CAPSTONE FINAL REPORT

Springboard.com Fundamentals of Data Science

Analysis of risk-adjusted cost outcomes for patients with hypertension

###### **Presented by**

Pierre Carpentier

Data Science Mentor: Graham Spinks

December 18th, 2016

CAPSTONE FINAL REPORT

Abstract

The ultimate goal of the Foundations of Data Science Capstone project is to demonstrate adaptability in taking a real-world issue, integrating and manipulating various data sets using the techniques learnt in the class, and outputting a solution to the issue. For this project, the US healthcare industry was chosen and the issue was very narrowly focused such that the data manipulation would be focused as well. This project integrated five data sets to first predict the probability of a hypertension diagnosis. It then allows to project the risk-adjusted cost of the hypertension potentially leading to an inpatient hospital event. Finally, it allows patients to verify their risk adjusted cost with base metrics from available insurance plans.

The data manipulation was mostly performed in R using techniques and packages from the course, with some minor manipulations using pivot tables in excel.

Table of Contents

1. Recap on the Project Proposal 5

1.1. Description of Problem 5

1.2. Future Work 5

2. DATA SETS AND BASIC HEALTH CARE KNOWLEDGE 6

2.1. US Healthcare Data Sets on Kaggle 6

2.1.1. Plan Attributes Data Set 6

2.1.2. Rate Data Set 6

2.1.3. Cross-Walk 2016 Data Set 6

2.2. Centers for Medicare and Medicaid Services (CMS) Provider Inpatient Charge Data Set 6

2.2.1. Inpatient, Outpatient, Emergency Room, and Office Based Events 6

2.3. National Health Records 2011 (NH11) Data Set 7

3. DATA MANIPULATION AND VISUALISATION 7

3.1. Insurance Plan Metrics 7

3.2. Severity and costing of cardiovascular events 8

3.3. Logistic Regression Model for Probability of Hypertension diagnosis 9

3.4. Putting it all together 11

4. Appendix A- R code used for data manipulation 14

List of Figures

Figure 1- Equation 1 by Age bucket and Metal Level using LM method 7

Figure 2- Average payment by DRG code by state, colored by Severity factor for inpatient services 8

Figure 3- NH11 Data Set, completeness of data objects 10

Figure 4- Effects plot for the logistic regression of hypertension diagnosis 11

Figure 5- Risk adjusted expected payment from hypertension leading to an inpatient event 11

Figure 6- Risk adjusted expected payment by state for hypertension leading to an inpatient event 12

Figure 7- Risk adjusted average payments for cardiology related inpatient events in Missouri 12

Figure 8- Various plan max OOP expenses for ages 60 plus 13

List of Tables

Table 1- DRG Descriptions chosen for hypertension outcome 9

1. Recap on the Project Proposal

The original proposal of the project was centered around comparing average insurance plan values pre and post implementation of the Affordable Care Act to test the hypothesis that given underlying pricing for services was not directly addressed in the 2008 legislation, the value of the plans did not significantly change and amended policies would be required to complete healthcare reform. In analyzing many data sets, however, it became evident that while the plans could be compared on many levels, it was difficult to find a direct comparison metric. The issue became more evident while attempting to find reasonable assumptions for outcome probabilities and their relevant costs. It was then decided to use the many data sets collected in order to focus on one specific outcome for which there were known, published, probabilities of outcomes and costs.

* 1. Description of Problem

About **1 of 3 U.S. adults**—or about **75 million people**—have high blood pressure.[[1]](#footnote-1) **Only about half (54%)** of these people have their high blood pressure under control. This common condition increases the risk for [heart disease](https://www.cdc.gov/heartdisease/index.htm) and [stroke](https://www.cdc.gov/stroke/index.htm), two of the leading causes of death for Americans. [[2]](#footnote-2) Many hospital inpatient cardiac events can be directly linked to patients diagnosed with hypertension that did not take action to keep their blood pressure in check. This project first attempts to model the probability of a hypertension diagnosis given statistically relevant physical attributes such as Body Mass Index (BMI), Age, Sex, and average sleep of a data set of patients. The probabilities for various types of test cases are then computed and compared to known hospital in-patient costs for low, medium and severe outcomes due to cardiac events. Finally, the various state-funded commercial plans for each state are used to give an idea of risk-adjusted Maximum Out-of-Pocket expenses (MOOP) that a patient can expect given their risk profile of developing a diagnosis of hypertension leading to an in-patient hospital event.

* 1. Future Work

The methods described herein can be replicated and summed for any combination of potential patient outcomes and could eventually become a tool for patients to understand their personal health risk profiles and estimate the next couple of years of their health expenses. This will help patients chose a coverage plan in their state that will allow them to get the maximum coverage of health services for the most affordable price to them. A front-end tool could potentially be developed to make the concept more user-friendly.

1. DATA SETS AND BASIC HEALTH CARE KNOWLEDGE

This section describes all the data sets used to produce the final result. The R code, as well as the manipulations to produce the results will be discussed in a further chapter. A basic description of health care specific knowledge required to understand the premise of this project will be provided as the issues arise.

* 1. US Healthcare Data Sets on Kaggle

Kaggle.com provides six different data sets that have information regarding various types of insurance plans currently available on the US healthcare market. Three of the six data sets were used for this project and are discussed in this section.

