Medical Insurance

Chandan Gowda

9/30/2020

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

1. BACKGROUND / CONTEXT
2. Domain The purpose of insurance is to provide protection against the risk of any financial loss. It is a contract between the insurer and the insured individual – where the insurer agrees to take the risk of the insured individual against either a future event or uncertain losses in return for a monetary compensation from the insured individual known as a premium. It is a form of risk management. In addition, insurance not only helps by mitigating risks but can also provide a financial “cushion” against financial burdens suffered. In short, insurance is a safeguard against the uncertainties of life.

Regarding insurance, there are several types: life, health, vehicle, general liability, property, worker’s compensation, and travel insurance.

In this use case, we will be focusing on health insurance. Health insurance provides coverage against the medical expenses of the insured. Among what is covered is medication, doctor visits, hospital stays and other medical expenses.

1. Brief Description of the Scenario Insurance is based largely on the understanding of risk. Due to this, data plays a crucial role in helping companies make better and well-informed business decisions. Patients should not be dealt with as one pool, rather each patient should be grouped with others that share similar characteristics. In order to determine and predict the insurance policy costs for individuals, we must assess the individual’s characteristics. These characteristics include age, sex, BMI, number of children, smoking or non-smoking, and the region they reside in. Do these characteristics create an impact on insurance premiums? If so, which characteristics create the most impact and which do not? If no, what else impacts the insurance premiums besides these characteristics?
2. Decisions of Interest To approach the question of how to determine insurance costs, we will be using two analytical models: clustering and regression.

Cluster analysis is one of the basic models that can be applied into analyzing large data sets. Insured individuals are distinguished into groups with certain, common characteristics. Once we segment the insured individuals into homogeneous groups, it can help calculate future policy rates. The data of each group can be explored, analyzed, and modeled. From there, we can build predictive models for each homogeneous group.

Regression is another model to predict a continuous or discrete outcome, and it allows you to examine the relationship between two or more variables. With the help of regression, we can determine which input variables (age, sex, BMI, number of children, smoking or non-smoking, and region) are influencing insurance costs. In addition, regression shall be applied per clustered groups in order to provide a higher accuracy.

1. BUSINESS UNDERSTANDING
2. Business Objective The business objective is to classify by segments, policy holders from the pool of data by clustering the characteristics which are maintained as age, sex, body mass index (BMI), number of children and smoking status. Studying those characteristics and properly identifying relevant and accurate segments of policy holders will provide a better understanding of the factors influencing the medical charge against each patient. In order to accurately define the relevant segments and factors, we must understand the business closely from the business domain expert, look for patterns in our data analysis and gain extensive insights to extract useful information out of the data. Finally, we will predict future medical charges for the produced segments through multiple linear regression using the gathered characteristics to assist our client (“Insurance Company X”) in the decision-making process of charging an accurate insurance premium to customers. This would be beneficial to both the insurance company and the beneficiary. The client would be able to plan its business, adjust pricing and target the right customers. The beneficiary would be able to find to a suitable plan that’s not one-size-fits-all.
3. Situation Assessment These are the research questions that we shall answer over the course of the project:
4. Data assessment (missing values/formats/sources), model of choice, method of analysis?
5. What is the input specification that have been used for segmenting policy holders?
6. How many clusters have been generated, and the basis for the segmentation?
7. What is the unsupervised learning method that have been used for segmenting policy holders?
8. What are the performance measures and evaluation results for segmenting policy holders?
9. What is the input specification that have been used for cost prediction?
10. What are the supervised learning methods that have been used for cost prediction?
11. What are the performance measures and evaluation results for cost prediction?
12. What are the key limitations that were faced while analyzing the data?
13. What are future improvements that are suggested based on current data and models results?
14. Data Mining Goals Data mining tasks can be classified into description methods and prediction methods. In this project we will use both unsupervised and supervised learning. Unsupervised learning (i.e. inputs with no given labels at the output) shall be used for data segmentation into meaningful sets or groups while supervised learning (i.e. input with a given label at the output) shall be used for predicting a single “target” or “outcome” variable.

First, we will use cluster analysis which is about finding groups of objects such that the objects in a group is similar (or related) to one another and different from (or unrelated to) the objects in other groups. This model uses descriptive analytics. The goal is to subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

Then, we will use regression which is about predicting a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency. This model uses predictive analytics. The goal is to predict numerical target (outcome) variable, which is the charges in our case against age, sex, BMI, number of children, region and smoking status.

1. DATA UNDERSTANDING
2. Data Requirements Business understanding determines the data gathering requirements. Based on domain knowledge and challenges, cross-functional meetings are held to govern and analyze all assumed data requirements, covering business scope of work and opportunity at hand.

During data requirements, consolidation of data sources, types and formats are pooled into a data lake.

So, to answer the question, what data do we need? We need to address the dimensionality of problem from dual angels (business and analytical perspectives) to reach a relevant model construction.

The above question is nicely summarized in the business understanding section of the project proposal. We will be using a dataset that contains data about age of primary beneficiary, sex, BMI (body mass index) providing understanding of the body regarding weight and height, region of a person where he/she lives. The data also provides information about the number of children covered by health insurance and whether the person has a smoking habit or not.

In this project, we will be predicting the premium insurance for policy holders based on the information provided and how these above categories mentioned would be responsible in the predictive analysis. After carefully looking through the data, we will make use of both unsupervised and supervised learning. To segment the data, we will be using clustering process and apply regression on each cluster formed to predict the target variable.

1. Describe Data We will be working on ‘insurance.csv’ data set, which is a cross sectional data. Variables: Description Age: Age of primary beneficiary Sex: Gender of Female or Male BMI: Body mass index, objective index of body weight using the ratio of height to weight Children: Number of children covered by health insurance Smoker: Status of Yes or No Region: The policyholder’s residential area in the USA which includes, northeast, southeast, southwest and northwest Charges: Individual medical costs billed by Health Insurance

Now, since we got a brief introduction about data understanding and requirements about the dataset, we will now begin with the coding to review the data in R.

if(!require("pacman"))   
install.packages("pacman")

## Loading required package: pacman

## Warning: package 'pacman' was built under R version 3.6.3

pacman::p\_load(ggplot2, caret, rmarkdown, corrplot, knitr, factoextra, forecast, NbClust, fpc, GGally, psych, tidyverse,colorspace, clustMixType, cluster, purrr)  
search()

## [1] ".GlobalEnv" "package:cluster" "package:clustMixType"  
## [4] "package:colorspace" "package:forcats" "package:stringr"   
## [7] "package:dplyr" "package:purrr" "package:readr"   
## [10] "package:tidyr" "package:tibble" "package:tidyverse"   
## [13] "package:psych" "package:GGally" "package:fpc"   
## [16] "package:NbClust" "package:forecast" "package:factoextra"   
## [19] "package:knitr" "package:corrplot" "package:rmarkdown"   
## [22] "package:caret" "package:lattice" "package:ggplot2"   
## [25] "package:pacman" "package:stats" "package:graphics"   
## [28] "package:grDevices" "package:utils" "package:datasets"   
## [31] "package:methods" "Autoloads" "package:base"

theme\_set(theme\_classic())  
options(digits = 3)  
insurance\_data<- read.csv("insurance.csv", header = TRUE)

1. Sources We got the description knowledge of the ‘Medical Cost Personal Datasets-Insurance Forecast’ by using Clustering and Regression from the Kaggle website.

<https://www.kaggle.com/mirichoi0218/insurance>

The data set ‘insurance.csv’ pertaining to the above domain is also obtained from the Kaggle website.

<https://www.kaggle.com/mirichoi0218/insurance#insurance.csv>

1. Quality The dataset ‘Insurance.csv’ obtained from the Kaggle website is not a real dataset i.e., this data was not collected using actual people. This is simulated data based on demographic statistics from the US Census Bureau.

After reviewing the data set in R, we were able to find that ‘insurance.csv’ dataset shows no missing values. We can see in the above summary() results that there is no NA value or missing value. Data accuracy is one of the components of data quality. The above data is very accurate and this can be confirmed with the above function str(), which provides the overall structure of the dataset. The data values stored in the object ‘insurance\_data’ are the right values which is represented in a consistent and unambiguous form.

The above data is very much relevant and enough for predicting premium insurance policy holders using clustering and regression with R. The person who will be charged more would be the premium policy holder and we can easily get that information from the variable ‘Charges’. With the help of this information, we will be able to predict which premium would be applied for current /new policy holders.

1. DATA PREPARATION Data preparation or pre- processing is the act of manipulating raw data which may come from different data sources into a form that can readily and accurately be analyzed for business purposes. Data preparation is the first step in data mining projects which can include many discrete tasks such as selecting the data, cleaning and preparing the data for building decision support models. Let us discuss each one, and how we implemented various analysis to prepare the insurance.csv dataset.
2. Data Selection Data Selection is the process where data relevant to the analysis task are retrieved from the database or source. Sometimes data transformation and consolidation are performed before the data selection process. The data set is obtained from Kaggle website, and selection of the dataset was purely based on the following CRISP selection specification requirements which are presented below.
3. Atomic- Singular, describing only one concept
4. Traceable- Indicating relationship among data points
5. Consistent- Steady and stable data
6. Accurate- Precise, correct and unambiguous
7. Clear- Readily understood by the user
8. Complete- Description and contextual references are correct
9. Concise- Brief but comprehensive Considering these above factors into account, data selection was carefully done in R by conducting various visualizations on the insurance.csv dataset. Visualization of the data was mainly done to understand the dataset and ensure whether the data selected, aptly solves the business problem indicated previously. Moreover, this is a simulated dataset containing hypothetical medical expenses for patients in the United States. The data was created using demographic statistics from the US Census Bureau, and thus, approximately reflect real-world conditions. The quality of the data set was in its pure form, with no NA or unwanted information present. Thus, the insurance.csv dataset is very much relevant and enough for predicting premium insurance policy holders using clustering and regression using R. It is important to give some thought to how the variables presented above in the insurancve.csv dataset may be related to billed medical expenses. For instance, we might expect that older people and smokers are at higher risk of large medical expenses. Unlike many other machine learning methods, in regression analysis, the relationships among the features are typically specified by the user rather than being detected automatically. Let us explore some of these potential relationships in the coming section.
10. Data Cleaning Data cleaning is a technique that is applied to remove the noisy data and correct the inconsistencies in data. Data cleaning involves transformations to correct the wrong data. Data cleaning is the second step in data preparation performed as a data pre-processing step while preparing the data for building a decision support model. Quality of the data is critical in getting to final analysis. Any data which tend to be incomplete, noisy and inconsistent can affect the results. Let us explore the data and clean it as needed. The first approach is to understand the structure of the data and see if any defects or error is present.

names(insurance\_data)

## [1] "age" "sex" "bmi" "children" "smoker" "region" "charges"

summary(insurance\_data) # --> Shows No missing values

## age sex bmi children smoker   
## Min. :18.0 female:662 Min. :16.0 Min. :0.00 no :1064   
## 1st Qu.:27.0 male :676 1st Qu.:26.3 1st Qu.:0.00 yes: 274   
## Median :39.0 Median :30.4 Median :1.00   
## Mean :39.2 Mean :30.7 Mean :1.09   
## 3rd Qu.:51.0 3rd Qu.:34.7 3rd Qu.:2.00   
## Max. :64.0 Max. :53.1 Max. :5.00   
## region charges   
## northeast:324 Min. : 1122   
## northwest:325 1st Qu.: 4740   
## southeast:364 Median : 9382   
## southwest:325 Mean :13270   
## 3rd Qu.:16640   
## Max. :63770

str(insurance\_data)

## 'data.frame': 1338 obs. of 7 variables:  
## $ age : int 19 18 28 33 32 31 46 37 37 60 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...  
## $ bmi : num 27.9 33.8 33 22.7 28.9 ...  
## $ children: int 0 1 3 0 0 0 1 3 2 0 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...  
## $ charges : num 16885 1726 4449 21984 3867 ...

class(insurance\_data)

## [1] "data.frame"

dim(insurance\_data)

## [1] 1338 7

rownames(insurance\_data)

## [1] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10"   
## [11] "11" "12" "13" "14" "15" "16" "17" "18" "19" "20"   
## [21] "21" "22" "23" "24" "25" "26" "27" "28" "29" "30"   
## [31] "31" "32" "33" "34" "35" "36" "37" "38" "39" "40"   
## [41] "41" "42" "43" "44" "45" "46" "47" "48" "49" "50"   
## [51] "51" "52" "53" "54" "55" "56" "57" "58" "59" "60"   
## [61] "61" "62" "63" "64" "65" "66" "67" "68" "69" "70"   
## [71] "71" "72" "73" "74" "75" "76" "77" "78" "79" "80"   
## [81] "81" "82" "83" "84" "85" "86" "87" "88" "89" "90"   
## [91] "91" "92" "93" "94" "95" "96" "97" "98" "99" "100"   
## [101] "101" "102" "103" "104" "105" "106" "107" "108" "109" "110"   
## [111] "111" "112" "113" "114" "115" "116" "117" "118" "119" "120"   
## [121] "121" "122" "123" "124" "125" "126" "127" "128" "129" "130"   
## [131] "131" "132" "133" "134" "135" "136" "137" "138" "139" "140"   
## [141] "141" "142" "143" "144" "145" "146" "147" "148" "149" "150"   
## [151] "151" "152" "153" "154" "155" "156" "157" "158" "159" "160"   
## [161] "161" "162" "163" "164" "165" "166" "167" "168" "169" "170"   
## [171] "171" "172" "173" "174" "175" "176" "177" "178" "179" "180"   
## [181] "181" "182" "183" "184" "185" "186" "187" "188" "189" "190"   
## [191] "191" "192" "193" "194" "195" "196" "197" "198" "199" "200"   
## [201] "201" "202" "203" "204" "205" "206" "207" "208" "209" "210"   
## [211] "211" "212" "213" "214" "215" "216" "217" "218" "219" "220"   
## [221] "221" "222" "223" "224" "225" "226" "227" "228" "229" "230"   
## [231] "231" "232" "233" "234" "235" "236" "237" "238" "239" "240"   
## [241] "241" "242" "243" "244" "245" "246" "247" "248" "249" "250"   
## [251] "251" "252" "253" "254" "255" "256" "257" "258" "259" "260"   
## [261] "261" "262" "263" "264" "265" "266" "267" "268" "269" "270"   
## [271] "271" "272" "273" "274" "275" "276" "277" "278" "279" "280"   
## [281] "281" "282" "283" "284" "285" "286" "287" "288" "289" "290"   
## [291] "291" "292" "293" "294" "295" "296" "297" "298" "299" "300"   
## [301] "301" "302" "303" "304" "305" "306" "307" "308" "309" "310"   
## [311] "311" "312" "313" "314" "315" "316" "317" "318" "319" "320"   
## [321] "321" "322" "323" "324" "325" "326" "327" "328" "329" "330"   
## [331] "331" "332" "333" "334" "335" "336" "337" "338" "339" "340"   
## [341] "341" "342" "343" "344" "345" "346" "347" "348" "349" "350"   
## [351] "351" "352" "353" "354" "355" "356" "357" "358" "359" "360"   
## [361] "361" "362" "363" "364" "365" "366" "367" "368" "369" "370"   
## [371] "371" "372" "373" "374" "375" "376" "377" "378" "379" "380"   
## [381] "381" "382" "383" "384" "385" "386" "387" "388" "389" "390"   
## [391] "391" "392" "393" "394" "395" "396" "397" "398" "399" "400"   
## [401] "401" "402" "403" "404" "405" "406" "407" "408" "409" "410"   
## [411] "411" "412" "413" "414" "415" "416" "417" "418" "419" "420"   
## [421] "421" "422" "423" "424" "425" "426" "427" "428" "429" "430"   
## [431] "431" "432" "433" "434" "435" "436" "437" "438" "439" "440"   
## [441] "441" "442" "443" "444" "445" "446" "447" "448" "449" "450"   
## [451] "451" "452" "453" "454" "455" "456" "457" "458" "459" "460"   
## [461] "461" "462" "463" "464" "465" "466" "467" "468" "469" "470"   
## [471] "471" "472" "473" "474" "475" "476" "477" "478" "479" "480"   
## [481] "481" "482" "483" "484" "485" "486" "487" "488" "489" "490"   
## [491] "491" "492" "493" "494" "495" "496" "497" "498" "499" "500"   
## [501] "501" "502" "503" "504" "505" "506" "507" "508" "509" "510"   
## [511] "511" "512" "513" "514" "515" "516" "517" "518" "519" "520"   
## [521] "521" "522" "523" "524" "525" "526" "527" "528" "529" "530"   
## [531] "531" "532" "533" "534" "535" "536" "537" "538" "539" "540"   
## [541] "541" "542" "543" "544" "545" "546" "547" "548" "549" "550"   
## [551] "551" "552" "553" "554" "555" "556" "557" "558" "559" "560"   
## [561] "561" "562" "563" "564" "565" "566" "567" "568" "569" "570"   
## [571] "571" "572" "573" "574" "575" "576" "577" "578" "579" "580"   
## [581] "581" "582" "583" "584" "585" "586" "587" "588" "589" "590"   
## [591] "591" "592" "593" "594" "595" "596" "597" "598" "599" "600"   
## [601] "601" "602" "603" "604" "605" "606" "607" "608" "609" "610"   
## [611] "611" "612" "613" "614" "615" "616" "617" "618" "619" "620"   
## [621] "621" "622" "623" "624" "625" "626" "627" "628" "629" "630"   
## [631] "631" "632" "633" "634" "635" "636" "637" "638" "639" "640"   
## [641] "641" "642" "643" "644" "645" "646" "647" "648" "649" "650"   
## [651] "651" "652" "653" "654" "655" "656" "657" "658" "659" "660"   
## [661] "661" "662" "663" "664" "665" "666" "667" "668" "669" "670"   
## [671] "671" "672" "673" "674" "675" "676" "677" "678" "679" "680"   
## [681] "681" "682" "683" "684" "685" "686" "687" "688" "689" "690"   
## [691] "691" "692" "693" "694" "695" "696" "697" "698" "699" "700"   
## [701] "701" "702" "703" "704" "705" "706" "707" "708" "709" "710"   
## [711] "711" "712" "713" "714" "715" "716" "717" "718" "719" "720"   
## [721] "721" "722" "723" "724" "725" "726" "727" "728" "729" "730"   
## [731] "731" "732" "733" "734" "735" "736" "737" "738" "739" "740"   
## [741] "741" "742" "743" "744" "745" "746" "747" "748" "749" "750"   
## [751] "751" "752" "753" "754" "755" "756" "757" "758" "759" "760"   
## [761] "761" "762" "763" "764" "765" "766" "767" "768" "769" "770"   
## [771] "771" "772" "773" "774" "775" "776" "777" "778" "779" "780"   
## [781] "781" "782" "783" "784" "785" "786" "787" "788" "789" "790"   
## [791] "791" "792" "793" "794" "795" "796" "797" "798" "799" "800"   
## [801] "801" "802" "803" "804" "805" "806" "807" "808" "809" "810"   
## [811] "811" "812" "813" "814" "815" "816" "817" "818" "819" "820"   
## [821] "821" "822" "823" "824" "825" "826" "827" "828" "829" "830"   
## [831] "831" "832" "833" "834" "835" "836" "837" "838" "839" "840"   
## [841] "841" "842" "843" "844" "845" "846" "847" "848" "849" "850"   
## [851] "851" "852" "853" "854" "855" "856" "857" "858" "859" "860"   
## [861] "861" "862" "863" "864" "865" "866" "867" "868" "869" "870"   
## [871] "871" "872" "873" "874" "875" "876" "877" "878" "879" "880"   
## [881] "881" "882" "883" "884" "885" "886" "887" "888" "889" "890"   
## [891] "891" "892" "893" "894" "895" "896" "897" "898" "899" "900"   
## [901] "901" "902" "903" "904" "905" "906" "907" "908" "909" "910"   
## [911] "911" "912" "913" "914" "915" "916" "917" "918" "919" "920"   
## [921] "921" "922" "923" "924" "925" "926" "927" "928" "929" "930"   
## [931] "931" "932" "933" "934" "935" "936" "937" "938" "939" "940"   
## [941] "941" "942" "943" "944" "945" "946" "947" "948" "949" "950"   
## [951] "951" "952" "953" "954" "955" "956" "957" "958" "959" "960"   
## [961] "961" "962" "963" "964" "965" "966" "967" "968" "969" "970"   
## [971] "971" "972" "973" "974" "975" "976" "977" "978" "979" "980"   
## [981] "981" "982" "983" "984" "985" "986" "987" "988" "989" "990"   
## [991] "991" "992" "993" "994" "995" "996" "997" "998" "999" "1000"  
## [1001] "1001" "1002" "1003" "1004" "1005" "1006" "1007" "1008" "1009" "1010"  
## [1011] "1011" "1012" "1013" "1014" "1015" "1016" "1017" "1018" "1019" "1020"  
## [1021] "1021" "1022" "1023" "1024" "1025" "1026" "1027" "1028" "1029" "1030"  
## [1031] "1031" "1032" "1033" "1034" "1035" "1036" "1037" "1038" "1039" "1040"  
## [1041] "1041" "1042" "1043" "1044" "1045" "1046" "1047" "1048" "1049" "1050"  
## [1051] "1051" "1052" "1053" "1054" "1055" "1056" "1057" "1058" "1059" "1060"  
## [1061] "1061" "1062" "1063" "1064" "1065" "1066" "1067" "1068" "1069" "1070"  
## [1071] "1071" "1072" "1073" "1074" "1075" "1076" "1077" "1078" "1079" "1080"  
## [1081] "1081" "1082" "1083" "1084" "1085" "1086" "1087" "1088" "1089" "1090"  
## [1091] "1091" "1092" "1093" "1094" "1095" "1096" "1097" "1098" "1099" "1100"  
## [1101] "1101" "1102" "1103" "1104" "1105" "1106" "1107" "1108" "1109" "1110"  
## [1111] "1111" "1112" "1113" "1114" "1115" "1116" "1117" "1118" "1119" "1120"  
## [1121] "1121" "1122" "1123" "1124" "1125" "1126" "1127" "1128" "1129" "1130"  
## [1131] "1131" "1132" "1133" "1134" "1135" "1136" "1137" "1138" "1139" "1140"  
## [1141] "1141" "1142" "1143" "1144" "1145" "1146" "1147" "1148" "1149" "1150"  
## [1151] "1151" "1152" "1153" "1154" "1155" "1156" "1157" "1158" "1159" "1160"  
## [1161] "1161" "1162" "1163" "1164" "1165" "1166" "1167" "1168" "1169" "1170"  
## [1171] "1171" "1172" "1173" "1174" "1175" "1176" "1177" "1178" "1179" "1180"  
## [1181] "1181" "1182" "1183" "1184" "1185" "1186" "1187" "1188" "1189" "1190"  
## [1191] "1191" "1192" "1193" "1194" "1195" "1196" "1197" "1198" "1199" "1200"  
## [1201] "1201" "1202" "1203" "1204" "1205" "1206" "1207" "1208" "1209" "1210"  
## [1211] "1211" "1212" "1213" "1214" "1215" "1216" "1217" "1218" "1219" "1220"  
## [1221] "1221" "1222" "1223" "1224" "1225" "1226" "1227" "1228" "1229" "1230"  
## [1231] "1231" "1232" "1233" "1234" "1235" "1236" "1237" "1238" "1239" "1240"  
## [1241] "1241" "1242" "1243" "1244" "1245" "1246" "1247" "1248" "1249" "1250"  
## [1251] "1251" "1252" "1253" "1254" "1255" "1256" "1257" "1258" "1259" "1260"  
## [1261] "1261" "1262" "1263" "1264" "1265" "1266" "1267" "1268" "1269" "1270"  
## [1271] "1271" "1272" "1273" "1274" "1275" "1276" "1277" "1278" "1279" "1280"  
## [1281] "1281" "1282" "1283" "1284" "1285" "1286" "1287" "1288" "1289" "1290"  
## [1291] "1291" "1292" "1293" "1294" "1295" "1296" "1297" "1298" "1299" "1300"  
## [1301] "1301" "1302" "1303" "1304" "1305" "1306" "1307" "1308" "1309" "1310"  
## [1311] "1311" "1312" "1313" "1314" "1315" "1316" "1317" "1318" "1319" "1320"  
## [1321] "1321" "1322" "1323" "1324" "1325" "1326" "1327" "1328" "1329" "1330"  
## [1331] "1331" "1332" "1333" "1334" "1335" "1336" "1337" "1338"

