PREDICTING FAILURE TO DIAGNOSE

ANALYSIS OF MALPRACTICE PAYMENT DATA USING A MACHINE LEARNING APPROACH

PROJECT GITHUB REPO: HTTPS://GITHUB.COM/TAIHUALI/NPDB_DATA

Trish Lugtu
M.S. in Predictive Analytics

Michael Marre
M.S. in Information System

Taihua Li
M.S. in Predictive Analytics

Zengyi Zhu
M.S. in Information System

CSC 424, SPRING 2016, DEPAUL UNIVERSITY

TABLE OF CONTENTS

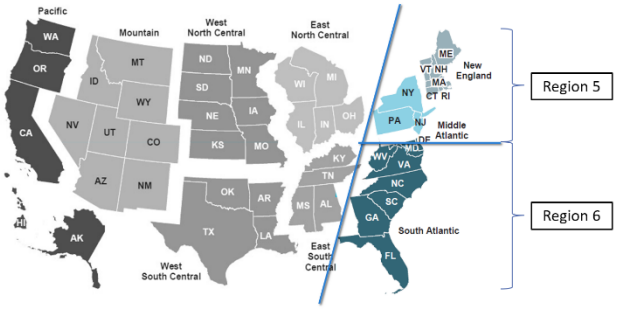
TABLE OF FIGURES

SECTION 1. INTRODUCTION (NON-TECHNICAL SUMMARY)

This project involved multiple attempts to create a meaningful, predictive model using medical malpractice payment records from the National Practitioner Databank (NPDB) Public Use File downloaded from www.npdb.hrsa.gov (accessed 1/31/2016). Our team was composed of two MSPA students - Taihua Li and Trish Lugtu; and two MSIS students - Michael Marre and Zengyi Zhu. Together, we brought a nice balance of different skillsets to the team.

"The NPDB Public Use Data File contains selected variables from medical malpractice payment and adverse licensure, clinical privileges, professional society membership, and Drug Enforcement Administration (DEA) reports (adverse actions) received by the NPDB concerning physicians, dentists, and other licensed health care practitioners. It also includes reports of Medicare and Medicaid exclusion actions taken by the Department of HHS Office of Inspector General." – npdb.hrsa.gov

After doing a thorough exploration of the 54 attributes in the data file, we decided to focus our analysis on a subset of medical malpractice payment reports collected after 2004 for physicians with diagnosis-related cases. We then further segmented this data to perform a comparison between two regions of the U.S. - the Northeast (Region 5) and Southeast (Region 6). We chose these regions because they included several states with the highest volumes of malpractice claims in the country, which ensured ample data for each subteam.



Map source: 2014 Physician Census. Accessed online: www.fsmb.org, 5/11/2016

Finally, our team split into two subteams to tackle analyses in parallel with each other. Each subteam included a MSPA-MSIS pairing - Taihua and Zengyi formed the Region 5 Subteam, and Trish and Mike formed the Region 6 Subteam. After two attempts with different dependent variables and a series of analyses, each subteam reached similar models with significant and meaningful results.

Our final three approaches (for grading) include the logistic regression, k-means clustering, and factor analysis with mixed data (FAMD).

- The logistic regression for each region produced a 92-93% fit predicting whether or not a diagnosis-related malpractice case was due to the physician's failure to diagnose, as opposed to the other reasons including delay in diagnosis, misdiagnosis, failure to order appropriate test, and radiology or imaging error.
- The k-means cluster analysis uncovered subtle differences between the two regions which could be practically explained through. Such differences involved differences in age and experience of physicians, different settings (i.e. inpatient vs. outpatient), and on different types of patients (i.e. age and gender).
- Finally, the FAMD confirmed as did PCA that malpractice payment data does not have hidden patterns.

SECTION 2. TECHNICAL REPORT

Note: All graphs and plots are included in an Appendix A.

EXPLORATORY DATA ANALYSIS

About the Data

Our final subsets for both region 5 (n=12,540) and region 6 (n=5,926) contain 19 variables (5 numerical, 14 variables). The data falls into the following groups of variables:

- *Physician Data:* Physician Age (MDage), Physician Experience (MDexp): malyear1 - grad
- *Patient Data:* Patient Age (ptage), Male Patients (ptMale), Female Patients (ptFemale); Inpatient (inpatient), Outpatient (outpatient)
- Malpractice Data:
  - Total Payment (totalPayment)
  - Major Allegations: Failure to Diagnose (a1failToDx), Delay in Diagnosis (a1delayInDx), Wrong/Misdiagnosis (a1wrongDx), Failure to Order Appropriate Test (a1failToOrder), Radiology or Imaging Error (a1radError)
  - Minor Allegations: Failure to Treat (a2failToTx), Delay in Treatment (a2delayInTx), Delay in Diagnosis (a2delayInDx), Failure to Diagnose (a2failToDx), Failure to Order Appropriate Test (a2failToTest)
  - Outcome (outcome): low injury severity (1) to high injury severity (9)

Our variable of interest in the first attempt was Total Payment (totalPayment), a numerical variable. When no meaningful models were discoverable, we shifted to the binary dependent variable, Failure to Diagnose (a1failToDx).

Data Preprocessing

When we started to dive into the data, we found that the supposed numerical variables were actually ordinal factor variables. For example, practitioner age with a value of 20 represented factor level “Ages 20 through 29.” Ideally, we wanted the physician and patient ages to be numerical, so we discussed whether treating them as numerical was viable. Ultimately, we decided to substitute their values with the bin midpoint and treat as numerical. We also created a new calculated variable representing physician experience by subtracting physician graduation year from malpractice event year. We needed to accommodate physician age, physician experience, patient age, and total payment in this way (Figure 4. Exploratory graphs – scatterplot & histograms – (totalPayment, MDexp)Figure 4).

Another major transformation required was creating dummy variables of the categorical data(Figure 4. Exploratory graphs – scatterplot & histograms – (totalPayment, MDexp)Figure 4). This was especially important for the allegation variables because the allegation1 and allegation2 factors have 91 levels each. We also needed to decide when values with NA needed to be omitted or when they should be set to 0. We relied on Trish’s understanding of what the variables meant to do this.

Exploratory Analysis

The National Practitioner Databank (NPDB) data includes 54 variables and over 1.2 million cases, which renders it unwieldy to process in its entirety. In order to focus the subset, we did an extensive exploratory analysis. We studied the data definitions file and discovered four major groupings of data in the file - Adverse Action Reports (< 2004), Adverse Action Reports (> 2004), Malpractice Payments (< 2004), and Malpractice Payments (> 2004) (Figure 1). Each major grouping involved different variables and distributions of attributes. Since the last group was the most current and involved more quantitative variables, we focused the next analysis on the remaining 167,027 cases available.

```
                                                    rectype
Adverse Action Report (Legacy Format)                     : 51301
Consolidated Adverse Action Report, 11/22/1999 and later:729253
Judgment or Conviction Report, 11/22/1999 and later       :     0
Malpractice Payment Report, 9/1/90 to 1/31/04             :250685
Malpractice Payment Report, 1/31/04 and later             :167027
```

Figure 1. NPDB Report Types (rectype)

Next we explored the different variables by looking at the shape of data distributions through histograms and plots, looked at correlations between variables through scatterplots and correlation plots, and examined other variables through descriptive statistics. We found that if we isolated the physician provider types (i.e. MD, MD residents, DO, DO residents), we still had 129,533 cases with which to work (Figure 2). We also had too many variables, so we looked to further subset the data.

```
                          licnfeld
Allopathic Physician (MD) :119501
Dentist                   : 18172
Osteopathic Physician (DO):  9092
Registered Nurse          :  4201
Podiatrist                :  3094
Chiropractor              :  2203
(Other)                   : 10764
```

Figure 2. NPDB Licensed Field (licnfeld)

To limit our dataset further, we considered focusing the subset on a specific allegation nature. Knowing that treatment-, diagnosis-, and surgical-related allegations compete for the most occurrence in claims, we weighed each choice, finally choosing diagnosis-related cases (Figure 3). Surgical-related claims would be less inclusive of the general physician population (only surgeons), and diagnosis-related cases are all-inclusive of the physician population. Diagnosis-related cases also naturally limit the levels of allegation1 and allegation2 factors. However, we still had too many allegation1 dummy variables, and so we decided to limit our subset to the top five levels of allegation1, which captured 90% of the cases.

```
                    algnnatr
Diagnosis Related :41268
Surgery Related   :34551
Treatment Related :25214
Obstetrics Related: 9952
Medication Related: 6709
Monitoring Related: 3929
(Other)           : 7930
```

Figure 3. NPDB Allegation Nature (algnnatr)

METHODOLOGIES (FIRST ATTEMPT)

Multivariate Regression

Preliminary analysis indicated that multivariate regression might be an acceptable method to predict total payments. There is a decent amount of correlation between total payments and other variables in the dataset, consequently a linear model like an ideal starting point for our analysis of the dataset.