* + 1. Plan Attributes Data Set

This data set contains 77353 observations of 176 variables. The variables describe various features of the different healthcare plans available by state. Examples of features are: dental only plans, if a plan is applicable for healthcare savings accounts, various disease management programs included in the plans, maximum deductibles and co-pays, etc… After several iterations of data manipulation and focusing on the most important plan attributes, the data set was reduced to 363 observations of 4 variables.

* + 1. Rate Data Set

This data set contains 12,694,445 observations of 24 variables. The variables describe the premium payments associated to various features such as Age, Tobacco, State Code, etc… After several iterations of data manipulation and focusing on the observations that were important for this project, the data set was reduced to 4,022,964 observations of 15 variables.

* + 1. Cross-Walk 2016 Data Set

This data set contains 150,005 observations of 21 variables. It is typically supposed to be used to convert 2015 plans to equivalent 2016 plans, but for the purpose of this project, it was used to associate the Plan’s ‘Metal Level’ to the information collected in the prior two data sets. As such, it was reduced to 3538 observations of 2 variables.

* 1. Centers for Medicare and Medicaid Services (CMS) Provider Inpatient Charge Data Set

The CMS data set for provider inpatient charges contains 202,656 observations of 13 variables. The inpatient data set was chosen because the procedures described are repeatable and have trust-worthy cost assessments as hospital health record systems are more reliable than physician and other out-of-network provider systems. After several data manipulation iterations, the data set was reduced to 1442 observations of 5 variables that focused on low, medium and, high severity cardiac specific outcomes. A brief explanation of the various types of health care usage events is needed to understand why this particular data set was chosen:

* + 1. Inpatient, Outpatient, Emergency Room, and Office Based Events

First and foremost, it is important to understand that patients of the US healthcare system have several choices when it comes to an insurance cost-incurring event. The main categories in which the system has been tracked are: inpatient services, outpatient services, Emergency room services, and office-based outpatient services. There are other types of services such as long-term acute care facilities (LTAC) and Skilled Nursing Facilities (SNF) that are not relevant to direct hypertension related events.

Inpatient Services: a hospital patient who receives lodging and food as well as treatment[[3]](#footnote-3)

Outpatient Services: a patient who is not hospitalized overnight but who visits a hospital, clinic, or associated facility for diagnosis or treatment[[4]](#footnote-4)

Emergency Room (ER) Services: are needed when a patient required immediate care

Office-Based Outpatient Services: scheduled office visits, typically with a primary care physician

As can be seen by the definitions, a very extreme event would likely trigger the need for ER services, as well as inpatient services. Given that the average ER services cost regardless of condition is in the order of $930, and inpatient costs for conditions related to hypertension can go as high as $250,000, one can speculate why the inpatient data set was chosen for this risk-based analysis.

* 1. National Health Records 2011 (NH11) Data Set

The NH11 data set includes 33,014 observations of 36 variables and comes directly from Sprinboard.com’s exercise section for logistic regression. This is the data set used to predict probabilities of hypertension diagnosis given certain dependent variables. The end result is a model data set of 12,160 observations of 7 variables.

By merging key data objects from these five data sets, one can solve the problem put forth in section 1. Further description of the data manipulations and visualizations used are offered in further chapters.

1. DATA MANIPULATION AND VISUALISATION



Equation 1- Patient Out-of-Pocket expense equation

The equation above represents the out-of-pocket (OOP) expense that a patient needs to pay for a healthcare related event. P(n) is the probability of the event occurring, cost is the cost of the event, and minOOP is the premium that the patient needs to pay in order to have insurance. Like car insurance plans, health insurance plans have deductible amounts that must be met on a yearly basis before the insurance plan starts to pay for services. Also like car insurance plans, a monthly or yearly premium (minOOP) must be paid in order to qualify for the insurance product. Therefore, the equation above represents OOP expenses for a given time period as the sum of the probability of events, times the cost of the events, plus the premium paid in the same time period.

The concept of maxOOP happens when the patient reaches the maximum out-of-pocket expenses as defined by the plan. Any health care expense over and above this amount will be covered by the insurance company. It is therefore very important for patients to understand not only the monthly premiums that they need to pay, but the maximum that they would go out-of-pocket should a highly severe event occur.

* 1. Insurance Plan Metrics

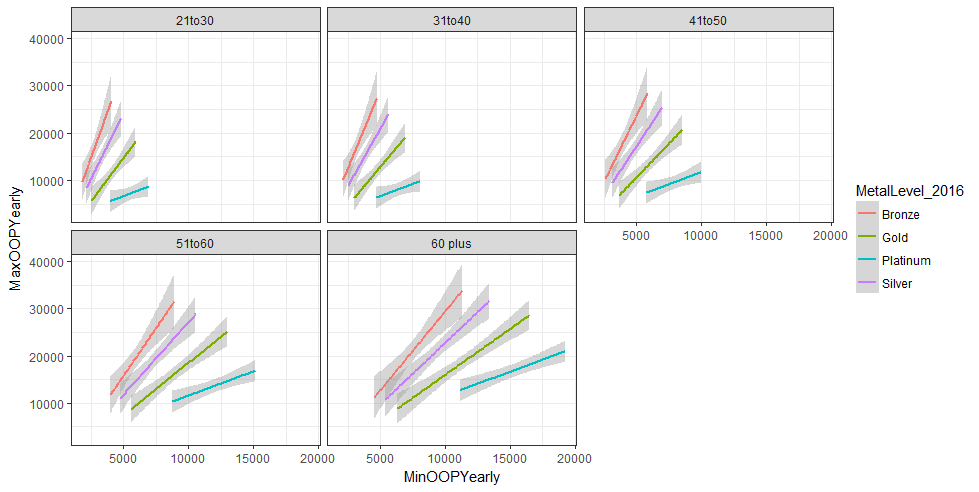


Figure 1- Equation 1 by Age bucket and Metal Level using LM method

Figure 1 shows the output of a geom\_smooth(method = “lm”) plot and has the 95% confidence interval band in grey. It has data split by age buckets, and colors split by plan Metal Level (measure of plan value). The x-axis shows the yearly premium amount that the patient in that age bucket is expected to pay. The y-axis shows the max out-of-pocket expenses associated to such plans.