head(insurance\_data)

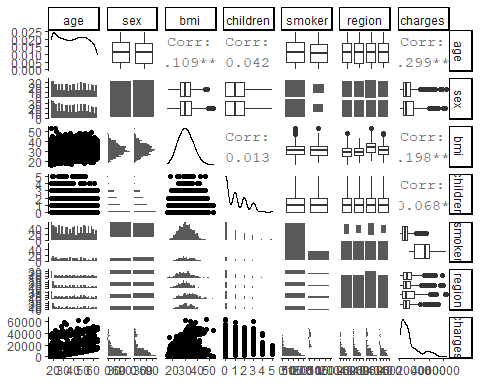
## age sex bmi children smoker region charges  
## 1 19 female 27.9 0 yes southwest 16885  
## 2 18 male 33.8 1 no southeast 1726  
## 3 28 male 33.0 3 no southeast 4449  
## 4 33 male 22.7 0 no northwest 21984  
## 5 32 male 28.9 0 no northwest 3867  
## 6 31 female 25.7 0 no southeast 3757

describe(insurance\_data)

## vars n mean sd median trimmed mad min max  
## age 1 1338 39.21 14.05 39.0 39.01 17.79 18 64.0  
## sex\* 2 1338 1.51 0.50 2.0 1.51 0.00 1 2.0  
## bmi 3 1338 30.66 6.10 30.4 30.50 6.20 16 53.1  
## children 4 1338 1.09 1.21 1.0 0.94 1.48 0 5.0  
## smoker\* 5 1338 1.20 0.40 1.0 1.13 0.00 1 2.0  
## region\* 6 1338 2.52 1.10 3.0 2.52 1.48 1 4.0  
## charges 7 1338 13270.42 12110.01 9382.0 11076.02 7440.81 1122 63770.4  
## range skew kurtosis se  
## age 46.0 0.06 -1.25 0.38  
## sex\* 1.0 -0.02 -2.00 0.01  
## bmi 37.2 0.28 -0.06 0.17  
## children 5.0 0.94 0.19 0.03  
## smoker\* 1.0 1.46 0.14 0.01  
## region\* 3.0 -0.04 -1.33 0.03  
## charges 62648.6 1.51 1.59 331.07

ggpairs(insurance\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



The above plot provides lot of information about the data. As we can see, Outliers are clearly shown among region, sex and smoker boxplots. The variable age is positively correlated with the charges and bmi is moderately positively correlated with the charges.

1. Removing Children variable No of children variable do not make any sense in the practical terms. No of children should not affect the persons premium charges. Suppose if one person has 4 children and another person have no children, then how will this information help in predicting the persons insurance premium charges? It makes least sense in considering this variable to carry out the further analysis and thus remove the children variable from the dataset.

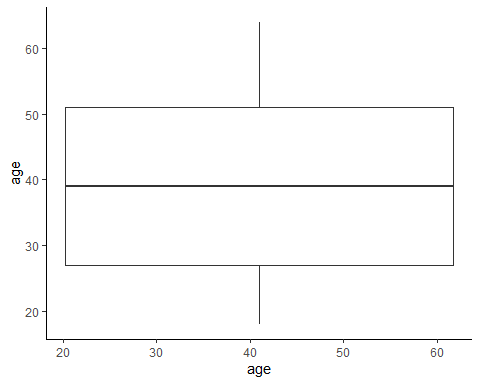
insurance\_data1<-insurance\_data[,-4]  
str(insurance\_data1)

## 'data.frame': 1338 obs. of 6 variables:  
## $ age : int 19 18 28 33 32 31 46 37 37 60 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...  
## $ bmi : num 27.9 33.8 33 22.7 28.9 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...  
## $ charges: num 16885 1726 4449 21984 3867 ...

1. Outliers detection

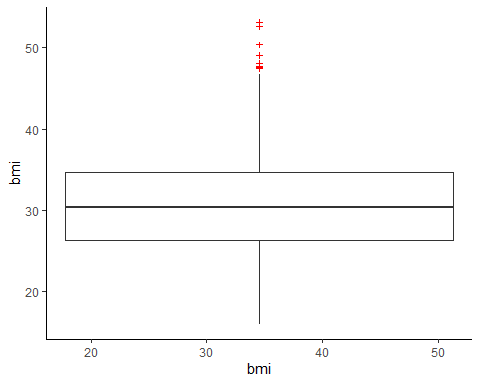
#age  
ggplot(data = insurance\_data1) +  
 (mapping = aes( x = age,y = age )) +  
 geom\_boxplot(outlier.colour= "red", outlier.shape=4,  
 outlier.size=1) #No Outliers

## Warning: Continuous x aesthetic -- did you forget aes(group=...)?



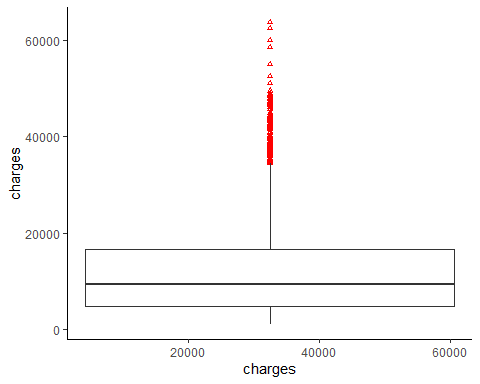
#bmi  
ggplot(data = insurance\_data1) +  
 (mapping = aes( x =bmi,y = bmi )) +  
 geom\_boxplot(outlier.colour= "red", outlier.shape=3,  
 outlier.size=1) #outliers are present

## Warning: Continuous x aesthetic -- did you forget aes(group=...)?



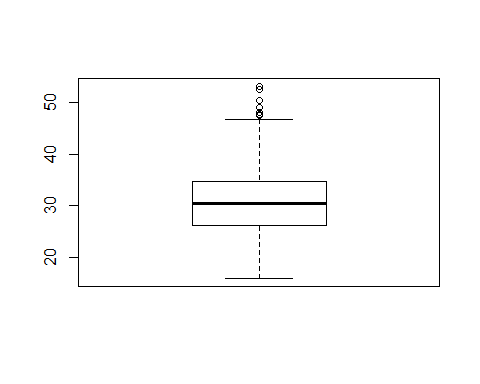
#charges  
ggplot(data = insurance\_data1) +  
 (mapping = aes( x = charges,y = charges )) +  
 geom\_boxplot(outlier.colour= "red", outlier.shape=2,  
 outlier.size=1) #Outliers are present

## Warning: Continuous x aesthetic -- did you forget aes(group=...)?

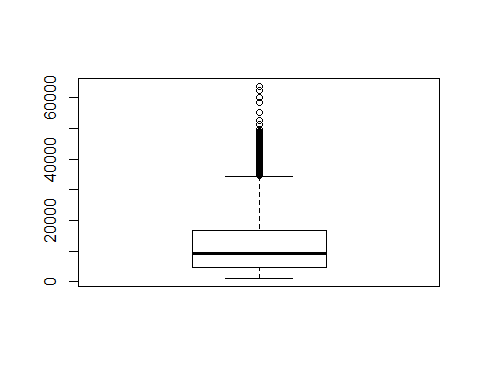


Outliers are detected among bmi and charges variables. Outliers can cause problems with certain types of models. In general, removing an outlier will help in model’s performance. The below code helps in removing the outliers of the dataset. Run this code untill all the outliers are removed (Three times)

Outliners1= boxplot(insurance\_data1$bmi)$out



index1<- which(insurance\_data1$bmi %in% Outliners1)  
Outliners2= boxplot(insurance\_data1$charges)$out

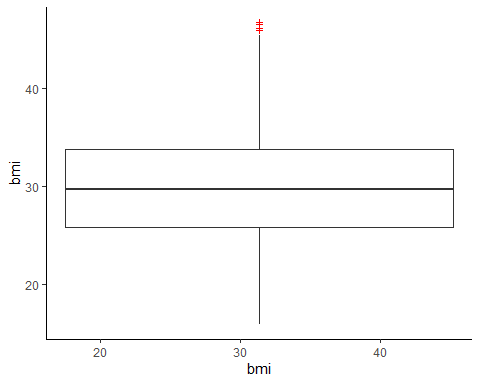


index2<- which(insurance\_data1$charges %in% Outliners2)  
index3<- unique(c(index1,index2))  
insurance\_data1<-insurance\_data1[-index3,]  
dim(insurance\_data1)

## [1] 1193 6

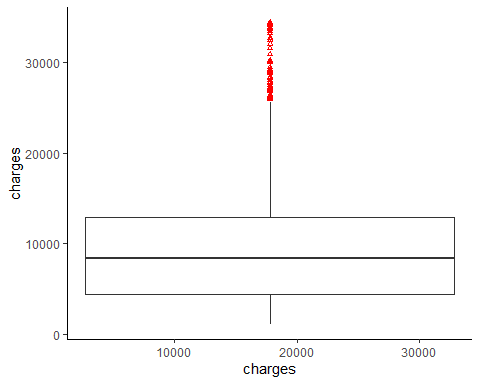
#no outliers  
#bmi  
ggplot(data = insurance\_data1) +  
 (mapping = aes( x =bmi,y = bmi )) +  
 geom\_boxplot(outlier.colour= "red", outlier.shape=3,  
 outlier.size=1)

## Warning: Continuous x aesthetic -- did you forget aes(group=...)?



#charges  
ggplot(data = insurance\_data1) +  
 (mapping = aes( x = charges,y = charges )) +  
 geom\_boxplot(outlier.colour= "red", outlier.shape=2,  
 outlier.size=1)

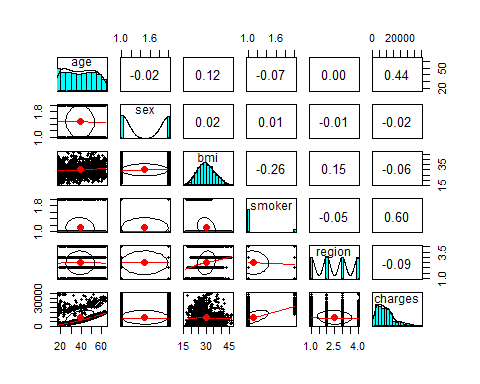
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?



Data vizualization

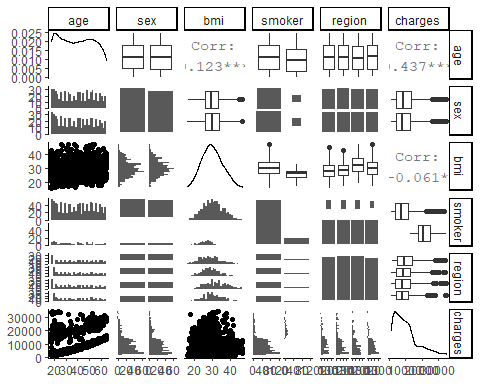
A super fancy way to show correlation is the use of pairs.panels() function from the psych package. The pairs.panels is splitting into three functional groups that demonstrated three kind of information. The bottom left portion off diagonal shows similar scatterplot matrices before, with additional information such as the mean of two features value given in a red dot; a lowest curve that show a more flexible relationship between two features, and a correlation ellipse (circle) that shows a correlation strength of higher correlation indicated by a more stretchy ellipse. The diagrams in diagonal show the histogram of each respective feature with a distribution curve. Numbers on the top right off diagonal are the correlation value between each two respective features.

pairs.panels(insurance\_data1)



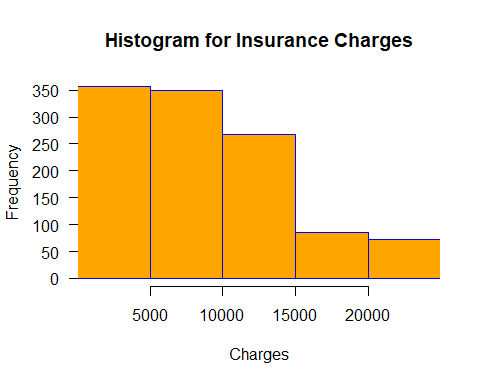
ggpairs(insurance\_data1)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



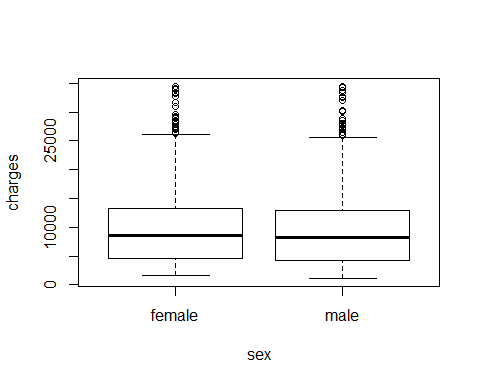
Distribution of Premium Charges Let us plot histogram of charges to understand the distribution.

hist(insurance\_data1$charges,   
 main="Histogram for Insurance Charges",   
 xlab="Charges",   
 border="blue",   
 col="orange",  
 xlim=c(1000,24000),  
 las=1,   
 breaks=5)

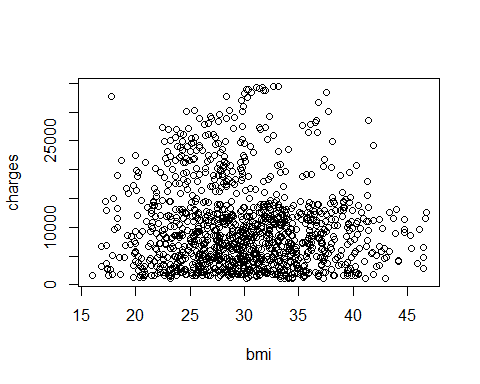


Type of Distribution: We have a right skewed distribution in which most patients are being charged between 1000− 15000. This indicates that most of the premium charges are around the value mentioned i.e., between 1000 – 15000.

