DS 6372 Project 1: [DATASET] Analysis

Audrene Tiakor, Spencer Fogelman, & Joseph Caguioa

6/15/2019

## Introduction

Lorem ipsum dolor sit amet, causae ornatus dolorem at vim, an quando melius epicurei quo, habemus lobortis cu quo. Ex quem animal vituperata sed. Semper alterum sed an, te dico virtute delectus usu. Ea rebum novum gloriatur vis, vix id prima errem interesset, duis nullam eum ut. Eam movet iuvaret te, mea suscipit convenire interesset te, at inermis philosophia ius.

## Data Description

For this project, we are working with the microsoft data science capstone data (<https://www.kaggle.com/nandvard/microsoft-data-science-capstone>), where the objective is to predict the variable heart\_disease\_mortality\_per\_100k which is a positive integer. Within this data set, there are 33 variables, 3 categorical variables, and 30 continuous variables. The categroical variables are: area\_\_rucc, area\_\_urban\_\_influence, and econ\_\_economic\_typology. The area\_\_rucc variable which is the Rural-Urban Continuum codes, forming a classification scheme that distinguishes metropolitan counties by degree of urbanization and adjacentty to a metro area. These counties are in the U.S. are assigned one of 9 codes, which can be found from the USDA Economic Research Service, <https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/>. The area\_\_urban\_\_influence variable represents the urban influence codes. The codes form a classificantion scheme that distinguishes metropolitan counties by means of population size of their metro area, and nonmetropolitan counties by size of the largest city or town and promixity to metro and micropolitan areas. These codes can be found from the USDA Economic Research Service, <https://www.ers.usda.gov/data-products/urban-influence-codes/>. Last, the econ\_\_economic\_typology variable is known as the county typology codes. These codes classify U.S. counties according to six mutulally exclusive categories of economic dependence and six overlapping categories of policy-relevant themes.These codes can be found at USDA Economic Research Service, <https://www.ers.usda.gov/data-products/county-typology-codes.aspx>. From the continuous variables there are split into different sections, economic: econ\_\_pct\_civilian\_labor, econ\_\_pct\_unemployment, econ\_\_pct\_uninsured\_adults, econ\_\_pct\_uninsured\_children, demographics: demo\_\_pct\_female, demo\_\_pct\_below\_18\_years\_of\_age, demo\_\_pct\_aged\_65\_years\_and\_older, demo\_\_pct\_hispanic, demo\_\_pct\_non\_hispanic\_african\_american, demo\_\_pct\_non\_hispanic\_white, demo\_\_pct\_american\_indian\_or\_alaskan\_native, demo\_\_pct\_asian, demo\_\_pct\_adults\_less\_than\_a\_high\_school\_diploma, demo\_\_pct\_adults\_with\_high\_school\_diploma, demo\_\_pct\_adults\_with\_some\_college, demo\_\_pct\_adults\_bachelors\_or\_higher, demo\_\_birth\_rate\_per\_1k, demo\_\_death\_rate\_per\_1k, health: health\_\_pct\_adult\_obesity, health\_\_pct\_adult\_smoking, health\_\_pct\_diabetes, health\_\_pct\_low\_birthweight, health\_\_pct\_excessive\_drinking, health\_\_pct\_physical\_inacticity, health\_\_air\_pollution\_particulate\_matter, health\_\_homicides\_per\_100k, health\_\_motor\_vehicle\_crash\_deaths\_per\_100k, health\_\_pop\_per\_dentist, health\_\_pop\_per\_primary\_care\_physician, and the year which is split into year a and year b.