The plot makes sense as the younger age bracket has the lowest premiums, and lowest spread of maxOOP, while the 60 plus age bracket requires much higher premiums for the same maxOOP. This suggest that an older population have a higher risk-profile, which makes sense.

Furthermore, the various metal colors overlap and allow patients differing options when it comes to choose the plan best suited for them. If a patient feels that they present a low health risk, the bronze plans have the cheapest premiums, but the highest maxOOP, which also makes sense. The variability of the premiums is likely due to inclusions of specific types of coverages depending on medical condition. The data manipulation to produce this graphic was performed by merging key data points from 2.1.1, 2.1.2, and 2.1.3. The R code can be found on line 1 to 95 in Appendix A. This graph was created as a lookup table for a patient who gets the output from the model below. It will help them chose the appropriate plan given their specific risk condition.

* 1. Severity and costing of cardiovascular events

Figure 2 was produced by understanding the underlying data in the CMS data set. The key variable is the ‘DRG Definition’, which describes the outcome linked to the associated cost in the data set. One key observation was that all DRG Definitions in the ‘200’to ‘299’ category were linked to cardiovascular events. A quick filtering of DRG Definitions in the 200’s with a number of states worth of data higher than 40 yields table 1 below.

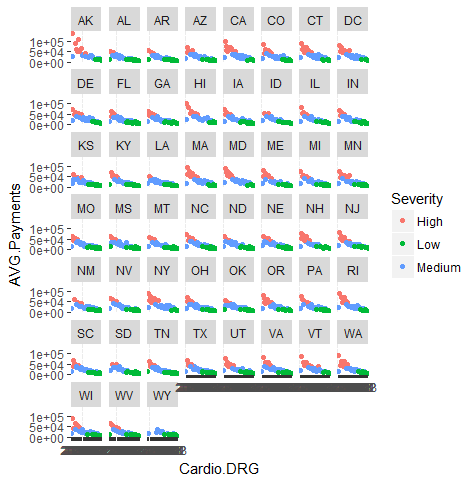


Figure 2- Average payment by DRG code by state, colored by Severity factor for inpatient services



Table 1- DRG Descriptions chosen for hypertension outcome

As can be seen in figure 2, the highest severity outcomes (red) are also the most expensive events. You can also see the variability by State. As an example, Arkansas looks to be the most expensive state for a high severity cardiovascular event. Table 1 shows the key cost data used to estimate OOP expenses from equation 1. The R code with the iterations of data manipulation on the CMS data set can be found on lines 97 to 116 in Appendix A. The last missing piece to solve the equation is the probability of hypertension diagnosis leading to an inpatient event.

* 1. Logistic Regression Model for Probability of Hypertension diagnosis

The last data set required was a model for predicting hypertension diagnosis. The NH11 data set was used to do this. First, we explore the data set to look for which variables are complete enough to perform a regression analysis. Figure 3 below was produced. I did this to ensure that the data set was complete enough for the data objects that interested me to perform the regression analysis. As can be seen, BMI (body mass index), sleep (# hours of sleep per night), hypev (diagnosed with hypertension 1 or 0), sex (M or F), and age\_p (age) are all complete data sets therefore we can proceed with the regression.

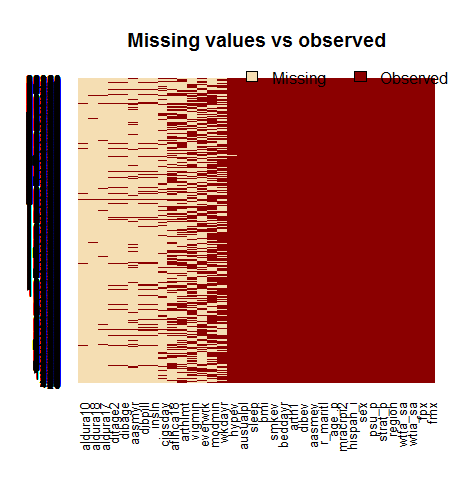
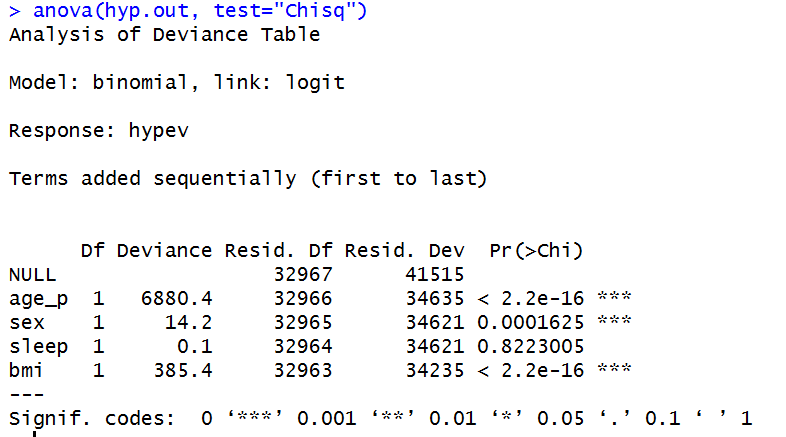