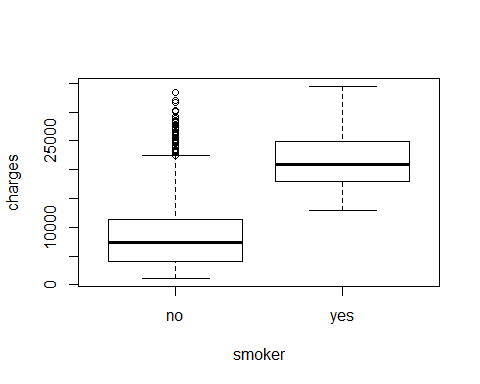
#plots  
plot(charges ~ sex, insurance\_data1)



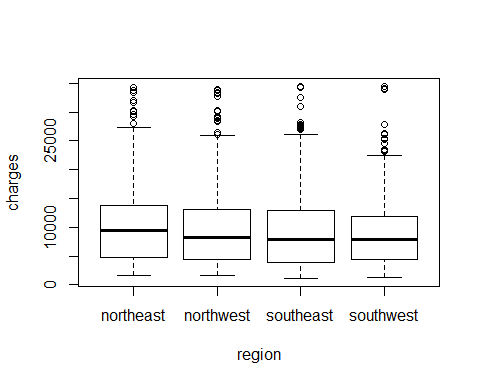
plot(charges ~ bmi, insurance\_data1)



plot(charges ~ smoker, insurance\_data1)



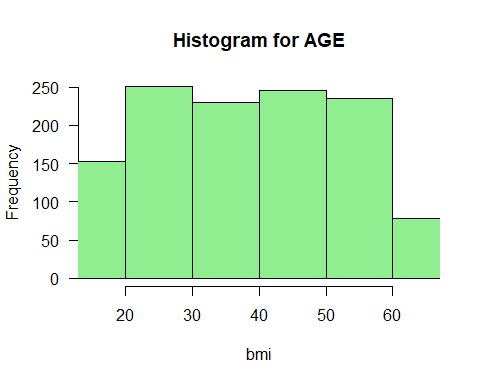
plot(charges ~ region, insurance\_data1)



Analyzing relationship among variables BMI frequency: Most of the BMI frequency is concentrated between 25 - 35.

age, bmi vs charges Hist of age Most of the AGE frequency is concentrated between 20 – 60

hist(insurance\_data1$age,   
 main="Histogram for AGE",   
 xlab="bmi",   
 border="black",   
 col="lightgreen",  
 xlim=c(15,65),  
 las=1,   
 breaks=5)

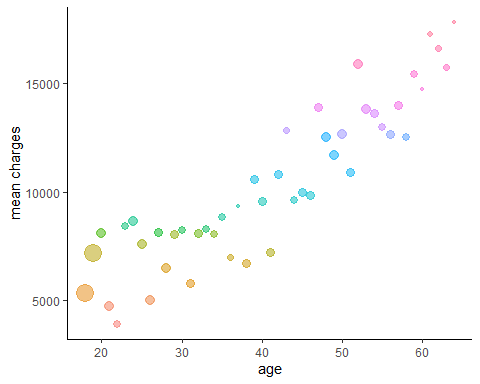


age(mean analysis of charges)

analysis <- group\_by(insurance\_data1, age)  
analysis1 <-summarise(analysis, count=n(), rate1 = mean(charges))

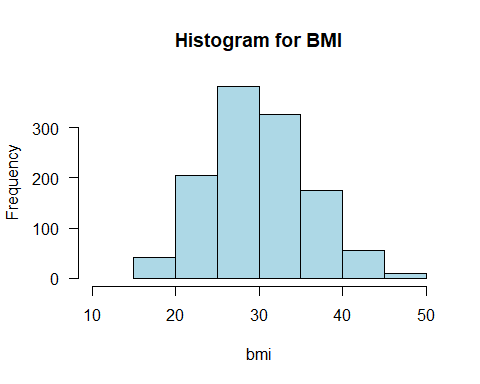
## `summarise()` ungrouping output (override with `.groups` argument)

ggplot(analysis1, aes(age, rate1))+geom\_point(aes(size = count, color = factor(rate1)), alpha=1/2)+  
 theme(legend.position = "none")+  
 labs(x = "age", y = "mean charges")



hist of bmi

hist(insurance\_data1$bmi,   
 main="Histogram for BMI",   
 xlab="bmi",   
 border="black",   
 col="lightblue",  
 xlim=c(10,50),  
 las=1,   
 breaks=5)

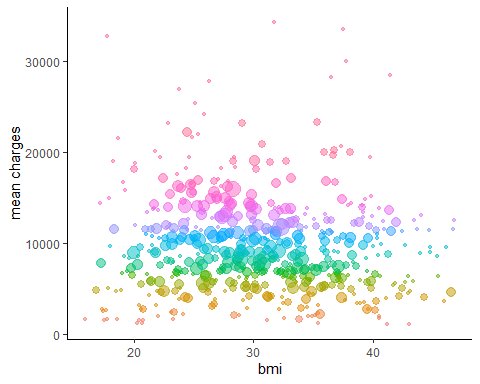


bmi(mean analysis of charges)

analysis <- group\_by(insurance\_data1, bmi)  
analysis1 <-summarise(analysis, count=n(), rate1 = mean(charges))

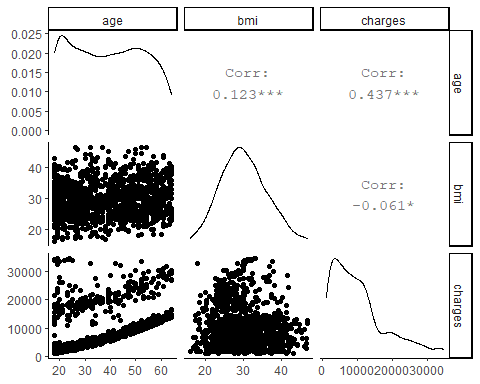
## `summarise()` ungrouping output (override with `.groups` argument)

ggplot(analysis1, aes(bmi, rate1))+geom\_point(aes(size = count, color = factor(rate1)), alpha=1/2)+  
 theme(legend.position = "none")+  
 labs(x = "bmi", y = "mean charges")



correlation plot Correlations: Age and charges have a correlation of 0.516 while bmi and charges have a correlation of -0.0566

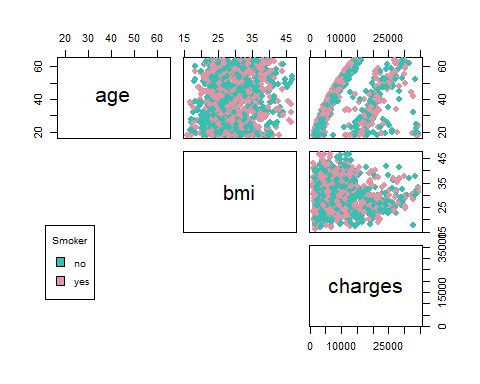
ggpairs(insurance\_data1[,-c(2,4,5,7,8)])

 Relationship between BMI and Age: The correlation for these two variables is 0.149 which is not that great. Therefore, we can disregard that age has a huge influence on BMI.

age, bmi Vs sex

The charges do increase with respect to age. There is no clear difference in charges for male vs female with respect to their age and bmi.

species\_col <- rev(rainbow\_hcl(2))[(insurance\_data1$sex)]  
pairs(insurance\_data1[,-c(2,4,5,7,8)], col = species\_col,  
 lower.panel = NULL,  
 cex.labels=2, pch=19, cex = 1.2)  
#add a legend  
par(xpd = TRUE)  
legend(x = 0.05, y = 0.4, cex = 0.6,  
 legend = as.character(levels(insurance\_data1$smoker)), title="Smoker",  
 fill = unique(species\_col))



par(xpd = NA)

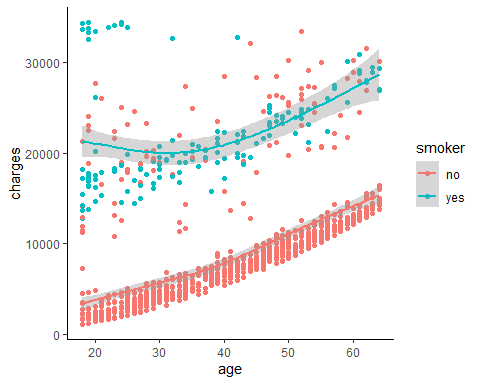
age, bmi Vs smoker age

Despite the BMI indicator is used to measure health risk for an individual, the feature is not as important as knowing whether the individual is a smoker or non-smoker.

Smoker tends to incur a much higher charge as compared to non-smoker. There is no smoker beyond BMI > 30 but when the BMI of a non- smoker goes beyond 30, the charges increases to a minimum of 20000 around. Smoker with BMI < 30 generally have charges incurred above 20000. The same goes with the age as well. As the age increases, the charges are more for smokers than for non-smokers.

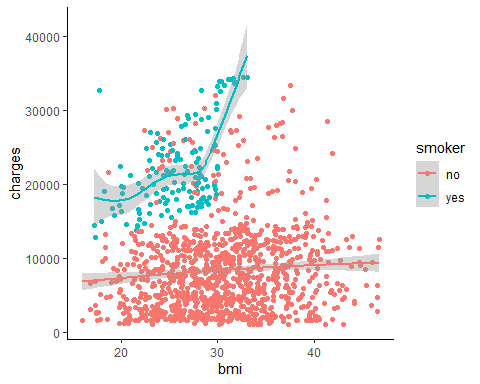
ggplot(data = insurance\_data1) +  
 (mapping = aes( x =age, y = charges, color = smoker)) +  
 geom\_point()+  
 geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



#bmi  
ggplot(data = insurance\_data1) +  
 (mapping = aes( x =bmi, y = charges, color = smoker)) +  
 geom\_point()+  
 geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

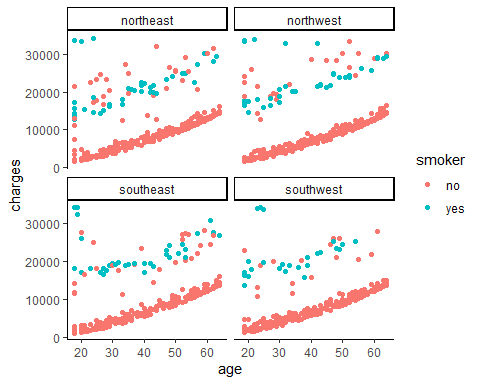


age, bmi, sex, smoker Vs region age

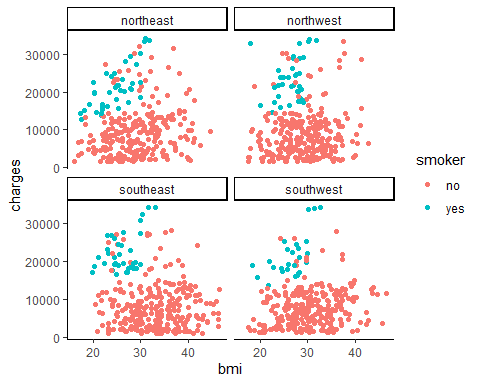
The charges do increase with respect to age and bmi. There is no clear difference in charges for smoker’s vs non-smoker with respect to their age and bmi across different regions.

The below boxplot shows that, across south west, north west and south east, the charges are rather similar. However, individuals from north east has a wider range of charges for female. It is the only distinguishing character regarding region. It seems that region too has very less influence over the charges.

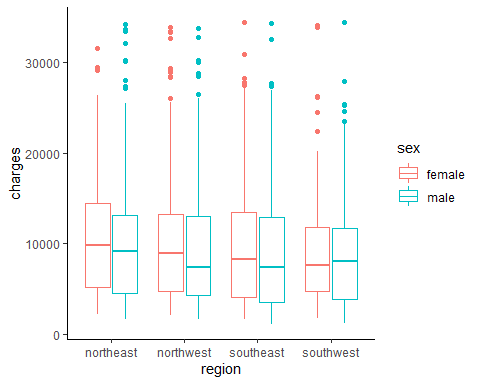
ggplot(data = insurance\_data1) +  
 (mapping = aes( x =age, y = charges, color=smoker)) +  
 geom\_point()+  
 facet\_wrap(~region , nrow = 2)



#bmi  
ggplot(data = insurance\_data1) +  
 (mapping = aes( x =bmi, y = charges, color=smoker)) +  
 geom\_point()+  
 facet\_wrap(~region , nrow = 2)



#sex  
ggplot(data = insurance\_data1) +  
 (mapping = aes( x =region, y = charges, color=sex)) +  
 geom\_boxplot()



According to CRISP-DM, the data preparation phase covers all activities to construct the final dataset from the initial raw data in order to prepare the data for further processing. After cleaning the data, its time to prepare the data accordingly to build models for predicting premium medical charges.

Age analysis The age analysis graph shows that there is positive correlation between age and charges. For some data values, the increase is exceptionally smooth and linear, and for some other data points the increase is not perfectly linear. The data can be divided approximately into two segments as shown in the graph.

Turning Age into Categorical Variables: Early Age: from 15 - 24 Prime Age: from 25 – 54 Mature Age: 55 or older

<https://www.indexmundi.com/united_states/age_structure.html>

DATA PREPARATION converting Age to Categorical variables

catAge<- cut(insurance\_data1$age,breaks=c(15,24,54,64),labels=c("Early\_Age","Prime\_Age","Mature\_Age"))  
insurance\_data1<-cbind(insurance\_data1,catAge)  
table(insurance\_data1$catAge)

##   
## Early\_Age Prime\_Age Mature\_Age   
## 253 728 212

aggregate(insurance\_data1$charges,list(insurance\_data1$catAge),mean)

## Group.1 x  
## 1 Early\_Age 6505  
## 2 Prime\_Age 9721  
## 3 Mature\_Age 14805

The mean charges of each category are shown above. It clearly indicates that the Mature\_age group of people has more charges and are at high risk.

BMI analysis Body mass index, or BMI, is a way to help figure out if you are at a healthy weight for your height. BMI is a number based on your weight and height. In general, the higher the number, the more body fat a person has. BMI is often used as a screening tool to decide if your weight might be putting you at risk for health problems such as heart disease, diabetes, and cancer.

BMI is used to broadly define different weight groups in adults 20 years old or older. The same groups apply to both men and women. Turning BMI into Categorical Variables: Under Weight: Body Mass Index (BMI) < 18.5 Normal Weight: Body Mass Index (BMI) ≥ 18.5 and Body Mass Index (BMI) < 24.9 Overweight: Body Mass Index (BMI) ≥ 25 and Body Mass Index (BMI) < 29.9 Obese: Body Mass Index (BMI) > 30 <https://www.cancer.org/cancer/cancer-causes/diet-physical-activity/body-weight-and-cancer-risk/adult-bmi.html>

#converting BMI to Categorical variables  
catBMI<- cut(insurance\_data1$bmi,breaks=c(0,18.499,24.999,29.999,60),labels=c("Underweight","Normal","Overweight","Obese"))  
insurance\_data1<-cbind(insurance\_data1,catBMI)  
table(insurance\_data1$catBMI)

##   
## Underweight Normal Overweight Obese   
## 20 224 382 567

aggregate(insurance\_data1$charges,list(insurance\_data1$catBMI),mean)

## Group.1 x  
## 1 Underweight 8852  
## 2 Normal 10299  
## 3 Overweight 10719  
## 4 Obese 9316

The mean charges of each category of BMI is shown above. It clearly indicates that the Normal and overweight group of people has more charges and are at high risk.

After doing the transformation of age and bmi to categorical data, the final set of data is obtained which is used for further analysis of modeling. Let our final data frame be named as “newInsuranceData”.

#Creating new Data Set  
newInsuranceData <- cbind(insurance\_data1[c(6,2,4,5)],catAge,catBMI)  
str(newInsuranceData)

## 'data.frame': 1193 obs. of 6 variables:  
## $ charges: num 16885 1726 4449 21984 3867 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 1 1 2 2 2 2 2 2 2 3 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 3 4 4 2 3 3 4 3 3 3 ...

summary(newInsuranceData)

## charges sex smoker region catAge   
## Min. : 1122 female:611 no :1055 northeast:295 Early\_Age :253   
## 1st Qu.: 4415 male :582 yes: 138 northwest:305 Prime\_Age :728   
## Median : 8410 southeast:302 Mature\_Age:212   
## Mean : 9942 southwest:291   
## 3rd Qu.:12979   
## Max. :34473   
## catBMI   
## Underweight: 20   
## Normal :224   
## Overweight :382   
## Obese :567   
##   
##

1. What type of decision-making model(s) is appropriate for the decision‐making tasks? The business objective is to classify by segments, policy holders from the pool of data by clustering the characteristics which are maintained as age, sex, body mass index (BMI), regions and smoking status. Then predict future medical charges for the produced segments through multiple linear regression using the gathered characteristics to assist our client in the decision-making process of charging an accurate insurance premium to customers. Unsupervised Learning:
2. Clustering: Clustering is a technique of data segmentation that partitions the data into several groups based on their similarity. Basically, we group the data through a statistical operation. These smaller groups that are formed from the bigger data are known as clusters. These cluster exhibit the following properties: They are discovered while carrying out the operation and the knowledge of their number is not known in advance. Clusters are the aggregation of similar objects that share common characteristics. Clustering is the most widespread and popular method of Data Analysis and Data Mining. It used in cases where the underlying input data has a colossal volume and we are tasked with finding similar subsets that can be analyzed in several ways. In the first stage we will do a cluster analysis on the dataset and see what we discovered through visualization so far match and how well the clusters form according to the risk class definition of the medical insurance charges. Supervised Learning:
3. Multiple Linear Regression: A multiple linear regression (MLR) model describes a dependent variable y by independent variables x1, x2, …, xP (P > 1) is expressed by the equation as follows, where the numbers α and βk (k = 1, 2, …, P) are the parameters, and ϵ is the error term.

