## [18:01:33] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:33] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:34] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:34] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:34] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:35] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:35] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:36] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:36] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:36] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:36] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:37] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:37] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:37] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:37] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:38] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:38] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:38] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:39] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:39] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:39] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:40] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:01:40] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de
```

[illegible]

[illegible]

[illegible]

[illegible]

```
precated in favor of reg:squarederror.  
## [18:02:09] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:09] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:09] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:09] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:09] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:10] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:10] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:10] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:10] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:11] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:11] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:11] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:11] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:11] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:12] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:12] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:12] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:13] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:13] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:13] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:13] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:13] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:02:14] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
precated in favor of reg:squarederror.  
## [18:04:45] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:04:46] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:04:46] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:04:46] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:04:46] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:04:47] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:04:47] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:04:47] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:04:47] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:04:48] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.  
## [18:04:48] WARNING: amalgamation/../src/objective/regression_obj.cu:170: reg:linear is now de  
precated in favor of reg:squarederror.
```

xgb

```

## eXtreme Gradient Boosting
##
## 400 samples
## 83 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 400, 400, 400, 400, 400, 400, ...
## Resampling results across tuning parameters:
##
##   eta  max_depth  colsample_bytree  subsample  nrounds  RMSE      Rsquared
##   0.3   1         0.6           0.50       50      18.94796  0.4052137
##   0.3   1         0.6           0.50      100      19.38501  0.3940234
##   0.3   1         0.6           0.50      150      19.59281  0.3900588
##   0.3   1         0.6           0.75       50      18.65452  0.4201367
##   0.3   1         0.6           0.75      100      18.95615  0.4123897
##   0.3   1         0.6           0.75      150      19.20493  0.4045784
##   0.3   1         0.6           1.00       50      18.55914  0.4222477
##   0.3   1         0.6           1.00      100      18.79256  0.4169608
##   0.3   1         0.6           1.00      150      18.92401  0.4141411
##   0.3   1         0.8           0.50       50      18.88343  0.4098330
##   0.3   1         0.8           0.50      100      19.21140  0.4013811
##   0.3   1         0.8           0.50      150      19.54423  0.3912959
##   0.3   1         0.8           0.75       50      18.63041  0.4223559
##   0.3   1         0.8           0.75      100      19.02423  0.4099619
##   0.3   1         0.8           0.75      150      19.24073  0.4034291
##   0.3   1         0.8           1.00       50      18.62808  0.4194252
##   0.3   1         0.8           1.00      100      18.82563  0.4159255
##   0.3   1         0.8           1.00      150      18.99230  0.4125945
##   0.3   2         0.6           0.50       50      19.67618  0.3782204
##   0.3   2         0.6           0.50      100      20.01380  0.3691968
##   0.3   2         0.6           0.50      150      20.07028  0.3676065
##   0.3   2         0.6           0.75       50      19.03961  0.4025731
##   0.3   2         0.6           0.75      100      