Figure 3- NH11 Data Set, completeness of data objects

A regression model was run in the logistic\_regression exercise from the Springboard Foundations of Data Science course:

hyp.out <- glm(hypev~age\_p+sex+sleep+bmi, data=NH11, family="binomial")

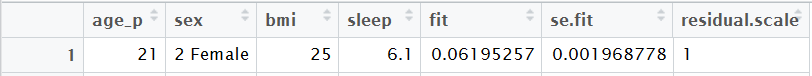
Testing the model for significance:



The model was then ran using a prediction set ‘testcases’ in which 80 observation with varying ages, sexes, BMI’s, and sleep are contained.

predset <- cbind(testcases, predict(hyp.out, type = "response",se.fit = TRUE, interval="confidence",newdata = testcases))

This yielded a data set where the column ‘fit’ corresponds to the probability of that specific observation being diagnosed with hypertension. As an example, the line item below represents a female of age 21, with a BMI of 25, who sleeps 6.1 hours per night. The model predicts that she has a 0.06195 probability of being diagnosed with hypertension, which makes sense given her young age and relatively low BMI. In contrast, a 69 year old male, with a BMI of 39, who sleeps 7 hours a night, has a 0.64424 probability of being diagnosed with hypertension.



Lastly, you can see the effects plot of the dependent factors on the probability of being diagnosed with hypertension below. The only variable for which there is a large confidence band is sleep, which has a minor effect on the overall probability.

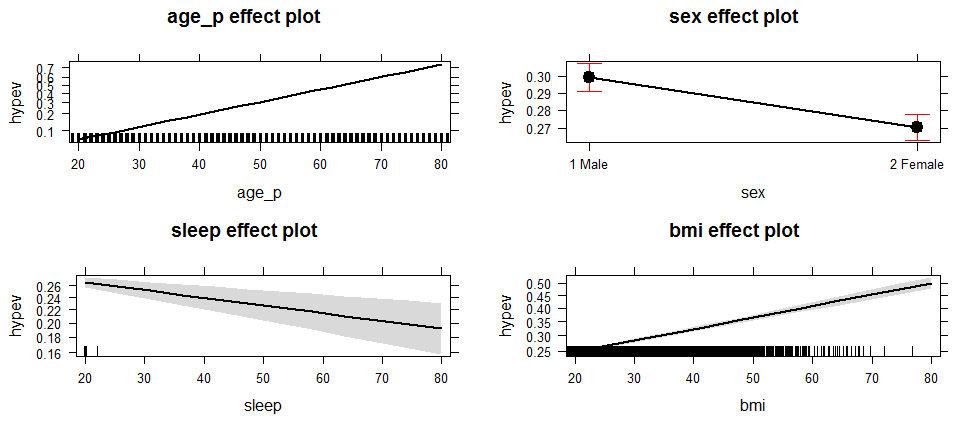


Figure 4- Effects plot for the logistic regression of hypertension diagnosis

* 1. Putting it all together

The final step was to merge the cost data from all the states from 3.2 with the prediction set model output from 3.3, which yields a data frame with various expected payments for the specific case of hypertension leading to an inpatient cardiovascular event. The figures 4 and 5 below show the aggregate and state-specific results.

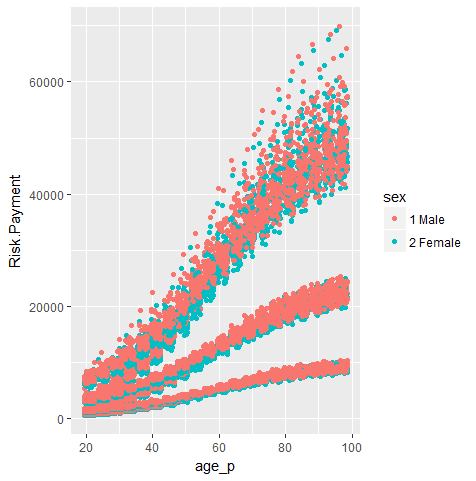


Figure 5- Risk adjusted expected payment from hypertension leading to an inpatient event