Region 5

Our initial domain knowledge suggested to conduct a multivariate regression analysis between totalpayment and other dependent variables. We first started with a full model with all variables and performed the regression to find model performance. The resulting R-squared value was extremely low (close to 0.04), which indicated a statistically insignificant result.

Next, we ran the forward selection function in R in order to reduce the model from full variables to a mode with minimum variables and maximum AIC scores. The results indicates that a1failToOrder, a2failToTx, a2delayInTx, MDexp, a1wrongdx, inpatient, a2delayInDx, and a1radError are significant in the dimensionally-reduced model. However, analysis of this reduced model showed a decreasing R-squared value of 0.00049, which indicated the model selection technique was inappropriate. We speculate that this occurred because we already reduced variables through Trish's domain knowledge. Further feature selection didn't help to improve overall model fitness, but rather decreased the overall variation of the sample, which may in turn reduced the R-squared value (Figure 7).

Further analyzing the data using logistic regression based on viewing a scatterplot between variables (Figure 8), the R-squared value increased to 0.0769 indicating a slightly better model, but still inadequate to claim a linear model between totalpayment and other variables.

Region 6

For the first attempt at producing a model all of the variables were used and no transformation was done to the variables. The initial model produced an extremely low R-Square of .0502, far from what would be considered a statistically significant model.

For the next attempt log transformation was used on the dependent variables, with the hopes finding a better linear relationship between the variables. This model slightly increased the R-Squared to .06, still far being an adequate predictor of Total Payments.

Finally we explored how model selection would impact the model. Using stepwise regression, the model was selected based on the lowest AIC score. While using model selection produced the best model, with an R-Square of .086, the model could not be considered an accurate predictor total payment (Figure 9).

After trying several different models that failed to produce adequate results, it's evident that multivariate regression is not an appropriate method for our analysis. It is interesting to note that all of the models had significantly F-statistics and extremely small P-values, mostly a results of the large number of samples associated with the dataset.

Principal Component Analysis (PCA)

Since the dataset we were working with had over twelve thousand instances, we thought the principal component analysis (PCA) would be applicable for both its purposes - dimensionality reduction and factor analysis to identify hidden patterns in the data.

Regions 5 & 6

The correlation plots for region 5 and 6 are nearly identical (Figure 5, Figure 6) and show that there are few strong correlations within the data. The strong correlations that do exist are expected. For example, physician age (MDage) should increase with physician experience (MDexp); inpatient and outpatient, although not exclusively one or the other, a patient is usually one type in the majority of cases; and failure to diagnose (a1failToDx) negatively correlates with a delay in diagnosis (a1delayInDx) simply because a delayed diagnosis isn't a failed diagnosis. Rather, it is just a late or untimely diagnosis.

The loadings for region 5 (Figure 11) show the first seven components, which capture at least 95% of variance (Figure 10). And the rotation for region 6 (Figure 13) shows four components which capture 93.6% of variance (Figure 12). As shown above, the Principal Component Analysis does not produce a meaningful result.

FINAL METHODOLOGIES

Based on our initial results, Taihua suggested moving to a binary model and Trish suggested the new dependent variable, major allegation type of failure to diagnose (a1failtoDx). We shifted our approaches to k-means clustering, logistic regression, and factor analysis with mixed data.

Approach One: K-Means Clustering

Since the Principal Component Analysis did not produce any meaningful result, we decided to apply the K-means clustering algorithm with the hope to group and discover latent patterns together that the Principal Component Analysis was not able to achieve. The normalization method applied to the data is min-max normalization, which scales each features into the interval between 0 and 1.

Region 5

As shown in Figure 15, the scree plot of within-cluster sum of squared of errors on different values of K indicated that 4 would be an ideal option for the K-means clustering. Therefore we applied the K-means algorithm with k equals to 4 and using euclidean distance as the similarity metric and the cluster centers are shown in Figure 14.

With Trish's domain knowledge, we interpreted the four clusters as the following:

- Young physicians with older male inpatients failed/delayed to treat
- Older physicians with female outpatients delayed/falsely diagnosed and made radiologic errors (e.g. misread a mammogram in the case of breast cancer)
- Average aged physicians with younger female outpatients failed to diagnose and therefore failed to treat (secondary allegation)
- Younger than average physicians with younger male outpatients failed to order proper test with minor injuries as outcomes

Region 6

As in Figure 17, the scree plot of within-cluster sum of squared of errors on different values of K indicated that 4 would be an ideal option for the K-mean clustering for region 6 as well. Therefore, we applied the K-means algorithm with the same parameters as those for region 5 and identified the following clusters,

- Young physicians with younger female outpatients failed to diagnose and treat
- Older physicians with older male outpatients failed to diagnose
- Average age physicians with younger male outpatients failed to order test, which led to failure to diagnosis
- Younger than average physicians with older female inpatients delayed to diagnose, which led to delay in treatment

The cluster definitions between two regions overlapped among average age, young and older physicians. It showed that in the east coast, if there is a medical malpractice or adverse action reported, cases involved with young physicians generally are related to failure in diagnosis and treatment. Cases involved with older and average age physicians are generally related to failure in diagnosis. Even though the clustering results seem interesting, we have not yet found practical explanations for these phenomena.

Approach Two: Logistic Regression

After our first attempt using the multivariate regression technique was failure, we selected variable "if the allegation one is failure to diagnose" as our new response variable to predict if there is a logarithmic relationship between aifailToDx ("if the allegation one is failure to diagnose") and certain independent variables.

Region 5

We continued to use the same region subsets, but for better model searching, we removed the variable "ptmale" and split the data into 80% training and 20% testing for cross validation. Next we used R to develop the logistic model for the response variable a1fialToDx and used a 10 fold validation technique to validate the model further. The results for the area under the ROC is 0.9269 (Figure 18), which showed a statistically significant model for predicting a1failToDx.

We then tested the new logistic model in the holdout data to predict a1failToDx. The AUC scores and plot are shown in Figure 19 and Figure 20. From the plot, The model showed an 0.9312965 on AUC score for the testing dataset, which could be an strong evidence providing that there is a logarithmic relationship between a1failToDx and outcomescale, inpatient, a2failToTx, a2delayInTx, a2failInDx.

Region 6

The region 6 subset was further subsetted into a randomly select 75% training group and 25% testing group. For this analysis we chose the response variable Failure to Diagnose, which is binary and seemed like a strong candidate for logistic regression. Out of the 15 variables used the output indicates that patientAge, inpatient, outpatient, and a2failToOrder contribute significantly to predicting failure to diagnosis (Figure 22). The model provides some interesting insight into the data. For example, the model indicates that patientAge and Failure to Diagnose have a negative correlation. Serious conditions are more often misidentified in younger patients resulting, these failure to diagnosis can results in large malpractice payments.

After we created the model in R, we determined the area under the curve (AUC) and the ROC. The results were very positive, for the training data and testing data. The training dataset produced an AUC score of .91 Plot (Figure 21) and the testing dataset .90 (Figure 23). The AUC score is close to 1, indicating the classifier is very good. Additionally, the training and the test datasets are very similar meaning that the data is not over- or underfitted.

Approach Three: Factor Analysis of Mixed Data (FAMD)

Our Principal Component Analysis did not produce meaningful result because most of the variables we have were categorical variables and the Principal Component Analysis is only suitable for numerical variables. Therefore, we applied Factor Analysis of Mixed Data (FAMD), which is a combined method of the Principal

Component Analysis and the Multiple Correspondence Analysis. Multiple Correspondence Analysis (MCA) is an extension of Correspondence Analysis, which is similar to the Principal Component Analysis except that instead of calculating eigenvalues and eigenvectors on the covariance or correlation matrices, MCA creates new dimensions, which is the same as principal components, based on the indicator matrix of categorical variables. However, MCA does not compromise the usage of numerical variables unless they are binned into intervals.

Therefore, with the goal to preserving numerical variables' information, we decide to apply the Factor Analysis of Mixed Data. Below is the result we have from the FAMD algorithm, which is available in the R library, FactoMineR,

Region 5

The graphs in Figure 24 include a biplot of all categorical variables projected on the first and second dimensions. The second graph is the biplot of all numerical variables projected on the first and second dimensions.

As shown on each graph's axes labels, the first dimension, which captures the greatest variance in the dataset, explains only 8.32% of the original dataset's variance while the second dimension explains 7.54%. This indicates that our dataset does not have significant explainable variance, or relationships of variables.

Region 6

Likewise for region 6, the significant correlations are similar as region 5 above and to the PCA results - MDage, MDexp, inpatient, outpatient, ptFemale, and ptMale.

As with Region 5, we can see in Figure 25 that the first dimension only captures 8.06% of the original dataset's variance and the second dimension captures 7.56%. Even though it shows that region 5 and region 6's analyses results are statistically consistent, it does not produce any meaningful result for us to interpret.