We create the regression model using the lm() function in R. The model determines the value of the coefficients using the input data. Next, we can predict the value of our response variable i.e., charges for a given set of predictor variables defined earlier using these coefficients. lm() Function: This function creates the relationship model between the predictor and the response variable. Syntax: The basic syntax for lm() function in multiple regression is − lm(y ~ x1+x2+x3…,data) Following is the description of the parameters used − 1. Formula is a symbol presenting the relation between the response variable and predictor variables. 2. Data is the vector on which the formula will be applied.

1. Provide rationale for choice of model(s) Here are the following results from the visualization analysis performed.
2. Age and Charges: We can see there is a slight increase in charges depending on the age of the patient.
3. BMI and Charges: We can see there is also a very slight increase in charges depending on the bmi of the patient.
4. Smoking: Smoker is the grand major factor for the patient’s habit having high charges.
5. Sex and Region: No major influence of these two variables against charges or any other predictor variables. Are these Variables in the form of Clusters? As seen in the charts, when smoker variable is plotted against charges with age and bmi factor, there is clear evidence of formation of segments among the data points. From the business point of view and as well from our data analysis, people with smoking habits are at high risk. Thus, those patients are usually charged high. Segmenting our data into clusters and dividing the clusters into the particular risk class of which most companies use to distinguish individuals. There are three varieties of risk class: preferred, standard, and smoker. The preferred health class is for those who are in superior health, posing the least risk to the insurer, and therefore deserve the lowest rates. The standard health class is where most individuals fall with a few minor health issues. They tend to pay normal prices for the same coverage. The smoker class is considered the riskiest one and has higher rates for individuals who have serious health issues. After segmenting into clusters and naming them according to their risk class category. For each cluster, we build specific models that is appropriate and provide the best results using Multiple Linear Regression. Why Multiple Linear Regression? Smoking is a huge factor as compared to other features. Just by being a smoker, the charge increases by quite a good amount regardless of age. Thereafter, the charges increase with age, bmi (higher health risk). Insurance is increasingly important with dependents for the regression model. Assumptions of Multiple Linear Regression Multiple linear regression analysis makes several key assumptions:
6. There must be a linear relationship between the outcome variable and the independent variables.
7. Multivariate Normality: Multiple regression assumes that the residuals are normally distributed.
8. No Multicollinearity: Multiple regression assumes that the independent variables are not highly correlated with each other.
9. Homoscedasticity: This assumption states that the variance of error terms is similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables. Multiple linear regression requires at least two independent variables, which can be nominal, ordinal, or interval/ratio level variables. A rule of thumb for the sample size is that regression analysis requires at least 20 cases per independent variable in the analysis. Our plots show that there is a linear or curvilinear relationship among age, bmi numerical variables. And after doing clear analysis and visualization on the data, all the assumptions of the Multiple Linear Regression are met. Thus, providing us a set of reasons or a logical basis for our course of action to predict the Insurance charges among the clusters formed.
10. Detail model development and output Cluster Analysis
11. Modelling – K-Prototype The “clustMixType” package provides a user-friendly way for clustering mixed-type data in R given by the k-prototypes algorithm. A record is allocated to the cluster which has the most similar looking prototype (mean) of a cluster. The algorithm iterates in a manner like the k-means algorithm where for the numeric variables the mean and the categorical variables the mode minimizes the total within cluster distance. The dissimilarity measure for numeric attributes is the square Euclidean distance whereas the similarity measure on categorical attributes is the number of matching attributes between objects and cluster prototypes. The steps of the algorithm are:
12. Initialization with random cluster prototypes
13. For each observation do:
14. Assign observations to its closest prototype according to d()
15. Update cluster prototypes by cluster-specific means/modes for all variables
16. If any observations have swapped their cluster assignment in 2 or the maximum number of iterations have not been reached: repeat from 2  
    Results: Elbow Plot The Elbow method looks at the total Within Sum of Square as a function of the number of clusters: One shoul

Silhouette Plot The Average Silhouette method measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering Thus, the optimal number of clusters are K=3.

#CLUSTER ANALYSIS  
#Modelling - K-Prototype  
#standardizing  
scaled.df <- sapply(newInsuranceData[c(1)], scale)  
unscaled.df <- newInsuranceData[,-c(1)]  
dim(scaled.df)

## [1] 1193 1

dim(unscaled.df)

## [1] 1193 5

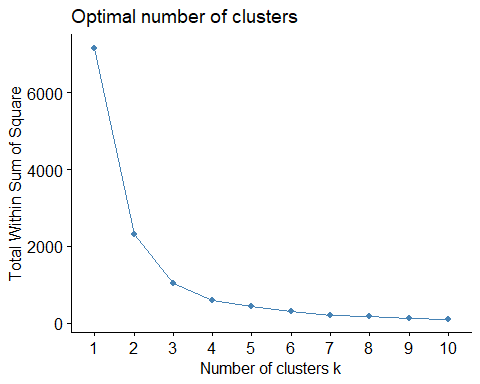
newData <- cbind(scaled.df,unscaled.df)  
str(newData)

## 'data.frame': 1193 obs. of 6 variables:  
## $ charges: num 0.957 -1.133 -0.757 1.661 -0.838 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 1 1 2 2 2 2 2 2 2 3 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 3 4 4 2 3 3 4 3 3 3 ...

# exploring no of clusters  
fviz\_nbclust(newData, FUN = hcut, method = "wss")

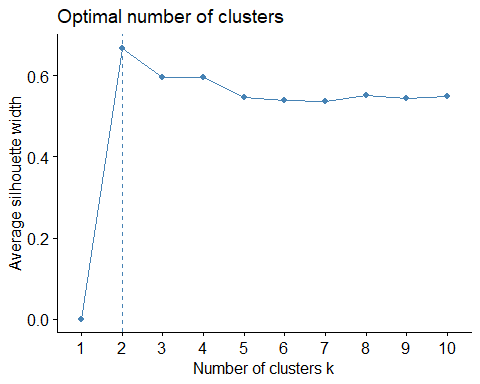
## Warning in stats::dist(x): NAs introduced by coercion

## Warning in stats::dist(x, method = method, ...): NAs introduced by coercion  
  
## Warning in stats::dist(x, method = method, ...): NAs introduced by coercion  
  
## Warning in stats::dist(x, method = method, ...): NAs introduced by coercion  
  
## Warning in stats::dist(x, method = method, ...): NAs introduced by coercion  
  
## Warning in stats::dist(x, method = method, ...): NAs introduced by coercion  
  
## Warning in stats::dist(x, method = method, ...): NAs introduced by coercion  
  
## Warning in stats::dist(x, method = method, ...): NAs introduced by coercion  
  
## Warning in stats::dist(x, method = method, ...): NAs introduced by coercion  
  
## Warning in stats::dist(x, method = method, ...): NAs introduced by coercion  
  
## Warning in stats::dist(x, method = method, ...): NAs introduced by coercion



fviz\_nbclust(newData, FUN = hcut, method = "silhouette")

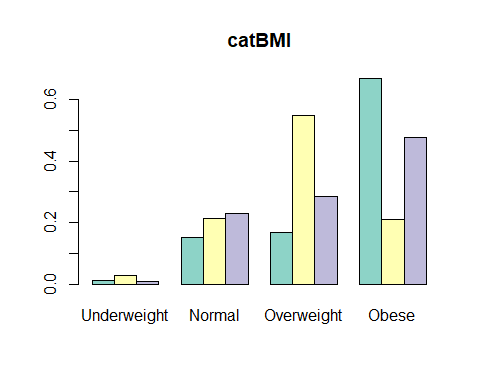
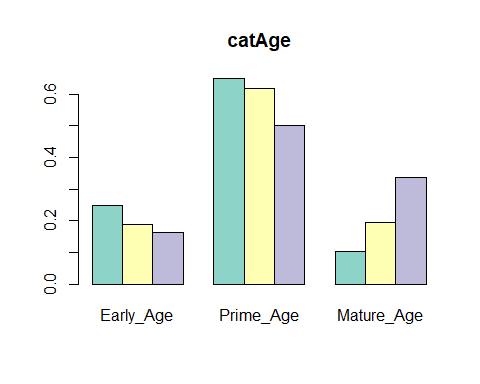
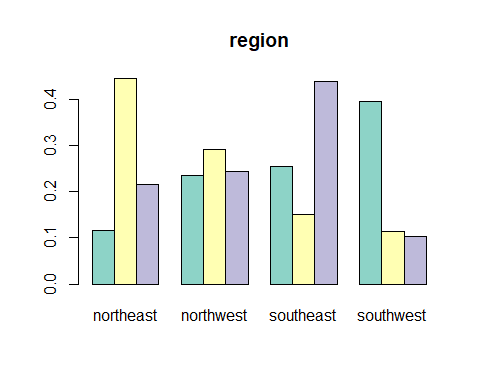
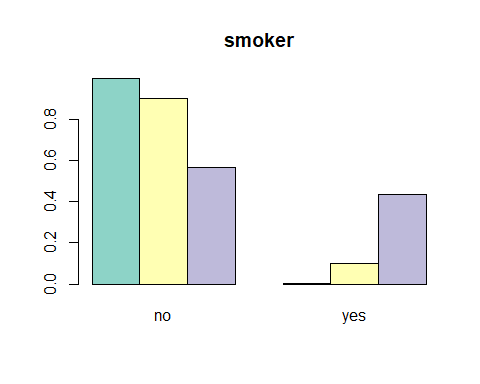
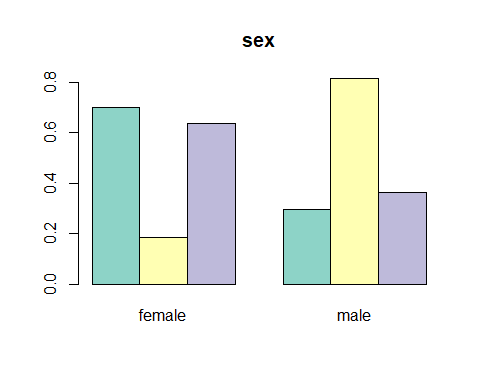
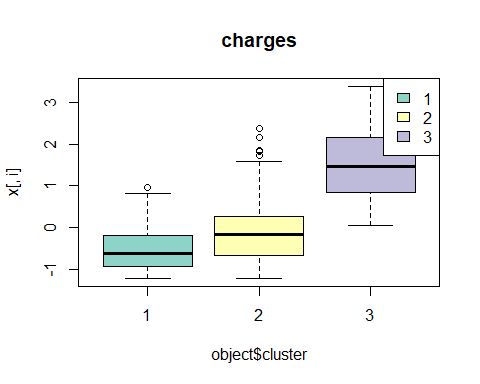
## Warning in stats::dist(x): NAs introduced by coercion  
  
## Warning in stats::dist(x): NAs introduced by coercion  
  
## Warning in stats::dist(x): NAs introduced by coercion  
  
## Warning in stats::dist(x): NAs introduced by coercion  
  
## Warning in stats::dist(x): NAs introduced by coercion  
  
## Warning in stats::dist(x): NAs introduced by coercion  
  
## Warning in stats::dist(x): NAs introduced by coercion  
  
## Warning in stats::dist(x): NAs introduced by coercion  
  
## Warning in stats::dist(x): NAs introduced by coercion  
  
## Warning in stats::dist(x): NAs introduced by coercion



set.seed(1)  
#The tested lambda values ranged from 0 to 5. A small lamnda value indicates that the clustering is  
#dominated by numeric attributes while a large lambda value implies that categorical attributes  
#dominate the clustering.  
  
# apply k prototyps  
kpres <- kproto(newData, 3)

## # NAs in variables:  
## charges sex smoker region catAge catBMI   
## 0 0 0 0 0 0   
## 0 observation(s) with NAs.  
##   
## Estimated lambda: 1.89

# Custering Visualisation  
clprofiles(kpres, newData, vars = NULL, col = NULL) # object$cluster and object$size are possible



test\_stats <- cluster.stats(dist(newData), kpres$cluster)

## Warning in dist(newData): NAs introduced by coercion

test\_stats

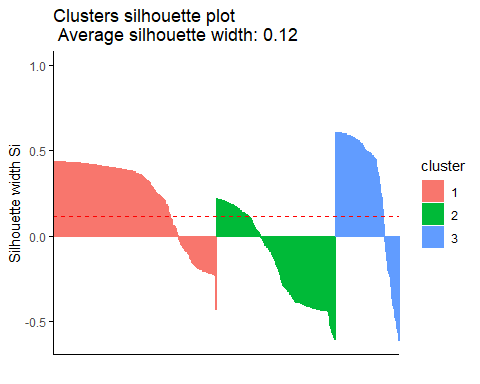
## $n  
## [1] 1193  
##   
## $cluster.number  
## [1] 3  
##   
## $cluster.size  
## [1] 561 410 222  
##   
## $min.cluster.size  
## [1] 222  
##   
## $noisen  
## [1] 0  
##   
## $diameter  
## [1] 5.32 8.82 8.15  
##   
## $average.distance  
## [1] 1.37 1.92 2.47  
##   
## $median.distance  
## [1] 1.15 1.61 2.15  
##   
## $separation  
## [1] 0.00e+00 7.35e-05 0.00e+00  
##   
## $average.toother  
## [1] 2.96 2.51 4.70  
##   
## $separation.matrix  
## [,1] [,2] [,3]  
## [1,] 0.00e+00 7.35e-05 0.000000  
## [2,] 7.35e-05 0.00e+00 0.000294  
## [3,] 0.00e+00 2.94e-04 0.000000  
##   
## $ave.between.matrix  
## [,1] [,2] [,3]  
## [1,] 0.00 1.82 5.05  
## [2,] 1.82 0.00 4.23  
## [3,] 5.05 4.23 0.00  
##   
## $average.between  
## [1] 3.22  
##   
## $average.within  
## [1] 1.76  
##   
## $n.between  
## [1] 445572  
##   
## $n.within  
## [1] 265456  
##   
## $max.diameter  
## [1] 8.82  
##   
## $min.separation  
## [1] 0  
##   
## $within.cluster.ss  
## [1] 3056  
##   
## $clus.avg.silwidths  
## 1 2 3   
## 0.234 -0.150 0.309   
##   
## $avg.silwidth  
## [1] 0.116  
##   
## $g2  
## NULL  
##   
## $g3  
## NULL  
##   
## $pearsongamma  
## [1] 0.337  
##   
## $dunn  
## [1] 0  
##   
## $dunn2  
## [1] 0.738  
##   
## $entropy  
## [1] 1.03  
##   
## $wb.ratio  
## [1] 0.548  
##   
## $ch  
## [1] 797  
##   
## $cwidegap  
## [1] 0.357 0.747 0.229  
##   
## $widestgap  
## [1] 0.747  
##   
## $sindex  
## [1] 0.000386  
##   
## $corrected.rand  
## NULL  
##   
## $vi  
## NULL

sil <- silhouette(kpres$cluster, dist(newData))

## Warning in dist(newData): NAs introduced by coercion

fviz\_silhouette(sil)

## cluster size ave.sil.width  
## 1 1 561 0.23  
## 2 2 410 -0.15  
## 3 3 222 0.31



1. Modelling – Hierarchal Clustering Algorithm: Agglomerative Agglomerative clustering will start with n clusters, where n is the number of observations, assuming that each of them is its own separate cluster. Then the algorithm will try to find most similar data points and group them, so they start forming clusters. Agglomerative clustering is better in discovering small clusters, and is used by most software; divisive clustering which is the opposite is better in discovering larger clusters Metric: Gower Gower is a distance measure that can be used to calculate distance between two entities whose attribute has a mix of categorical and numerical values. Gower’s dissimilarity measure is a weighted average of the distances computed for each variable, after scaling each variable to a [0,1] scale. We used daisy function form the Cluster package to run Gower metric Also we used package Purr to determine the best distance (“Average”,” Single”, ”Complete”, ”Ward”). Ward’s method gets us the highest agglomerative coefficient

#H-Clust  
#clustering H-clust by using metric = 'GOWER'  
insuranceDist<-daisy(newInsuranceData,metric="gower")  
summary(insuranceDist)

## 711028 dissimilarities, summarized :  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.349 0.510 0.479 0.583 0.997   
## Metric : mixed ; Types = I, N, N, N, N, N   
## Number of objects : 1193

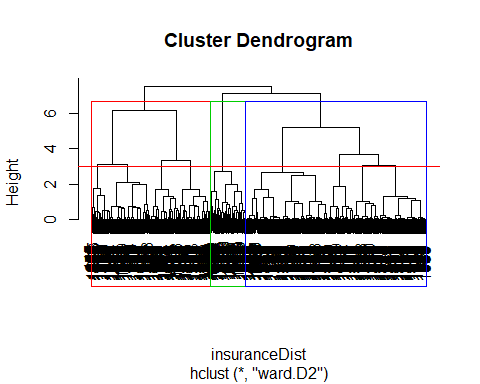
str(insuranceDist)

## 'dissimilarity' num [1:711028] 0.742 0.895 0.859 0.732 0.566 ...  
## - attr(\*, "Labels")= chr [1:1193] "1" "2" "3" "4" ...  
## - attr(\*, "Size")= int 1193  
## - attr(\*, "Metric")= chr "mixed"  
## - attr(\*, "Types")= chr [1:6] "I" "N" "N" "N" ...

