Analyzing the Availability of High Speed Internet in New York

Janet Wei
Chirag Kashyap
Jingyang Chen
Yuji Mori

Introduction

New York State has just completed a broadband mapping program as part of the national broadband mapping program funded by the National Telecommunication and Information Administration in the US Department of Commerce. Information about the availability of high-speed Internet services, commonly called Broadband. The data was updated every six months for five years, and is shown on the NYS Broadband Map. In this project, we will analyze a dataset regarding high-speed Internet availability in the state of New York. Information about high-speed Internet was collected from different telecommunication companies based on the different types of technology utilized (fiber, satellite, cable, dsl, wireless).

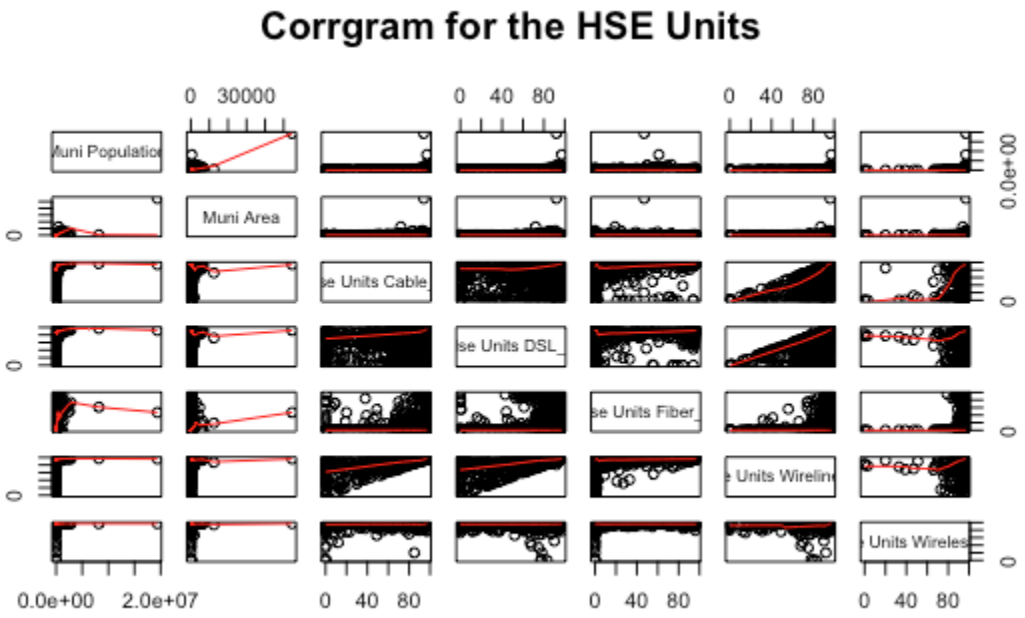
We want to identify how demographical metrics and size of a municipality impact internet availability, graphically representing New York on a map with respect to different variables will solve this problem; and also to understand how the categorization of Regional Economic Development Council (REDC) and municipalities affects the number of wireline providers and other indicators of internet availability. We may use Linear and logistic regression utilizing stepwise procedures and other selection criterion to differentiate between models which depending on the data. In addition, we still need to figure out how companies choose where to introduce fiber technology based on the metrics and data provided, which we can plot different variable to compare different regions' internet available with the possibility of comparing that with national average.

Data and Material

The data comes from OpenDataNetwork and was last updated on 22 Feb 2016. It has information regarding the availability of high speed Internet for 1635 municipalities in New York. For each of the municipalities, some interesting characteristics include municipality area, number of cable providers, amount houses with access to cable, percentage of houses with access to cable, etc. In order to analyze this data we primarily used linear and multinomial logistic regression with stepwise procedures and confusion matrices.

Results and Conclusion

Hse Units Relationship



First, plotting a corrgram could give us the relations between each Hse units against Muni population and area at a glance. Looking at the upper panel, there is a trend of positive correlation across most of the variables. For instance, the larger Muni area is, the higher population will live in the area. Next, we see that Hse Units Wireline has a strong correlation with Hse Units Cable and DSL. This is essentially correct since Cable and DSL are included and parts of Wireline. Since then, we decide to use Hse Units Wireline as the response variable for our performance in linear models; by doing so, we can extract the most information between the variables.