After applying the Factor Analysis of Mixed Data, even though we found that physicians' age and experience tend to have a moderately positive relationship with patient gender and the patient type, it does not explain any both statistically significant and practical result, or in another word, it does not indicate that there is a latent factor in the dataset. With that being said, the result of FAMD is statistically and strongly consistent with the result of PCA.

CONCLUSION

In conclusion, we discovered that total payments are difficult to predict. We speculate that this is probably due to the differing torte reform per state and nature of lawsuits, but have not delved into its causality at any depth. However, focusing on a more specific question such as whether or not a diagnosis-related allegation is due to failure to diagnose by a physician is more feasible. The logistic regression produced the best model with significant and meaningful results.

APPENDIX A. ANALYSIS-RELATED GRAPHS

Exploratory Analysis Plots

Exploratory Graphs

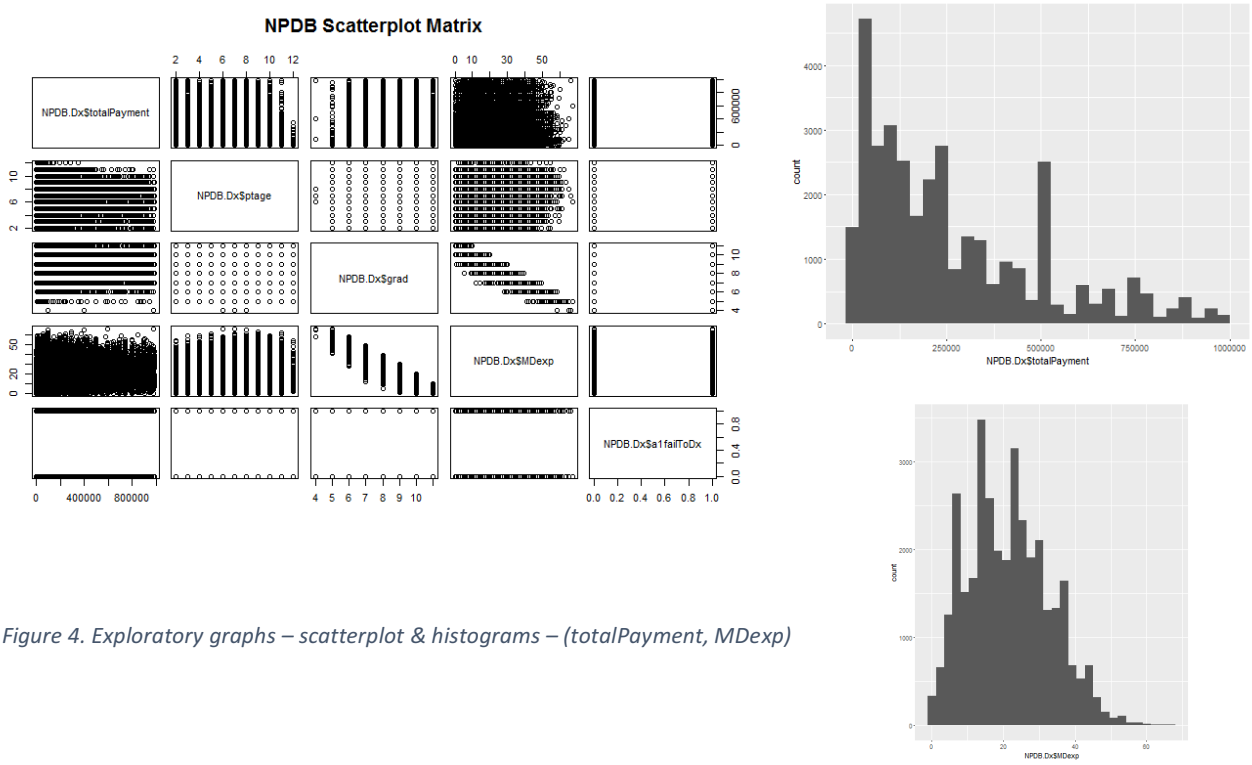


Figure 4. Exploratory graphs – scatterplot & histograms – (totalPayment, MDexp)

Correlation Plots

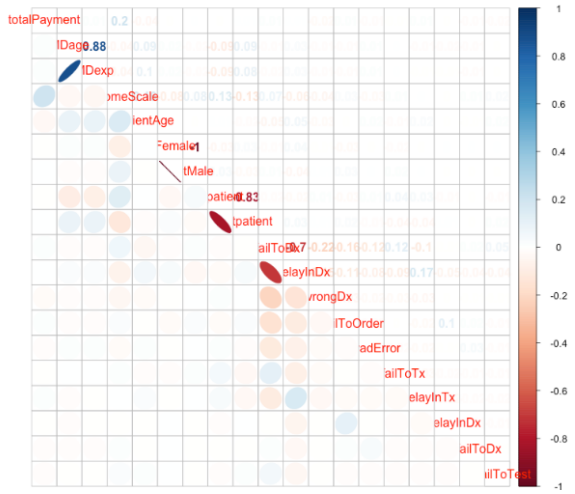


Figure 5. Correlation Plot for Region 5

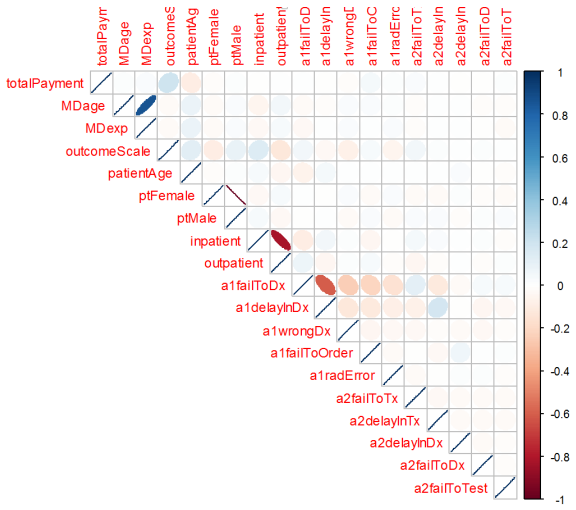


Figure 6. Correlation Plot for Region 6

Multivariate Regression

Region 5

```
Call:
lm(formula = logtotalPayment ~ ., data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4476 -0.5393  0.2103  0.6986  2.0849

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.2386052  0.0813994 138.067  < 2e-16 ***
MDage         0.0023854  0.0017004   1.403 0.160684
MDexp        -0.0003822  0.0016520  -0.231 0.817025
outcomeScale  0.1439985  0.0045429  31.697  < 2e-16 ***
patientAge   -0.0035131  0.0004557  -7.709 1.37e-14 ***
ptFemale      0.0291554  0.0177854   1.639 0.101178
ptMale               NA         NA      NA       NA
inpatient     0.0131726  0.0333720   0.395 0.693056
outpatient    0.0877090  0.0358541   2.446 0.014448 *
a1failToDx    0.0870536  0.0325677   2.673 0.007527 **
a1delayInDx   0.1252475  0.0340779   3.675 0.000239 ***
a1wrongDx    -0.0052352  0.0519179  -0.101 0.919682
a1failToOrder 0.1640710  0.0653720   2.510 0.012092 *
a1radError   -0.0074174  0.0838382  -0.088 0.929502
a2failToTx   -0.1323231  0.0629148  -2.103 0.035467 *
a2delayInTx   0.0486534  0.0519087   0.937 0.348629
a2delayInDx  -0.0121157  0.0740812  -0.164 0.870091
a2failToDx   -0.2337096  0.1037146  -2.253 0.024252 *
a2failToTest -0.1291033  0.0941385  -1.371 0.170269
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 0.9892 on 12522 degrees of freedom
Multiple R-squared:  0.07815,   Adjusted R-squared:  0.0769
F-statistic: 62.44 on 17 and 12522 DF,  p-value: < 2.2e-16
```

Figure 7. Region 5 Regression Output

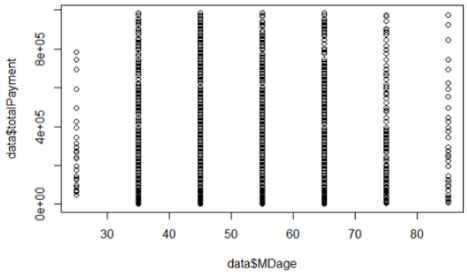


Figure 8. Scatterplot totalPaymet x MDage

Region 6

	Dependent variable:
	log.totalPayments
MDexp	0.004***
	(0.001)
outcomeScale	0.162***
	(0.007)
patientAge	-0.005***
	(0.001)
inpatient	-0.062**
	(0.029)
a1delayInDx	0.024
	(0.034)
a1failToOrder	0.131*
	(0.076)
a1radError	0.142
	(0.101)
Constant	10.932***
	(0.064)
Observations	5,926
R^2	0.087
Adjusted R^2	0.086
Residual Std. Error	1.102 (df = 5918)
F Statistic	80.850*** (df = 7; 5918)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 9. Region 6 Regression Output

