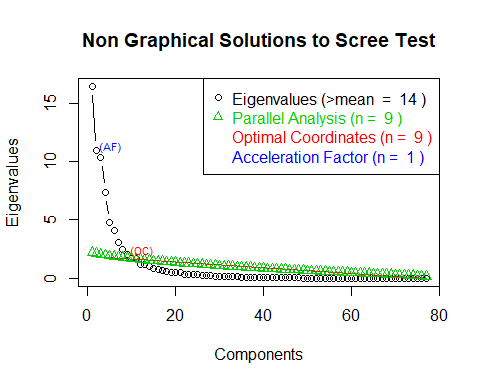
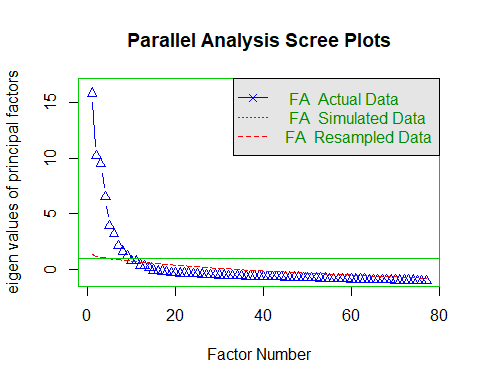
Austen\_Corey\_BCA\_Assignment\_2

Based on the Scree tests, the optimal number of factors is 9 to 11. Having run through the 3 different configurations, I decided to go with **9** factors because the variable combinations made the most sense. 

o **F1** - This is an obvious income-based factor, variables that positively impact are “household income over 100k”, “per capita”, “household income 75-100k”, “median value”. Variables that have a negative effect are low income variables, like “household income 20-30k”, “household income below 20k”, and “9th to 12th grade education”.

o **F2** - This factor is age based. The most influential are higher age range variables that have negative influence (householder 65-74 years, age 65-74 years), and the most positively influential factors are lower age range variables, like “householder 25-34 years”, “age 25-34 years”, and “never married.”

o **F3** - Variables seem to be transportation related, specifically no vehicles and public transportation use. The variables that have negative scores are centered on multiple vehicle ownership and private transportation use.

o **F4** - Factor includes variables related to household size, with positive scores on variables of household of 5 people, household of 6 or more people. It also positive scores for “Hispanic”, “education less than 9th grade”, and “unemployed”. The negative scores are smaller households or 1 or 2 people.

o **F5** - Factor appears to be related to travel time, with positive scores on travel times of 30-44 and 45-59, as well as low scores from travel times of 5-14. Interestingly, this also seems to be linked to race, with positive scores from black residents and negative scores from white residents.

o **F6** - This factor is interesting in that the variables don’t seem to make sense together, such as “divorced”, “some college education”, “work in residence” “single fathers”, “work in central city”, and “Indian”.

o **F7** - Factor is also related to income, with the only positive score linked to “below poverty line” and negative scores from income brackets of 40-50k and 50-60k.

o **F8** - Age related factor, with positive scores of ages 35-44, 45-54, and householder 45-54. Interestingly, this also has negative scores from household and population growth.

o **F9** - Factor that seems to be a collection of outliers, such as “work at home”, “male”, and “traveltime over 90 minutes.”

# Rename factors  
colnames(retail.score)[colnames(retail.score)=="F1"] <- "high.income"  
colnames(retail.score)[colnames(retail.score)=="F2"] <- "young.single"  
colnames(retail.score)[colnames(retail.score)=="F3"] <- "public.transit"  
colnames(retail.score)[colnames(retail.score)=="F4"] <- "large.hisp.fam"  
colnames(retail.score)[colnames(retail.score)=="F5"] <- "black.commuters"  
colnames(retail.score)[colnames(retail.score)=="F6"] <- "single.parents"  
colnames(retail.score)[colnames(retail.score)=="F7"] <- "poverty"  
colnames(retail.score)[colnames(retail.score)=="F8"] <- "adults"  
colnames(retail.score)[colnames(retail.score)=="F9"] <- "travel.other"

**high.income** - This matches what I would expect, MSAs with high scores on “high.income” such as Norwalk, CT, have a high average income. MSAs with low scores, like Danville, VA, have a lower average income.