19.14765  0.4014763
##   0.3   2         0.6           0.75      150      19.19331  0.4001603
##   0.3   2         0.6           1.00       50      18.75274  0.4156204
##   0.3   2         0.6           1.00      100      18.89743  0.4124225
##   0.3   2         0.6           1.00      150      18.96594  0.4100977
##   0.3   2         0.8           0.50       50      19.49059  0.3857508
##   0.3   2         0.8           0.50      100      19.74933  0.3785317
##   0.3   2         0.8           0.50      150      19.81515  0.3772668
##   0.3   2         0.8           0.75       50      18.87871  0.4152274
##   0.3   2         0.8           0.75      100      19.00540  0.4130644
##   0.3   2         0.8           0.75      150      19.03060  0.4136085
##   0.3   2         0.8           1.00       50      18.71630  0.4170761
##   0.3   2         0.8           1.00      100      18.80830  0.4152382
##   0.3   2         0.8           1.00      150      18.84215  0.4148400
##   0.3   3         0.6           0.50       50      19.63425  0.3801806
##   0.3   3         0.6           0.50      100      19.64943  0.3808997
##   0.3   3         0.6           0.50      150      19.65263  0.3810312
##   0.3   3         0.6           0.75       50      19.02905  0.4003739
##   0.3   3         0.6           0.75      100      19.05823  0.4001343
##   0.3   3         0.6           0.75      150      19.05796  0.4004250

```

##	0.3	3	0.6	1.00	50	19.15106	0.3922245
##	0.3	3	0.6	1.00	100	19.17711	0.3917522
##	0.3	3	0.6	1.00	150	19.18128	0.3918035
##	0.3	3	0.8	0.50	50	19.48625	0.3895147
##	0.3	3	0.8	0.50	100	19.56379	0.3885584
##	0.3	3	0.8	0.50	150	19.56675	0.3886550
##	0.3	3	0.8	0.75	50	19.05858	0.4003116
##	0.3	3	0.8	0.75	100	19.10718	0.3994839
##	0.3	3	0.8	0.75	150	19.10701	0.3997405
##	0.3	3	0.8	1.00	50	18.85253	0.4137200
##	0.3	3	0.8	1.00	100	18.87391	0.4139873
##	0.3	3	0.8	1.00	150	18.87565	0.4141066
##	0.4	1	0.6	0.50	50	19.60291	0.3817954
##	0.4	1	0.6	0.50	100	19.99331	0.3771060
##	0.4	1	0.6	0.50	150	20.33050	0.3667343
##	0.4	1	0.6	0.75	50	19.30503	0.3939481
##	0.4	1	0.6	0.75	100	19.77590	0.3812976
##	0.4	1	0.6	0.75	150	20.01827	0.3762378
##	0.4	1	0.6	1.00	50	18.84621	0.4134630
##	0.4	1	0.6	1.00	100	19.07955	0.4104031
##	0.4	1	0.6	1.00	150	19.26107	0.4066709
##	0.4	1	0.8	0.50	50	19.67780	0.3852858
##	0.4	1	0.8	0.50	100	20.17999	0.3699231
##	0.4	1	0.8	0.50	150	20.52211	0.3600507
##	0.4	1	0.8	0.75	50	19.30294	0.3964569
##	0.4	1	0.8	0.75	100	19.54023	0.3942751
##	0.4	1	0.8	0.75	150	19.79454	0.3867804
##	0.4	1	0.8	1.00	50	18.85728	0.4135323
##	0.4	1	0.8	1.00	100	19.16763	0.4061728
##	0.4	1	0.8	1.00	150	19.33731	0.4034150
##	0.4	2	0.6	0.50	50	20.18074	0.3688294
##	0.4	2	0.6	0.50	100	20.38277	0.3637412
##	0.4	2	0.6	0.50	150	20.41612	0.3641216
##	0.4	2	0.6	0.75	50	19.48962	0.3879546
##	0.4	2	0.6	0.75	100	19.66408	0.3851415
##	0.4	2	0.6	0.75	150	19.68373	0.3859104
##	0.4	2	0.6	1.00	50	19.24173	0.3941779
##	0.4	2	0.6	1.00	100	19.39895	0.3894086
##	0.4	2	0.6	1.00	150	19.44480	0.3884964
##	0.4	2	0.8	0.50	50	20.10749	0.3637379
##	0.4	2	0.8	0.50	100	20.33995	0.3580607
##	0.4	2	0.8	0.50	150	20.37675	0.3577068
##	0.4	2	0.8	0.75	50	19.65310	0.3791726
##	0.4	2	0.8	0.75	100	19.79232	0.3756311
##	0.4	2	0.8	0.75	150	19.80273	0.3757228
##	0.4	2	0.8	1.00	50	18.91779	0.4132195
##	0.4	2	0.8	1.00	100	18.97364	0.4132271
##	0.4	2	0.8	1.00	150	19.03782	0.4108164
##	0.4	3	0.6	0.50	50	20.38851	0.3538991
##	0.4	3	0.6	0.50	100	20.40310	0.3550747
##	0.4	3	0.6	0.50	150	20.41938	0.3544892
##	0.4	3	0.6	0.75	50	19.84625	0.3633218
##	0.4	3	0.6	0.75	100	19.90483	0.3607623
##	0.4	3	0.6	0.75	150	19.90600	0.3607846

##	0.4	3	0.6	1.00	50	19.48021	0.3805840
##	0.4	3	0.6	1.00	100	19.49510	0.3809089
##	0.4	3	0.6	1.00	150	19.49843	0.3807777
##	0.4	3	0.8	0.50	50	20.35298	0.3526640
##	0.4	3	0.8	0.50	100	20.38552	0.3521196
##	0.4	3	0.8	0.50	150	20.38680	0.3521928
##	0.4	3	0.8	0.75	50	20.05363	0.3611767
##	0.4	3	0.8	0.75	100	20.07321	0.3613189
##	0.4	3	0.8	0.75	150	20.07181	0.3614522
##	0.4	3	0.8	1.00	50	19.28585	0.3976157
##	0.4	3	0.8	1.00	100	19.28289	0.3986994
##	0.4	3	0.8	1.00	150	19.28956	0.3984693
##	MAE						
##	13.63213						
##	13.90478						
##	14.11348						
##	13.35211						
##	13.53568						
##	13.69648						
##	13.15152						
##	13.26648						
##	13.37049						
##	13.54951						
##	13.79353						
##	14.05111						
##	13.33254						
##	13.53930						
##	13.70807						
##	13.20223						
##	13.33391						
##	13.43930						
##	14.08482						
##	14.29778						
##	14.37339						
##	13.59466						
##	13.69686						
##	13.74007						
##	13.24187						
##	13.33543						
##	13.39294						
##	13.82234						
##	14.04323						
##	14.09990						
##	13.40003						
##	13.51199						
##	13.54522						
##	13.28946						
##	13.34799						
##	13.39019						
##	14.19837						
##	14.23287						
##	14.23926						
##	13.54939						
##	13.58694						

##	13.58438
##	13.48076
##	13.52556
##	13.53017
##	13.96998
##	14.08447
##	14.08265
##	13.48906
##	13.53062
##	13.53465
##	13.30017
##	13.29186
##	13.29178
##	14.20632
##	14.51618
##	14.72792
##	13.78492
##	14.06288
##	14.26605
##	13.41783
##	13.59987
##	13.75246
##	14.10350
##	14.48060
##	14.76361
##	13.76987
##	13.91551
##	14.15620
##	13.41424
##	13.65299
##	13.83968
##	14.56383
##	14.72565
##	14.75558
##	13.99396
##	14.12118
##	14.15361
##	13.76810
##	13.92180
##	13.96972
##	14.50133
##	14.66694
##	14.68557
##	14.10614
##	14.18569
##	14.19261
##	13.58022
##	13.63230
##	13.67238
##	14.69632
##	14.71332
##	14.72739
##	14.20060
##	14.25326