Does the REDC or type of municipality have any affect on a it's broadband availability?

To determine whether REDC or type of municipality have any effect, we do a linear regression estimator model. The dependent variable is Hse Units Wirline_1, the variable is the percentage of 2010 Census Housing Units with access to wireline (DSL, cable, or fiber) broadband in the Municipality with an interval of 3% bias factor. The predictor variables are REDC region and municipality type (Town, City, Village, etc). The hypothesis test is as following:

$\square_0$:Type of municipality does not have effect on its broadband availability, beta= 0

$\square_\square$:Type of municipality has effect on its broadband availability, beta<0 or beta>0

At level of significance = 0.05
We will test two linear regression model: (1) examining the linear regression of “type of municipality” on broadband availability; (2) examining the linear regression of “REDC” on broadband availability.

(1) examining the linear regression of “type of municipality” on broadband availability

```
Call:
lm(formula = y ~ x1, data = Projectdata)
Residuals:
      Min       1Q   Median       3Q      Max
-85.475  -0.475   1.800   7.525  19.222
Coefficients:
                                              Estimate      Std. Error  t
value    Pr(>|t|)
(Intercept)                                          97.0000   6.2768
      15.454   <2e-16 ***
x1City                                         -0.5246   6.5290        -
0.080   0.9360
x1County                                  -4.5484      6.5250      -0.697
0.4859
x1Economic Development Region -2.8000    7.6875       -0.364   0.7157
x1Statewide                                -1.0000  15.3751        -0.065
0.9481
x1Town                                      -11.5247       6.2937        -
1.831   0.0673 .
x1Tribal Community                         -19.2222  7.8286        -2.455
0.0142 *
x1Village                                  -2.0342      6.3050     -0.323
0.7470
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 14.04 on 1627 degrees of freedom
Multiple R-squared:  0.1034,        Adjusted R-squared:  0.09953
F-statistic:  26.8 on 7 and 1627 DF,  p-value: < 2.2e-16
```

From the information above, we can interpret the following:
We look at the p-value of the F-test to see if the model is significant. Since the p-value is extremely small, we can reject the null hypothesis, stating that the type of municipality has effect on its broadband availability.
The R-squared is 0.09953, meaning that approximately 9.953% of the variability of “broadband availability” is accounted for by the variables in the model.
Consider the variable “type of municipality,” we should expect a less broadband availability in tribal community and town.

(2) examining the linear regression of “REDC” on broadband availability

```
Call:
lm(formula = y ~ x2, data = Projectdata)
Residuals:
      Min       1Q   Median       3Q      Max
-87.198  -0.328   2.727   5.923  15.237
Coefficients:
                              Estimate    Std. Error  t value
      Pr(>|t|)
(Intercept)                   90.0769    1.0924        82.461      < 2e-16 ***
x2Central NY Region     1.1257     1.5847       0.710             0.477591
x2Finger Lakes Region   2.1260     1.4804       1.436             0.151157
x2Long Island Region    5.2507     1.7122       3.067             0.002201 **
x2Mid-Hudson Region     4.4425     1.4738       3.014             0.002616 **
x2Mohawk Valley Region  -2.8792      1.5273     -1.885      0.059585 .
x2New York City                      6.9231     4.4188       1.567
      0.117374
x2North Country Region    -8.3143    1.4872     -5.591      2.65e-08 ***
x2Southern Tier Region     0.1958    1.4872     0.132             0.895269
x2Western NY Region  -5.6750    1.4771  -       3.842             0.000127
***
x2n/a                              5.9231    14.2427        0.416
      0.677561
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 14.2 on 1624 degrees of freedom
Multiple R-squared:  0.08384,       Adjusted R-squared:  0.0782
F-statistic: 14.86 on 10 and 1624 DF,  p-value: < 2.2e-16
```

From the information above, we can interpret the following:

We look at the p-value of the F-test to see if the model is significant. Since the p-value is extremely small, we can reject the null hypothesis, stating that the REDC has effect on its broadband availability.