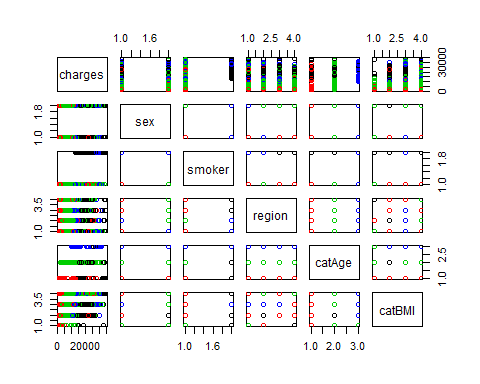
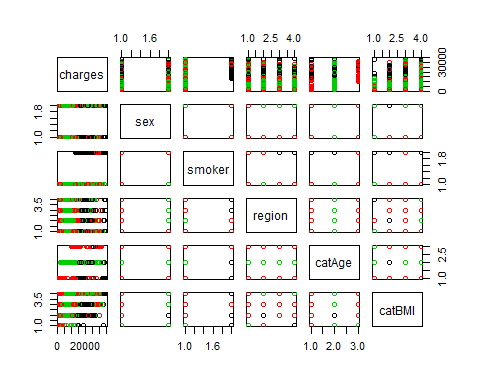
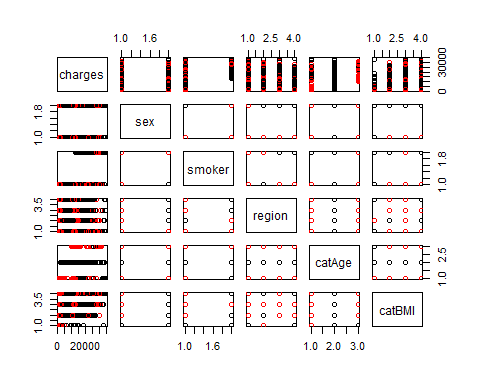
#Selecting hClustering method  
  
m <- c( "average", "single", "complete", "ward")  
names(m) <- c( "average", "single", "complete", "ward")  
ac <- function(x) {  
 agnes(newInsuranceData, method = x)$ac  
}  
map\_dbl(m, ac)

## average single complete ward   
## 0.999 0.981 1.000 1.000

#hClustering using ward.D2  
fitH<- hclust(insuranceDist,"ward.D2")  
clusters<-cutree(fitH,3)  
plot(fitH)  
rect.hclust(fitH, k=3, border= 2:6, )  
abline(h = 3, col = 'red')



for(i in 2:4){  
 plot(newInsuranceData, col=hclusters<- cutree(fitH, i))  
}



newInsuranceData<-cbind(newInsuranceData,clusters)  
aggregate(newInsuranceData[,c(1)],list(newInsuranceData$clusters),mean)

## Group.1 x  
## 1 1 21383  
## 2 2 9120  
## 3 3 8245

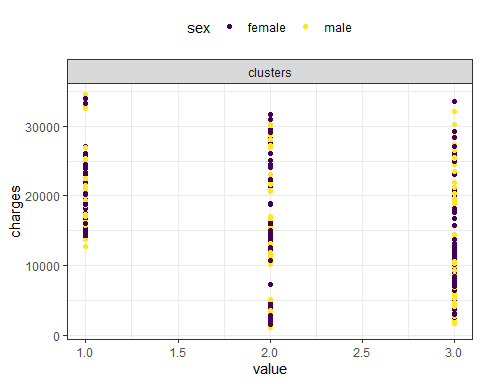
table(newInsuranceData$clusters)

##   
## 1 2 3   
## 126 422 645

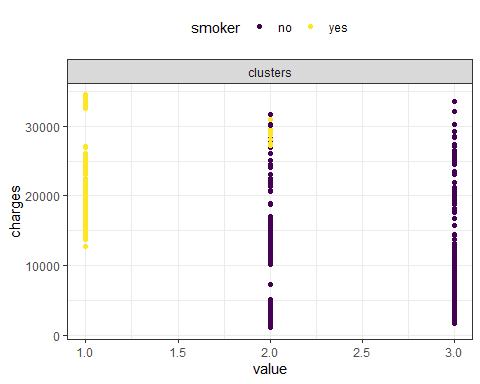
#Cluster Visualization  
theme\_set(  
 theme\_bw() +  
 theme(legend.position = "top")  
)  
  
insurance.gathered <- newInsuranceData %>%  
 as\_data\_frame() %>%  
 gather(key = "variable", value = "value",  
 -sex,-smoker, -region, -catAge, -catBMI,-charges)

## Warning: `as\_data\_frame()` is deprecated as of tibble 2.0.0.  
## Please use `as\_tibble()` instead.  
## The signature and semantics have changed, see `?as\_tibble`.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.

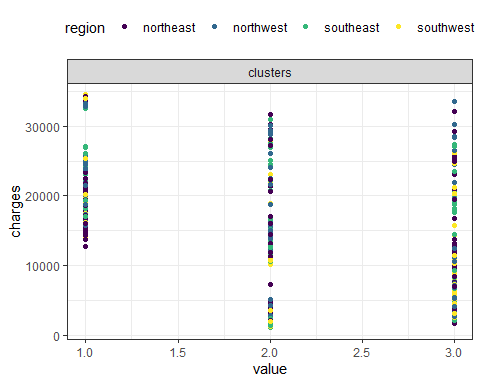
ggplot(insurance.gathered, aes(x = value, y = charges)) +  
 geom\_point(aes(color =sex)) +  
 facet\_wrap(~variable)+  
 scale\_color\_viridis\_d()



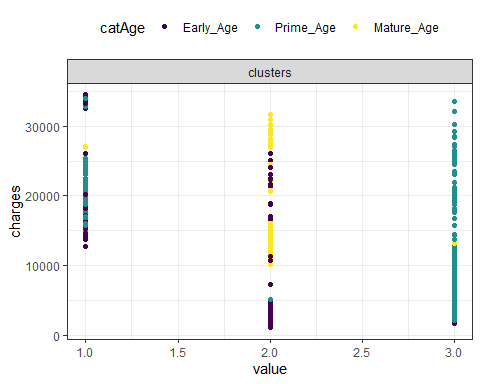
ggplot(insurance.gathered, aes(x = value, y = charges)) +  
 geom\_point(aes(color =smoker)) +  
 facet\_wrap(~variable)+  
 scale\_color\_viridis\_d()



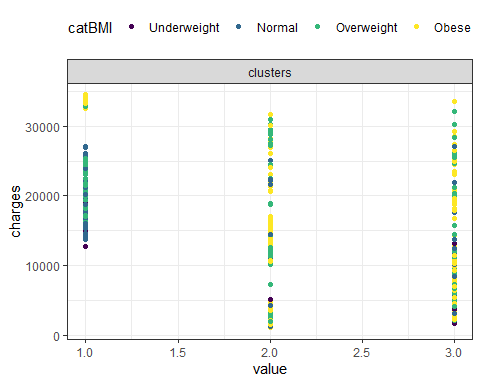
ggplot(insurance.gathered, aes(x = value, y = charges)) +  
 geom\_point(aes(color =region)) +  
 facet\_wrap(~variable)+  
 scale\_color\_viridis\_d()



ggplot(insurance.gathered, aes(x = value, y = charges)) +  
 geom\_point(aes(color =catAge)) +  
 facet\_wrap(~variable)+  
 scale\_color\_viridis\_d()



ggplot(insurance.gathered, aes(x = value, y = charges)) +  
 geom\_point(aes(color =catBMI)) +  
 facet\_wrap(~variable)+  
 scale\_color\_viridis\_d()



On comparing the clusters formed from each of the cluster analysis modelling. The clusters formed from H-clust is very apt as it clearly segments smokers and non-smokers. The first cluster formed from H-clust contains only smokers. The results of the K-prototype were only used to get clear vision of the cluster formation. The result of each cluster is tabulated below with specific name to each cluster.

Cluster 1 Smoker n: 94 m: 18864.738 Smoker Prime Age Overweight BMI Highest Risk Highest Charges

Cluster 2 Standard N: 396 m: 7876.261 Non Smoker Early & Mature Age Obesity & Overweight BMI Medium Risk Medium Charges

Cluster 3 Preferred N: 615 m: 7546.257 Non Smoker Normal BMI Prime Age Lowest Risk Lowest Charges

These results are further used to build multi linear regression models on each cluster and choose the best model to predict the charges.

Multiple Linear Regression

#Multi Linear Regression Analysis for premium charges calculations  
#Cluster 1  
Reg\_Cluster1 <- newInsuranceData %>%  
 filter(clusters==1)  
str(Reg\_Cluster1)

## 'data.frame': 126 obs. of 7 variables:  
## $ charges : num 16885 23568 34303 23245 14712 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 1 1 2 1 2 2 2 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 4 3 3 2 3 3 2 1 4 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 1 2 1 2 1 2 2 2 3 2 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 3 3 4 2 2 2 2 2 2 2 ...  
## $ clusters: int 1 1 1 1 1 1 1 1 1 1 ...

dim(Reg\_Cluster1)

## [1] 126 7

Reg\_Cluster1<-Reg\_Cluster1[,-c(3,7)] #Smoker column contains only 'Yes' values for all the data, thus it will not have any effect over prediction  
  
#model build  
#without outliers  
m1 <- lm(charges ~ catAge + catBMI ,data = Reg\_Cluster1)  
summary(m1)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI, data = Reg\_Cluster1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13637 -1890 -302 1588 13256   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16137 1528 10.56 < 2e-16 \*\*\*  
## catAgePrime\_Age 3341 692 4.83 4.1e-06 \*\*\*  
## catAgeMature\_Age 9150 1756 5.21 7.9e-07 \*\*\*  
## catBMINormal 227 1501 0.15 0.88   
## catBMIOverweight 1989 1484 1.34 0.18   
## catBMIObese 16303 1734 9.40 4.6e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3180 on 120 degrees of freedom  
## Multiple R-squared: 0.659, Adjusted R-squared: 0.645   
## F-statistic: 46.5 on 5 and 120 DF, p-value: <2e-16

coef(m1)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 16137 3341 9150 227   
## catBMIOverweight catBMIObese   
## 1989 16303

m2 <- lm(formula= charges ~ catAge + catBMI + sex ,data = Reg\_Cluster1)  
summary(m2)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + sex, data = Reg\_Cluster1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13425 -1881 -347 1642 13092   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16217 1525 10.63 < 2e-16 \*\*\*  
## catAgePrime\_Age 3426 693 4.94 2.6e-06 \*\*\*  
## catAgeMature\_Age 9277 1754 5.29 5.7e-07 \*\*\*  
## catBMINormal 389 1503 0.26 0.80   
## catBMIOverweight 2287 1499 1.53 0.13   
## catBMIObese 16664 1753 9.51 2.8e-16 \*\*\*  
## sexmale -738 582 -1.27 0.21   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3180 on 119 degrees of freedom  
## Multiple R-squared: 0.664, Adjusted R-squared: 0.647   
## F-statistic: 39.2 on 6 and 119 DF, p-value: <2e-16

coef(m2)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 16217 3426 9277 389   
## catBMIOverweight catBMIObese sexmale   
## 2287 16664 -738

m3 <- lm(formula= charges ~ catAge + catBMI + sex + region ,data = Reg\_Cluster1)  
summary(m3)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + sex + region, data = Reg\_Cluster1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12868 -1835 -478 1759 11920   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15961.0 1535.3 10.40 < 2e-16 \*\*\*  
## catAgePrime\_Age 3375.4 691.6 4.88 3.4e-06 \*\*\*  
## catAgeMature\_Age 9025.2 1755.1 5.14 1.1e-06 \*\*\*  
## catBMINormal 159.5 1504.4 0.11 0.916   
## catBMIOverweight 1933.6 1515.2 1.28 0.204   
## catBMIObese 16410.8 1764.4 9.30 1.0e-15 \*\*\*  
## sexmale -665.0 578.1 -1.15 0.252   
## regionnorthwest 1477.8 781.4 1.89 0.061 .   
## regionsoutheast 662.6 791.1 0.84 0.404   
## regionsouthwest -70.5 839.8 -0.08 0.933   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3150 on 116 degrees of freedom  
## Multiple R-squared: 0.678, Adjusted R-squared: 0.653   
## F-statistic: 27.1 on 9 and 116 DF, p-value: <2e-16

coef(m3)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 15961.0 3375.4 9025.2 159.5   
## catBMIOverweight catBMIObese sexmale regionnorthwest   
## 1933.6 16410.8 -665.0 1477.8   
## regionsoutheast regionsouthwest   
## 662.6 -70.5

#Comparing the models  
anova(m1, m2) #the result shows a Df of 1

## Analysis of Variance Table  
##   
## Model 1: charges ~ catAge + catBMI  
## Model 2: charges ~ catAge + catBMI + sex  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 120 1.22e+09   
## 2 119 1.20e+09 1 16240229 1.61 0.21

#(indicating that the 2nd model has one additional parameter),   
 #and a very large p-value (0.14 > 0.1).   
 #This means that adding sex variable to the model   
 #did not lead to a significantly improved fit over the model 1.  
 #Choose model 1  
  
anova(m1, m3) #the result shows a Df of 4

## Analysis of Variance Table  
##   
## Model 1: charges ~ catAge + catBMI  
## Model 2: charges ~ catAge + catBMI + sex + region  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 120 1.22e+09   
## 2 116 1.15e+09 4 64639982 1.63 0.17

#(indicating that the 3rd model has 4 additional parameter),   
 #and a very large p-value (0.6409 > 0.1).   
 #This means that adding sex, region variable to the model   
 #did not lead to a significantly improved fit over the model 1.  
 #Choose model 1  
  
# Compute the analysis of variance for model 1A  
res.aov <- aov(charges ~ catAge + catBMI, data = Reg\_Cluster1)  
# Summary of the analysis  
summary(res.aov) #As the p-value is less than the significance level 0.1, we can conclude that there are significant differences between the groups highlighted with “\*" in the model summary.

## Df Sum Sq Mean Sq F value Pr(>F)   
## catAge 2 9.65e+07 4.82e+07 4.76 0.01 \*   
## catBMI 3 2.26e+09 7.53e+08 74.30 <2e-16 \*\*\*  
## Residuals 120 1.22e+09 1.01e+07   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Model 1A  
#Partition data  
set.seed(1) # set seed for reproducing the partition  
train.rows <- sample(row.names(Reg\_Cluster1), dim(Reg\_Cluster1)[1]\*0.6)   
train.data <- Reg\_Cluster1[train.rows,]   
colnames(train.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(train.data)

## [1] 75 5

str(train.data)

## 'data.frame': 75 obs. of 5 variables:  
## $ charges: num 14283 21196 19595 16885 24394 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 1 1 2 2 1 1 2 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 1 3 1 4 2 2 3 2 2 1 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 1 2 2 1 2 2 2 2 2 2 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 2 2 2 3 3 2 3 1 2 2 ...

valid.rows <- sample(setdiff(rownames(Reg\_Cluster1), train.rows))  
valid.data <- Reg\_Cluster1[valid.rows,]  
colnames(valid.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(valid.data)

## [1] 51 5

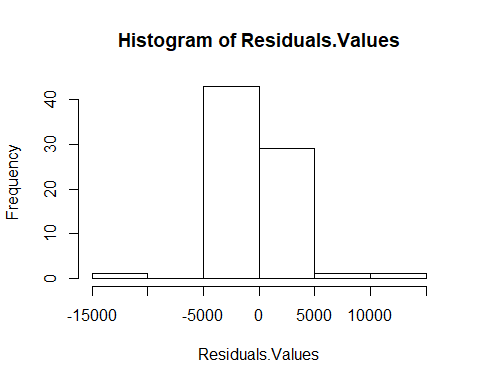
#Linear Regression on Charges (model 1A)  
charges.lm <- lm(charges ~ catAge + catBMI, data = train.data)   
options(scipen = 999)  
summary(charges.lm)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI, data = train.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12181 -2039 -390 1999 13339   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16469 1753 9.40 0.000000000000057 \*\*\*  
## catAgePrime\_Age 2926 949 3.08 0.0030 \*\*   
## catAgeMature\_Age 9893 3669 2.70 0.0088 \*\*   
## catBMINormal 564 1702 0.33 0.7412   
## catBMIOverweight 1979 1706 1.16 0.2500   
## catBMIObese 14930 2125 7.03 0.000000001193708 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3530 on 69 degrees of freedom  
## Multiple R-squared: 0.572, Adjusted R-squared: 0.541   
## F-statistic: 18.4 on 5 and 69 DF, p-value: 0.0000000000137

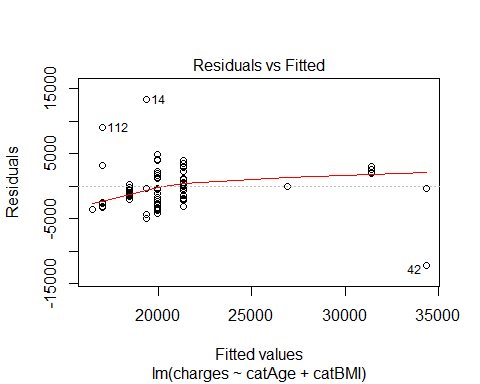
coef(charges.lm)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 16469 2926 9893 564   
## catBMIOverweight catBMIObese   
## 1979 14930

#Residuals and fitted plots  
Residuals.Values <- resid(charges.lm)  
Fitted.Values <- fitted(charges.lm)  
hist(Residuals.Values)



plot(charges.lm,1)



#Since the data set of cluster 1 is very less,  
#during the sampling, if the valid data contains Mature\_Age and Obese categories   
#which is of only two records, has to be removed to carryout the prediction.   
#(This step is necessary only if the valid.data contains any of the above two factors)  
valid.data<- valid.data %>%  
 filter(catAge!= 'Mature\_Age' & catBMI!= 'Obese')  
  