Principal Component Analysis

Region 5

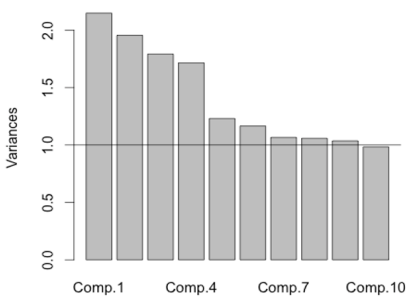


Figure 10. Region 5 PCA Scree Plot

	PC2	PC1	PC4	PC3	PC8	PC6	PC7
ptFemale	−0.999						
ptMale	0.999						
MDage		0.967					
MDexp		0.967					
inpatient			−0.953				
outpatient			0.954				
a1failToDx				−0.911			
a1delayInDx				0.927			
a1wrongDx					0.993		
a1failToOrder						0.994	
a1radError							0.997
a2delayInDx							
totalPayment							
a2failToDx							
a2failToTest							
a2failToTx							
a2delayInTx							
patientAge							
outcomeScale							

Figure 11. Region 5 PCA Loadings (princomp)

Region 6

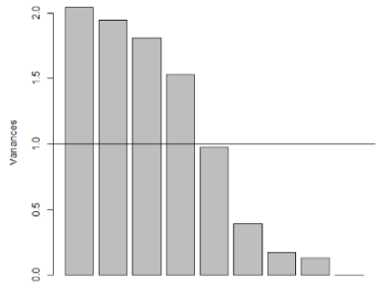


Figure 12. Region 6 PCA Scree Plot

Rotation:

	PC1	PC2	PC3	PC4	PC5
MDage	-0.4178	0.010282	-0.005598	0.01214	-0.21545
MDexp	-0.3917	0.004202	-0.008205	0.02552	-0.59630
patientAge	-0.3953	-0.020540	0.006105	-0.01252	-0.04688
ptFemale	-0.2919	0.349292	0.614220	-0.11289	0.02546
ptMale	-0.2988	-0.331570	-0.605673	0.10534	0.03834
inpatient	-0.2764	-0.288567	0.032095	-0.75594	0.04764
outpatient	-0.3447	0.215811	-0.014377	0.54312	0.04846
a1failToDx	-0.3152	0.461804	-0.283401	-0.15723	0.63392
a1delayInDx	-0.2162	-0.651462	0.417380	0.28992	0.43269

Figure 13. Region 6 PCA Loadings (prcomp)

K-means Clustering

Region 5

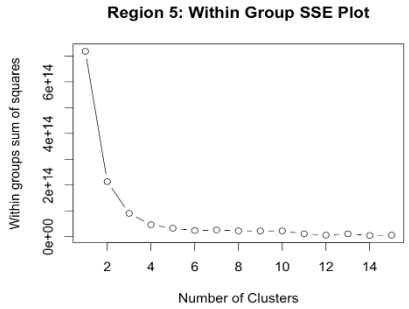


Figure 15. Region 5 K-means Scree Plot

	totalPayment	MDage	MDexp	outcomeScale	patientAge	ptFemale	ptMale	inpatient	outpatient	a1failToDx
1	-0.006043572	-0.009847685	-0.001714924	0.03442854	0.05829353	-1.0601387	1.0601387	0.03916754	-0.01712796	-1.019935530
2	-0.021552555	-0.618602641	-0.602770631	0.01400171	-0.07549368	0.9431976	-0.9431976	0.25137869	-0.23778776	-0.006724268
3	0.024271764	0.936957172	0.912908459	-0.22037808	0.11873843	0.9322464	-0.9322464	-0.42972097	0.40870724	-0.023424677
4	0.011385887	-0.041347339	-0.047552660	0.14821069	-0.06429987	-1.0601387	1.0601387	0.02814532	-0.04713758	0.977754301

	a1delayInDx	a1wrongDx	a1failToOrder	a1radError	a2failToTx	a2delayInTx	a2delayInDx	a2failToDx	a2failToTest
1	0.663879456	0.247203463	0.22799422	0.13128045	-0.124741079	0.11504343	0.001092140	-0.01988189	-0.06495629
2	0.008303036	-0.019328728	0.01137215	-0.02246240	0.003483331	0.02726779	0.024845596	0.02800035	-0.03779962
3	0.069967103	0.009466111	-0.07988434	0.02269241	-0.044333397	-0.04147590	-0.031103966	-0.03899336	0.05613327
4	-0.690647349	-0.213556320	-0.15466841	-0.11367955	0.151353836	-0.10455557	-0.004856115	0.01769507	0.05830092

Figure 14. Region 5 K-means Cluster Centers

Region 6

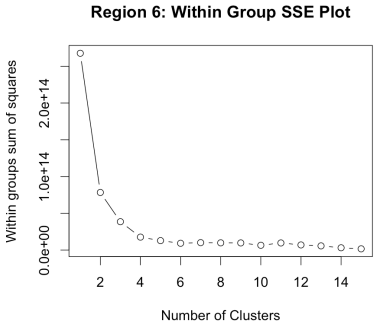


Figure 17. Region 6 K-means Scree Plot

	totalPayment	MDage	MDexp	outcomeScale	patientAge	ptFemale	ptMale	inpatient	outpatient	a1failToDx	a1delayInDx	a1wrongDx
1	-0.0257374610	-0.02529560	-0.024406778	-0.07442568	-0.03062656	1.013506672	-1.013506672	-0.06365930	0.0552234417	0.3400823	-0.5652203	0.11274379
2	0.0254956458	0.04111974	0.021803842	0.10449917	-0.02039058	-0.986506828	0.986506828	-0.01029957	0.0008983703	0.3462399	-0.5652203	0.04766752
3	-0.0001234501	-0.01781571	0.004235003	-0.05608658	0.08028946	-0.006163132	0.006163132	0.10678707	-0.0744033360	-1.0742118	1.7640056	-0.22599408
4	-0.0025769008	-0.08546591	-0.010137338	0.08433801	-0.02306770	-0.083274924	0.083274924	0.07067732	-0.1275306018	0.2193171	-0.3016880	-0.22599408

	a1failToOrder	a1radError	a2failToTx	a2delayInTx	a2delayInDx	a2failToDx	a2failToTest
1	0.02123008	0.07927791	0.01401607	-0.11453624	0.02971261	-0.1461791	-0.01604340
2	0.10308909	0.01194554	0.08339188	-0.09737826	-0.01520605	-0.1461791	0.05687310
3	-0.19726502	-0.14617912	-0.13508144	0.34106090	-0.00915110	-0.1461791	-0.05309814
4	0.05752847	0.07917402	-0.17949628	-0.17849230	-0.14375252	6.8397680	-0.12276446

Figure 16. Region 6 K-means cluster centers

Logistic Regression

Region 5

```
Call:
glm(formula = a1failToDx ~ ., family = "binomial", data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1994  -0.0001   0.1563   0.5270   1.1225

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.722e+00  3.110e-01   5.538 3.06e-08 ***
totalPayment   2.610e-07  1.620e-07   1.611   0.1071
MDage          4.926e-03  7.181e-03   0.686   0.4928
MDexp         -4.195e-03  6.942e-03  -0.604   0.5456
outcomeScale   4.455e-02  1.900e-02   2.344   0.0191 *
patientAge    -3.307e-03  1.870e-03  -1.768   0.0770 .
ptFemale       5.737e-02  7.405e-02   0.775   0.4385
inpatient     -7.620e-01  1.262e-01  -6.037 1.57e-09 ***
outpatient    -1.658e-01  1.318e-01  -1.258   0.2084
a1delayInDx   -2.148e+01  1.868e+02  -0.115   0.9085
a1wrongDx     -2.137e+01  5.094e+02  -0.042   0.9665
a1failToOrder -2.162e+01  6.760e+02  -0.032   0.9745
a1radError    -2.128e+01  9.486e+02  -0.022   0.9821
a2failToTx     2.910e+00  7.125e-01   4.085 4.41e-05 ***
a2delayInTx    9.424e-01  4.665e-01   2.020   0.0434 *
a2delayInDx   -9.976e-01  2.129e-01  -4.685 2.79e-06 ***
a2failToDx    -3.432e-01  3.269e-01  -1.050   0.2937
a2failToTest   8.018e-01  4.285e-01   1.871   0.0613 .
---
Signif. codes:  0 ?***?0.001 ?*?0.01 ??0.05 ??0.1 ??1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13901.8  on 10031  degrees of freedom
Residual deviance:  4873.1  on 10014  degrees of freedom
AIC: 4909.1
```

