Predicting the Capture rate of Eye Care Practices

# Abstract

Optometry practices across the United States want to better understand what aspects of their practices influence their financial health, and from there to improve the way they offer eye care to both existing and prospect patients. There are very few studies have been done on the optometry field, and even fewer that touch on the business side of optometry practices. This study aims to explore and further understand this aspect of optometry. The original sample goes back 2 years; it consists of 9000 observations, each being one optometry practice performance for a quarter of a fiscal year. The analysis methodology includes Principal component analysis, Factor analysis and Correspondence analysis to determine the unobserved relationships between variables, and to identify which variables should be included into Model building and Multivariate Regression. Cluster analysis was used to group samples into clusters, and explore their intra/inter-similarities. One of the findings indicate that there are significantly more patients aged between 30 and 59 than patients aged younger than 29, or older than 60. This suggests that optometry practices might find it beneficial to focus on their largest patient segment. However, they could also attempt to attain more younger patients, as millennials are becoming more aware of, and involved with, their personal health. On the other hand, we found that Medical management ratio, New patient percentage, and Production booked are the prevalent predictors of capture rate. Lastly, future studies of this topic could potentially include more non-metric variables, or more details of the individual practices, which would help to classify optometry practices, based on sales and receipts, more clearly.

# I. Introduction

According to the “Health United States 2016” report published by the Centers for Disease Control and Prevention, “In 2015, 17.8% of the U.S. Gross Domestic Product (GDP) was spent on national health care—more than twice the percentage in 1975 (7.9% of GDP). More is spent on healthcare in the United States, in terms of a percentage of GDP, than any other developed country for which data are collected by the Organization of Economic Cooperation and Development (OECD).” But yet, by 2015, in the United States, myopia (nearsightedness) rates have almost doubled among the millennials. This indicates the importance of keeping optometry practices to be more sustainable, in order to provide high quality and affordable services. The objective of this study is to identify which aspects of eye care practices influence their financial health across the United States. In particular, accounts receivables, eyeglasses capture rate, and patient ages are some of the variables that this study analyzed and interpreted their influence on the practices’ finance.

## Data source

The data used in this study comes from aggregated practice performance metrics provided by RevolutionEHR – a cloud-based EHR that is specialized for optometry and ophthalmology. All of the data in this source is aggregated by quarter during the 2016 and 2017 fiscal years. In order to use this data source, some redaction was performed at the request of RevolutionEHR. Practice identifiers and quarterly time periods have been removed from all observations. A total of 9000 observations are present in the dataset.

There is a variety of discrete and continuous quantitative variables in the dataset. Patient age rates and capture rates are among some of the most interesting variables to work with. There are only a few units used in the data: dollar amounts, business event counts, and business event rates. Business event counts refer to variables such as male patients seen, or the number of eyeglass prescriptions filled. Business event rates refer to the percentage of patients seen that were male, or the percentage of eyeglass prescriptions filled.

Significant steps needed to be taken in our initial analysis to limit the impact of outliers in the dataset. Without any formal sampling criteria, the data samples from all RevolutionEHR users: small, large, fully-implemented, and partially implemented optometry practices. We chose to eliminate the very large and very small practices by requiring a quarterly revenue between $125,000 and $1,500,000. We chose to eliminate the partially implemented practices by requiring non-zero capture rates for the business events and accounts receivable amounts significantly exceeding sales and receipt amounts. This process reduced our available data from 9,000 observations to 4262 observations.

# II. Literature Review

There are many broadly accepted optometry practice performance metrics (Gerber 2013). The 2013 study expanded on many of the variables in our dataset to include practice operational details like employee hours and patient satisfaction. It further expanded on the individual strategy behind the performance metrics and how they impact the practice.

The impact and benefits of EHR adoption on solo and small group optometry practices have been researched (Miller 2003). This research was done very early in the adoption of EHR technology and had anticipated the efforts that were undertaken by RevolutionEHR to include the coding rates for medical issues and patient refractions.

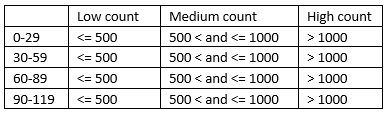
# III. Methods

## Correspondence Analysis

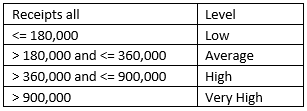
Data transformation:

Age group are first combined. Initially, there were 12 age group classifications. After grouping, we have 4 groups: 0 to 29; 30 to 59; 60 to 89; and 90 to 119.

Further transformation was needed, as each variable still contains patient counts. Based on the frequency distribution of each age group, we transformed Age variables to have 3 main categorical values: Low, Medium and High.



The next variable that we transformed into categorical was **Receiptsbyproviderall**, based on the average amount of this variable, which does not take into consideration of the size of the optometry practice (small, large, fully-implemented, and partially implemented).



## Cluster Analysis

A cluster analysis is performed on three subsets of the practice data: Revenue Amounts, Clinical Events, and Patient Demographics. The CLARA (Clustering for Large Applications) algorithm was used in conjunction with Manhattan distances. The number of clusters was determined by examining the within sum of squares values of the clustering via the elbow method. Hierarchical Clustering on Principal Components was used for de-noising, and leading to more stable clustering. Also, the method provides v.test scores to determine which variable is associated to which cluster.

## Principal Component Analysis and Common Factor Analysis

As the dataset has a more number of variables, Model is optimized by using PCA dimension reduction focusing on the good interpretations. The driving factors of the rotated components are interpreted by using Latent factor detection and the Common factor analysis.

## Canonical Correlation Analysis

A canonical correlation analysis is performed on the resulting components created during the principal component analysis and Common factor analysis. This analysis will focus on showing the correlation between the components and the capture rates.

## Model Building and Multivariate Regression

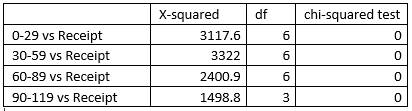
Applying LASSO regression technique, an initial model building is performed in order to predict eye glasses capture rate. With a 60% training and 40% testing split of the data, a cross-validation has returned minimized prediction error while selecting prior variables: Age groups 10 – 19, 30 – 39, 40 – 49, 60 – 69, 100 – 109, Medical management ratio, New patient percentage, and Production booked. This model would be beneficial to eye care providers as it presents the important approaches to achieving higher spectacles capture rate, which is directly related to financial health.