## high.income  
## Stamford--Norwalk, CT PMSA 4.224408  
## San Jose, CA PMSA 3.788482  
## Danbury, CT PMSA 3.493702  
## Nassau--Suffolk, NY PMSA 3.335431  
## Middlesex--Somerset--Hunterdon, NJ PMSA 3.072129  
## San Francisco, CA PMSA 2.843753  
## Washington, DC--MD--VA--WV PMSA 2.684000  
## Bergen--Passaic, NJ PMSA 2.370862  
## Boulder--Longmont, CO PMSA 2.137474  
## Newark, NJ PMSA 2.132677

## high.income  
## Merced, CA MSA -1.385622  
## Jamestown, NY MSA -1.398582  
## Mansfield, OH MSA -1.410712  
## Houma, LA MSA -1.430390  
## Longview--Marshall, TX MSA -1.521725  
## Fort Smith, AR--OK MSA -1.549187  
## Huntington--Ashland, WV--KY--OH MSA -1.607431  
## Florence, SC MSA -1.664848  
## Killeen--Temple, TX MSA -1.700594  
## Danville, VA MSA -1.723920

**young.single** - Matches what I would expect. MSAs with high “young.single” scores are all college towns, where as those with low scores are towns with a large retirement community.

## young.single  
## Bryan--College Station, TX MSA 2.965981  
## Ann Arbor, MI PMSA 2.910730  
## Charlottesville, VA MSA 2.596331  
## Bloomington, IN MSA 2.485507  
## Fayetteville, NC MSA 2.238879  
## Kenosha, WI PMSA 2.186188  
## Memphis, TN--AR--MS MSA 2.114150  
## Santa Cruz--Watsonville, CA PMSA 2.110619  
## Bremerton, WA PMSA 2.095445  
## Columbia, MO MSA 2.060132

## young.single  
## Steubenville--Weirton, OH--WV MSA -2.323887  
## Wheeling, WV--OH MSA -2.527486  
## Chico--Paradise, CA MSA -2.801235  
## Monmouth--Ocean, NJ PMSA -2.877232  
## Fort Myers--Cape Coral, FL MSA -3.107074  
## Ocala, FL MSA -3.148924  
## West Palm Beach--Boca Raton, FL MSA -3.191332  
## Naples, FL MSA -3.625615  
## Fort Pierce--Port St. Lucie, FL MSA -4.290128  
## Sarasota--Bradenton, FL MSA -4.361969

**public.transit** - Matches what I would expect. MSAs with high scores here are New York City and Jersey City. Those with low scores are resort towns or have very little in terms of public transit

## public.transit  
## New York, NY PMSA 8.418320  
## Jersey City, NJ PMSA 8.039815  
## San Francisco, CA PMSA 5.938126  
## Lowell, MA--NH PMSA 5.448840  
## Eugene--Springfield, OR MSA 4.093404  
## Bangor, ME MSA 3.966971  
## Madison, WI MSA 3.647953  
## Santa Rosa, CA PMSA 3.355939  
## Chico--Paradise, CA MSA 3.122255  
## State College, PA MSA 3.092684

## public.transit  
## Bremerton, WA PMSA -3.561255  
## Olympia, WA PMSA -3.580054  
## San Jose, CA PMSA -3.616150  
## Akron, OH PMSA -4.217254  
## Huntsville, AL MSA -4.344243  
## Medford--Ashland, OR MSA -4.461509  
## Elkhart--Goshen, IN MSA -4.615655  
## Ann Arbor, MI PMSA -5.196852  
## Kenosha, WI PMSA -5.890050  
## Bellingham, WA MSA -7.734373

**large.hisp.fam** - Matches what I would expect. MSAs with high scores are towns along the US/Mexico border. MSAs with low scores are located in the Midwest or further from the Mexico border.

## large.hisp.fam  
## McAllen--Edinburg--Mission, TX MSA 5.468697  
## Brownsville--Harlingen--San Benito, TX MSA 4.199836  
## Merced, CA MSA 3.664534  
## Visalia--Tulare--Porterville, CA MSA 3.523092  
## El Paso, TX MSA 3.506467  
## Modesto, CA MSA 2.956231  
## Provo--Orem, UT MSA 2.862051  
## Stockton--Lodi, CA MSA 2.521673  
## Riverside--San Bernardino, CA PMSA 2.272164  
## Yakima, WA MSA 2.263048

## large.hisp.fam  
## Lewiston--Auburn, ME MSA -1.431039  
## Bellingham, WA MSA -1.469776  
## Ann Arbor, MI PMSA -1.610864  
## Gainesville, FL MSA -1.632469  
## Charleston, WV MSA -1.639255  
## Columbia, MO MSA -1.711966  
## Springfield, MO MSA -1.784866  
## Tuscaloosa, AL MSA -1.789241  
## Champaign--Urbana, IL MSA -1.941045  
## Bloomington, IN MSA -1.958587

**black.commuters** - This matches what I expect. The MSAs with high scores have large black populations, whereas those with low scores have high white populations.