Figure 20. Region 5 Logistic Regression

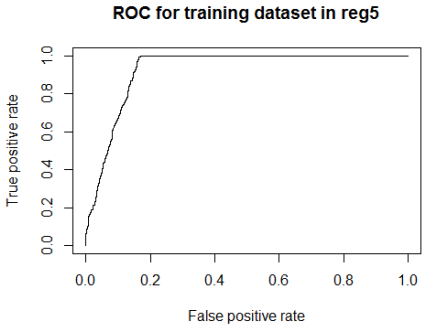


Figure 18. Region 5 ROC Curve Training Dataset

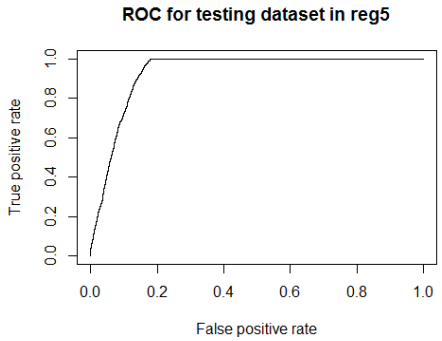


Figure 19. Region 5 ROC Curve Testing Dataset

Region 6

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.68098  -0.00009   0.45894   0.58458   1.09448

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.519e+00  3.724e-01   4.080 4.51e-05 ***
totalPayment   2.846e-07  2.217e-07   1.283   0.1994
MDage          1.516e-03  8.234e-03   0.184   0.8539
MDexp         -1.178e-02  7.880e-03  -1.495   0.1349
outcomeScale   2.633e-02  2.175e-02   1.210   0.2261
patientAge    -4.698e-03  2.162e-03  -2.173   0.0297 *
ptFemale       4.235e-02  8.628e-02   0.491   0.6236
inpatient     -3.533e-01  1.602e-01  -2.205   0.0275 *
outpatient     4.134e-01  1.616e-01   2.558   0.0105 *
a1delayInDx   -2.113e+01  2.799e+02  -0.075   0.9398
a1wrongDx     -2.112e+01  6.265e+02  -0.034   0.9731
a1failToOrder -2.131e+01  7.065e+02  -0.030   0.9759
a1radError    -2.109e+01  9.536e+02  -0.022   0.9824
a2failToTx     1.601e+00  3.669e-01   4.363 1.28e-05 ***
a2delayInTx    7.558e-02  3.944e-01   0.192   0.8480
a2delayInDx   -3.962e-01  2.803e-01  -1.414   0.1575
a2failToDx    -3.631e-02  2.582e-01  -0.141   0.8881
a2failToTest   4.301e-01  3.824e-01   1.125   0.2606
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8184.8  on 5925  degrees of freedom
Residual deviance: 3480.1  on 5908  degrees of freedom
AIC: 3516.1

Number of Fisher Scoring iterations: 18
```

Figure 22. Region 6 Logistic Model Output

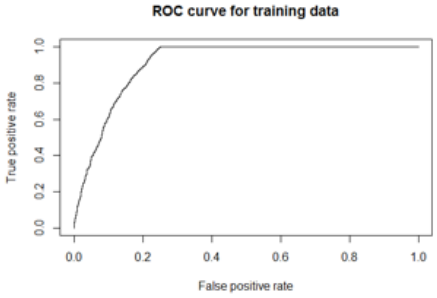


Figure 21. Region 6 ROC curve for training data

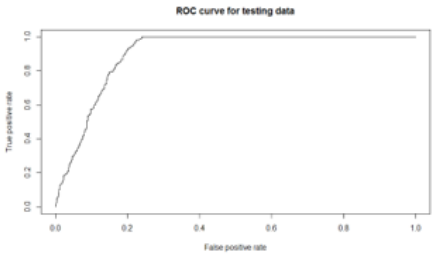


Figure 23. Region 6 ROC curve for testing data

Factor Analysis of Mixed Type

Region 5

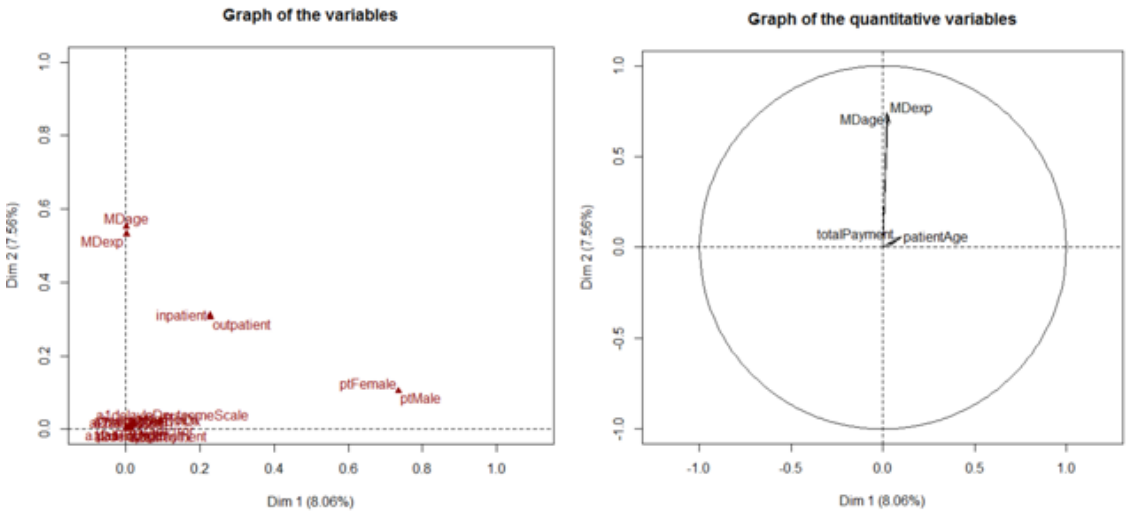


Figure 24. Region 5 FAMD Biplots

Region 6

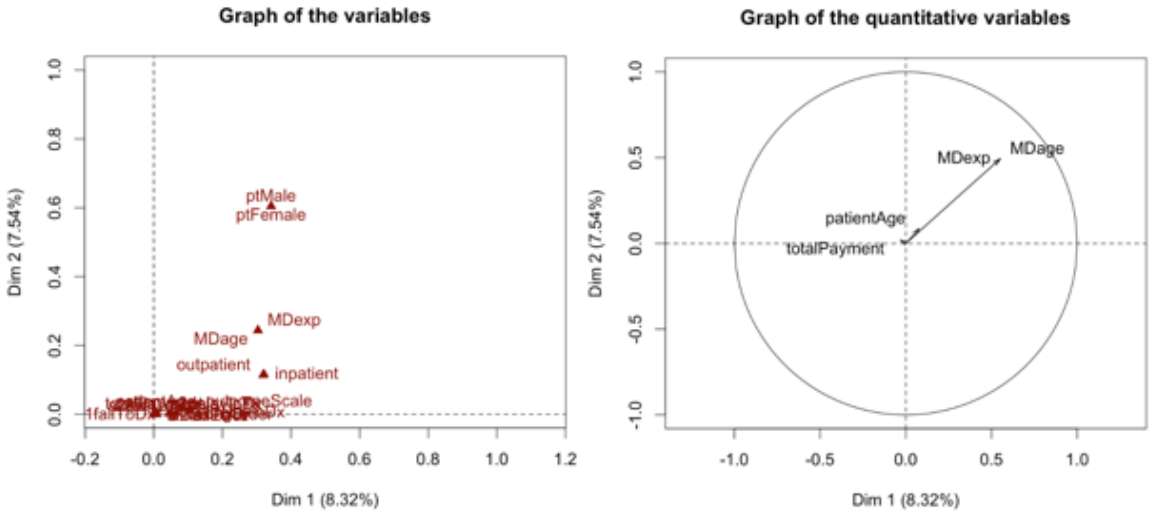


Figure 25. Region 6 FAMD Biplots

APPENDIX B. R CODE

Exploratory Data Analysis

```
# LOAD LIBRARIES
library(foreign)
library(stringr)

# LOAD DATA (.POR file)
NPDB.full <- read.spss("NPDB1510.POR", to.data.frame = TRUE)
names(NPDB.full) <- str_to_lower(names(NPDB.full))

# EXPLORE DATA
head(NPDB.full, n=10)
sort(summary(NPDB.full$rectype))

# How many of each rectype?
#  Judgment or Conviction Report, 11/22/1999 and later      = 0
#  Adverse Action Report (Legacy Format)                    = 51301
#  Malpractice Payment Report, 1/31/04 and later            = 167027    (INCLUDE)
#  Malpractice Payment Report, 9/1/90 to 1/31/04            = 250685
#  Consolidated Adverse Action Report, 11/22/1999 and later = 729253

# Explore adverse action reports
NPDB.AAR <- subset(NPDB.full, NPDB.full$RECTYPE %in% c("Adverse Action Report (Legacy
Format)","Consolidated Adverse Action Report, 11/22/1999 and later"))
summary(NPDB.AAR)

# Explore all malpractice payment reports
NPDB.MP <- subset(NPDB.full, NPDB.full$rectype %in% c("Malpractice Payment Report, 9/1/90 to
1/31/04","Malpractice Payment Report, 1/31/04 and later"))
summary(NPDB.MP)