# IV. Discussion and Results

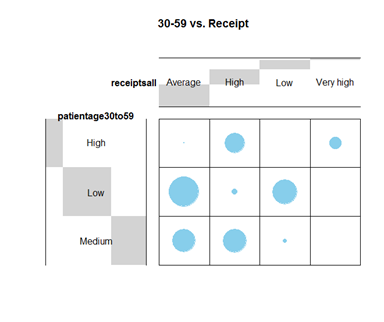
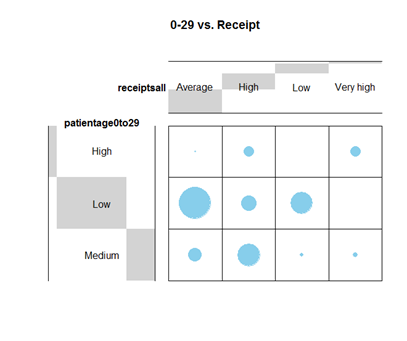
## **Correspondence Analysis**

Running Correspondence analysis on each Age variable vs. Receipt variable, we found:

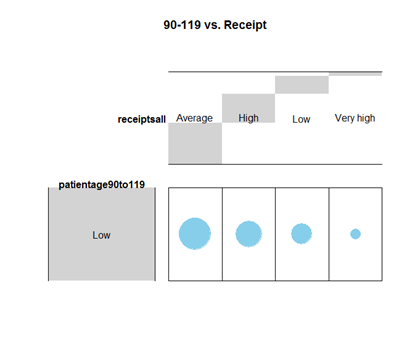
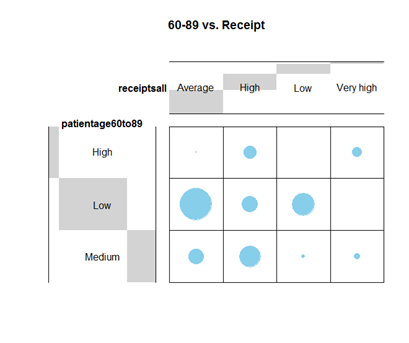
1. **Chi-squared test** for each of the Correspondence analysis indicates significant independence between row and column categories.



**CA Visualizations**



**Figure 1A Figure 1B**



**Figure 1C Figure 1D**

1. Interpretations:

* Age groups **0-29** and **60-89** are very similar in terms of their correlation to the total receipt of an eye care practice. (Figures 1A and 1C). In other words, practices that have a low number of patients (under 500 patients) within these 2 age groups are most likely to have an Average total receipt for that quarter ($180,000 - $360,000).
* Age group **30-59** is slightly different compared to age groups **0-29** and **60-89**. An optometry practice that has a low number of patients aged 30-59 are much less likely to have a High total quarterly receipt (360,000 - 900,000). Additionally, this age group has the strongest correlation between High number of patients and High total receipt, compared to other age groups. (Figure 1B)
* **90-119** is the only age group that has correlation between Low number of patients and Very High total receipt, even though it is not a strong correlation. (Figure 1D)

1. Conclusion:

It would be most interesting to use age variables that are between age groups 30-59, as we want to further understand the largest group of patients, who contribute the majority of the total receipt by optometry practices.

## **Cluster Analysis - Clinical events**

### Clinical Events

The cluster analysis led to five clusters with the following medoids:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Size | AR Sld | CL Rx | CL Sld | EG Rx | EG Sld | Medical | New Pat | 99214 | 99215 |
| 1. High Capture, Up-Coding | 17.9% | 526 | 459 | 317 | 1723 | 734 | 793 | 442 | 44 | 244 |
| 2. Up-Coding | 10.9% | 258 | 367 | 263 | 1141 | 293 | 825 | 289 | 276 | 376 |
| 3. Low Capture, Up-Coding | 27.3% | 243 | 152 | 88 | 636 | 191 | 371 | 207 | 89 | 174 |
| 4. Low Capture | 29.0% | 217 | 156 | 120 | 495 | 185 | 178 | 213 | 13 | 45 |
| 5. Low Cap. Under-Coding | 14.9% | 413 | 316 | 200 | 839 | 231 | 374 | 260 | 4 | 105 |

High Capture Up-Coding — 17.9% of the sample

These practices have higher than average capture rates. They code and bill with a higher than average rate of 99214(up-coded) examinations.

Up-Coding — 10.9% of the sample

These practices have average capture rates. They code and bill with a higher than average rate of 99214(up-coded) examinations.

Low Capture Up-Coding — 27.3% of the sample

These practices have lower than average capture rates. They code and bill with a higher than average rate of 99214(up-coded) examinations.

Low Capture — 29.0% of the sample

These practices have lower than average capture rates. They code and bill with an average rate of 99214 examinations.

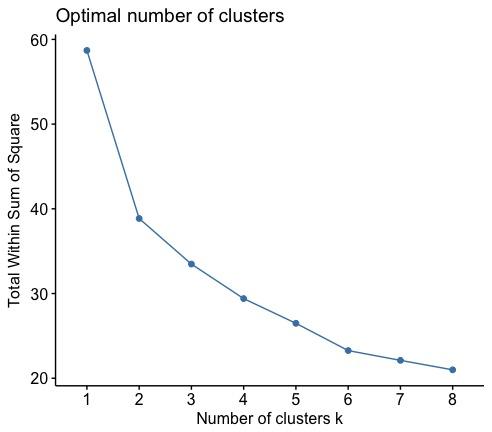
Low Capture Under-Coding — 14.9% of the sample

These practices have lower than average capture rates. They code and bill with a lower than average rate of 99214 examinations.

## **Cluster Analysis – Patients**

There are 10 age variables used: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89 and 90-99.

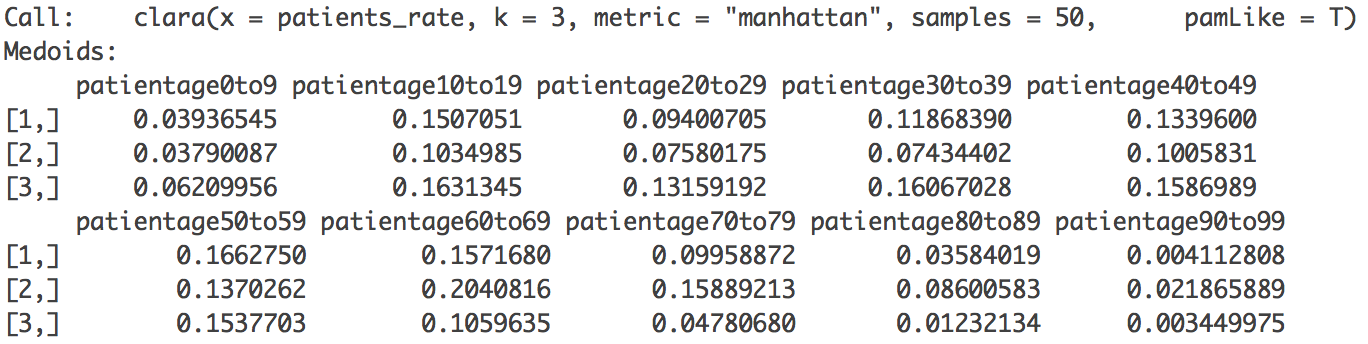
We performed additional transformation for the dataset from number to percentage. It is important for normalization, since the K-mean method is performed.