The R-squared is 0.0782, meaning that approximately 7.782% of the variability of “broadband availability” is accounted for by the variables in the model.

Consider the variable “REDC”, we would expect a high broadband availability in long island region, mid-Hudson region and less broadband availability in north country region, western NY regions.

What factors influence broadband availability the most?

Compare the correlation between various variables such as type of municipality, Muni population, Muni Housing Units, Muni Area, and REDC, each with the % of Hse Units Wireline, find the variable with the strongest influence on broadband availability.

Adjusted R-squared:

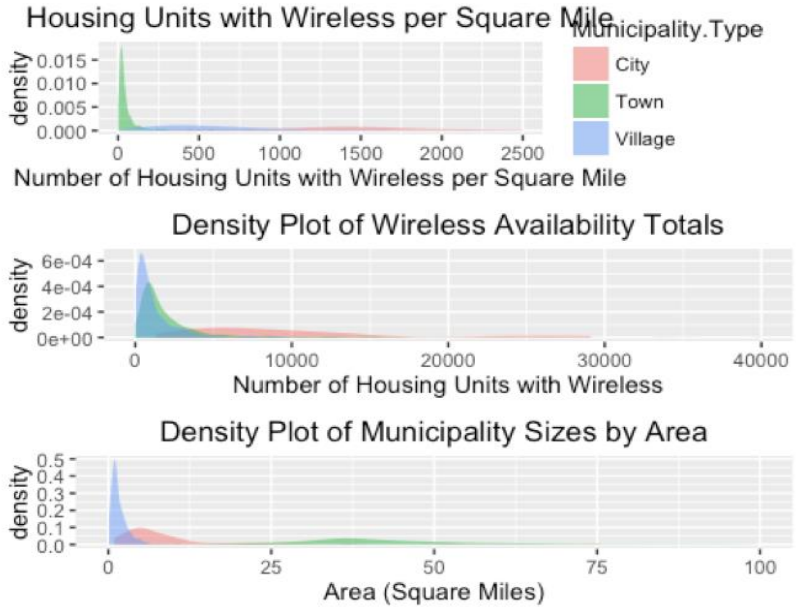
type of municipality:	0.09953
Muni population:	0.0008972
Muni Housing Units:	0.000844
Muni Area:	-0.0004998
REDC:	0.0782

Base on the results, type of municipality has the highest Adjusted R-squared, which is closed to 0.1. We can conclude one important factor that affects the broadband availability is the type of municipality.

Constructing a Classification Model

The objective of this part is to use Wireless Availability to predict the type of any municipality. Among all the 8 distinct municipality types found in the data, 3 categories were used for analysis: City, Town, and Village. These levels were chosen because they should be independent of each other (as opposed to using a category like County, which is a grouping of many smaller municipality types). These categories also accounted for the majority of the observations.

To select an appropriate predictor variable, each continuous variable was considered and analyzed by their distribution. However, it became clear that the variables needed to be transformed to acquire distinct distributions for each of the 3 levels. The figure below illustrates this problem, and ultimately, a new variable was considered for the classification model: Transformed Variable = # of Housing Units with Wireless / Area in Square Miles.



It can be seen that with the transformation, the distribution for each municipality types spreads out, so it may be used in the following multiple regression model. The results are found below:

Confusion Matrix:

muni_predictions	City	Town	Village
City	9	0	0
Town	6	20	1
Village	5	0	19

Misclassification rate:
[1] 0.2

The purpose of the multinomial model is to classify New York municipalities by their Wireless Availability per Square Mile. 20 observations were selected from each of the 3 municipality types of interest (City, Town, Village) to comprise the test set, while the remaining 1488 observations formed the training set. Once the model was established, probabilities were calculated for each of the test points and sorted. The confusion matrix shows that all 20 Town and 19/20 Village observations in the test set were correctly categorized, so the model may be suitable for these categories. However, the matrix also shows that the City observations were poorly classified, being responsible for 11/12 misclassifications. Overall, the 20% misclassification indicates that the model is not necessarily reliable, but if the City class is omitted (binomial model), the rate reduces to a mere 2.5%.