# use predict() to make predictions on a new set.  
charges.lm.pred <- predict(charges.lm, valid.data)  
  
# use accuracy() to compute common accuracy measures.  
accuracy(charges.lm.pred, valid.data$charges)

## ME RMSE MAE MPE MAPE  
## Test set -390 2801 2132 -3.9 10.7

#Predicting results  
results.df <- data.frame(cbind(actuals = valid.data$charges, predicted = charges.lm.pred))  
results.df <- results.df %>%  
 mutate(error = results.df$actuals - results.df$predicted) %>%  
 round(., 2)  
results.df <- results.df %>%  
 mutate(percerror = paste0(round(results.df$error/results.df$actuals\*100,2),"%"))  
  
kable(head(results.df))

|  |  |  |  |
| --- | --- | --- | --- |
| actuals | predicted | error | percerror |
| 19041 | 21374 | -2333 | -12.25% |
| 18034 | 18449 | -415 | -2.3% |
| 18972 | 21374 | -2402 | -12.66% |
| 18768 | 19959 | -1192 | -6.35% |
| 23307 | 21374 | 1932 | 8.29% |
| 24667 | 21374 | 3293 | 13.35% |

#Our model will be able to predict the insurance premium for   
#policy holders under cluster1 with only having smoking habit   
#with a mean difference of the result from the below output  
sprintf("The mean percent error is: %s%%", round(mean(results.df$error/results.df$actuals\*100), 2))

## [1] "The mean percent error is: -3.9%"

#predicting with new data using model 1A  
m1.lm.pred <- predict(m1, newdata = data.frame(sex='female', region='northwest', catAge='Prime\_Age', catBMI='Obese'))  
m1.lm.pred

## 1   
## 35781

#Cluster 2  
Reg\_Cluster2 <- newInsuranceData %>%  
 filter(clusters==2)  
str(Reg\_Cluster2)

## 'data.frame': 422 obs. of 7 variables:  
## $ charges : num 1726 28923 27809 1827 11091 ...  
## $ sex : Factor w/ 2 levels "female","male": 2 1 1 2 1 2 2 2 1 2 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 3 2 3 4 3 4 1 4 1 3 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 1 3 3 1 3 1 1 3 3 1 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 4 3 3 4 4 2 2 4 4 4 ...  
## $ clusters: int 2 2 2 2 2 2 2 2 2 2 ...

dim(Reg\_Cluster2)

## [1] 422 7

Reg\_Cluster2<-Reg\_Cluster2[,-c(3,7)] #Smoker column contains only 'No' values for all the data, thus it will not have any effect over prediction  
  
#Model Build  
#Without outliers  
  
m1 <- lm(formula= charges ~ catAge + catBMI ,data = Reg\_Cluster2)  
summary(m1)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI, data = Reg\_Cluster2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5125 -2435 -1617 -441 24137   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2177 2581 0.84 0.40   
## catAgePrime\_Age 1721 4470 0.39 0.70   
## catAgeMature\_Age 10673 513 20.82 <0.0000000000000002 \*\*\*  
## catBMINormal 1546 2660 0.58 0.56   
## catBMIOverweight 2489 2637 0.94 0.35   
## catBMIObese 1410 2618 0.54 0.59   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5160 on 416 degrees of freedom  
## Multiple R-squared: 0.526, Adjusted R-squared: 0.52   
## F-statistic: 92.2 on 5 and 416 DF, p-value: <0.0000000000000002

coef(m1)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 2177 1721 10673 1546   
## catBMIOverweight catBMIObese   
## 2489 1410

m2 <- lm(formula= charges ~ catAge + catBMI + sex ,data = Reg\_Cluster2)  
summary(m2)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + sex, data = Reg\_Cluster2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4800 -2334 -1655 -543 24439   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2477 2592 0.96 0.34   
## catAgePrime\_Age 2021 4475 0.45 0.65   
## catAgeMature\_Age 10661 512 20.80 <0.0000000000000002 \*\*\*  
## catBMINormal 1580 2658 0.59 0.55   
## catBMIOverweight 2476 2636 0.94 0.35   
## catBMIObese 1407 2616 0.54 0.59   
## sexmale -600 504 -1.19 0.24   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5160 on 415 degrees of freedom  
## Multiple R-squared: 0.527, Adjusted R-squared: 0.521   
## F-statistic: 77.2 on 6 and 415 DF, p-value: <0.0000000000000002

coef(m2)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 2477 2021 10661 1580   
## catBMIOverweight catBMIObese sexmale   
## 2476 1407 -600

m3 <- lm(formula= charges ~ catAge + catBMI + sex + region ,data = Reg\_Cluster2)  
summary(m3) #Choose model 3 with the help of summary function( high r-square value)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + sex + region, data = Reg\_Cluster2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5168 -2406 -1687 -363 24311   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2997 2653 1.13 0.259   
## catAgePrime\_Age 1110 4466 0.25 0.804   
## catAgeMature\_Age 10674 511 20.90 <0.0000000000000002 \*\*\*  
## catBMINormal 1376 2653 0.52 0.604   
## catBMIOverweight 2209 2634 0.84 0.402   
## catBMIObese 1176 2619 0.45 0.654   
## sexmale -580 502 -1.15 0.249   
## regionnorthwest 371 713 0.52 0.603   
## regionsoutheast -180 723 -0.25 0.803   
## regionsouthwest -1431 723 -1.98 0.048 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5130 on 412 degrees of freedom  
## Multiple R-squared: 0.536, Adjusted R-squared: 0.525   
## F-statistic: 52.8 on 9 and 412 DF, p-value: <0.0000000000000002

coef(m3)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 2997 1110 10674 1376   
## catBMIOverweight catBMIObese sexmale regionnorthwest   
## 2209 1176 -580 371   
## regionsoutheast regionsouthwest   
## -180 -1431

m4 <- lm(formula= charges ~ catAge + region ,data = Reg\_Cluster2)  
summary(m4)

##   
## Call:  
## lm(formula = charges ~ catAge + region, data = Reg\_Cluster2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4746 -2400 -1695 -577 23837   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4253 576 7.38 0.00000000000087 \*\*\*  
## catAgePrime\_Age -696 3675 -0.19 0.850   
## catAgeMature\_Age 10746 502 21.41 < 0.0000000000000002 \*\*\*  
## regionnorthwest 342 712 0.48 0.631   
## regionsoutheast -366 714 -0.51 0.608   
## regionsouthwest -1523 722 -2.11 0.035 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5140 on 416 degrees of freedom  
## Multiple R-squared: 0.53, Adjusted R-squared: 0.524   
## F-statistic: 93.8 on 5 and 416 DF, p-value: <0.0000000000000002

coef(m4)

## (Intercept) catAgePrime\_Age catAgeMature\_Age regionnorthwest   
## 4253 -696 10746 342   
## regionsoutheast regionsouthwest   
## -366 -1523

m5 <- lm(formula= charges ~ catAge + sex ,data = Reg\_Cluster2)  
summary(m5)

##   
## Call:  
## lm(formula = charges ~ catAge + sex, data = Reg\_Cluster2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4212 -2312 -1723 -682 24158   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4186 435 9.62 <0.0000000000000002 \*\*\*  
## catAgePrime\_Age 333 3678 0.09 0.93   
## catAgeMature\_Age 10730 504 21.27 <0.0000000000000002 \*\*\*  
## sexmale -620 504 -1.23 0.22   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5160 on 418 degrees of freedom  
## Multiple R-squared: 0.523, Adjusted R-squared: 0.519   
## F-statistic: 153 on 3 and 418 DF, p-value: <0.0000000000000002

coef(m5)

## (Intercept) catAgePrime\_Age catAgeMature\_Age sexmale   
## 4186 333 10730 -620

# Compute the analysis of variance for model 3  
res.aov <- aov(charges ~ catAge + catBMI + sex + region, data = Reg\_Cluster2)  
# Summary of the analysis  
summary(res.aov) #As the p-value is less than the significance level 0.1, we can conclude that there are significant differences between the groups highlighted with “\*" in the model summary.

## Df Sum Sq Mean Sq F value Pr(>F)   
## catAge 2 12178939461 6089469730 231.13 <0.0000000000000002 \*\*\*  
## catBMI 3 107086402 35695467 1.35 0.256   
## sex 1 37619719 37619719 1.43 0.233   
## region 3 191008692 63669564 2.42 0.066 .   
## Residuals 412 10854756680 26346497   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Model 2A(Without outliers, lets call model 3 as model 2A)  
#Partition data  
set.seed(1) # set seed for reproducing the partition  
train.rows <- sample(row.names(Reg\_Cluster2), dim(Reg\_Cluster2)[1]\*0.6)   
train.data <- Reg\_Cluster2[train.rows,]   
colnames(train.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(train.data)

## [1] 253 5

str(train.data)

## 'data.frame': 253 obs. of 5 variables:  
## $ charges: num 2730 1749 16586 10796 29331 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 1 2 1 1 1 1 2 1 2 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 2 4 3 4 2 3 3 2 1 4 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 1 1 1 1 3 1 1 1 3 3 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 4 4 4 4 3 4 4 3 3 4 ...

valid.rows <- sample(setdiff(rownames(Reg\_Cluster2), train.rows))  
valid.data <- Reg\_Cluster2[valid.rows,]  
colnames(valid.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(valid.data)

## [1] 169 5

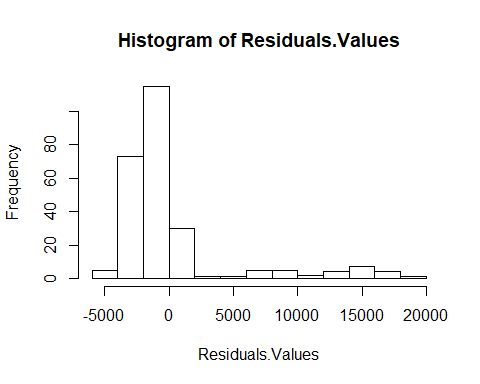
#Linear Regression on Charges(Model 2A)  
charges.lm <- lm(charges ~ catAge + catBMI + sex + region, data = train.data)   
options(scipen = 999)  
summary(charges.lm)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + sex + region, data = train.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5087 -2256 -1476 -91 19857   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2834 2523 1.12 0.26   
## catAgePrime\_Age 1362 4162 0.33 0.74   
## catAgeMature\_Age 10616 619 17.14 <0.0000000000000002 \*\*\*  
## catBMINormal 906 2528 0.36 0.72   
## catBMIOverweight 2269 2495 0.91 0.36   
## catBMIObese 1242 2474 0.50 0.62   
## sexmale 165 606 0.27 0.79   
## regionnorthwest -462 856 -0.54 0.59   
## regionsoutheast -905 854 -1.06 0.29   
## regionsouthwest -1015 879 -1.16 0.25   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4770 on 243 degrees of freedom  
## Multiple R-squared: 0.577, Adjusted R-squared: 0.561   
## F-statistic: 36.8 on 9 and 243 DF, p-value: <0.0000000000000002

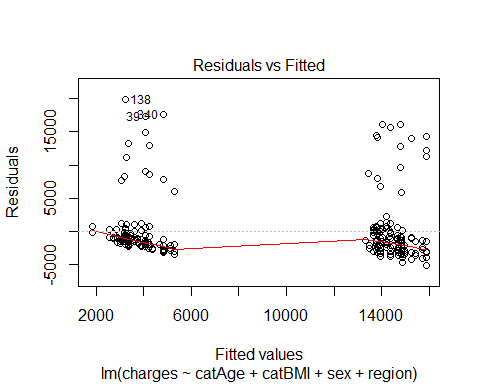
coef(charges.lm)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 2834 1362 10615 906   
## catBMIOverweight catBMIObese sexmale regionnorthwest   
## 2269 1242 165 -462   
## regionsoutheast regionsouthwest   
## -905 -1015

#Residuals and fitted plots  
Residuals.Values <- resid(charges.lm)  
Fitted.Values <- fitted(charges.lm)  
hist(Residuals.Values)



plot(charges.lm,1)



# use predict() to make predictions on a new set.  
charges.lm.pred <- predict(charges.lm, valid.data)  
  
# use accuracy() to compute common accuracy measures.  
accuracy(charges.lm.pred, valid.data$charges)

## ME RMSE MAE MPE MAPE  
## Test set 378 5713 3306 -42.3 58.9

#Predicting results  
results.df <- data.frame(cbind(actuals = valid.data$charges, predicted = charges.lm.pred))  
results.df <- results.df %>%  
 mutate(error = results.df$actuals - results.df$predicted) %>%  
 round(., 2)  
results.df <- results.df %>%  
 mutate(percerror = paste0(round(results.df$error/results.df$actuals\*100,2),"%"))  
  
kable(head(results.df))

|  |  |  |  |
| --- | --- | --- | --- |
| actuals | predicted | error | percerror |
| 1532 | 3336 | -1803 | -117.68% |
| 12630 | 13450 | -820 | -6.49% |
| 12957 | 13505 | -548 | -4.23% |
| 12431 | 14691 | -2260 | -18.18% |
| 3044 | 5102 | -2058 | -67.6% |
| 28950 | 15255 | 13695 | 47.31% |

#Our model will be able to predict the insurance premium for   
#policy holders under cluster 2 wtih not having smoking habit with a mean difference of the result from the below output  
sprintf("The mean percent error is: %s%%", round(mean(results.df$error/results.df$actuals\*100), 2))

## [1] "The mean percent error is: -42.29%"

#Cluster 3  
library(tidyverse)  
Reg\_Cluster3 <- newInsuranceData %>%  
 filter(clusters==3)  
str(Reg\_Cluster3)

## 'data.frame': 645 obs. of 7 variables:  
## $ charges : num 4449 21984 3867 3757 8241 ...  
## $ sex : Factor w/ 2 levels "female","male": 2 2 2 1 1 1 2 2 1 1 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 3 2 2 3 3 2 1 1 1 4 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 4 2 3 3 4 3 3 3 4 4 ...  
## $ clusters: int 3 3 3 3 3 3 3 3 3 3 ...

dim(Reg\_Cluster3)

## [1] 645 7

Reg\_Cluster3<-Reg\_Cluster3[,-c(3,7)] #Smoker column contains only 'No' values for all the data, thus it will not have any effect over prediction  
  
#Model Build  
m1 <- lm(charges ~ catAge ,data = Reg\_Cluster3)  
summary(m1)

##   
## Call:  
## lm(formula = charges ~ catAge, data = Reg\_Cluster3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6113 -3301 -1049 1291 25221   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2431 3639 0.67 0.504   
## catAgePrime\_Age 5820 3644 1.60 0.111   
## catAgeMature\_Age 9938 5146 1.93 0.054 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5150 on 642 degrees of freedom  
## Multiple R-squared: 0.00594, Adjusted R-squared: 0.00285   
## F-statistic: 1.92 on 2 and 642 DF, p-value: 0.148

coef(m1)

## (Intercept) catAgePrime\_Age catAgeMature\_Age   
## 2431 5820 9938

m2 <- lm(formula= charges ~ catAge + catBMI ,data = Reg\_Cluster3)  
summary(m2)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI, data = Reg\_Cluster3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6149 -3222 -1028 1251 25000   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2431 3643 0.67 0.505   
## catAgePrime\_Age 4946 4311 1.15 0.252   
## catAgeMature\_Age 9938 5152 1.93 0.054 .  
## catBMINormal 672 2358 0.29 0.776   
## catBMIOverweight 639 2333 0.27 0.784   
## catBMIObese 1095 2322 0.47 0.637   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5150 on 639 degrees of freedom  
## Multiple R-squared: 0.008, Adjusted R-squared: 0.000242   
## F-statistic: 1.03 on 5 and 639 DF, p-value: 0.398

coef(m2)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 2431 4946 9938 672   
## catBMIOverweight catBMIObese   
## 639 1095

m3 <- lm(formula= charges ~ catAge + catBMI + sex ,data = Reg\_Cluster3)  
summary(m3)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + sex, data = Reg\_Cluster3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6020 -3297 -1030 1223 24819   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2611 3650 0.72 0.475   
## catAgePrime\_Age 4838 4313 1.12 0.262   
## catAgeMature\_Age 9938 5153 1.93 0.054 .  
## catBMINormal 763 2361 0.32 0.747   
## catBMIOverweight 732 2336 0.31 0.754   
## catBMIObese 1204 2325 0.52 0.605   
## sexmale -359 407 -0.88 0.378   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5150 on 638 degrees of freedom  
## Multiple R-squared: 0.00921, Adjusted R-squared: -0.000105   
## F-statistic: 0.989 on 6 and 638 DF, p-value: 0.432

coef(m3)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 2611 4838 9938 763   
## catBMIOverweight catBMIObese sexmale   
## 732 1204 -359

m4 <- lm(formula= charges ~ catAge + catBMI + sex + region ,data = Reg\_Cluster3)  
summary(m4)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + sex + region, data = Reg\_Cluster3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6754 -3102 -1059 1255 24525   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2641 3624 0.73 0.4664   
## catAgePrime\_Age 5202 4289 1.21 0.2256   
## catAgeMature\_Age 9938 5117 1.94 0.0526 .   
## catBMINormal 1303 2349 0.55 0.5795   
## catBMIOverweight 1464 2329 0.63 0.5298   
## catBMIObese 2058 2323 0.89 0.3758   
## sexmale -420 405 -1.04 0.2993   
## regionnorthwest -955 576 -1.66 0.0978 .   
## regionsoutheast -1736 584 -2.97 0.0031 \*\*  
## regionsouthwest -1756 579 -3.03 0.0025 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5120 on 635 degrees of freedom  
## Multiple R-squared: 0.0277, Adjusted R-squared: 0.0139   
## F-statistic: 2.01 on 9 and 635 DF, p-value: 0.0361

coef(m4)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 2641 5202 9938 1303   
## catBMIOverweight catBMIObese sexmale regionnorthwest   
## 1464 2059 -420 -955   
## regionsoutheast regionsouthwest   
## -1736 -1756

m5 <- lm(formula= charges ~ catAge + catBMI + region ,data = Reg\_Cluster3)  
summary(m5)#Choose model 5

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + region, data = Reg\_Cluster3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6948 -3126 -963 1201 24745   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2431 3619 0.67 0.5019   
## catAgePrime\_Age 5326 4288 1.24 0.2147   
## catAgeMature\_Age 9938 5118 1.94 0.0526 .   
## catBMINormal 1189 2347 0.51 0.6126   
## catBMIOverweight 1346 2326 0.58 0.5631   
## catBMIObese 1919 2319 0.83 0.4082   
## regionnorthwest -949 576 -1.65 0.0999 .   
## regionsoutheast -1716 584 -2.94 0.0034 \*\*  
## regionsouthwest -1733 579 -2.99 0.0029 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5120 on 636 degrees of freedom  
## Multiple R-squared: 0.026, Adjusted R-squared: 0.0138   
## F-statistic: 2.12 on 8 and 636 DF, p-value: 0.0317

coef(m5)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 2431 5326 9938 1189   
## catBMIOverweight catBMIObese regionnorthwest regionsoutheast   
## 1346 1919 -949 -1716   
## regionsouthwest   
## -1733

m6 <- lm(formula= charges ~ catAge + region ,data = Reg\_Cluster3)  
summary(m6)

##   
## Call:  
## lm(formula = charges ~ catAge + region, data = Reg\_Cluster3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6557 -3111 -1026 1189 25094   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2431 3618 0.67 0.5019   
## catAgePrime\_Age 6847 3641 1.88 0.0605 .   
## catAgeMature\_Age 9938 5117 1.94 0.0525 .   
## regionnorthwest -901 575 -1.57 0.1176   
## regionsoutheast -1546 574 -2.69 0.0072 \*\*  
## regionsouthwest -1625 574 -2.83 0.0048 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5120 on 639 degrees of freedom  
## Multiple R-squared: 0.0217, Adjusted R-squared: 0.014   
## F-statistic: 2.83 on 5 and 639 DF, p-value: 0.0154

coef(m6)

## (Intercept) catAgePrime\_Age catAgeMature\_Age regionnorthwest   
## 2431 6847 9938 -901   
## regionsoutheast regionsouthwest   
## -1546 -1625

# Compute the analysis of variance for model 5  
res.aov <- aov(charges ~ catAge + catBMI + region, data = Reg\_Cluster3)  
# Summary of the analysis  
summary(res.aov) #As the p-value is less than the significance level 0.1, we can conclude that there are significant differences between the groups highlighted with “\*" in the model summary.