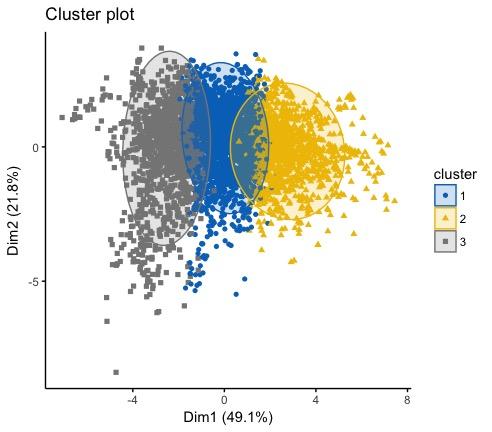
Using the Optimal number of clusters chart, we chose k = 3.

K-means sample size of each clusters:

|  |  |
| --- | --- |
| K-means clustering with 3 clusters’ sizes | |
| Cluster 1 | 2060 |
| Cluster 2 | 1042 |
| Cluster 3 | 1160 |

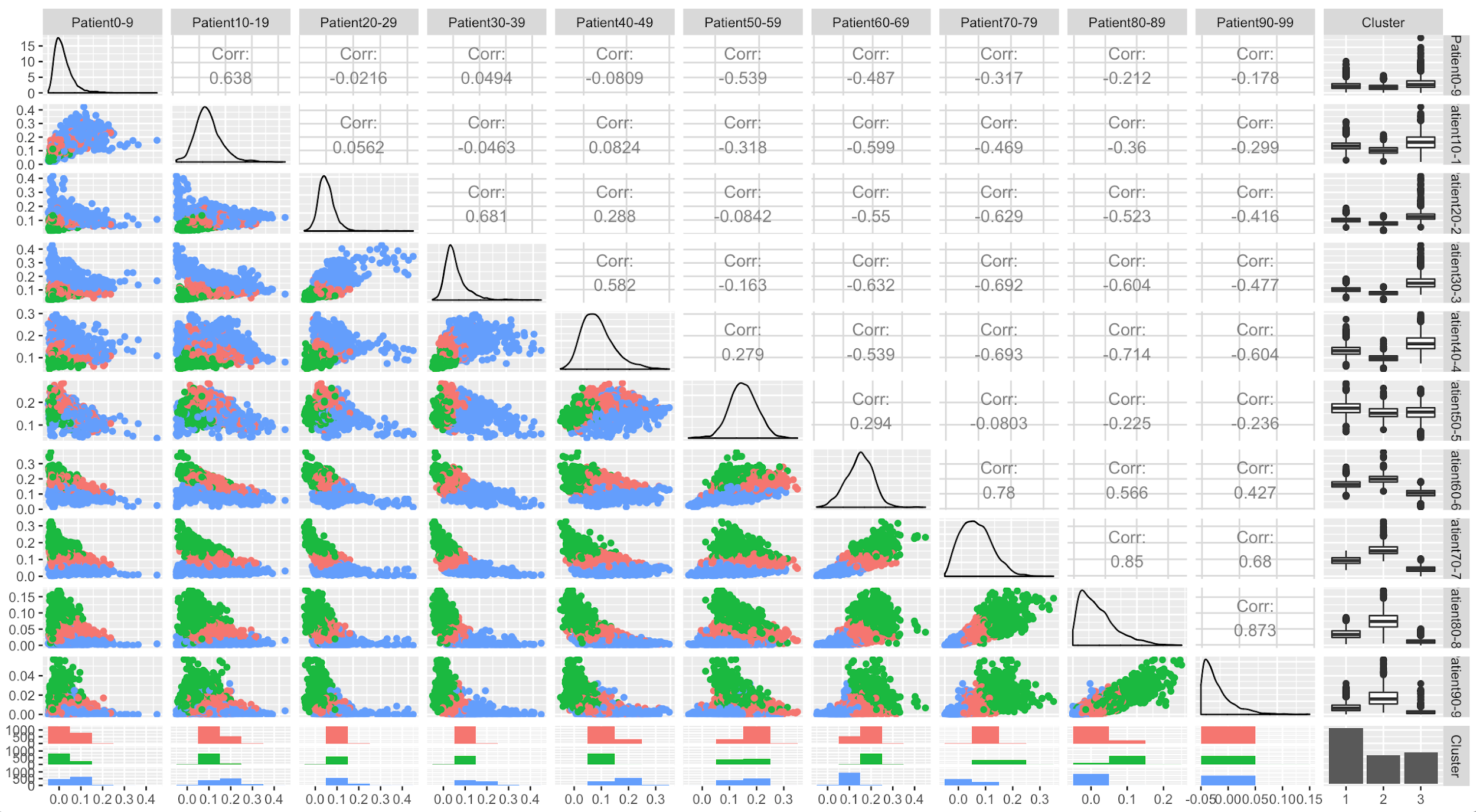


Scatter plot is then performed to visualize the distribution of 3 clusters:



From here, we have more insight of how 3 clusters distributed, but there is no clear pattern for each cluster. Therefore, we applied ggpairs() function to see the correlation of each variable toward each cluster.

**Correlation matrix:**



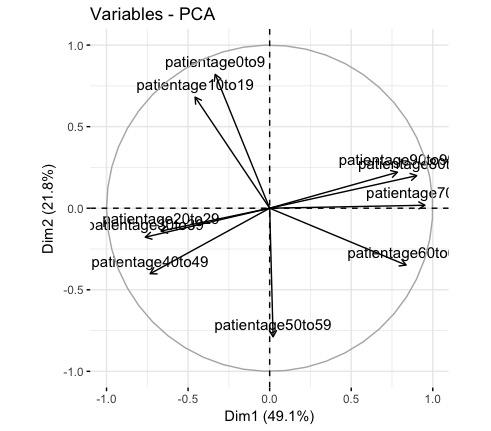
Focus on the cluster column, we see:

* Cluster 1: Balance - practices that do not have a skewed population of patients’ age.
* Cluster 2: Old Patients - practices that have a population of patients that are skewed toward younger ages.
* Cluster 3: Young patients - practices that have a population of patients that are skewed toward older ages.