Constructing a Classification Model Using Broadband Availability Data

Using the methods from above, we tried to predict type of municipality and REDC Region by simply using the broadband availability data (HSE Units, Number of Providers, and Percentage of HSE Units). By using a forward stepwise procedure and multinomial logistic regression with multiple explanatory variables, we found that Municipality Type can be predicted with decent certainty. Like before, we focused on the municipality types of interest (City, Town, Village). Like above, we utilized test data, but instead of having 20 of each type of municipality, we simply used the first 200 observations out of 1538 total, in order to have a better representation of the data (since were were only . We found that the model below accurately predicted municipality type.

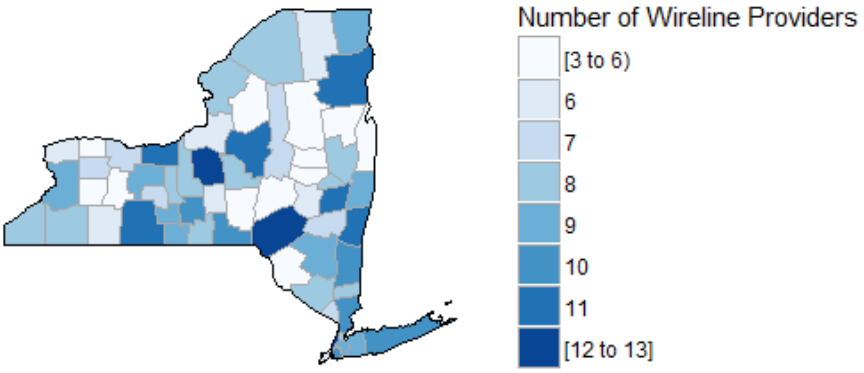
```
multinom(formula = Municipality.Type ~ X..Hse.Units.Cable.1 +
    X..Hse.Units.Wireless + X..Hse.Units.Fiber + X..Hse.Units.DSL +
    X..of.DSL.Providers, data = muni_train)
```

The resulting confusion matrix below had a misclassification rate of 7.5%.

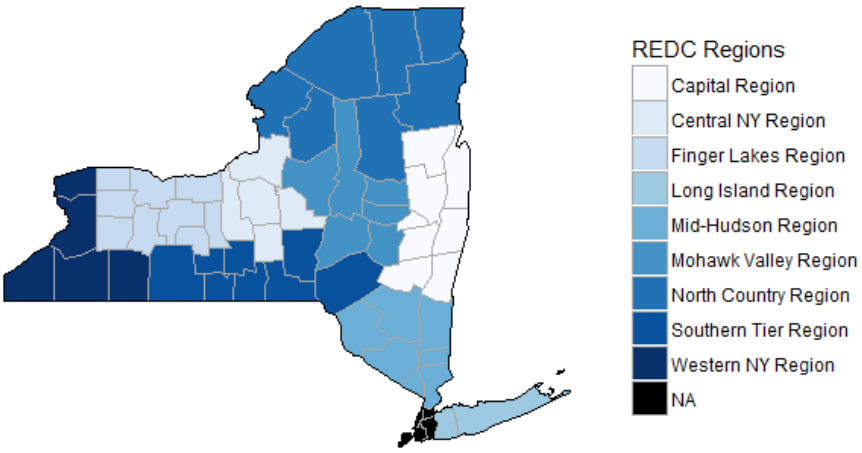
	true		
model	City	Town	Village
City	3	0	0
Town	5	113	4
Village	0	6	69

Although the in-class misclassification rate for City is 62.5%, because of the way the state of New York classifies what a City is, it is really hard for any model to correctly predict the smaller cities. The in-class misclassification rate for Town and Village were a much more reasonable 5.04% and 7.25%.

Because the model is able to differentiate between towns and villages with great accuracy, it is reasonable that towns and villages have different value to internet providers. Towns would have more value to providers, because of a greater population and greater density of people, while villages could be better for small providers to serve a niche area. The heatmap below shows the number of wireline providers (by county) in the state of New York. In some the more suburban areas of New York, there are the most wireline providers, probably because they serve a niche market, unlike someone like Comcast or AT&T who serve a nationwide market. Likewise, in the rural areas, there are not many providers because there is less money to be made out in the middle of nowhere.