## Df Sum Sq Mean Sq F value Pr(>F)   
## catAge 2 101646997 50823498 1.94 0.1445   
## catBMI 3 35227035 11742345 0.45 0.7185   
## region 3 308107212 102702404 3.92 0.0086 \*\*  
## Residuals 636 16655797697 26188361   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Model 3A (without outliers, naming model 5 from the above as model 3A)  
#Partition data  
set.seed(1) # set seed for reproducing the partition  
train.rows <- sample(row.names(Reg\_Cluster3), dim(Reg\_Cluster3)[1]\*0.6)   
train.data <- Reg\_Cluster3[train.rows,]   
colnames(train.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(train.data)

## [1] 387 5

str(train.data)

## 'data.frame': 387 obs. of 5 variables:  
## $ charges: num 7742 8978 4519 18806 6356 ...  
## $ sex : Factor w/ 2 levels "female","male": 2 2 2 2 2 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 3 3 1 3 3 3 4 3 3 1 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 3 4 4 4 4 2 4 3 3 1 ...

valid.rows <- sample(setdiff(rownames(Reg\_Cluster3), train.rows))  
valid.data <- Reg\_Cluster3[valid.rows,]  
colnames(valid.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(valid.data)

## [1] 258 5

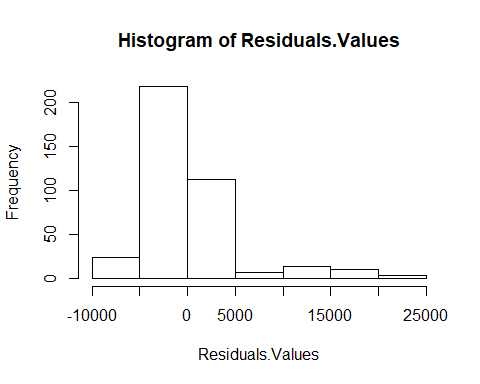
#Linear Regression on Charges(model 3A)  
charges.lm <- lm(charges ~ catAge + catBMI + region, data = train.data)   
options(scipen = 999)  
summary(charges.lm)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + region, data = train.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7478 -2934 -922 1225 24847   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3168 5112 0.62 0.5359   
## catAgePrime\_Age 4226 5908 0.72 0.4748   
## catAgeMature\_Age 8367 7230 1.16 0.2479   
## catBMINormal 1389 3039 0.46 0.6480   
## catBMIOverweight 1773 3004 0.59 0.5553   
## catBMIObese 2811 2990 0.94 0.3477   
## regionnorthwest -1580 740 -2.14 0.0334 \*   
## regionsoutheast -2038 752 -2.71 0.0071 \*\*  
## regionsouthwest -2352 740 -3.18 0.0016 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5110 on 378 degrees of freedom  
## Multiple R-squared: 0.0424, Adjusted R-squared: 0.0221   
## F-statistic: 2.09 on 8 and 378 DF, p-value: 0.0359

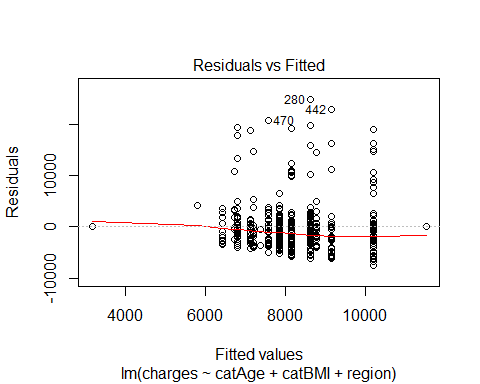
coef(charges.lm)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 3167 4226 8367 1389   
## catBMIOverweight catBMIObese regionnorthwest regionsoutheast   
## 1773 2811 -1580 -2038   
## regionsouthwest   
## -2352

#Residuals and fitted plots  
Residuals.Values <- resid(charges.lm)  
Fitted.Values <- fitted(charges.lm)  
hist(Residuals.Values)



plot(charges.lm,1)



# use predict() to make predictions on a new set.  
library(forecast)  
charges.lm.pred <- predict(charges.lm, valid.data)  
  
# use accuracy() to compute common accuracy measures.  
accuracy(charges.lm.pred, valid.data$charges)

## ME RMSE MAE MPE MAPE  
## Test set 335 5164 3363 -24.3 47.3

#Predicting results  
results.df <- data.frame(cbind(actuals = valid.data$charges, predicted = charges.lm.pred))  
results.df <- results.df %>%  
 mutate(error = results.df$actuals - results.df$predicted) %>%  
 round(., 2)  
results.df <- results.df %>%  
 mutate(percerror = paste0(round(results.df$error/results.df$actuals\*100,2),"%"))  
  
kable(head(results.df))

|  |  |  |  |
| --- | --- | --- | --- |
| actuals | predicted | error | percerror |
| 2302 | 6431 | -4128 | -179.32% |
| 3471 | 8167 | -4696 | -135.27% |
| 26467 | 8625 | 17842 | 67.41% |
| 5478 | 7129 | -1651 | -30.14% |
| 7345 | 7854 | -508 | -6.92% |
| 3206 | 8782 | -5576 | -173.9% |

#Our model will be able to predict the insurance premium for   
#policy holders under cluster3 wtih not having smoking habit with a mean difference of the result from the below output  
sprintf("The mean percent error is: %s%%", round(mean(results.df$error/results.df$actuals\*100), 2))

## [1] "The mean percent error is: -24.32%"

#---  
  
  
#WITH OUTLIERS  
#MLR ON 3 CLUSTERS WITH OUTLIERS  
#remove children variable from the dataset  
insurance\_data1<-insurance\_data[,-4]  
str(insurance\_data1)

## 'data.frame': 1338 obs. of 6 variables:  
## $ age : int 19 18 28 33 32 31 46 37 37 60 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...  
## $ bmi : num 27.9 33.8 33 22.7 28.9 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...  
## $ charges: num 16885 1726 4449 21984 3867 ...

#DATA PREPARATION  
#converting Age to Categorical variables  
catAge<- cut(insurance\_data1$age,breaks=c(15,24,54,64),labels=c("Early\_Age","Prime\_Age","Mature\_Age"))  
insurance\_data1<-cbind(insurance\_data1,catAge)  
table(insurance\_data1$catAge)

##   
## Early\_Age Prime\_Age Mature\_Age   
## 278 818 242

aggregate(insurance\_data1$charges,list(insurance\_data1$catAge),mean)

## Group.1 x  
## 1 Early\_Age 9011  
## 2 Prime\_Age 13167  
## 3 Mature\_Age 18513

#converting BMI to Categorical variables  
catBMI<- cut(insurance\_data1$bmi,breaks=c(0,18.499,24.999,29.999,60),labels=c("Underweight","Normal","Overweight","Obese"))  
insurance\_data1<-cbind(insurance\_data1,catBMI)  
table(insurance\_data1$catBMI)

##   
## Underweight Normal Overweight Obese   
## 20 225 386 707

aggregate(insurance\_data1$charges,list(insurance\_data1$catBMI),mean)

## Group.1 x  
## 1 Underweight 8852  
## 2 Normal 10409  
## 3 Overweight 10988  
## 4 Obese 15552

#Creating new Data Set  
newInsuranceData <- cbind(insurance\_data1[c(6,2,4,5)],catAge,catBMI)  
str(newInsuranceData)

## 'data.frame': 1338 obs. of 6 variables:  
## $ charges: num 16885 1726 4449 21984 3867 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 1 1 2 2 2 2 2 2 2 3 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 3 4 4 2 3 3 4 3 3 3 ...

summary(newInsuranceData)

## charges sex smoker region catAge   
## Min. : 1122 female:662 no :1064 northeast:324 Early\_Age :278   
## 1st Qu.: 4740 male :676 yes: 274 northwest:325 Prime\_Age :818   
## Median : 9382 southeast:364 Mature\_Age:242   
## Mean :13270 southwest:325   
## 3rd Qu.:16640   
## Max. :63770   
## catBMI   
## Underweight: 20   
## Normal :225   
## Overweight :386   
## Obese :707   
##   
##

#H-Clust  
#clustering H-clust by using metric = 'GOWER'  
insuranceDist<-daisy(newInsuranceData,metric="gower")  
summary(insuranceDist)

## 894453 dissimilarities, summarized :  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.346 0.511 0.489 0.609 1.000   
## Metric : mixed ; Types = I, N, N, N, N, N   
## Number of objects : 1338

str(insuranceDist)

## 'dissimilarity' num [1:894453] 0.707 0.866 0.847 0.701 0.535 ...  
## - attr(\*, "Size")= int 1338  
## - attr(\*, "Metric")= chr "mixed"  
## - attr(\*, "Types")= chr [1:6] "I" "N" "N" "N" ...

#hClustering using ward.D2  
fitH<- hclust(insuranceDist,"ward.D2")  
clusters<-cutree(fitH,3)  
  
newInsuranceData<-cbind(newInsuranceData,clusters)  
aggregate(newInsuranceData[,c(1)],list(newInsuranceData$clusters),mean)

## Group.1 x  
## 1 1 32086  
## 2 2 9210  
## 3 3 8014

table(newInsuranceData$clusters)

##   
## 1 2 3   
## 273 386 679

#Multi Linear Regression Analysis for premium charges calculations  
#Cluster 1  
Reg\_Cluster1 <- newInsuranceData %>%  
 filter(clusters==1)  
str(Reg\_Cluster1)

## 'data.frame': 273 obs. of 7 variables:  
## $ charges : num 16885 27809 39612 36837 37702 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 1 2 2 1 2 2 2 2 2 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 4 1 4 4 4 1 4 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 1 3 2 2 2 2 1 2 2 3 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 3 3 4 4 4 4 4 4 4 4 ...  
## $ clusters: int 1 1 1 1 1 1 1 1 1 1 ...

dim(Reg\_Cluster1)

## [1] 273 7

Reg\_Cluster1<-Reg\_Cluster1[,-c(3,7)] #Smoker column contains only 'Yes' values for all the data, thus it will not have any effect over prediction  
  
#model build  
#with outliers  
m1 <- lm(charges ~ catAge + catBMI ,data = Reg\_Cluster1)  
summary(m1)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI, data = Reg\_Cluster1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19580 -2476 -609 1260 22047   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14770 2113 6.99 0.0000000000219 \*\*\*  
## catAgePrime\_Age 5050 687 7.35 0.0000000000024 \*\*\*  
## catAgeMature\_Age 10534 926 11.37 < 0.0000000000000002 \*\*\*  
## catBMINormal 561 2142 0.26 0.79   
## catBMIOverweight 3015 2112 1.43 0.15   
## catBMIObese 21904 2080 10.53 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4560 on 267 degrees of freedom  
## Multiple R-squared: 0.847, Adjusted R-squared: 0.844   
## F-statistic: 295 on 5 and 267 DF, p-value: <0.0000000000000002

coef(m1)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 14770 5050 10534 561   
## catBMIOverweight catBMIObese   
## 3015 21904

# Compute the analysis of variance for model 1B  
res.aov <- aov(charges ~ catAge + catBMI, data = Reg\_Cluster1)  
# Summary of the analysis  
summary(res.aov) #As the p-value is less than the significance level 0.1, we can conclude that there are significant differences between the groups highlighted with “\*" in the model summary.

## Df Sum Sq Mean Sq F value Pr(>F)   
## catAge 2 3779146662 1889573331 90.8 <0.0000000000000002 \*\*\*  
## catBMI 3 26937247378 8979082459 431.5 <0.0000000000000002 \*\*\*  
## Residuals 267 5555978676 20808909   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Model 1B  
#Partition data  
set.seed(1) # set seed for reproducing the partition  
train.rows <- sample(row.names(Reg\_Cluster1), dim(Reg\_Cluster1)[1]\*0.6)   
train.data <- Reg\_Cluster1[train.rows,]   
colnames(train.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(train.data)

## [1] 163 5

str(train.data)

## 'data.frame': 163 obs. of 5 variables:  
## $ charges: num 37270 24667 18766 38999 38126 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 1 1 2 2 2 2 2 1 2 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 1 3 2 4 3 2 2 4 1 3 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 2 2 2 2 1 3 2 2 2 2 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 4 3 2 4 4 4 3 3 2 2 ...

valid.rows <- sample(setdiff(rownames(Reg\_Cluster1), train.rows))  
valid.data <- Reg\_Cluster1[valid.rows,]  
colnames(valid.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(valid.data)

## [1] 110 5

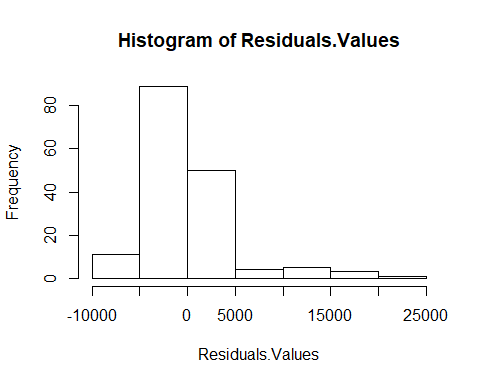
#Linear Regression on Charges (model 1B)  
charges.lm <- lm(charges ~ catAge + catBMI, data = train.data)   
options(scipen = 999)  
summary(charges.lm)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI, data = train.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8115 -2466 -621 1237 21755   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14972 2235 6.70 0.000000000355558 \*\*\*  
## catAgePrime\_Age 4797 922 5.20 0.000000609830847 \*\*\*  
## catAgeMature\_Age 10121 1184 8.55 0.000000000000011 \*\*\*  
## catBMINormal 200 2272 0.09 0.93   
## catBMIOverweight 3582 2229 1.61 0.11   
## catBMIObese 22246 2190 10.16 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4720 on 157 degrees of freedom  
## Multiple R-squared: 0.847, Adjusted R-squared: 0.842   
## F-statistic: 173 on 5 and 157 DF, p-value: <0.0000000000000002

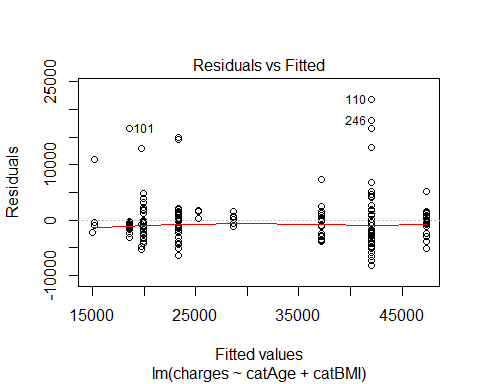
coef(charges.lm)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 14972 4797 10121 200   
## catBMIOverweight catBMIObese   
## 3582 22246

#Residuals and fitted plots  
Residuals.Values <- resid(charges.lm)  
Fitted.Values <- fitted(charges.lm)  
hist(Residuals.Values)



plot(charges.lm,1)



#Since the data set of cluster 1 is very less,  
#during the sampling, if the valid data contains Mature\_Age and Obese categories   
#which is of only two records, has to be removed to carryout the prediction.   
#(This step is necessary only if the valid.data contains any of the above two factors)  
valid.data<- valid.data %>%  
 filter(catAge!= 'Mature\_Age' & catBMI!= 'Obese')  
  
# use predict() to make predictions on a new set.  
charges.lm.pred <- predict(charges.lm, valid.data)  
  
# use accuracy() to compute common accuracy measures.  
accuracy(charges.lm.pred, valid.data$charges)

## ME RMSE MAE MPE MAPE  
## Test set -531 3961 2954 -5 13.8

#Predicting results  
results.df <- data.frame(cbind(actuals = valid.data$charges, predicted = charges.lm.pred))  
results.df <- results.df %>%  
 mutate(error = results.df$actuals - results.df$predicted) %>%  
 round(., 2)  
results.df <- results.df %>%  
 mutate(percerror = paste0(round(results.df$error/results.df$actuals\*100,2),"%"))  
  
kable(head(results.df))

|  |  |  |  |
| --- | --- | --- | --- |
| actuals | predicted | error | percerror |
| 21224 | 19970 | 1254 | 5.91% |
| 24520 | 23351 | 1169 | 4.77% |
| 19350 | 23351 | -4000 | -20.67% |
| 21349 | 23351 | -2002 | -9.38% |
| 14572 | 15172 | -600 | -4.12% |
| 15818 | 19970 | -4152 | -26.25% |

#Our model will be able to predict the insurance premium for   
#policy holders under cluster1 with only having smoking habit   
#with a mean difference of the result from the below output  
sprintf("The mean percent error is: %s%%", round(mean(results.df$error/results.df$actuals\*100), 2))

## [1] "The mean percent error is: -5%"

#predicting with new data using model 1B  
m1.lm.pred <- predict(m1, newdata = data.frame(sex='female', region='northwest', catAge='Prime\_Age', catBMI='Obese'))  
m1.lm.pred

## 1   
## 41724

#Cluster 2  
Reg\_Cluster2 <- newInsuranceData %>%  
 filter(clusters==2)  
str(Reg\_Cluster2)

## 'data.frame': 386 obs. of 7 variables:  
## $ charges : num 1726 28923 1827 11091 10602 ...  
## $ sex : Factor w/ 2 levels "female","male": 2 1 2 1 2 1 2 1 1 1 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 3 2 4 3 4 1 3 3 1 2 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 1 3 1 3 3 3 1 3 3 3 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 4 3 4 4 4 4 4 3 2 4 ...  
## $ clusters: int 2 2 2 2 2 2 2 2 2 2 ...

dim(Reg\_Cluster2)

## [1] 386 7

Reg\_Cluster2<-Reg\_Cluster2[,-c(3,7)] #Smoker column contains only 'No' values for all the data, thus it will not have any effect over prediction  
  
#Model Build  
#With outliers  
  
m3 <- lm(formula= charges ~ catAge + catBMI + sex + region ,data = Reg\_Cluster2)  
summary(m3)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + sex + region, data = Reg\_Cluster2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3504 -2291 -1605 -431 24135   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2717 3655 0.74 0.46   
## catAgeMature\_Age 10142 519 19.54 <0.0000000000000002 \*\*\*  
## catBMINormal 2265 3700 0.61 0.54   
## catBMIOverweight 1483 3654 0.41 0.69   
## catBMIObese 1896 3644 0.52 0.60   
## sexmale -694 524 -1.33 0.19   
## regionnorthwest 11 748 0.01 0.99   
## regionsoutheast -330 742 -0.45 0.66   
## regionsouthwest -560 747 -0.75 0.45   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5060 on 377 degrees of freedom  
## Multiple R-squared: 0.509, Adjusted R-squared: 0.499   
## F-statistic: 48.9 on 8 and 377 DF, p-value: <0.0000000000000002

coef(m3)

## (Intercept) catAgeMature\_Age catBMINormal catBMIOverweight   
## 2717 10141 2265 1483   
## catBMIObese sexmale regionnorthwest regionsoutheast   
## 1896 -694 11 -330   
## regionsouthwest   
## -560

# Compute the analysis of variance for model 3  
res.aov <- aov(charges ~ catAge + catBMI + sex + region, data = Reg\_Cluster2)  
# Summary of the analysis  
summary(res.aov) #As the p-value is less than the significance level 0.1, we can conclude that there are significant differences between the groups highlighted with “\*" in the model summary.