## black.commuters  
## Atlanta, GA MSA 3.248300  
## Orlando, FL MSA 2.679012  
## Birmingham, AL MSA 2.293659  
## Washington, DC--MD--VA--WV PMSA 2.233968  
## Mobile, AL MSA 2.196595  
## Memphis, TN--AR--MS MSA 2.060404  
## Baltimore, MD PMSA 2.028530  
## Dallas, TX PMSA 1.974208  
## Miami, FL PMSA 1.825781  
## Richmond--Petersburg, VA MSA 1.758488

## black.commuters  
## Scranton--Wilkes-Barre--Hazleton, PA MSA -2.087531  
## Williamsport, PA MSA -2.105728  
## Lafayette, IN MSA -2.168537  
## Bellingham, WA MSA -2.263804  
## Lincoln, NE MSA -2.287865  
## Eau Claire, WI MSA -2.334246  
## Waterloo--Cedar Falls, IA MSA -2.412225  
## Cedar Rapids, IA MSA -2.419437  
## Champaign--Urbana, IL MSA -2.483375  
## State College, PA MSA -2.540993

**single.parents** - This doesn’t quite show what I expected. The factor seems to be a bit random anyway, so I am not surprised by this overall. The MSAs with high and low scores don’t seem to have an underlying theme as far as I can tell.

## single.parents  
## Bellingham, WA MSA 5.439140  
## Gainesville, FL MSA 4.182266  
## McAllen--Edinburg--Mission, TX MSA 3.790628  
## San Jose, CA PMSA 3.486589  
## Tuscaloosa, AL MSA 3.398515  
## Stamford--Norwalk, CT PMSA 3.189370  
## Bloomington, IN MSA 3.134944  
## Waco, TX MSA 3.066855  
## Jackson, MS MSA 3.014139  
## Ann Arbor, MI PMSA 2.952433

## single.parents  
## Reading, PA MSA -2.804506  
## Modesto, CA MSA -2.976104  
## Hagerstown, MD PMSA -2.997144  
## Grand Rapids--Muskegon--Holland, MI MSA -3.065439  
## Killeen--Temple, TX MSA -3.150608  
## York, PA MSA -3.216101  
## Fort Walton Beach, FL MSA -3.248577  
## Janesville--Beloit, WI MSA -3.272157  
## Appleton--Oshkosh--Neenah, WI MSA -3.460579  
## Lowell, MA--NH PMSA -5.126728

**poverty** - This lines up somewhat with how I expected. The amount of people living below the poverty line in MSAs in both the high and low ranges are very similar.

## poverty  
## Bellingham, WA MSA 5.400480  
## Kenosha, WI PMSA 5.169299  
## Santa Cruz--Watsonville, CA PMSA 4.635794  
## Bremerton, WA PMSA 4.417362  
## Medford--Ashland, OR MSA 3.997652  
## Huntsville, AL MSA 3.164930  
## Olympia, WA PMSA 3.075279  
## Akron, OH PMSA 2.912266  
## Memphis, TN--AR--MS MSA 2.842105  
## Nashua, NH PMSA 2.775960

## poverty  
## Fort Walton Beach, FL MSA -3.493865  
## Boise City, ID MSA -3.580163  
## Lafayette, IN MSA -3.620029  
## Sarasota--Bradenton, FL MSA -3.674370  
## Naples, FL MSA -3.715763  
## Fort Myers--Cape Coral, FL MSA -3.792635  
## Madison, WI MSA -3.883462  
## Lowell, MA--NH PMSA -3.964265  
## Provo--Orem, UT MSA -4.253591  
## Fort Pierce--Port St. Lucie, FL MSA -4.603805

**adults** - The MSAs with high scores have high concentrations of adults, whereas those with low scores have more young people.

## adults  
## Anchorage, AK MSA 3.651729  
## Albuquerque, NM MSA 2.840198  
## Reno, NV MSA 2.713304  
## Redding, CA MSA 2.583695  
## Las Vegas, NV--AZ MSA 2.314735  
## Boise City, ID MSA 2.222407  
## Eugene--Springfield, OR MSA 2.201309  
## Tucson, AZ MSA 2.177119  
## Santa Rosa, CA PMSA 2.129292  
## Chico--Paradise, CA MSA 1.983609

## adults  
## Scranton--Wilkes-Barre--Hazleton, PA MSA -1.729493  
## Tuscaloosa, AL MSA -1.757701  
## Lawrence, MA--NH PMSA -1.956423  
## Ann Arbor, MI PMSA -1.973433  
## Nassau--Suffolk, NY PMSA -1.993043  
## Florence, SC MSA -2.062410  
## Johnstown, PA MSA -2.128873  
## Charlottesville, VA MSA -2.274651  
## Middlesex--Somerset--Hunterdon, NJ PMSA -2.300082  
## State College, PA MSA -3.074504

**travel.other** - This what I would expect. The MSAs with high scores are more than an hour and a half form a metropolitan area, so the commute would be longer, and MSAs with low scores are either close to or are metropolitan areas themselves.