Next, we tried to predict the REDC (Regional Economic Development Councils) Region by using the broadband availability data. This would help us to see if there was any real economic difference between being in one REDC region compared to another. To get a better representation of what and REDC region constitutes there is a map below. The NA region constitutes the five boroughs of New York City, because of technical limitations in R.



By using stepwise multinomial logistic regression, we ended up with the model below.

```
multinom(formula = REDC.Region ~ X..Hse.Units.Fiber.1 + X..Wireless.Providers +
    X..Wireline.Providers + X..Hse.Units.Cable.1 + X..Cable.Providers,
    data = muni_train)
```

This is the confusion matrix obtained from the model above. It has a misclassification rate of 68%.

```
                true
model                Capital Region Central NY Region Finger Lakes Region Long Island Region Mid-Hudson Region Mohawk Valley Region North Country Region
  Capital Region                  7                 2                   9                   1                  0                    1                     0
  Central NY Region               0                 0                   0                   0                  0                    0                     0
  Finger Lakes Region             6                 1                  18                   1                  2                    5                     4
  Long Island Region              1                 4                   0                   8                  2                    0                     0
  Mid-Hudson Region               1                 0                   0                   3                  8                    0                     1
  Mohawk Valley Region            3                 0                   2                   0                  1                    5                     4
  North Country Region            1                 5                   2                   0                  1                    8                    12
  Southern Tier Region            1                 1                   1                   0                  0                    2                     3
  Western NY Region               0                 0                   0                   0                  0                    1                     0
                true
model                Southern Tier Region Western NY Region
  Capital Region                        3                 4
  Central NY Region                     0                 0
  Finger Lakes Region                   6                10
  Long Island Region                    2                 2
  Mid-Hudson Region                     2                 1
  Mohawk Valley Region                  3                 6
  North Country Region                  5                 8
  Southern Tier Region                  3                 3
  Western NY Region                     1                 3
```

Because of the very high over misclassification rate, the only two REDC regions with decent in-class misclassification rates were the Long Island, Mid-Hudson, and Finger Lakes Regions, with rates of 25%, 42.9%, and 43.8%, respectively. The confusion matrix tells that REDC region has minimal impact on the broadband availability of a town, city, or village.

Resources Used Outside of Class/Piazza/Notes

1. http://www.ats.ucla.edu/stat/r/dae/mlogit.htm
2. http://rstudio-pubs-static.s3.amazonaws.com/140202_529bec3c57004e3da55f3df889b59c62.html