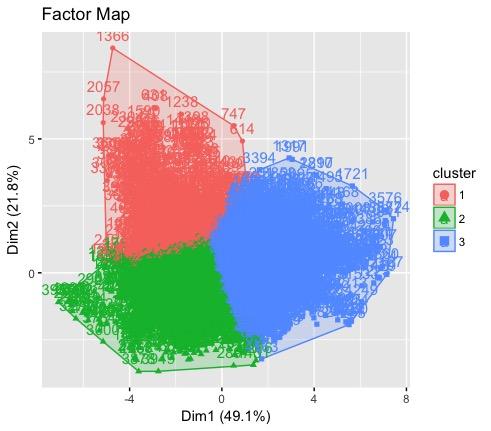
Limitation of the method is that we do not know which variable is associated with which cluster. Therefore, Hierarchical Clustering on Principal Components is applied for more insight.

**Hierarchical Clustering on Principal Components:**

Using PCA, we reduce the dimension of the data into a few continuous variables that contain the most information of the dataset. This step is for de-noising which can lead to more stable clustering.



The bi-plot shows the distribution of each variables, and how variables are grouped. Based on PCA dimension, we can perform cluster analysis and draw out the factor map of three clusters:



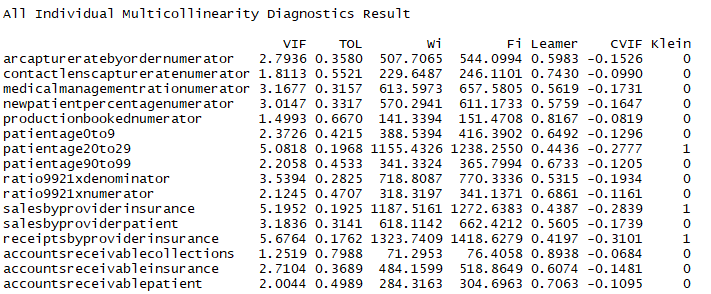
V.test or Kuiper’s test (fitted distribution of data) table:

|  |  |  |
| --- | --- | --- |
| Variables | V.test | Cluster |
| patientage0-9 | 44.11 | Cluster 1: associated with Young Aged Patients (0-19) |
| patientage10-19 | 43.03 |
| patientage20-29 | 26.48 | Cluster 2: associated with Balance Aged Patients (20-59) |
| patientage30-39 | 28.17 |
| patientage40-49 | 37.62 |
| patientage50-59 | 30.85 |
| patientage60-69 | 39.85 | Cluster 3: associated with Old Aged Patients (60-99) |
| patientage70-79 | 50.96 |
| patientage80-89 | 49.75 |
| patientage90-99 | 42.89 |

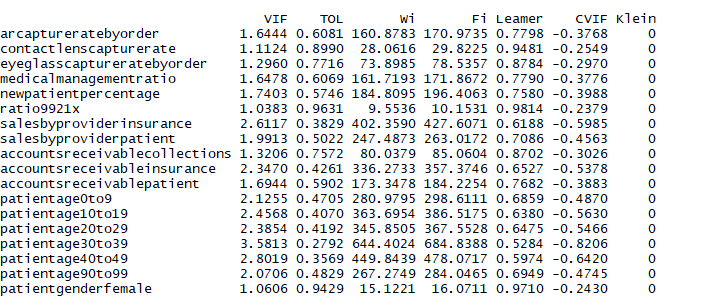
## **PCA and Common Factor Analysis**

The PCA and Factor analysis are performed on the Rate Performance subset to find the factors in the data which helps in interpretation between the Ages and the other features. The VIF is performed on the subset, and the variables with high VIF values (greater than 5) are removed.

**VIF of the Initial dataset**



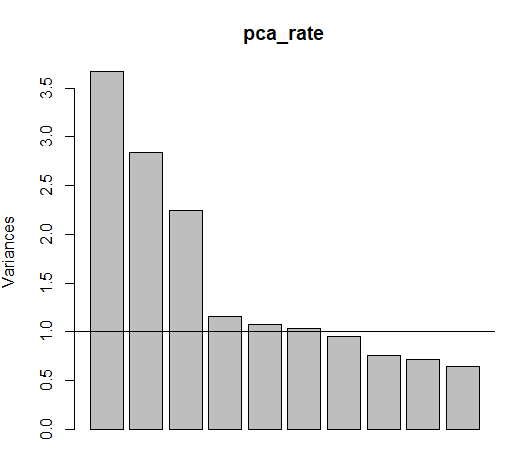
**Rate Performance Final list of variables with low VIF**

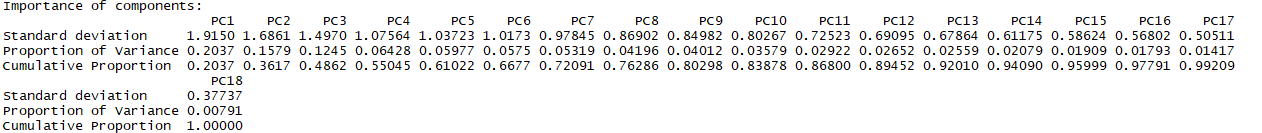


In the Rate Performance subset, variables that have VIFscores greater than 5 are excluded only after checking correlations with the other variables, so that the variance of the variables which are not highly correlated with other can’t be explained when excluded and eventually they are not removed.

The PCA and CFA assumptions are attained by performing Log transformation, including scaling to normalize the data. The P-values of **Bartlett test** is less than **0.05**, indicating the data is statistically significant for finding the factors. The data adequacy is tested by using **KMO** test and the MSA values for Rate Performance is **0.71**, which is greater than the required value 0.5.

**PCA and CFA on Rate Performance:**

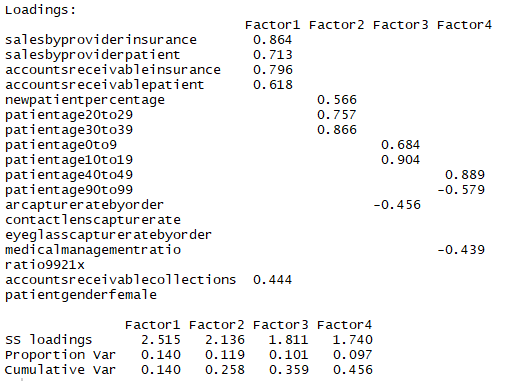
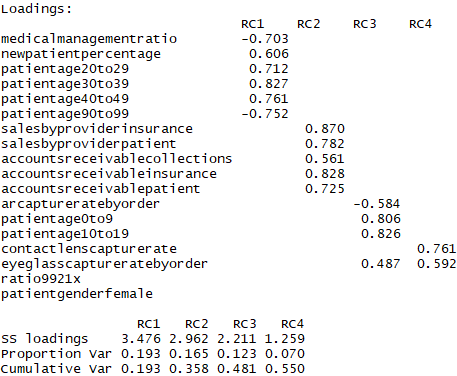
****

****

By using Scree plot and Kaiser Method on the Scree plot we can state that 3 and 6 components are required for the data. But 4 or 5 components are also valid and we can consider the results for all of the cases.