# Load two files, Training\_values and Training\_labels, as data.frames.  
dfValues <- read.csv(file='microsoft-data-science-capstone/Training\_values.csv', header=T, sep=",")  
dfLabels <- read.csv(file='microsoft-data-science-capstone/Training\_labels.csv', header=T, sep=",")  
  
# Add response variable to left side of Training\_values and shorten name.  
dfFull <- cbind(dfLabels$heart\_disease\_mortality\_per\_100k, dfValues)  
names(dfFull)[names(dfFull) == 'dfLabels$heart\_disease\_mortality\_per\_100k'] <- 'heart\_disease\_mortality\_per\_100k'  
  
# Remove row\_id.  
dfFull <- subset(dfFull, select=-row\_id)

# Optional shorten column names by removing \*\_\_ category descriptors.  
# names(dfFull) = sub("area\_\_|econ\_\_|demo\_\_|health\_\_", "", names(dfFull))  
str(dfFull)

## 'data.frame': 3198 obs. of 34 variables:  
## $ heart\_disease\_mortality\_per\_100k : int 312 257 195 218 355 288 283 315 426 309 ...  
## $ area\_\_rucc : Factor w/ 9 levels "Metro - Counties in metro areas of 1 million population or more",..: 3 3 1 6 7 8 6 2 7 5 ...  
## $ area\_\_urban\_influence : Factor w/ 12 levels "Large-in a metro area with at least 1 million residents or more",..: 12 12 1 7 10 3 5 12 4 11 ...  
## $ econ\_\_economic\_typology : Factor w/ 6 levels "Farm-dependent",..: 3 4 5 5 5 2 5 3 5 1 ...  
## $ econ\_\_pct\_civilian\_labor : num 0.408 0.556 0.541 0.5 0.471 0.501 0.462 0.425 0.313 0.371 ...  
## $ econ\_\_pct\_unemployment : num 0.057 0.039 0.057 0.061 0.05 0.048 0.088 0.077 0.111 0.044 ...  
## $ econ\_\_pct\_uninsured\_adults : num 0.254 0.26 0.07 0.203 0.225 0.212 0.18 0.252 0.275 0.241 ...  
## $ econ\_\_pct\_uninsured\_children : num 0.066 0.143 0.023 0.059 0.103 0.055 0.039 0.079 0.063 0.11 ...  
## $ demo\_\_pct\_female : num 0.516 0.503 0.522 0.525 0.511 0.516 0.507 0.499 0.473 0.489 ...  
## $ demo\_\_pct\_below\_18\_years\_of\_age : num 0.235 0.272 0.179 0.2 0.237 0.207 0.221 0.247 0.248 0.347 ...  
## $ demo\_\_pct\_aged\_65\_years\_and\_older : num 0.176 0.101 0.115 0.164 0.171 0.121 0.167 0.134 0.11 0.121 ...  
## $ demo\_\_pct\_hispanic : num 0.109 0.41 0.202 0.013 0.025 0.022 0.034 0.053 0.015 0.014 ...  
## $ demo\_\_pct\_non\_hispanic\_african\_american : num 0.039 0.07 0.198 0.049 0.008 0.046 0.002 0.224 0.719 0 ...  
## $ demo\_\_pct\_non\_hispanic\_white : num 0.829 0.493 0.479 0.897 0.953 0.903 0.942 0.681 0.257 0.426 ...  
## $ demo\_\_pct\_american\_indian\_or\_alaskan\_native : num 0.004 0.008 0.013 0.007 0.003 0.002 0.004 0.004 0.008 0.546 ...  
## $ demo\_\_pct\_asian : num 0.011 0.015 0.085 0.001 0 0.006 0 0.023 0.005 0.004 ...  
## $ demo\_\_pct\_adults\_less\_than\_a\_high\_school\_diploma: num 0.194 0.164 0.159 0.182 0.122 ...  
## $ demo\_\_pct\_adults\_with\_high\_school\_diploma : num 0.424 0.234 0.238 0.407 0.413 ...  
## $ demo\_\_pct\_adults\_with\_some\_college : num 0.227 0.342 0.186 0.249 0.307 ...  
## $ demo\_\_pct\_adults\_bachelors\_or\_higher : num 0.154 0.259 0.417 0.163 0.157 ...  
## $ demo\_\_birth\_rate\_per\_1k : int 12 19 12 11 14 11 10 12 12 21 ...  
## $ demo\_\_death\_rate\_per\_1k : int 12 7 6 12 12 8 10 10 10 10 ...  
## $ health\_\_pct\_adult\_obesity : num 0.297 0.288 0.212 0.285 0.284 0.283 0.305 0.317 0.417 0.332 ...  
## $ health\_\_pct\_adult\_smoking : num 0.23 0.19 0.156 NA 0.234 0.22 0.205 0.264 0.208 0.341 ...  
## $ health\_\_pct\_diabetes : num 0.131 0.09 0.084 0.104 0.137 0.112 0.11 0.131 0.127 0.101 ...  
## $ health\_\_pct\_low\_birthweight : num 0.089 0.082 0.098 0.058 0.07 0.089 0.07 0.092 0.139 0.073 ...  
## $ health\_\_pct\_excessive\_drinking : num NA 0.181 0.195 NA 0.194 0.067 0.193 0.13 0.106 0.214 ...  
## $ health\_\_pct\_physical\_inacticity : num 0.332 0.265 0.209 0.238 0.29 0.272 0.282 0.277 0.373 0.333 ...  
## $ health\_\_air\_pollution\_particulate\_matter : int 13 10 10 13 9 13 12 13 12 9 ...  
## $ health\_\_homicides\_per\_100k : num 2.8 2.3 9.31 NA NA 3.8 NA 5.6 13.6 NA ...  
## $ health\_\_motor\_vehicle\_crash\_deaths\_per\_100k : num 15.09 19.79 3.14 NA 29.39 ...  
## $ health\_\_pop\_per\_dentist : int 1650 2010 629 1810 3489 2439 3100 2800 3990 NA ...  
## $ health\_\_pop\_per\_primary\_care\_physician : int 1489 2480 690 6630 2590 1540 3689 1790 3549 NA ...  
## $ yr : Factor w/ 2 levels "a","b": 1 1 2 2 1 1 2 1 1 1 ...