```
## 14.25481
## 13.82949
## 13.83931
## 13.84346
## 14.81591
## 14.85226
## 14.84785
## 14.29798
## 14.30197
## 14.30078
## 13.61699
## 13.63309
## 13.63580
##
## Tuning parameter 'gamma' was held constant at a value of 0
## Tuning
## parameter 'min_child_weight' was held constant at a value of 1
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were nrounds = 50, max_depth = 1, eta
## = 0.3, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample
## = 1.
```

```
xgbPred <- predict(xgb, newdata = test_X)
xgb.results <- postResample(pred = xgbPred, obs = test_y)
xgb.results
```

```
##      RMSE    Rsquared      MAE
## 20.0079475  0.3574677 14.1354080
```

The RMSE for this model is 20.01 with a RM^2 of 0.357.

We considered Basic Regression, Random Forest and XGBoost tree models, and Random Forest performed the best out of the three in predicting AffordableHousingShortfall as it had the smallest RMSE value at 18.82 with a RM^2 of 0.416.

Conclusion

Finally, we will choose our best model for predicting AffordableHousingShortfall among the chosen Linear (GLM Net), Non-Linear (Neural Network) and Tree Models (Random Forest) that we have gone over in this analysis. Their RMSE metrics are summarized here:

```
lin_model_perf <- getTrainPerf(glmnet_train)
print(lin_model_perf)
```

```
## TrainRMSE TrainRsquared TrainMAE method
## 1 17.60362 0.4802538 12.55867 glmnet
```

```
print(nn.rmse)
```

```
## [1] 31.54765
```

```
rf_perf <- as.data.frame(as.list(rf.results))
print(rf_perf)
```

```
##          RMSE  Rsquared      MAE
## 1 18.82115 0.4164222 13.31316
```

The RMSE for the chosen Linear Model (GLM Net) was 17.604. The RMSE for the chosen Non-Linear Model (Neural Net) was 31.54. And lastly, the RMSE for the chosen Tree model (Random Forest) was 18.82. So, we chose the GLM Net Model as our final model to predict the AffordableHousingShortfall of our counties given the predictors on hand.

In this final chosen GLM Net model, the top 5 predictors that we found to influence AffordableHousingShortfall, or gentrification, for counties are HouseholdsSpendingatLeast30.ofIncomeonHousing, SegregationIndexScore, JobDiversityIndexScore, PopulationWithAdvancedDegree, and UnemploymentRate.

```
varImp(glmnet_train)
```

```
## glmnet variable importance
##
##   only 20 most important variables shown (out of 83)
##
##                                     Overall
## HouseholdsSpendingatLeast30.ofIncomeonHousing 100.0000
## SegregationIndexScore                        70.5762
## JobDiversityIndexScore                        49.8749
## PopulationWithAdvancedDegree                  39.3378
## UnemploymentRate                             27.1655
## PovertyRate                                  25.2956
## CancerIncidenceRate                          16.5135
## Air.QualityHazard                            16.2184
## YouthWithin5MilesOfaPublicSchool              14.9089
## NeighborhoodDisparityinPoverty                14.3494
## MedicareBeneficiariesWithRecentPrimaryCareVisit 14.2138
## HouseholdsWithNoVehicle                       11.2255
## PopulationWithoutAccessToLargeGroceryStore      9.3253
## MedicalDebtinCollections                       8.5635
## NonprofitsRate                                8.3418
## PopulationWithNoHealthInsurance                8.1551
## PoorMentalHealthDaysperMonth                   5.4384
## PrimaryCareDoctorAvailabilityRate               4.0731
## EvictionRate                                   1.3699
## ChildCareFacilitiesRate                       0.4047
```

We should pay close attention to these predictors when trying to learn more about and control for the gentrification of counties in the U.S.