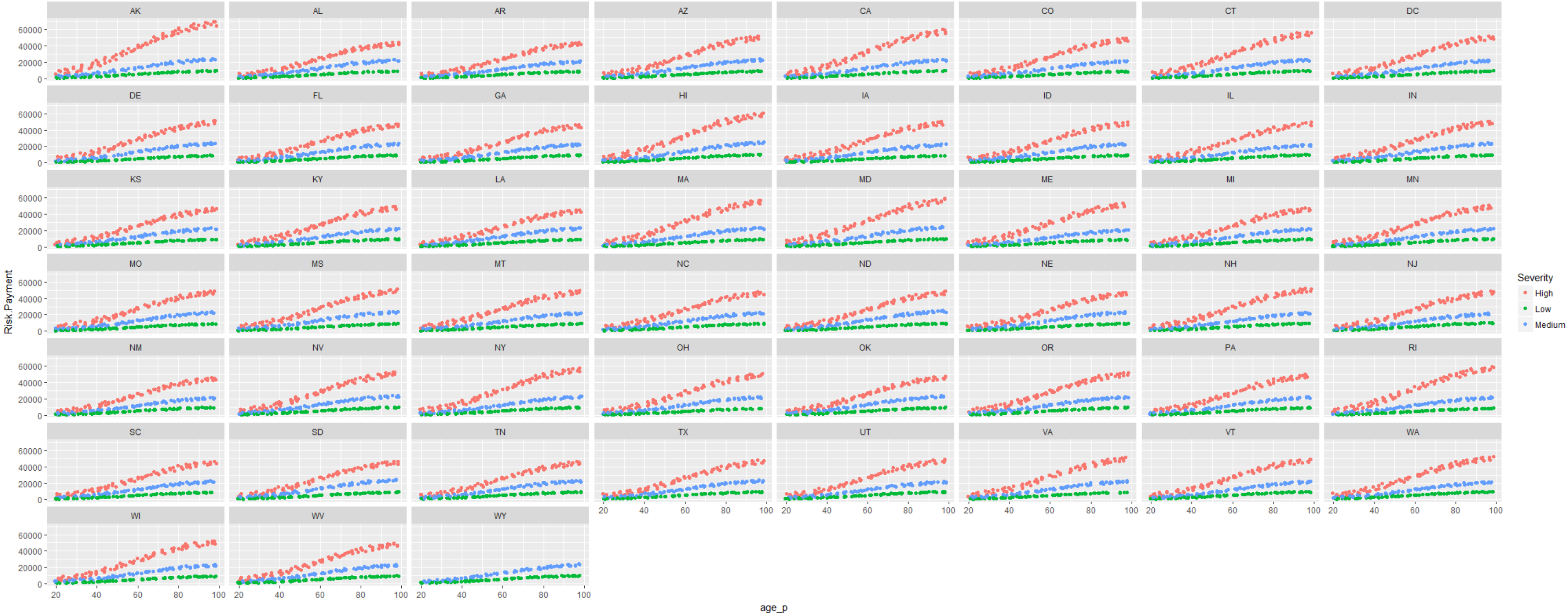


Figure 6- Risk adjusted expected payment by state for hypertension leading to an inpatient event

It can be seen that males present more of a risk than females, which is expected given the coefficients of our regression equation. The 3 severity buckets can clearly be seen as well in terms of their associated costs. In conclusion, with the data manipulation of five disparate data sets, a logistic regression model was produced that allows patients to understand their risk of being diagnosed with hypertension. Furthermore, if that hypertension should lead to an inpatient event, the patient can understand their personal risk-adjusted expected payment for such an event. Even more useful is the ability to select a specific-targeted insurance plan based on the modelling information.

An example would best describe how this model is intended to be used. Let’s take the case of the 69 year old male from section 3.3. Recall that he had a BMI of 39 and slept 7 hours a night. This yielded a fit of 0.64424. Therefore, the probability P(n) with a 95% CI of this person being diagnosed with hypertension is 0.64424. Let’s say the patient happens to live in Missouri, the patient could then derive his risk.adjusted cost for a high (~$35,000), medium (~16,000), and low (~7,500) severity inpatient event as shown by where the red vertical line intercepts the graph lines in figure 7.

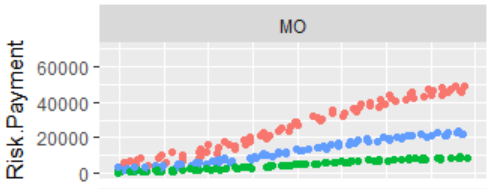


Figure 7- Risk adjusted average payments for cardiology related inpatient events in Missouri

Let’s say that patient can only afford a $625 a month premium payment and is very concerned about not being able to afford the out-of-pocket expense of a high severity event. That patient could then refer to figure 8 and see that he has three plan options: Bronze, Silver, and Gold, each with a different maximum out-of-pocket for the premium he can afford. He would then need to look deeper into the plan’s inclusions to make sure that they cover high severity cardiology events such as the ones in table 1.