* Note that demo\_\_pct\_male can be derived as 1 - demo\_\_pct\_female.
* Adding all 5 demo\_*pct*[race] columns produces number between .990 and 1.000, as expected. Race proportions explicitly provided.
* Adding all 4 demo\_*pct*[education] columns produces number between .990 and 1.000, as expected. Education proportions explicitly provided.
* This dataset is already presplit from a test set. It is unclear whether the available data points are considered repeated measures based on year, as there is no way to identify individual counties.

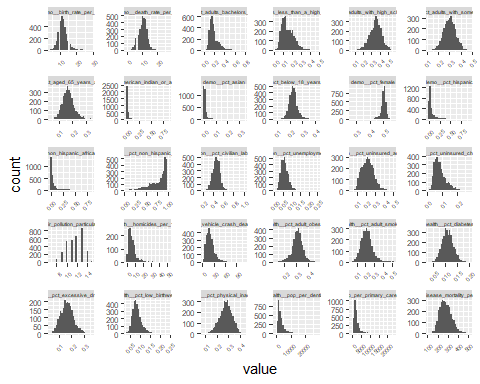
# Exploratory Data Analysis

For the exploratory data analysis portion of this project, tools such as histograms, heatmaps, and variance of inflation factors will be used. During the exploratory data analysis phase, the distribution of our continuous variables will be address, as well as multicolinearity.

# Quick exploratory histograms of any numeric variables.  
dfFull %>%  
 keep(is.numeric) %>%  
 gather() %>%  
 ggplot(aes(value)) +  
 facet\_wrap(~ key, scales="free") +  
 geom\_histogram() +  
 theme(axis.text.x=element\_text(size=4, angle=45, vjust=1, hjust=1),  
 axis.text.y=element\_text(size=6),  
 strip.text=element\_text(size=4, margin=margin()))

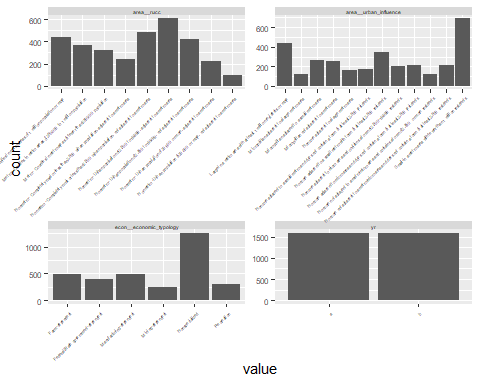
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 4536 rows containing non-finite values (stat\_bin).