## travel.other  
## Bellingham, WA MSA 9.566949  
## Bremerton, WA PMSA 6.745886  
## Kenosha, WI PMSA 6.475019  
## Ann Arbor, MI PMSA 6.248335  
## Medford--Ashland, OR MSA 6.007158  
## Santa Cruz--Watsonville, CA PMSA 5.955533  
## Charlottesville, VA MSA 5.475212  
## Brazoria, TX PMSA 4.735305  
## Olympia, WA PMSA 3.957730  
## Salem, OR PMSA 3.936975

## travel.other  
## Baton Rouge, LA MSA -3.457523  
## El Paso, TX MSA -3.603497  
## Santa Rosa, CA PMSA -3.905590  
## Springfield, IL MSA -3.910945  
## Bangor, ME MSA -4.525469  
## Columbus, GA--AL MSA -4.568284  
## Roanoke, VA MSA -4.594174  
## Kansas City, MO--KS MSA -4.730573  
## Wichita, KS MSA -5.011576  
## Lowell, MA--NH PMSA -6.929683

d.var <- retail[, c("Share100\_4445\_72","Share4451\_722","Groc\_non\_food","Groc\_food","S100\_H","S120\_H","pq\_g","pq\_r","pqr\_nonfood","pqr\_food","Groc\_non\_food1","Groc\_food1","Sr4451\_100","Sr722\_120","Nr4451\_100","Nr722\_120")]

factors <- cbind(retail.score,d.var)

factors[] <- lapply(factors, function(x) {   
 x[is.na(x)] <- mean(x, na.rm = TRUE)  
 x  
})

**Share100\_4445\_72**:“Sales of ML100 / (Sales of ML100 and ML120)”

Based on this model, the percentage of grocery sales compared to that of resturants increases in areas of higher poverty rates and large Hispanic families, but drops in areas of **high.income**, **young.single**, and **adults**. This makes sense since people with less disposible income are going to spend more of their income at a grocery store.

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'family' will be disregarded

##   
## Call:  
## lm(formula = Share100\_4445\_72 ~ high.income + young.single +   
## public.transit + large.hisp.fam + black.commuters + single.parents +   
## poverty + adults + travel.other, data = factors, family = binomial)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.199778 -0.021485 -0.001985 0.023193 0.112757   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.6143061 0.0024294 252.862 < 2e-16 \*\*\*  
## high.income -0.0116408 0.0024058 -4.839 2.33e-06 \*\*\*  
## young.single -0.0179342 0.0022511 -7.967 6.38e-14 \*\*\*  
## public.transit -0.0005898 0.0020039 -0.294 0.768758   
## large.hisp.fam 0.0135222 0.0023708 5.704 3.41e-08 \*\*\*  
## black.commuters -0.0127235 0.0022245 -5.720 3.14e-08 \*\*\*  
## single.parents -0.0071946 0.0020455 -3.517 0.000521 \*\*\*  
## poverty 0.0072497 0.0021739 3.335 0.000987 \*\*\*  
## adults -0.0106678 0.0023825 -4.478 1.16e-05 \*\*\*  
## travel.other 0.0054093 0.0019072 2.836 0.004952 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.03857 on 242 degrees of freedom  
## Multiple R-squared: 0.4529, Adjusted R-squared: 0.4325   
## F-statistic: 22.26 on 9 and 242 DF, p-value: < 2.2e-16

**pq\_r**:“Restaurant sales per household”

The outcome from this model definitely makes sense, as more wealthy people would go out to eat mroe often. Age does not seem to make a difference, as both **young.single** and **adults** add to this. **Poverty** and **large.hisp.fam** bring this down, since populations with less disposible income are bound to spend less at resturants.