# Explore all malpractice payment reports < 2004
NPDB.MPold <- subset(NPDB.full, NPDB.full$rectype == "Malpractice Payment Report, 9/1/90 to
1/31/04")
summary(NPDB.MPold)

# Explore all malpractice payment reports > 1/31/2004
NPDB.MPnew <- subset(NPDB.full, NPDB.full$rectype == "Malpractice Payment Report, 1/31/04 and
later")
summary(NPDB.MPnew)

# Physicians only
NPDB.MPnew.MDs <- subset(NPDB.MPnew, NPDB.MPnew$licnfeld %in% c("Allopathic Physician
(MD)","Physician Resident (MD)","Osteopathic Physician (DO)","Osteopathic Physician Resident
(DO)"))
summary(NPDB.MPnew.MDs)

sort(summary(NPDB.MPnew.MDs$licnfeld))
# Doctors (MD, DO)
#  Allopathic Physician (MD)           = 119,501 (EXPLORE)
#  Physician Resident (MD)             = 825
#  Osteopathic Physician (DO)          = 9,092
#  Osteopathic Physician Resident (DO) = 135

# Diagnosis Related only
NPDB.Dx <- subset(NPDB.MPnew.MDs, NPDB.MPnew.MDs$algnnatr == "Diagnosis Related")
summary(NPDB.Dx)


sort(summary(NPDB.MPnew$practage))
```

Data Preprocessing

```
### Data transformations ###

# remove unknowns
NPDB.Dx <- NPDB.Dx[!is.na(NPDB.Dx$practage),]
NPDB.Dx <- NPDB.Dx[!is.na(NPDB.Dx$malyear1),]
NPDB.Dx <- NPDB.Dx[!is.na(NPDB.Dx$grad),]
NPDB.Dx <- NPDB.Dx[!is.na(NPDB.Dx$ptage),]
NPDB.Dx <- NPDB.Dx[NPDB.Dx$ptage != "fetus",]
NPDB.Dx <- NPDB.Dx[!is.na(NPDB.Dx$ptgender),]
NPDB.Dx <- NPDB.Dx[NPDB.Dx$ptgender != "U",]
NPDB.Dx <- NPDB.Dx[!is.na(NPDB.Dx$pttype),]
NPDB.Dx <- NPDB.Dx[NPDB.Dx$pttype != "U",]
NPDB.Dx <- NPDB.Dx[!is.na(NPDB.Dx$outcome),]
NPDB.Dx <- NPDB.Dx[!is.na(NPDB.Dx$totalpmt),]
NPDB.Dx <- NPDB.Dx[as.character(NPDB.Dx$outcome) != "Cannot Be Determined from Available
Records",]
summary(NPDB.Dx)

# totalPayment - use totalpmt bin mean as value
unique(NPDB.Dx$totalpmt)
NPDB.Dx$strTotPmt = as.character(NPDB.Dx$totalpmt)

tpstr1 = substr((NPDB.Dx$strTotPmt), 2, str_locate(NPDB.Dx$strTotPmt, " ")[,1]-1)
tpnum1 = as.numeric(str_replace(tpstr1, ",", ""))

tpstr2 = substr((NPDB.Dx$strTotPmt), str_locate(NPDB.Dx$strTotPmt, " ")+10,
str_length(NPDB.Dx$strTotPmt))
tpnum2 = as.numeric(str_replace(tpstr2, ",", ""))

NPDB.Dx$totalPayment = round((tpnum1 + tpnum2)/2, digits=0)
NPDB.Dx <- NPDB.Dx[!is.na(NPDB.Dx$totalPayment),]


# MDage (new variable) - use practage bin mean as value
sort(summary(NPDB.Dx$practage))
NPDB.Dx$MDage = as.numeric(substr(as.character(NPDB.Dx$practage),6,7)) + 5
summary(NPDB.Dx$MDage)
levels(NPDB.MPnew$alegatn1)
levels(NPDB.MPnew$alegatn2)

levels(NPDB.Dx$grad)
levels(NPDB.Dx$alegatn1)
levels(NPDB.Dx$alegatn2)
unique(NPDB.Dx$alegatn1)
unique(NPDB.Dx$alegatn2)
sort(summary(NPDB.Dx$alegatn1))

sort(summary(NPDB.Dx$alegatn2))

# MDexp (new variable) = malyear1 - (use grad bin min)
NPDB.Dx$MDexp <- NPDB.Dx$malyear1 - (as.numeric(substr(as.character(NPDB.Dx$grad),1,4))+5)
sort(unique(NPDB.Dx$malyear1))

head(NPDB.Dx)
NPDB.Dx <- NPDB.Dx[NPDB.Dx$MDexp >= 0,]


# outcome - outcomes are an ordered list representing harm scale, we can use as numerical values
summary(NPDB.Dx$outcome)
NPDB.Dx$outcomeScale <- as.integer(NPDB.Dx$outcome)
head(NPDB.Dx)

# alegatn1 dummy variables
```

```
NPDB.Dx$a1failToDx <- ifelse(NPDB.Dx$alegatn1 == "Failure to Diagnose",1,0)
NPDB.Dx$a1delayInDx <- ifelse(NPDB.Dx$alegatn1 == "Delay in Diagnosis",1,0)
NPDB.Dx$a1wrongDx <- ifelse(NPDB.Dx$alegatn1 == "Wrong or Misdiagnosis (e.g. Original Diagnosis
is Incorrect)",1,0)
NPDB.Dx$a1failToOrder <- ifelse(NPDB.Dx$alegatn1 == "Failure to Order Appropriate Test",1,0)
NPDB.Dx$a1radError <- ifelse(NPDB.Dx$alegatn1 == "Radiology or Imaging Error",1,0)

# alegatn2 dummy variables (tx is a medical abbreviation for "treatment")
NPDB.Dx$a2failToTx <- ifelse(NPDB.Dx$alegatn2 == "Failure to Treat",1,0)
NPDB.Dx$a2delayInTx <- ifelse(NPDB.Dx$alegatn2 == "Delay in Treatment",1,0)
NPDB.Dx$a2delayInDx <- ifelse(NPDB.Dx$alegatn2 == "Delay in Diagnosis",1,0)
NPDB.Dx$a2failToDx <- ifelse(NPDB.Dx$alegatn2 == "Failure to Diagnose",1,0)
NPDB.Dx$a2failToTest <- ifelse(NPDB.Dx$alegatn2 ==  "Failure to Order Appropriate Test",1,0)

NPDB.Dx$a2delayInTx[is.na(NPDB.Dx$a2delayInTx)] <- 0
NPDB.Dx$a2failToDx[is.na(NPDB.Dx$a2failToDx)] <- 0
NPDB.Dx$a2failToTx[is.na(NPDB.Dx$a2failToTx)] <- 0
NPDB.Dx$a2failToTest[is.na(NPDB.Dx$a2failToTest)] <- 0
NPDB.Dx$a2delayInDx[is.na(NPDB.Dx$a2delayInDx)] <- 0

summary(NPDB.Dx)

# ptage
summary(NPDB.Dx$ptage)

patientAge <- as.character(NPDB.Dx$ptage)
patientAge[patientAge=="Age under 1 year"] <- 0
patientAge[patientAge=="Ages 1 through 9"] <- 5
patientAge[patientAge=="Ages 10 through 19"] <- 15
patientAge[patientAge=="Ages 20 through 29"] <- 25
patientAge[patientAge=="Ages 30 through 39"] <- 35
patientAge[patientAge=="Ages 40 through 49"] <- 45
patientAge[patientAge=="Ages 50 through 59"] <- 55
patientAge[patientAge=="Ages 60 through 69"] <- 65
patientAge[patientAge=="Ages 70 through 79"] <- 75
patientAge[patientAge=="Ages 80 through 89"] <- 85
patientAge[patientAge=="Ages 90 through 99"] <- 95
patientAge <- as.numeric(patientAge)
NPDB.Dx$patientAge <- patientAge

summary(NPDB.Dx$patientAge)

# ptgender
summary(NPDB.Dx$ptgender)
NPDB.Dx$ptFemale <- ifelse(NPDB.Dx$ptgender == "F",1,0)
NPDB.Dx$ptMale <- ifelse(NPDB.Dx$ptgender == "M",1,0)

# pttype
summary(NPDB.Dx$pttype)
NPDB.Dx$inpatient <- ifelse(NPDB.Dx$pttype %in% c("I","B"),1,0)
NPDB.Dx$outpatient <- ifelse(NPDB.Dx$pttype %in% c("O","B"),1,0)

fields = c("totalPayment", "MDage", "MDexp", "outcomeScale", "patientAge", "ptFemale",
"ptMale", "inpatient", "outpatient", "a1failToDx", "a1delayInDx", "a1wrongDx", "a1failToOrder",
"a1radError", "a2failToTx", "a2delayInTx", "a2delayInDx", "a2failToDx", "a2failToTest")
summary(NPDB.Dx)