**Loadings with Best Factors**

**PCA CFA**

****

Based on the results of various number of factors, the interpretations are reasonable for 3 and 4 components. But the cumulative variance explained in 4 components (55%) is higher than 3 Components (48%), and also it is close to the 60% which is required for Latent Factor Detection.

**Naming the Factors:**

**Adults** = -0.7 (MedicalManagementratio) + 0.6 (newpatientpercentage) + 0.7 (patientage20to29) + 0.8 (patientage30to39) + 0.76 (patientage40to49) - 0.75 (patientage90to99)

**AR and Sales** = 0.87(salesbyproviderinsurance) + 0.78(salesbyproviderpatient) + 0.56 (accountsreceivablecollection) + 0.83 (accountsreceivableinsurance) + 0.73 (accountsreceivablepatient)

**Youth** = -0.6 (arcapturebyorder) + 0.8 (patientage0to9) + 0.8 (patientage10to19)

**Capturerates** = 0.76 (contactlenscapturerate) + 0.6 (eyeglasscapturertebyorder)

**Conclusion:** The noteworthy results from the factors are that there is clear separation between the ages in the factors as one consisting of younger ages and other have older ages. This is evident that, as in reality, the younger people are less prone to health issues than the adults.

Factor 1 (Adults) comprises of various age variables and the **newpatientpercentage**. As all age variables except **patientage90to99** doesn’t have inverse relation with Factor and from this we can conclude that the new patients are mostly adults below the age 90.

To conclude, even though the interpretations are reasonable in CFA, we believe there are influential outliers even after Data Cleaning. So it is good to have the factors which can explain all variance including error variance in the data. As a result, the Latent Factors are good for the interpretation instead of Common Factors.

## **Canonical Correlation Analysis**

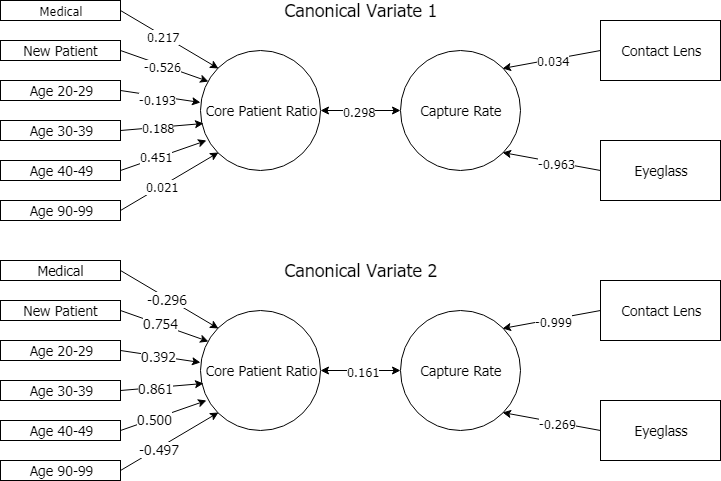
For this analysis we have chosen to focus on the Adults and Capture Rates components created in our principal component analysis. The Adults component accounts for the greatest variation in our data, and the Capture Rates component contains our dependent variable of interest.

|  |  |  |  |
| --- | --- | --- | --- |
| Independent Variables | * Medical Management Ratio * New Patient Ratio * Age 20-29 Ratio * Age 30-39 Ratio * Age 40-49 Ratio * Age 90-99 Ratio | Dependent Variables | * Contact Lens Capture Rate * Eyeglass Capture Rate |

The analysis resulted in two statistically significant canonical variates. These canonical variates do not show a strong correlation between the Adult and Capture Rate components.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Correlation | F-Test | p-value |
| Canonical Variate 1 | 0.298 | 43.3 | 0.000 |
| Canonical Variate 2 | 0.160 | 22.5 | 0.000 |

The individual loadings for the canonical variates can be shown in the following visualization.



## 

## **Lasso Regression**

Using the log transformed Count Performance subset data for predicting eyeglasses capture rate, after removing missing values, we have 4298 observations in this subset. A total of 23 predictors include patient demographics, clinical details, and revenue details. The reason for predicting eye glasses capture rate is because it’s directly related to practices financial health. Most of the profits are generated through glasses sales. Also, glasses is more significant concerning patients’ age groups, based on our previous analyses.

The data is split into 80% and 20% for training and testing, cross validation is performed with lasso, as well as k-fold method. Since the dataset is on count variables, “Poisson” family is used.

Below is the plot of an initial fit, known as the coefficients path:

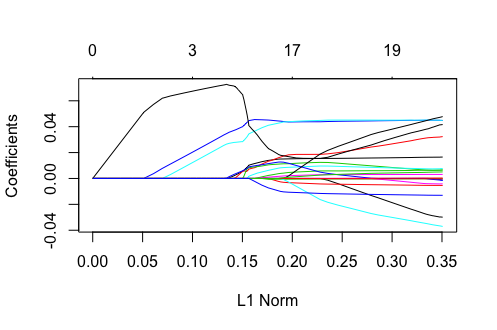


   Figure 1

As shown in the above plot (figure 1), each curve corresponds to a variable. It shows the path of its coefficient against the ℓ1-norm of the whole coefficient vector as λ varies. The axis above indicates the number of nonzero coefficients at the current λ, which is the effective degrees of freedom (df) for the lasso.

The function glmnet returns a sequence of models for us to choose from, and cross-validation chooses the best one for us, with the help of minimum lambda and lambda with 1 standard error. With cv.glmnet function and plotting, it returns the plot below:

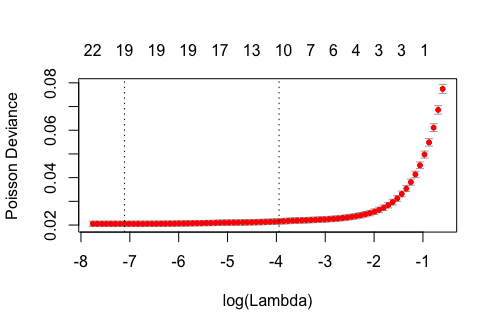


Figure 2

The figure 2 above is the cross-validation curve. The upper and lower standard deviation curves along the lambda sequence (error bars). The two vertical lines are lambda minimum (with minimum mean cross-validation error and lambda 1 standard error (with most regularized model - error is within 1 standard error of the minimum), looks like LASSO is selecting 10 variables.