Appendix:

```
# Import data
  setwd('~/Documents/FALL2016/sta141a_project_datasets/')
  broadband.df =
read.csv('Broadband_Availability_By_Municipality.csv')
#################################################################
# Analysis of muni type
      vtc_df = broadband.df[which(broadband.df$Municipality.Type %in%
c('Village','City','Town')),]
      vtc_df = droplevels(vtc_df)   # Only considering 3 levels for
this analysis.
# First question: What variable makes Muni types distinguishable?
    vtc_list = split(vtc_df,vtc_df$Municipality.Type)
  # Looking at Population and Area:
    # basic summary statistics
    lapply(vtc_list, function(type){summary(type$Muni.Area..sq.mi.)})
    lapply(vtc_list,
function(type){summary(type$X..Hse.Units.Wireless)})
# density plot for Area
library(ggplot2)
area_plot = ggplot() + stat_density(aes(x= Muni.Area..sq.mi. , fill =
Municipality.Type), alpha = I(1/2), trim=TRUE,
            data=vtc_df) + xlim(0,100) + xlab('Area (Square Miles)') +
            ggtitle('Density Plot of Municipality Sizes by Area')
       ## highest 69 observations not displayed on plot.
       ## It is worth noting that they are all from the "Town"
category.

summary(vtc_df$Municipality.Type[which(vtc_df$Muni.Area..sq.mi.>100)])
# density plot for total wireless availability:
wireless_plot =  ggplot() + stat_density(aes(x= X..Hse.Units.Wireless
, fill = Municipality.Type), alpha = I(1/2), trim=TRUE,
                 data=vtc_df) + xlim(0,40000) + xlab('Number of
Housing Units with Wireless') +
                 ggtitle('Density Plot of Wireless Availability
Totals')
        ## highest 30 observations not displayed on plot.

summary(vtc_df$Municipality.Type[which(vtc_df$X2010.Muni.Population>55
000)])
# Ratio of the two:
wireless_ratio_plot = ggplot() + stat_density(aes(x=
X..Hse.Units.Wireless/Muni.Area..sq.mi., fill = Municipality.Type),
```

```
alpha = I(1/2), trim=TRUE, data=vtc_df) + xlim(0,2500) +
ggtitle('Housing Units with Wireless per Square Mile') + xlab('Number
of Housing Units with Wireless per Square Mile')
library(gridExtra)
    grid.arrange(wireless_ratio_plot,
                 wireless_plot+theme(legend.position="none"),
                 area_plot+theme(legend.position="none"),
                nrow=3,ncol=1)

# Import and subset data (only looking at 3 types of municipalities)
broadband.df = read.csv('Broadband_Availability_By_Municipality.csv')
vtc_df = broadband.df[which(broadband.df$Municipality.Type %in%
c('Village','City','Town')),]
vtc_df = droplevels(vtc_df)
vtc_list = split(vtc_df,vtc_df$Municipality.Type)
list2env(vtc_list,.GlobalEnv)   # automatically convert list to
separate df's.

# Create Training and Testing sets.
muni_testing =
data.frame(rbind(tail(Town,n=20),tail(City,n=20),tail(Village,n=20)))
muni_training = setdiff(vtc_df,muni_testing)

# Multinomial Logistic Regression Model
library(nnet)
muni_model = multinom(Municipality.Type~ (X..Hse.Units.Wireless /
Muni.Area..sq.mi.), muni_training)
muni_predictions = predict(muni_model,muni_testing)
# Confusion Matrix
table(muni_predictions,muni_testing$Municipality.Type)
# Misclassification Rate
mean(as.character(muni_predictions) !=
as.character(muni_testing$Municipality.Type))  # 20% misclassification

##lm models and corrgram##
##regression
y = Projectdata$`Hse Units Wireline_1`
y1= as.numeric(as.factor(Y))
x1 = Projectdata$`Municipality Type`
x2 = Projectdata$`REDC Region`
x3 = Projectdata$`Muni Housing Units`
x4 = Projectdata$`Muni Area`
x5 = Projectdata$`Muni Population`
x6 = Projectdata$`Hse Units Wireless_1`
```

```
reg1 = lm(y~x1,data=Projectdata)
summary(reg1)

reg2 = lm(y~x2,data=Projectdata)
summary(reg2)

reg3 = lm(y~x3,data=Projectdata)
summary(reg3)

reg4 = lm(y~x4,data=Projectdata)
summary(reg4)

reg5 = lm(y~x5, data=Projectdata)
summary(reg5)

pairs(~ `Muni Population` + `Muni Area` + `Hse Units Cable_1` + `Hse