# Get Regions
sort(summary(NPDB.Dx$licnstat))

region1 = c("Hawaii","Alaska","Washington","California","Oregon")
region2 = c("Montana", "Idaho", "Wyoming", "Nevada", "Utah", "Colorado", "Arizona", "New Mexico")
region3 = c("North Dakota", "South Dakota", "Minnesota", "Iowa", "Nebraska", "Kansas",
"Missouri")
region4 = c("Texas", "Oklahoma", "Arkansas", "Louisiana", "Kentucky", "Tennessee",
"Mississippi", "Alabama")
```

```
region5 = c("Maine","Vermont","New Hampshire","Massachusetts","Connecticut","Rhode Island","New
York","Pennsylvania","New Jersey")
region6 = c("West Virginia", "Maryland", "Delaware", "Virginia", "North Carolina", "South
Carolina", "Georgia", "Florida")

NPDB.reg1 <- subset(NPDB.Dx, NPDB.Dx$licnstat %in% region1, select = fields)
NPDB.reg2 <- subset(NPDB.Dx, NPDB.Dx$licnstat %in% region2, select = fields)
NPDB.reg3 <- subset(NPDB.Dx, NPDB.Dx$licnstat %in% region3, select = fields)
NPDB.reg4 <- subset(NPDB.Dx, NPDB.Dx$licnstat %in% region4, select = fields)
NPDB.reg5 <- subset(NPDB.Dx, NPDB.Dx$licnstat %in% region5, select = fields)
NPDB.reg6 <- subset(NPDB.Dx, NPDB.Dx$licnstat %in% region6, select = fields)

# Region Counts
# region 1: 3372
# region 2: 1901
# region 3: 1631
# region 4: 4147
# region 5: 12540
# region 6: 5926

head(NPDB.reg5)
summary(NPDB.reg5)
nrow(NPDB.reg5)

head(NPDB.reg6)
summary(NPDB.reg6)
nrow(NPDB.reg6)

pairs(NPDB.reg6)
```

Multivariate Regression

REGION 5

```
# select data
data <- NPDB.reg5

# initial multi regression analysis
fit_MRegression <- lm(totalPayment ~ MDage + MDexp + outcomeScale + patientAge + ptFemale + ptMale + inpatient + outpatient + a1failToDx + a1delayInDx +
a1wrongDx + a1radError + a2failToTx + a2delayInTx + a2delayInDx + a2failToDx + a2failToTest, data = data)
summary(fit_MRegression)

anova(fit_MRegression, update(fit_MRegression, ~1), test = "Chisq")

## predict(fit_MRegression, data=Housing, interval = "confidence", level = 0.95)

summary((fit_MRegression))

# detect correlation between variables
cor.data = cor(data)
round(cor.data, 3)

library(corrplot)
corrplot(cor.data, method = "ellipse")

plot(data$MDage,data$totalPayment)
plot(data$MDexp,data$totalPayment)

# begin feature selection
library(MASS)
null = lm(totalPayment ~ 1, data = data)
full = lm(totalPayment ~ ., data = data)
Forwardselection = step(null, scope = list(lower = null, upper = full), direction = "forward")

# second multi-regerssion model
fit_selectedmodel <- lm(totalPayment ~ a2failToDx + a2failToTx + a2delayInTx + MDexp + inpatient + a2delayInDx + a1wrongDx + a1radError, data = data)
summary(fit_selectedmodel)

# test logistic regrssion on totalpayment
# failed fit_logtotalpay <- glm(totalPayment ~ ., data = data, family = "binomial")
# failed summary(fit_logtotalpay)


# test logistic regression on precentage increase of the totalpayment
data$logtotalPayment = log(data$totalPayment)
data2 = data[, c(2:20)]
fit_logptotalpay <- lm(logtotalPayment ~ ., data = data2)
summary(fit_logptotalpay) # the result is slightly better than the normal multicollinear regression model
# the multicollinear regression is not perfect fit for this case
```

REGION 6

```
#############################################################################################################
# Multivariate Regression on Region 6
#############################################################################################################
library(car)

#Reg 6 linear model with all variables
reg6.mod1 = lm(NPDB.reg6$totalPayment~., data = NPDB.reg6);summary(reg6.mod1)
#Stepwise selection
step(reg6.mod1, direction = "both")
#stepwise selected model
step.model =lm(formula = NPDB.reg6$totalPayment ~ MDexp + outcomeScale +  patientAge +
                inpatient + a1delayInDx + a1failToOrder + a1radError, data = NPDB.reg6); summary(step.model)
###residuals Plot
plot(step.model)

##try log scale on total payments
log.totalPayments = log(NPDB.reg6$totalPayment)

log.model = lm(formula = log.totalPayments ~ MDexp + outcomeScale +  patientAge +
                inpatient + a1delayInDx + a1failToOrder + a1radError, data = NPDB.reg6);summary(log.model)

###residuals Plot
plot(log.model)
```

Principal Component Analysis

REGION 5

```
library(corrplot)
dat <- NPDB.reg5
var_cor <- cor(dat)
corrplot.mixed(var_cor, upper = "number", lower= "ellipse")

norm.dat <- scale(dat)

pca.fit <- princomp(norm.dat, scale=FALSE)
summary(pca.fit)
plot(pca.fit)
abline(1, 0)
print(pca.fit$loadings, cutoff = 0.4)

pca.roated <- psych::principal(norm.dat, rotate="varimax", nfactors=15, scores=TRUE)
print(pca.roated$loadings, cutoff=0.4, sort=T)
```

REGION 6

```
#####################################################################
# PCA/CFA on Region 6 dataset
#####################################################################
library(car)
library(stats)
library(corrplot)
library(psych)
library(MASS)

source("PCA_Plot.r")
mar.default = c(5, 4, 4, 2) + 0.1
par(mfrow = c(1, 1), mar=mar.default) # RESET PLOT MARGINS

reg6 = NPDB.reg6
head(reg6)
```

```
IV = reg6[1:19]

# ANALYZE CORRELATIONS
IV.cor = cor(IV)
IV.cor
corrplot(IV.cor, method="ellipse", type="upper")

IV.cov = cov(IV)
IV.cov

# TRY PCA WITH RAW DATA X CORRELATION MATRIX
pC1 = prcomp(IV, scale = T, center=T, scores = T)
print(pC1)
summary(pC1)
pC1$rotation
pC1$x
plot(pC1, main="pC1 Raw Data x Corr")
abline(h=1)

# TRY PCA WITH RAW DATA X COVARIANCE MATRIX
pC2 = prcomp(IV, scale = F, center = T, scores = T)
print(pC2)
summary(pC2)
pC2$rotation
pC2$x
plot(pC2, main="pC2 Raw Data x Cov")
abline(h=1)

options("scipen"=100, "digits"=5)
IV.cor.test = corr.test(IV.cor, adjust="none", alpha=.1)
print(IV.cor.test, short=T)
M = IV.cor.test$p
MTest = ifelse(M < .01, T, F)
MTest
colSums(MTest) - 1

# not correlated with anything: totalPayment, outcomeScale, patientAge, a1wrongDx,
a1failToOrder, a1radError, a2failToTx, a2delayInTx, a2delayInDx, a2failToDx, a2failToTest
# no variables are highly correlated - i.e. w/75% of vars

IV.subset = IV[,c(1:2,4:10)]
head(IV.subset)
IV.subset.cor = cor(IV.subset)
IV.subset.cov = cov(IV.subset)
corrplot(IV.subset.cor, method="ellipse")
```

```
# MAKE SUBSET OF SCALED VARIABLES
IV.scaled = scale(IV.subset, center=F)
head(IV.scaled)
IV.scaled.cor = cor(IV.scaled)
IV.scaled.cov = cov(IV.scaled)
corrplot(IV.scaled.cor, method="ellipse")
princomp(IV)
options(digits=4)

# pC3 TRY PCA ON SUBSET X CORRELATION MATRIX
pC3 = prcomp(IV.subset, scale = T, center = T, scores = T)
print(pC3)
summary(pC3)
pC3$rotation
pC3$x
plot(pC3, main="pC3 Subset x Corr")
abline(h=1)

PCA_Plot_Psyc(pC3)
PCA_Plot_Psyc_Secondary(pC3)

# TRY ROTATING DATA pC3
R = as.matrix(pC3$rotation)
Mtmp = as.matrix(IV.subset)
Mtmp.rot = Mtmp %*% R
head(Mtmp.rot)

plot(Mtmp.rot, col="red", pch=16, asp=1, main="Mtmp.rot")
Mtmp.rot.cov = cov(Mtmp.rot)
pC3.rot = prcomp(Mtmp.rot, scale = F, center = T, scores = T)
print(pC3.rot)
summary(pC3.rot)
pC3.rot$rotation
par(mfrow = c(1, 1), mar=mar.default)
plot(pC3.rot, main="pC3.rot") # scree plot
abline(h=1, col=1, lty=1, lwd=2)