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'family' will be disregarded

##   
## Call:  
## lm(formula = pq\_r ~ high.income + young.single + public.transit +   
## large.hisp.fam + black.commuters + single.parents + poverty +   
## adults + travel.other, data = factors, family = binomial)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7477 -0.2192 -0.0526 0.1577 2.2963   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.55660 0.02325 109.941 < 0.0000000000000002 \*\*\*  
## high.income 0.23132 0.02303 10.045 < 0.0000000000000002 \*\*\*  
## young.single 0.17318 0.02155 8.037 0.0000000000000406 \*\*\*  
## public.transit -0.01597 0.01918 -0.832 0.406008   
## large.hisp.fam -0.07869 0.02269 -3.468 0.000621 \*\*\*  
## black.commuters 0.02524 0.02129 1.186 0.236978   
## single.parents 0.06142 0.01958 3.137 0.001919 \*\*   
## poverty -0.06661 0.02081 -3.201 0.001551 \*\*   
## adults 0.11737 0.02280 5.147 0.0000005487611888 \*\*\*  
## travel.other -0.04705 0.01826 -2.577 0.010558 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3691 on 242 degrees of freedom  
## Multiple R-squared: 0.4732, Adjusted R-squared: 0.4536   
## F-statistic: 24.15 on 9 and 242 DF, p-value: < 0.00000000000000022

**Nr4451\_100**:“Number of grocery stores (handling ML100) per household”

This would makes sense because a lack public transit, commuting, and travel, along with **high.income**, more young single people, and adults imply a suburban area, a college town, or a resort/retirement community. Areas like this tend to have more sprawl or more areas for shopping.

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'family' will be disregarded

##   
## Call:  
## lm(formula = Sr4451\_100 ~ high.income + young.single + public.transit +   
## large.hisp.fam + black.commuters + single.parents + poverty +   
## adults + travel.other, data = factors, family = binomial)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1901.32 -431.78 -35.71 403.57 2567.76   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3385.79 46.52 72.788 < 0.0000000000000002 \*\*\*  
## high.income 286.19 46.06 6.213 0.000000002253747 \*\*\*  
## young.single 217.58 43.10 5.048 0.000000878150945 \*\*\*  
## public.transit -302.37 38.37 -7.881 0.000000000000111 \*\*\*  
## large.hisp.fam -62.36 45.39 -1.374 0.170785   
## black.commuters -319.51 42.59 -7.501 0.000000000001197 \*\*\*  
## single.parents -188.91 39.17 -4.823 0.000002497457999 \*\*\*  
## poverty -137.47 41.62 -3.303 0.001102 \*\*   
## adults 212.48 45.62 4.658 0.000005275592210 \*\*\*  
## travel.other -139.92 36.52 -3.832 0.000162 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 738.4 on 242 degrees of freedom  
## Multiple R-squared: 0.4567, Adjusted R-squared: 0.4365   
## F-statistic: 22.6 on 9 and 242 DF, p-value: < 0.00000000000000022

**Sr722\_120**:“ML120 sales per restaurant”

The outcome from this one is surprising. I would have thought income would play a bigger role here, but the model didn’t find high.income to be significant. It did however, find that poverty was significant in that it increases ML120 sales per restaurant as it drops. We also see it increases as **young.single**, **single.parents**, **black.commuters**, and adults increase in score. I interpret this to mean populations with middle class incomes increase sales per resturant.

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'family' will be disregarded

##   
## Call:  
## lm(formula = Sr722\_120 ~ black.commuters + young.single + public.transit +   
## large.hisp.fam + high.income + single.parents + poverty +   
## adults + travel.other, data = factors, family = binomial)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -150.991 -38.494 -1.397 37.161 199.380   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 481.553 3.583 134.414 < 0.0000000000000002 \*\*\*  
## black.commuters 33.832 3.280 10.313 < 0.0000000000000002 \*\*\*  
## young.single 29.662 3.320 8.935 < 0.0000000000000002 \*\*\*  
## public.transit -11.453 2.955 -3.876 0.000137 \*\*\*  
## large.hisp.fam -4.749 3.496 -1.358 0.175621   
## high.income 3.860 3.548 1.088 0.277722   
## single.parents 19.157 3.016 6.351 0.0000000010506 \*\*\*  
## poverty -22.106 3.206 -6.896 0.0000000000464 \*\*\*  
## adults 9.622 3.513 2.739 0.006628 \*\*   
## travel.other -11.601 2.813 -4.125 0.0000510411943 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 56.87 on 242 degrees of freedom  
## Multiple R-squared: 0.5501, Adjusted R-squared: 0.5333   
## F-statistic: 32.87 on 9 and 242 DF, p-value: < 0.00000000000000022

**Conclusion**:

Based on the outcome from the models run, a company would want to open a grocery store in a suburban area with average to low income as they would likely see more income overall than if they put a grocery store into a wealthier MSA. However, opening a resturant would be a wise choice in a wealthy MSA, areas where there are large populations of commuters traveling in and out of the area, and areas where there are more young people.