Units DSL_1` + `Hse Units Fiber_1` + `Hse Units Wireline_1` + `Hse
Units Wireless_1`,
      data = Projectdata, lower.panel = panel.smooth, upper.panel =
panel.smooth,
      main="Corrgram for the HSE Units")

#County/REDC
bycounty = ny_broadband[which(ny_broadband$Municipality.Type ==
"County"),]
county_redc = bycounty[,c("County","REDC.Region")]
colnames(county_redc)[2] = "value"
county_redc = county_redc[-which(county_redc$value == "New York
City"),]
county_redc$value = droplevels(county_redc$value)
county_redc$County = droplevels(county_redc$County)
county_redc$value = as.character(county_redc$value)

data(county.regions)
county.regions = filter(county.regions, state.name == "new york")
county_redc$County = tolower(county_redc$County)
county_redc = left_join(county_redc, county.regions, by = c("County" =
"county.name"))
county_redcmap = county_choropleth(county_redc, legend = "REDC
Regions", state_zoom = "new york")
county_redcmap

#Predict Municipality Type
muni = ny_broadband
muni = muni[muni$Municipality.Type %in% c("Town", "City", "Village"),]
```

```
muni$X..Hse.Units.Cable.1 =
as.numeric(sub("%","",muni$X..Hse.Units.Cable.1))/100 #convert char
percentages to decimals
muni$X..Hse.Units.DSL.1 =
as.numeric(sub("%","",muni$X..Hse.Units.DSL.1))/100 #convert char
percentages to decimals
muni$X..Hse.Units.Fiber.1 =
as.numeric(sub("%","",muni$X..Hse.Units.Fiber.1))/100 #convert char
percentages to decimals
muni$X..Hse.Units.Wireline.1 =
as.numeric(sub("%","",muni$X..Hse.Units.Wireline.1))/100 #convert char
percentages to decimals
muni$X..Hse.Units.Wireless.1 =
as.numeric(sub("%","",muni$X..Hse.Units.Wireless.1))/100 #convert char
percentages to decimals
muni = muni[complete.cases(muni), ]
muni$Municipality.Type = droplevels(muni$Municipality.Type)
muni$REDC.Region = droplevels(muni$REDC.Region)
muni$County = droplevels(muni$County)

muni_test = muni[1:200,]
muni_train = muni[201:1548,]

library(nnet)
muni.null = multinom(Municipality.Type ~ 1, data = muni_train)
muni.full = multinom(Municipality.Type ~ X..Cable.Providers +
X..Hse.Units.Fiber.1 + X..Hse.Units.Fiber + X..Hse.Units.DSL.1 +
X..Hse.Units.DSL + X..Hse.Units.Cable.1 + X..Hse.Units.Cable +
X..of.DSL.Providers + X..Fiber.Providers + X..Hse.Units.Wireline.1 +
X..Hse.Units.Wireline + X..Wireline.Providers +
X..Hse.Units.Wireless.1 + X..Hse.Units.Wireless +
X..Wireless.Providers + X..Satellite.Providers, data = muni_train)
muni_model = stepAIC(muni.null, scope = list(upper=muni.full),
direction="forward", data=muni_train)
muni_model = multinom(Municipality.Type ~ X..Hse.Units.Cable.1 +
X..Hse.Units.Wireless + X..Hse.Units.Fiber + X..Hse.Units.DSL +
X..of.DSL.Providers, data = muni_train )
summary(muni_model)
muni_predict = predict(muni_model,muni_test)
table(model = muni_predict,true = muni_test$Municipality.Type )
mean(as.character(muni_predict) !=
as.character(muni_test$Municipality.Type ))

#Predict REDC REgion
redc.null = multinom(REDC.Region ~ 1, data = muni_train)
```

```
redc.full = multinom(REDC.Region ~ X..Cable.Providers +
X..Hse.Units.Fiber.1 + X..Hse.Units.Fiber + X..Hse.Units.DSL.1 +
X..Hse.Units.DSL + X..Hse.Units.Cable.1 + X..Hse.Units.Cable +
X..of.DSL.Providers + X..Fiber.Providers + X..Hse.Units.Wireline.1 +
X..Hse.Units.Wireline + X..Wireline.Providers +
X..Hse.Units.Wireless.1 + X..Hse.Units.Wireless +
X..Wireless.Providers + X..Satellite.Providers, data = muni_train)
redc_model = stepAIC(redc.null, scope = list(upper=redc.full),
direction="forward", data=muni_train)
redc_model = multinom(REDC.Region ~ X..Hse.Units.Fiber.1 +
X..Wireless.Providers + X..Wireline.Providers + X..Hse.Units.Cable.1 +
X..Cable.Providers, data = muni_train)
summary(redc_model)
redc_predict = predict(redc_model,muni_test)
table(model = redc_predict,true = muni_test$REDC.Region )
mean(as.character(redc_predict) != as.character(muni_test$REDC.Region
))
```