## Df Sum Sq Mean Sq F value Pr(>F)   
## catAge 1 9920062387 9920062387 387.47 <0.0000000000000002 \*\*\*  
## catBMI 3 33638634 11212878 0.44 0.73   
## sex 1 46922738 46922738 1.83 0.18   
## region 3 21502260 7167420 0.28 0.84   
## Residuals 377 9652087022 25602353   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Model 2B(With outliers, lets call model 3 as model 2B)  
#Partition data  
set.seed(1) # set seed for reproducing the partition  
train.rows <- sample(row.names(Reg\_Cluster2), dim(Reg\_Cluster2)[1]\*0.6)   
train.data <- Reg\_Cluster2[train.rows,]   
colnames(train.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(train.data)

## [1] 231 5

str(train.data)

## 'data.frame': 231 obs. of 5 variables:  
## $ charges: num 15020 1708 2255 3280 3443 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 2 1 1 2 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 2 1 1 1 4 3 2 3 2 2 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 3 1 1 1 1 1 1 1 1 3 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 4 3 4 3 4 4 2 3 3 3 ...

valid.rows <- sample(setdiff(rownames(Reg\_Cluster2), train.rows))  
valid.data <- Reg\_Cluster2[valid.rows,]  
colnames(valid.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(valid.data)

## [1] 155 5

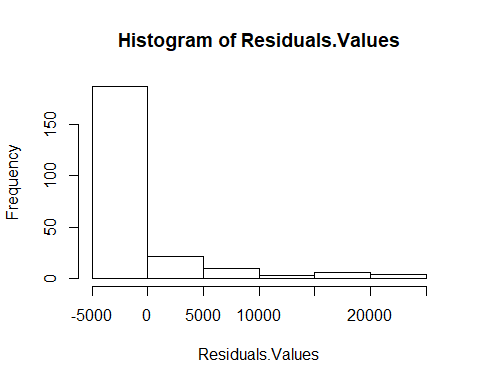
#Linear Regression on Charges(Model 2B)  
charges.lm <- lm(charges ~ catAge + catBMI + sex + region, data = train.data)   
options(scipen = 999)  
summary(charges.lm)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + sex + region, data = train.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4090 -2228 -1374 -308 23770   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2323.6 5015.8 0.46 0.64   
## catAgeMature\_Age 10730.4 656.9 16.33 <0.0000000000000002 \*\*\*  
## catBMINormal 1760.0 5074.5 0.35 0.73   
## catBMIOverweight 874.0 5024.0 0.17 0.86   
## catBMIObese 1699.7 5003.7 0.34 0.73   
## sexmale -115.6 661.6 -0.17 0.86   
## regionnorthwest 350.1 923.3 0.38 0.70   
## regionsoutheast 46.5 923.8 0.05 0.96   
## regionsouthwest -595.8 913.3 -0.65 0.51   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4930 on 222 degrees of freedom  
## Multiple R-squared: 0.553, Adjusted R-squared: 0.537   
## F-statistic: 34.3 on 8 and 222 DF, p-value: <0.0000000000000002

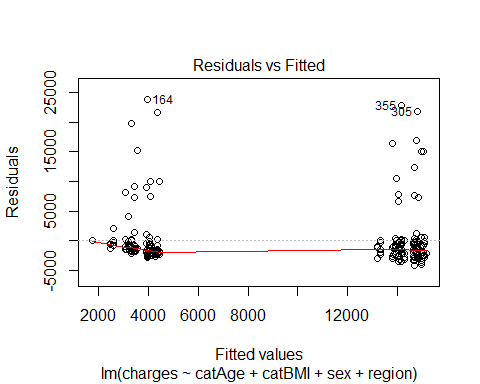
coef(charges.lm)

## (Intercept) catAgeMature\_Age catBMINormal catBMIOverweight   
## 2323.6 10730.4 1760.0 874.0   
## catBMIObese sexmale regionnorthwest regionsoutheast   
## 1699.7 -115.6 350.1 46.5   
## regionsouthwest   
## -595.8

#Residuals and fitted plots  
Residuals.Values <- resid(charges.lm)  
Fitted.Values <- fitted(charges.lm)  
hist(Residuals.Values)



plot(charges.lm,1)



# use predict() to make predictions on a new set.  
charges.lm.pred <- predict(charges.lm, valid.data)  
  
# use accuracy() to compute common accuracy measures.  
accuracy(charges.lm.pred, valid.data$charges)

## ME RMSE MAE MPE MAPE  
## Test set -6.46 5299 3129 -35.2 49.9

#Predicting results  
results.df <- data.frame(cbind(actuals = valid.data$charges, predicted = charges.lm.pred))  
results.df <- results.df %>%  
 mutate(error = results.df$actuals - results.df$predicted) %>%  
 round(., 2)  
results.df <- results.df %>%  
 mutate(percerror = paste0(round(results.df$error/results.df$actuals\*100,2),"%"))  
  
kable(head(results.df))

|  |  |  |  |
| --- | --- | --- | --- |
| actuals | predicted | error | percerror |
| 1880 | 4070 | -2189.8 | -116.47% |
| 16456 | 15104 | 1351.9 | 8.22% |
| 1622 | 4070 | -2448.0 | -150.94% |
| 3538 | 3432 | 105.6 | 2.99% |
| 12929 | 14685 | -1755.8 | -13.58% |
| 14255 | 14278 | -23.4 | -0.16% |

#Our model will be able to predict the insurance premium for   
#policy holders under cluster 2 wtih not having smoking habit with a mean difference of the result from the below output  
sprintf("The mean percent error is: %s%%", round(mean(results.df$error/results.df$actuals\*100), 2))

## [1] "The mean percent error is: -35.17%"

#Cluster 3  
library(tidyverse)  
Reg\_Cluster3 <- newInsuranceData %>%  
 filter(clusters==3)  
str(Reg\_Cluster3)

## 'data.frame': 679 obs. of 7 variables:  
## $ charges : num 4449 21984 3867 3757 8241 ...  
## $ sex : Factor w/ 2 levels "female","male": 2 2 2 1 1 1 2 2 2 1 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 3 2 2 3 3 2 1 1 4 1 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 2 2 2 2 2 2 2 2 1 2 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 4 2 3 3 4 3 3 3 2 4 ...  
## $ clusters: int 3 3 3 3 3 3 3 3 3 3 ...

dim(Reg\_Cluster3)

## [1] 679 7

Reg\_Cluster3<-Reg\_Cluster3[,-c(3,7)] #Smoker column contains only 'No' values for all the data, thus it will not have any effect over prediction  
  
#Model Build  
  
m5 <- lm(formula= charges ~ catAge + catBMI + region ,data = Reg\_Cluster3)  
summary(m5)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + region, data = Reg\_Cluster3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6875 -3073 -946 1128 24685   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2325 1688 1.38 0.1687   
## catAgePrime\_Age 4751 1032 4.60 0.000005 \*\*\*  
## catAgeMature\_Age 10044 3956 2.54 0.0113 \*   
## catBMINormal 1789 1604 1.12 0.2650   
## catBMIOverweight 1955 1624 1.20 0.2291   
## catBMIObese 2526 1613 1.57 0.1178   
## regionnorthwest -816 550 -1.48 0.1386   
## regionsoutheast -1593 563 -2.83 0.0048 \*\*   
## regionsouthwest -1696 556 -3.05 0.0024 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5060 on 670 degrees of freedom  
## Multiple R-squared: 0.0659, Adjusted R-squared: 0.0547   
## F-statistic: 5.91 on 8 and 670 DF, p-value: 0.000000253

coef(m5)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 2325 4751 10044 1789   
## catBMIOverweight catBMIObese regionnorthwest regionsoutheast   
## 1955 2526 -816 -1593   
## regionsouthwest   
## -1696

# Compute the analysis of variance for model 5  
res.aov <- aov(charges ~ catAge + catBMI + region, data = Reg\_Cluster3)  
# Summary of the analysis  
summary(res.aov) #As the p-value is less than the significance level 0.1, we can conclude that there are significant differences between the groups highlighted with “\*" in the model summary.

## Df Sum Sq Mean Sq F value Pr(>F)   
## catAge 2 849494747 424747374 16.59 0.000000092 \*\*\*  
## catBMI 3 57170442 19056814 0.74 0.5258   
## region 3 302551025 100850342 3.94 0.0084 \*\*   
## Residuals 670 17148811108 25595240   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Model 3B (with outliers, naming model 5 from the above as model 3B)  
#Partition data  
set.seed(1) # set seed for reproducing the partition  
train.rows <- sample(row.names(Reg\_Cluster3), dim(Reg\_Cluster3)[1]\*0.6)   
train.data <- Reg\_Cluster3[train.rows,]   
colnames(train.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(train.data)

## [1] 407 5

str(train.data)

## 'data.frame': 407 obs. of 5 variables:  
## $ charges: num 10601 7266 8269 11535 3559 ...  
## $ sex : Factor w/ 2 levels "female","male": 2 2 1 2 1 2 2 2 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 2 2 4 1 1 4 1 1 1 2 ...  
## $ catAge : Factor w/ 3 levels "Early\_Age","Prime\_Age",..: 2 2 2 3 2 2 2 2 2 2 ...  
## $ catBMI : Factor w/ 4 levels "Underweight",..: 4 4 2 1 3 2 4 3 4 4 ...

valid.rows <- sample(setdiff(rownames(Reg\_Cluster3), train.rows))  
valid.data <- Reg\_Cluster3[valid.rows,]  
colnames(valid.data)

## [1] "charges" "sex" "region" "catAge" "catBMI"

dim(valid.data)

## [1] 272 5

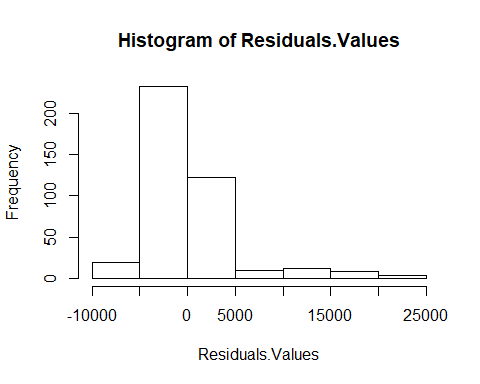
#Linear Regression on Charges(model 3B)  
charges.lm <- lm(charges ~ catAge + catBMI + region, data = train.data)   
options(scipen = 999)  
summary(charges.lm)

##   
## Call:  
## lm(formula = charges ~ catAge + catBMI + region, data = train.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6381 -2832 -991 1083 24557   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2902 2166 1.34 0.1811   
## catAgePrime\_Age 3587 1312 2.73 0.0065 \*\*  
## catAgeMature\_Age 9468 4040 2.34 0.0196 \*   
## catBMINormal 1792 2064 0.87 0.3858   
## catBMIOverweight 2256 2082 1.08 0.2793   
## catBMIObese 3026 2068 1.46 0.1442   
## regionnorthwest -599 668 -0.90 0.3704   
## regionsoutheast -1563 704 -2.22 0.0271 \*   
## regionsouthwest -2076 680 -3.06 0.0024 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4820 on 398 degrees of freedom  
## Multiple R-squared: 0.0694, Adjusted R-squared: 0.0507   
## F-statistic: 3.71 on 8 and 398 DF, p-value: 0.000337

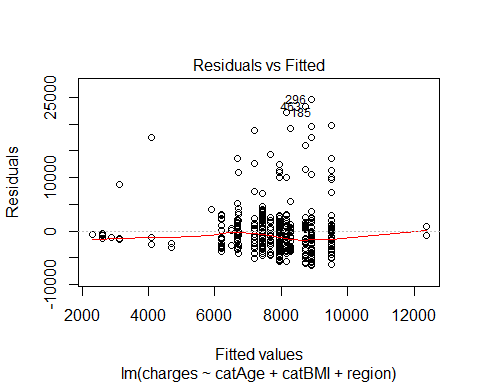
coef(charges.lm)

## (Intercept) catAgePrime\_Age catAgeMature\_Age catBMINormal   
## 2902 3587 9468 1792   
## catBMIOverweight catBMIObese regionnorthwest regionsoutheast   
## 2255 3026 -599 -1563   
## regionsouthwest   
## -2076

#Residuals and fitted plots  
Residuals.Values <- resid(charges.lm)  
Fitted.Values <- fitted(charges.lm)  
hist(Residuals.Values)



plot(charges.lm,1)



# use predict() to make predictions on a new set.  
library(forecast)  
charges.lm.pred <- predict(charges.lm, valid.data)  
  
# use accuracy() to compute common accuracy measures.  
accuracy(charges.lm.pred, valid.data$charges)

## ME RMSE MAE MPE MAPE  
## Test set 580 5422 3524 -25.6 50.5

#Predicting results  
results.df <- data.frame(cbind(actuals = valid.data$charges, predicted = charges.lm.pred))  
results.df <- results.df %>%  
 mutate(error = results.df$actuals - results.df$predicted) %>%  
 round(., 2)  
results.df <- results.df %>%  
 mutate(percerror = paste0(round(results.df$error/results.df$actuals\*100,2),"%"))  
  
kable(head(results.df))

|  |  |  |  |
| --- | --- | --- | --- |
| actuals | predicted | error | percerror |
| 26392 | 9514 | 16878 | 63.95% |
| 5970 | 6668 | -698 | -11.69% |
| 5003 | 7181 | -2179 | -43.55% |
| 23563 | 7952 | 15611 | 66.25% |
| 3214 | 8744 | -5530 | -172.09% |
| 5125 | 7681 | -2556 | -49.86% |

#Our model will be able to predict the insurance premium for   
#policy holders under cluster3 wtih not having smoking habit with a mean difference of the result from the below output  
sprintf("The mean percent error is: %s%%", round(mean(results.df$error/results.df$actuals\*100), 2))

## [1] "The mean percent error is: -25.55%"

1. DSM EVALUATION Refer to the document
2. DISCUSSION
3. Based on the DSM what would your decision/recommendation be? Why? Preferred: This is the lowest-risk category. People in this risk class are in excellent health, are typically younger, and have no other immediate cause for concern. Also, nonsmokers would reside in this group. Standard: This means typical risk, and for life insurers it means an average life expectancy. You may have some health issues in your family or in your past, which keeps you out of a preferred risk group. Age would ideally be following the early or mature group. Also, nonsmokers would reside in this group. Smoker: Classified as a higher risk than standard. Hence, premiums will also be higher. This class would include smokers, people with health issues or a risky past. Age would ideally be following the prime group.
4. What are the limitations of the DSM you have used? Clustering 1
5. Based on K-prototype gave the visualization across variables but not the stability.
6. 1Models work between with numerical predictors than categorical predictors. Clustering 2
7. Hierarchal clustering was greatly impacted by outliers, even though we removed them, still clusters were not completely accurate. The reason behind that was for the data as BMI did not correlate with charges, which was expected to be a main contributor.
8. Another problem with clustering that they have not provided a clear cut between clusters, so charges values will overlap between clusters but still maintain a different mean that shows that the three levels of charges based on risk classes. Regression 1 After removing outliers, correlation between BMI and Charges drops which affects significance. Regression 2
9. After keeping outliers, adjusted R-square increases but errors increase as well
10. Models work better with numerical predictors than categorical predictors
11. What would you expect (most likely) to influence the decision‐making process? How does decision support mitigate some/all of these?
12. The risk classes will affect the marketing 4 Ps (price, product, promotion and place)
13. The regression will affect the both the finance department which works on budgeting and the marketing department as well
14. It is critical to regularly update the models to ensure accuracy of the risk classes and relevance to the business as well as the predictions
15. What enhancements would you aim for to enable better decision support for this task?
16. Updating models with more data points, because current data is not enough
17. Collecting more variables as: family history, diseases, driving history, blood pressure, cholesterol, profession and lifestyle habits, as the current variables are not explaining well the variability
18. Enhancing the existing variables such as increasing the granularity of region
19. Exploring more risk classes
20. Conducting tests for heteroskedasticity, multicollinearity, linearity, normality
21. Detecting and closely diagnosing the influence of outliers