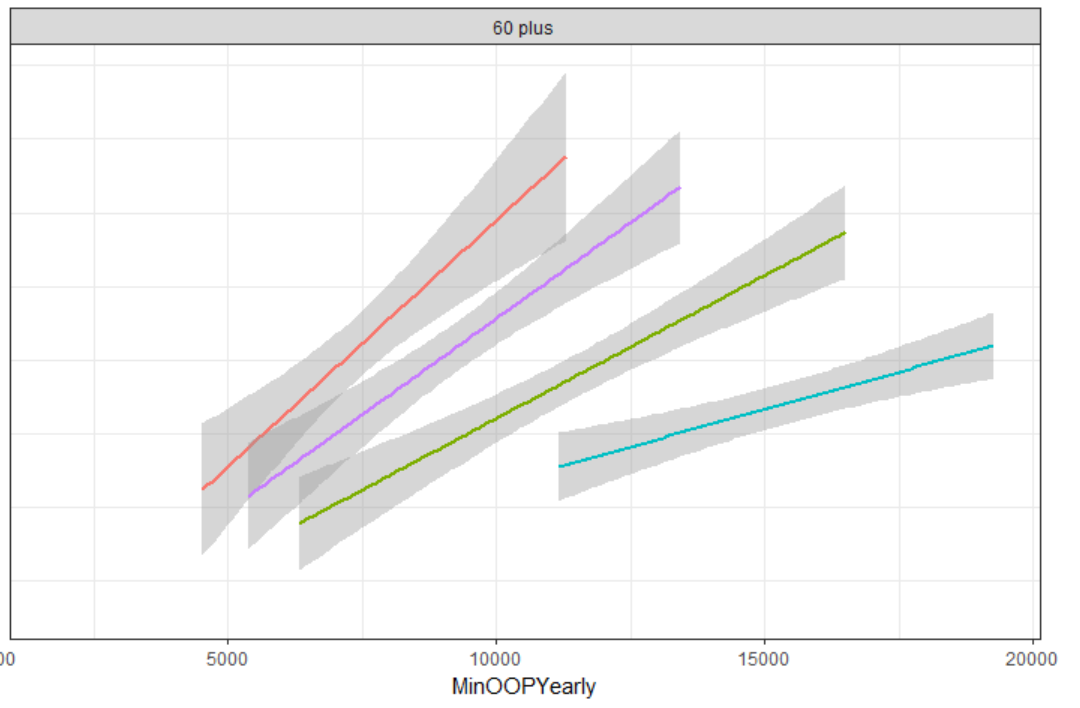


Figure 8- Various plan max OOP expenses for ages 60 plus

It can be seen that integrating many disparate data sources can help patients understand and navigate the very complex world of US healthcare insurance and to fully understand the risks associated to signing up for a plan given their current health status. As suggested in future work, an elaborated model for more risk factors can be integrated using concepts from this model and a front-end, easy to use tool could be developed to help patients understand their health expense risks.

1. Appendix A- R code used for data manipulation

# December 2016- Springboard.com Capstone script for Pierre Carpentier

# Fundamentals of Data Science

# load kaggle data sets for US healthcare policies.

library(readr)

BenefitsCostSharing <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/health-insurance-marketplace-release-2016-01-20-15-52-37/health-insurance-marketplace/BenefitsCostSharing.csv")

BusinessRules <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/health-insurance-marketplace-release-2016-01-20-15-52-37/health-insurance-marketplace/BusinessRules.csv")

Network <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/health-insurance-marketplace-release-2016-01-20-15-52-37/health-insurance-marketplace/Network.csv")

PlanAttributes <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/health-insurance-marketplace-release-2016-01-20-15-52-37/health-insurance-marketplace/PlanAttributes.csv")

Rate <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/health-insurance-marketplace-release-2016-01-20-15-52-37/health-insurance-marketplace/Rate.csv")

ServiceArea <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/health-insurance-marketplace-release-2016-01-20-15-52-37/health-insurance-marketplace/ServiceArea.csv")

Crosswalk2016 <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/health-insurance-marketplace-release-2016-01-20-15-52-37/health-insurance-marketplace/Crosswalk2016.csv")

# Exploratory merging of the data

# Take plans from current business Year as the other years are expired plans and no longer available

library(dplyr)

library(tidyr)

Rate1 <- Rate %>% filter(BusinessYear == "2016") %>% arrange(PlanId)

# Keep only interesting columns and sort Age by groups of 10 year buckets

Rate1 <- data.frame(Rate1$StateCode, Rate1$IssuerId, Rate1$PlanId, Rate1$RatingAreaId, Rate1$Tobacco, Rate1$Age, Rate1$IndividualRate)

Rate1 <- arrange(Rate1, Rate1.Age)

Rate2 <- Rate1 %>% filter(Rate1.Age != "0-20")

Rate2$Rate1.Age <- as.numeric(as.character(Rate2$Rate1.Age))

#Expand age buckets using the mutate function

Rate3 <- Rate2 %>% mutate("21-30" = between(Rate1.Age, 21, 30)) %>% mutate("31-40" = between(Rate1.Age, 31, 40)) %>% mutate ("41-50" = between(Rate1.Age, 41, 50)) %>% mutate("51-60" = between(Rate1.Age, 51, 60)) %>% mutate("60+" = Rate1.Age > 60)

#Now to add basic benefits data to the Rates

Benefits1 <- BenefitsCostSharing %>% filter(BusinessYear == 2016)

Attributes1 <- PlanAttributes %>% filter(BusinessYear == 2016 & DentalOnlyPlan == "No")

MetalLevel <- Crosswalk2016 %>% group\_by(PlanID\_2016, MetalLevel\_2016) %>% summarise()

names(MetalLevel)[1] <- paste("Rate1.PlanId")

Rate4 <- left\_join(Rate3, MetalLevel, by = "Rate1.PlanId")

# Get Plan attributes in order to setup cost calculation

Attributes2 <- Attributes1 %>% group\_by(HIOSProductId, AVCalculatorOutputNumber, TEHBCombInnOonIndividualMOOP, TEHBDedCombInnOonIndividual, BenefitPackageId, FormularyId) %>% summarise()

names(Attributes2)[1] <- paste("Rate1.PlanIdshort")

Rate4 <- Rate4 %>% mutate(Rate1.PlanIdshort = strtrim(Rate4$Rate1.PlanId, 10))

# Split Rates by age buckets and Metal Level to setup yearly max spending for individuals

Rate21to30 <- Rate4 %>% filter(Rate4[8] == TRUE) %>% group\_by(Rate1.PlanIdshort, MetalLevel\_2016) %>% mutate(AVG.IndRate = mean(Rate1.IndividualRate)) %>% arrange(Rate1.PlanId)