Selected variables are shown with the coefficients calling function. The final equation is as follows:

*LogEyeGlassesCapptureRate = 1.204377e-02 \* (NewPatientNumber) + 4.543061e-02 \*(Age10-19) + 3.750349e-02 \*(Age 20-29) + 9.244941e-03\*(Age 50-59) + 9.168723e-03 \*(Age 60-69) + 4.149451e-03 \*(Age 80-89) + 1.273625e-05 \*(Age 90-99) + 3.355240e-02 \*(Female) + 1.008777e-02 \*(SalesByInsurance) -3.335736e-03 \* (AccountsReceiviblePatient)*

All the parameter coefficients are small, which is the exact purchase of LASSO. Here we get an inverse relationship: Accounts Receivable Patient is inversely related with eye glasses capture rate, so is practices financial health. This aligns with our group’s domain knowledge, as well as other research findings: practices don’t want to leave accounts unpaid because it is more likely that they are never going to be paid.

With RPtest package, p-value of 0.02 for the LASSO regression was obtained, which shows the statistical significance. A validation was performed on the 20% testing set with the trained model. The metric used to evaluate the model performance is Root Mean Square Error (RMSE). The Model performs well with low RMSE scores for both Training and Generalization error. The difference between the Training and Generalization error is less than 10%, which shows that the model has low bias and low variance.

## **Principal Component Regression**

Although LASSO is a great technique for avoiding model overfitting with “penalty” functions, it eliminate some features in the model regardless of any measure of significance. Sometimes those can be necessary control variables. In order to take more variables into account, while concerning multi-collinearity at the same time, principal component regression was conducted on the same subset to predicting eye glasses capture rate.

Firstly, with the PCR package, PCA is automatically conducted in a regression. The initial step for this would be determining number of components. We first tried 4 components, based on the PCA we’ve already done, the Root Mean Square Error of Prediction (RMSEP) by components plot (figure 3) was obtained:

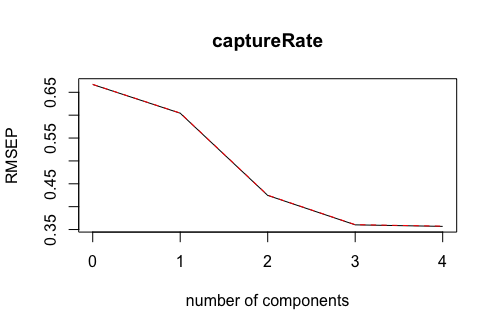


Figure 3

Looks like using 3 components or 4 components would generate about the same prediction error.  We went ahead with 3 components, and the R-squared plot (figure 4) was returned with more than 60% of the variation captured:

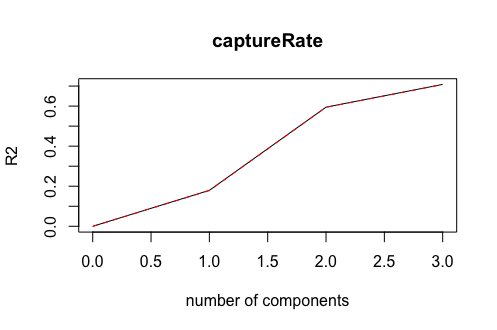


Figure 4

Below are the coefficient plot (figure 5) and prediction error plot (figure 6):

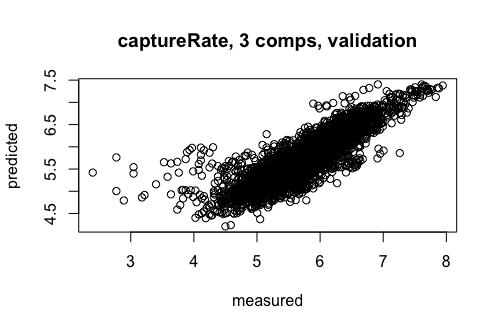
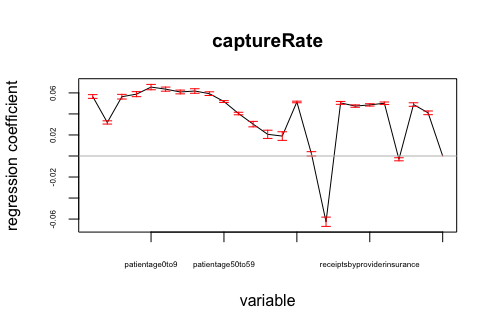
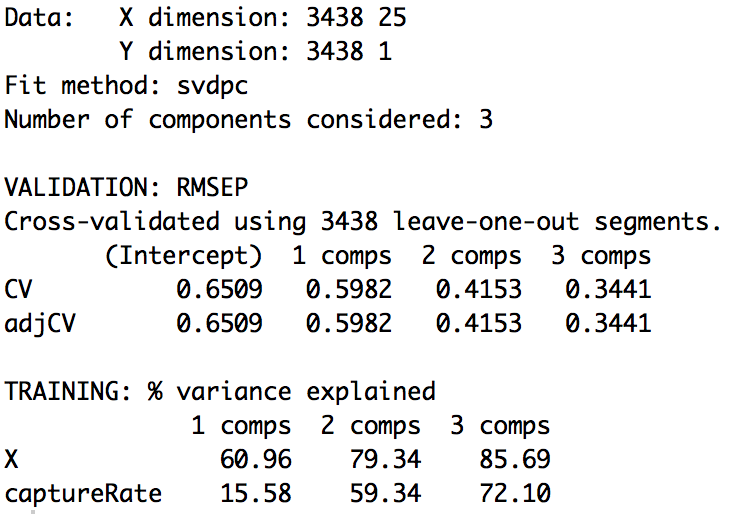
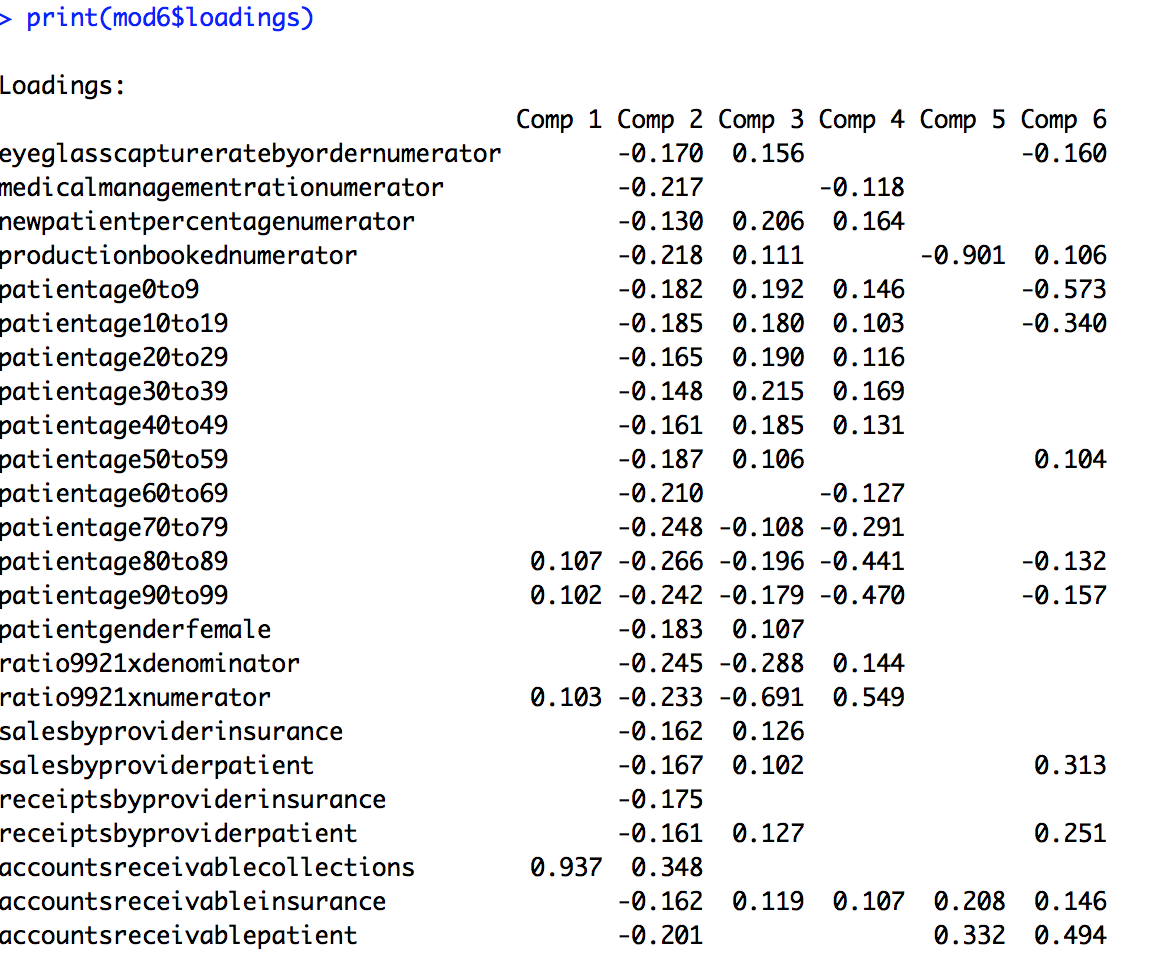