# CFA WITH pC3.rot
cfa3.rot = psych::principal(Mtmp.rot, rotate="varimax", nfactors=4, scores=TRUE)
print(cfa3.rot$loadings, cutoff=.4, sort=T)
summary(cfa3.rot)
PCA_Plot_Psyc(cfa3.rot)
PCA_Plot_Psyc_Secondary(cfa3.rot)

# pC4 TRY PCA ON SUBSET X COVARIANCE MATRIX
pC4 = prcomp(IV.subset, scale = F, center = F, scores = T)
print(pC4)
summary(pC4)
pC4$rotation
pC4$x
plot(pC4, main="pC4 Subset x Cov")
abline(h=1)

# pC5 TRY PCA ON CENTERED SUBSET X COVARIANCE MATRIX
pC5 = prcomp(IV.subset, scale = F, center = T, scores = T)
print(pC5)
summary(pC5)
pC5$rotation
plot(pC5, main="pC5 Subset x Cov")
abline(h=1)

# TRY ROTATING DATA pC5
R = as.matrix(pC5$rotation)
Mtmp = as.matrix(IV.subset)
Mtmp.rot = Mtmp %*% R
head(Mtmp.rot)
```

```
plot(Mtmp.rot, col="red", pch=16, asp=1, main="Mtmp.rot")
Mtmp.rot.cov = cov(Mtmp.rot)
pC5.rot = prcomp(Mtmp.rot, scale = F, center = T, scores = T)
print(pC5.rot)
summary(pC5.rot)
pC5.rot$rotation
par(mfrow = c(1, 1), mar=mar.default)
plot(pC5.rot, main="pC5.rot") # scree plot
abline(h=1, col=1, lty=1, lwd=2)

par(mfrow = c(1, 2), mar=mar.default)
biplot(pC5, main="before rot")
abline(h=0, v=0, col="pink", lty=3)
biplot(pC5.rot, main="after rot")
abline(h=0, v=0, col="pink", lty=3)

# CFA WITH pC5.rot
cfa5.rot = psych::principal(Mtmp.rot, rotate="varimax", nfactors=4, scores=TRUE)
print(cfa5.rot$loadings, cutoff=.4, sort=T)
summary(cfa5.rot)
PCA_Plot_Psyc(cfa5.rot)
PCA_Plot_Psyc_Secondary(cfa5.rot)

# pC6 TRY PCA ON scaled SUBSET X COVARIANCE MATRIX
pC6 = prcomp(IV.scaled, scale = F, center = F, scores = T)
print(pC6)
summary(pC6)
pC6$rotation
plot(pC6, main="pC6 scaled Subset x Cov")
abline(h=1)


# TRY ROTATING DATA pC6
R = as.matrix(pC6$rotation)
IVtmp = as.matrix(IV.scaled)
IVtmp.rot = IVtmp %*% R
head(IVtmp.rot)

plot(IVtmp.rot, col="red", pch=16, asp=1, main="Mtmp.rot")
pC6.rot = prcomp(IVtmp.rot, scale = F, center = F, scores = T)
print(pC6.rot)
summary(pC6.rot)
pC6.rot$rotation
par(mfrow = c(1, 1), mar=mar.default)
plot(pC6.rot, main="pC6.rot") # scree plot
abline(h=1, col=1, lty=1, lwd=2)

par(mfrow = c(1, 2), mar=mar.default)
biplot(pC6, main="before rot")
abline(h=0, v=0, col="pink", lty=3)
biplot(pC6.rot, main="after rot")
abline(h=0, v=0, col="pink", lty=3)
```

K-Means Clustering

Region 5

```
mydata <- NPDB.reg5

wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata,
                                     centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares", main="Region 5: Within Group SSE Plot")

reg5.fit <- kmeans(scale(NPDB.reg5), 4)
reg5.fit$centers
```

Region 6

```
mydata <- NPDB.reg6

wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata,
                                     centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares", main="Region 6: Within Group SSE Plot")

reg6.fit <- kmeans(scale(NPDB.reg6), 4)
reg6.fit$centers
```

Logistic Regression

Region 5

```
# next step analysis
# we predict an potential logistic regression on a1failtodox

# preprocessing the dataset
data3 = data[, c(1:19)]
data3$ptMale = NULL

# divide the dataset into training and testing
data4 <- sample(1:nrow(data3), 0.80*nrow(data3))
train.data <- data3[data4,]       # 80% of data based on random select.MDATA sample
test.data <- data3[-data4,]     # 20% of data based on leftovers

# develop the model for logistic regression
fit_loga1failToDx <- glm(a1failToDx~., data = train.data, family = "binomial")
summary(fit_loga1failToDx)

# run 1o fold cross validation on training dataset
library(boot)
val.10.fold <- cv.glm(data = train.data, glmfit = fit_loga1failToDx, K = 10)
val.10.fold

# plot ROC curve on both the training and testing set

library(ROCR)
library(AUC)

predict.test = test.data
predict.test$a1failToDx <- NULL

predict.train = train.data
predict.train$a1failToDx <- NULL

p2 <- predict(fit_loga1failToDx, newdata = predict.train)
pr2 <- prediction(p2, train.data$a1failToDx)
prf2 <- performance(pr2, measure = "tpr", x.measure = "fpr")
plot(prf2, main="ROC for testing dataset in reg5")
auc2 <- performance(pr2, measure = "auc")
auc2 <- auc2@y.values[[1]]
auc2

p <- predict(fit_loga1failToDx, newdata = predict.test)
pr <- prediction(p, test.data$a1failToDx)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf, main="ROC for training dataset in reg5")
auc <- performance(pr, measure = "auc")
auc<- auc@y.values[[1]]
auc
# ROC curve shows a high accuaryc rate in the both training and testing set
```

Region 6

```
library(pscl)
library(ROCR)
library(AUC)

#logit regression
##reg6.glm = glm(NPDB.reg6$a1failToDx~., family = binomial, data = NPDB.reg6); summary(reg6.glm)
#---------------------------------------------
# PREPARE PARTITIONS: TRAIN VS. TEST SUBSETS
#---------------------------------------------

NPDB.reg6$ptMale = NULL
select.NPDB.reg6 <- sample(1:nrow(NPDB.reg6 ), 0.75*nrow(NPDB.reg6 ))
TRAIN <- NPDB.reg6 [select.NPDB.reg6,]       # 75% of data based on random select.MDATA sample
TEST <-   NPDB.reg6 [-select.NPDB.reg6,]     # 25% of data based on leftovers

reg6.glm = glm(NPDB.reg6$a1failToDx~., family = binomial, data = NPDB.reg6); summary(reg6.glm)


##reg6.glm = glm(TRAIN$a1failToDx~TRAIN$totalPayment+TRAIN$MDage
   ##       +TRAIN$MDexp+TRAIN$outcomeScale+TRAIN$patientAge+TRAIN$ptFemale+TRAIN$ptMale
   ##       +TRAIN$inpatient+TRAIN$outpatient, family = binomial (link = "logit"), data = TRAIN)


#anova

reg6.anova = anova(TRAIN, test = "Chisq");

#plot
glm.plot = glm.diag(reg6.glm)
glm.diag.plots(glm.plot,reg6.glm)
glm.diag.plots(reg6.glm,glm.plot)

##predict Value
fitted.results <- predict(reg6.glm,newdata=subset(TEST,select=c(2,3,4,5,6,7,8,9)),type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != TEST$a1failToDx)
print(paste('Accuracy',1-misClasificError))
newdata=subset(TEST,select=c(2,3,4,5,6,7,8), type="response")

Predict.Set = TEST
Predict.Set$a1failToDx <- NULL



Predict.Set = TEST
p <- predict(reg6.glm, newdata = Predict.Set)
pr <- prediction(p, TEST$a1failToDx)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc

Predict.Set = TRAIN
Predict.Set$a1failToDx <- NULL

p <- predict(reg6.glm, newdata = Predict.Set)
pr <- prediction(p, TRAIN$a1failToDx)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
title("ROC curve for training data")
```

Factor Analysis of Mixed Data

Region 5

```
library(FactoMineR)

dat <- NPDB.reg5
dat <- as.data.frame(lapply(dat, as.factor))
dat$totalPayment <- as.numeric(dat$totalPayment)
dat$MDage <- as.numeric(dat$MDage)
dat$MDexp <- as.numeric(dat$MDexp)
dat$patientAge <- as.numeric(dat$patientAge)

famd.fit <- FAMD(dat)

summary(famd.fit)
```

Region 6

```
library(FactoMineR)

dat6 <- NPDB.reg6[,c(10,1:9,11:19)]
dat6 <- as.data.frame(lapply(dat6, as.factor))
dat6$totalPayment <- as.numeric(dat6$totalPayment)
dat6$MDage <- as.numeric(dat6$MDage)
dat6$MDexp <- as.numeric(dat6$MDexp)
dat6$patientAge <- as.numeric(dat6$patientAge)

famd.fit6 <- FAMD(dat6)

summary(dat6)
```