Rate31to40 <- Rate4 %>% filter(Rate4[9] == TRUE) %>% group\_by(Rate1.PlanIdshort, MetalLevel\_2016) %>% mutate(AVG.IndRate = mean(Rate1.IndividualRate)) %>% arrange(Rate1.PlanId)

Rate41to50 <- Rate4 %>% filter(Rate4[10] == TRUE) %>% group\_by(Rate1.PlanIdshort, MetalLevel\_2016) %>% mutate(AVG.IndRate = mean(Rate1.IndividualRate)) %>% arrange(Rate1.PlanId)

Rate51to60 <- Rate4 %>% filter(Rate4[11] == TRUE) %>% group\_by(Rate1.PlanIdshort, MetalLevel\_2016) %>% mutate(AVG.IndRate = mean(Rate1.IndividualRate)) %>% arrange(Rate1.PlanId)

Rate60plus <- Rate4 %>% filter(Rate4[12] == TRUE) %>% group\_by(Rate1.PlanIdshort, MetalLevel\_2016) %>% mutate(AVG.IndRate = mean(Rate1.IndividualRate)) %>% arrange(Rate1.PlanId)

#Bind them together to get the final Individual RAte file

RateFinal <- bind\_rows(Rate21to30, Rate31to40, Rate41to50, Rate51to60, Rate60plus)

# Group by Short ID and Metal LEvel to get short table versions of the data by age group

Rate21to30G <- Rate21to30 %>% group\_by(Rate1.StateCode, Rate1.PlanIdshort, MetalLevel\_2016, AVG.IndRate) %>% summarise() %>% arrange(Rate1.PlanIdshort)

Rate21to30G$Age <- "21to30"

Rate31to40G <- Rate31to40 %>% group\_by(Rate1.StateCode, Rate1.PlanIdshort, MetalLevel\_2016, AVG.IndRate) %>% summarise() %>% arrange(Rate1.PlanIdshort)

Rate31to40G$Age <- "31to40"

Rate41to50G <- Rate41to50 %>% group\_by(Rate1.StateCode, Rate1.PlanIdshort, MetalLevel\_2016, AVG.IndRate) %>% summarise() %>% arrange(Rate1.PlanIdshort)

Rate41to50G$Age <- "41to50"

Rate51to60G <- Rate51to60 %>% group\_by(Rate1.StateCode, Rate1.PlanIdshort, MetalLevel\_2016, AVG.IndRate) %>% summarise() %>% arrange(Rate1.PlanIdshort)

Rate51to60G$Age <- "51to60"

Rate60plusG <- Rate60plus %>% group\_by(Rate1.StateCode, Rate1.PlanIdshort, MetalLevel\_2016, AVG.IndRate) %>% summarise() %>% arrange(Rate1.PlanIdshort)

Rate60plusG$Age <- "60 plus"

#Bind them together to get shortened version for data exploration

RateFinalG <- bind\_rows(Rate21to30G, Rate31to40G, Rate41to50G, Rate51to60G, Rate60plusG)

#Now to do some visualizations

library(ggplot2)

ggplot(RateFinalG, aes(x = Age, y = AVG.IndRate, col = MetalLevel\_2016)) + geom\_point() + facet\_wrap(~Rate1.StateCode)

ggplot(RateFinalG, aes(x = Age, y = AVG.IndRate, col = MetalLevel\_2016)) + geom\_jitter()

ggplot(RateFinalG, aes(x = Age, y = AVG.IndRate, fill = MetalLevel\_2016)) + geom\_bar(position = "dodge", stat = "identity") + coord\_flip()

ggplot(RateFinalG, aes(x = Age, y = AVG.IndRate, fill = MetalLevel\_2016)) + geom\_bar(position = "dodge", stat = "identity") + coord\_flip() + facet\_wrap(~Rate1.StateCode)

#Need to add more attributes plus remove NA's and group Low, High and Catastrophic, and Bronze, Silver, GOld, Platinum together

#Note: Performed Data manipualtion for Attributes 2 in excel, explained in final report

Attributes2 <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/Attributes2.csv",

+ col\_types = cols(TEHBCombInnOonIndividualMOOP = col\_number(),

+ TEHBDedCombInnOonIndividual = col\_number()))

Attributes3 <- Attributes2 %>% group\_by(Rate1.PlanIdshort, MetalLevel\_2016) %>% summarise(AVG.AV = mean(AVCalculatorOutputNumber), AVG.MOOP = mean(TEHBCombInnOonIndividualMOOP))

RateFinaljoin <- left\_join(RateFinalG, Attributes3, by = c("Rate1.PlanIdshort", "MetalLevel\_2016"))

#Note: Perform some data manipulation in Excel, explained in final report

RateFinaljoin <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/RateFinaljoin.csv")

RateFinaljoinfilter <- RateFinaljoin %>% filter(MaxOOPYearly != "NA")

#Some more visual exploration

ggplot(RateFinaljoinfilter, aes(x = MinOOPYearly, y = MaxOOPYearly, col = MetalLevel\_2016)) + geom\_point()

ggplot(RateFinaljoinfilter, aes(x = MinOOPYearly, y = MaxOOPYearly, col = MetalLevel\_2016)) + geom\_smooth(method = "lm") + facet\_wrap(~Age) + theme\_bw()

#creating a map

library(mapdata)

library(maps)

RateFinaljoinfilter$region <- state.name[match(RateFinaljoinfilter$Rate1.StateCode, state.abb)]

RateFinaljoinfilter$region <- tolower(RateFinaljoinfilter$region)