Figure 5 Figure 6

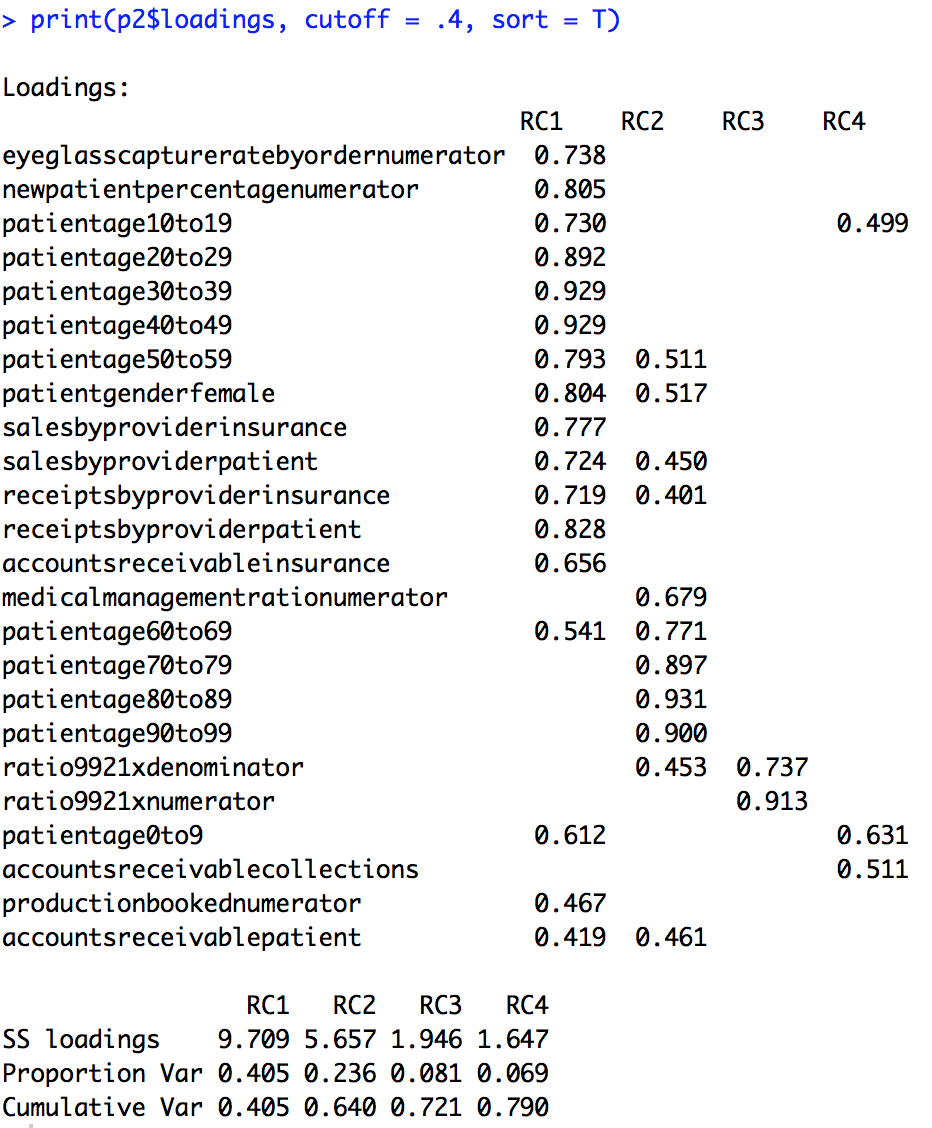
Cross-validation was also performed using leave-one-out segments, below is the summary:



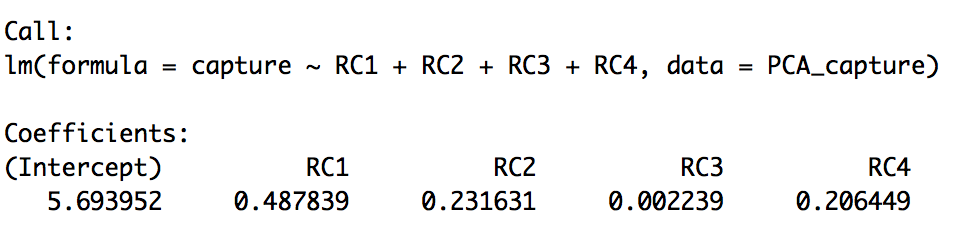
It seems that the regression using 3 components is pretty accurate. However, when we tried to look at what items are in each component in order to compare with the results of our PCA, the loadings returned were hard to interpret, even when we tried different cutoff points. We then decided to look at the loadings for 6 components. With almost the same prediction error and R-squared, the loadings for 6 components looked less confusing:



We took a step further to conduct a PCA on this subset and then manually use the loadings of number of 4 components in the regression based on our scree plot and the findings from our previous analysis, below is the loadings of the PCA to be used in our linear model:



We took the scores of each component as variables for the linear model, the model fit coefficient and diagnostics are as following:



\*\*\*The diagnostic plots are in the Appendix

# V. Conclusion

## **Limitations**

There are several significant limitations in this analysis, much of which is imposed by the nature of the data source: redaction, aggregation, and significant outliers.

The removal of practice identifiers and reporting time periods has restricted our ability to analyze any practice trends over time: we cannot determine if practice performance is improving over time; we cannot determine if there are any seasonal trends that affect the practice population.

The aggregation of the data shifts the analysis away from a traditional healthcare data source to more of a business data source.

Properly handling outliers was a significant difficulty as we initially worked with the data source. Instead of developing selection criteria before sampling the practices, we had to infer the appropriate selection criteria based on the individual observations. This is made especially difficult due to the differing levels of EHR adoption in the practices. A significant assumption in our initial data analysis was that if a practice had significant levels of missing data, they were not fully utilizing the EHR and could be excluded from our analysis.

## **Future Work**

Possible future work would include some of the currently unavailable practice details in the data source. These would allow for the practices themselves to be differentiated more clearly. RevolutionEHR provides their services to customers large and small. One practice might be a single provider operating out of a single location to a practice group with dozens of locations and providers. Details that could be included would be number of locations, location addresses, number of providers, number of employees, etc. Including these details would also be an opportunity to develop and implement a sampling criteria that would hopefully eliminate many of the troublesome outliers that are present in the current data source.

RevolutionEHR has been certified for participation in the Medicare Electronic Health Records Incentive program (American Optometric Association 2011). This program, now known as MIPS, requires participating providers to attest their compliance by submitting clinical quality measures (CQM) summary data on their patient population. This data is composed of incident and outcome counts that have been separated from the individual patient details, and can thus be analyzed without working directly with patient protected health information. Including these details would allow researchers to look at the relationships between patient population health and practice financial health directly.

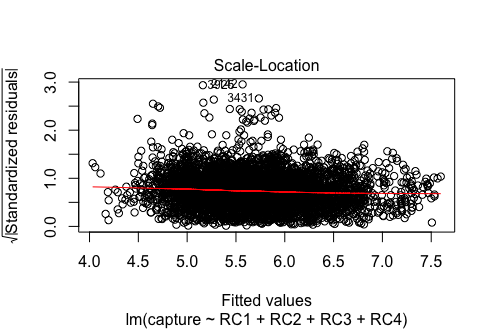
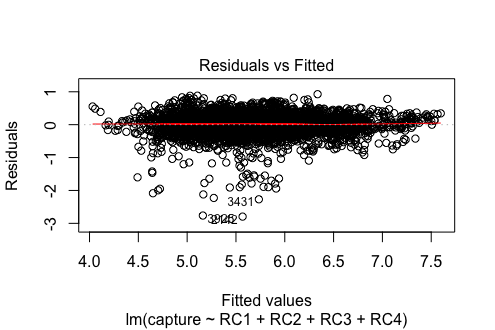
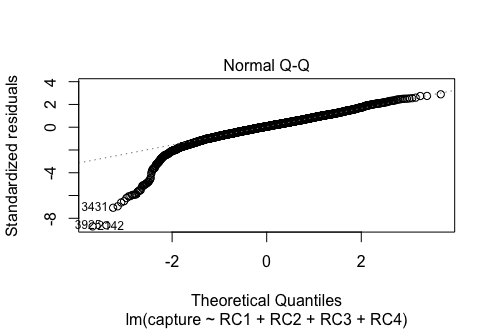
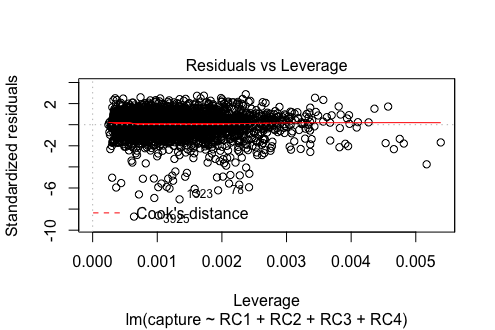
## **Overall conclusion**

Combining the findings of cluster analysis, correspondence analysis and lasso regression, we come to realization that our project needs a focus on age variables. Based on the results in cluster analysis, these practices have a population of patients that are skewed towards being 50 years old or older. Moreover, correspondence analysis found that younger (0-29) and older (60-89) individuals don’t prescribe to and pay for eyeglasses as often as middle-aged (30-59) individuals. Additionally, it is even more rare for individuals older than 90 to visit eye doctors. We conclude that focusing on middle-aged patients could prove more beneficial to this study.

On the other hand, the three very important variables have been selected for predicting capture rate: medical management ratio, new patient percentage and production booked, which aligns with our assumptions. Future eye care practices should strategize based on these results to obtain and maintain high spectacles capture rate.

# VI. Appendix

**Visualizations - Multiple Regression Diagnostic Plots using PCA**



**R Codes** - submitted separately in PDF and RMD files

# **References**

U.S. Department of Health and Human Services (2017). *Health, United States, 2016: With Chartbook on Long-term Trends in Health*. Hyattsville, MD. 2017

American Optometric Association. (2011). Practical hints on meaningful use. *Optometry-Journal of the American Optometric Association*, *82*(12), 757-765.

Gerber, G. (2013). Essential elements of tracking practice performance: these key metrics can reveal areas of concern--and opportunity--to assist in strategic planning. *Review of Optometry*, *150*(10), 74-82.

Miller, R. H., West, C., Brown, T. M., Sim, I., & Ganchoff, C. (2005). The value of electronic health records in solo or small group practices. *Health Affairs*, *24*(5), 1127-1137.

Soroka, M., Krumholz, D., & Bennett, A. (2006). The practice of optometry: National Board of Examiners in Optometry survey of optometric patients. *Optometry-Journal of the American Optometric Association*, *77*(9), 427-437.

# 