TotalMap <- merge(states, RateFinaljoinfilter, by = "region")

ggplot() + geom\_polygon(data = TotalMap, aes(x=long, y=lat, group=group, fill=TotalMap$MaxOOPYearly), colour = "white") + scale\_fill\_continuous(low="thistle2",high="darkred", guide="colorbar") + theme\_bw()

#plot showing linear regression models for each metal level and age group

ggplot(RateFinaljoinfilter, aes(x = MinOOPYearly, y = MaxOOPYearly, col = MetalLevel\_2016)) + stat\_summary(fun.data = mean\_cl\_normal) + geom\_smooth(method = 'lm') + facet\_wrap(~Age) + theme\_bw()

# LoadHospital inpatient data set and manipulate to get Profiles

Medicare\_Provider\_Charge\_Inpatient\_DRGALL\_FY2014 <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/health-insurance-marketplace-release-2016-01-20-15-52-37/health-insurance-marketplace/Medicare\_Provider\_Charge\_Inpatient\_DRGALL\_FY2014.csv")

Inpatient1 <- Medicare\_Provider\_Charge\_Inpatient\_DRGALL\_FY2014 %>% group\_by(`Provider State`, `DRG Definition`) %>% mutate(AVG.Payments = mean(`AVG Total Payments 2016`)) %>% arrange(`Provider State`, `DRG Definition`)

Inpatient2 <- Inpatient1 %>% group\_by(`Provider State`, `DRG Definition`, AVG.Payments) %>% summarise()

colnames(Inpatient2) <- c("Provider.State", "DRG.Definitions", "AVG.Payments")

# Selected cost buckets related to cardiology outcomes, with data for over 45 states, performed exploration with excel using pivot tables.

Cardio.DRG <- data.frame(c(216, 219, 233, 217, 235, 237, 227, 220, 239, 236, 242, 252, 238, 246, 248, 264, 253, 243, 208, 286, 251, 249, 283, 254, 280, 291, 287, 281, 292, 293))

colnames(Cardio.DRG) <- "Cardio.DRG"

Cardio.DRG$Cardio.DRG <- as.character(Cardio.DRG$Cardio.DRG)

Inpatient2 <- separate(Inpatient2, DRG.Definitions, c("x","y"), sep = " - ")

colnames(Inpatient2) <- c("Provider.State", "Cardio.DRG", "Definitions", "AVG.Payments")

Inpatient3 <- semi\_join(Inpatient2, Cardio.DRG, by = "Cardio.DRG")

Inpatient3$Cardio.DRG <- as.factor(Inpatient3$Cardio.DRG)

ggplot(Inpatient3, aes(x=Cardio.DRG, y=AVG.Payments, col=Provider.State)) + geom\_jitter()

ggplot(Inpatient3, aes(x=Cardio.DRG, y=AVG.Payments)) + geom\_jitter() + facet\_wrap(~Provider.State)

#Did some data manipulation in excel to get Severity factors in Inpatient3

Inpatient3 <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/Inpatient3.csv")

Inpatient3$Cardio.DRG <- as.factor(Inpatient3$Cardio.DRG)

ggplot(Inpatient3, aes(x=Cardio.DRG, y=AVG.Payments, col= Severity)) + geom\_jitter() + facet\_wrap(~Provider.State)

# test the model hypev from Logistic regression exercise

anova(hyp.out, test="Chisq")

#Generate test cases to get probabilities for model output

testcases <- read\_csv("C:/Users/greed/OneDrive/Documents/Springboard/Capstone/testcases.csv")

predset <- cbind(testcases, predict(hyp.out, type = "response",se.fit = TRUE, interval="confidence",newdata = testcases))

# use probability of hypertension leading to an in-patient procedure to calculate out of pocket expenses for the predset

Payments <- Inpatient3 %>% group\_by(Provider.State, Severity) %>% mutate(Payment = mean(AVG.Payments))

Payments <- Payments %>% group\_by(Provider.State, Severity, Payment) %>% summarise

predsetexpand <- expandRows(predset, count = 152, count.is.col = FALSE)

Paymentsexpand <- expandRows(Payments, count = 80, count.is.col = FALSE)

Fullpredset <- bind\_cols(predsetexpand, Paymentsexpand)

Fullpredset <- Fullpredset %>% mutate(Risk.Payment = Payment\*fit)

ggplot(Fullpredset, aes(x = age\_p, y = Risk.Payment, col = sex)) + geom\_jitter()

ggplot(Fullpredset, aes(x = age\_p, y = Risk.Payment, col = Severity)) + geom\_jitter() + facet\_wrap(~Provider.State)

1. Merai R, Siegel C, Rakotz M, Basch P, Wright J, Wong B; DHSc., Thorpe P. CDC Grand Rounds: A Public Health Approach to Detect and Control Hypertension. MMWR Morb Mortal Wkly Rep. 2016 Nov 18;65(45):1261-1264 [↑](#footnote-ref-1)
2. Merai R, Siegel C, Rakotz M, Basch P, Wright J, Wong B; DHSc., Thorpe P. CDC Grand Rounds: A Public Health Approach to Detect and Control Hypertension. MMWR Morb Mortal Wkly Rep. 2016 Nov 18;65(45):1261-1264 [↑](#footnote-ref-2)
3. https://www.merriam-webster.com/dictionary/inpatient [↑](#footnote-ref-3)
4. https://www.merriam-webster.com/dictionary/outpatient [↑](#footnote-ref-4)