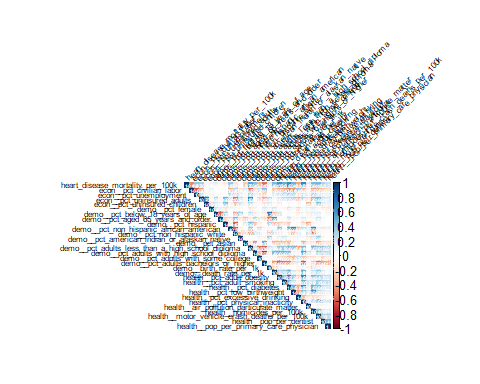


# Quick exploratory histograms of any factors.  
dfFull %>%  
 keep(is.factor) %>%  
 gather() %>%  
 ggplot(aes(value)) +  
 facet\_wrap(~ key, scales="free") +  
 geom\_bar() +  
 theme(axis.text.x=element\_text(size=3, angle=45, vjust=1, hjust=1),  
 axis.text.y=element\_text(size=6),  
 strip.text=element\_text(size=4, margin=margin()))

## Warning: attributes are not identical across measure variables;  
## they will be dropped

 \* I think we should drop the scatterplots and just work with the heatmap and the vif to address multicolinearity. \* Some collinearity shown between race variables because they are proportions. \* Maybe air\_pollution\_particulate\_matter should be redefined as factor.

# Correlation plot only takes numeric data. Remove factor variables.  
dfNumeric <- dfFull %>% keep(is.numeric)  
  
# Correlation plot on numeric variables. pairwise.complete.obs ignoring any pairs with NA.  
corrplot(cor(dfNumeric, use="pairwise.complete.obs"), type = "upper", tl.col = "black", tl.srt = 45, tl.cex = 0.5, number.cex = 0.3, number.digits=2, method="color", addCoef.col="white")



An initial attempt at obtaining VIFs results in an error on aliased coefficients, indicating that some variables are likely linearly dependent.

fullModel <- lm(heart\_disease\_mortality\_per\_100k~., data=dfFull, singular.ok=T)  
ld.vars <- attributes(alias(fullModel)$Complete)$dimnames[[1]]  
ld.vars

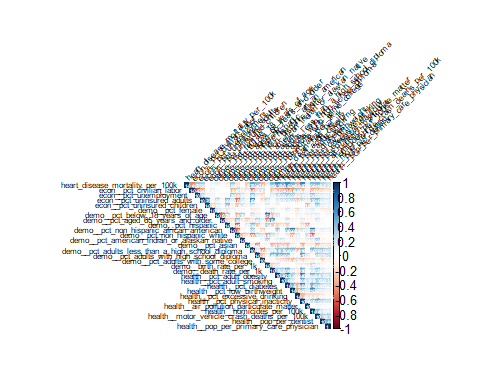
## [1] "area\_\_urban\_influenceNoncore adjacent to a small metro with town of at least 2,500 residents"   
## [2] "area\_\_urban\_influenceNoncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents"  
## [3] "area\_\_urban\_influenceSmall-in a metro area with fewer than 1 million residents"   
## [4] "demo\_\_pct\_adults\_bachelors\_or\_higher"

#plot(fullModel)  
#summary(fullModel)

According to alias(), a few levels of area\_\_urban\_influence and the variable demo\_\_pct\_adults\_bachelors\_or\_higher show linear dependence.

#removing the two variables that show linear dependence   
drop <- c('demo\_\_pct\_adults\_bachelors\_or\_higher',   
 'area\_\_urban\_influence')  
dfPartial <- dfFull[, !(names(dfFull) %in% drop)]

#using heatmap visual to see which variables have multicolinarity  
# Correlation plot only takes numeric data. Remove factor variables.  
dfNumeric2 <- dfPartial %>% keep(is.numeric)  
  
# Correlation plot on numeric variables. pairwise.complete.obs ignoring any pairs with NA.  
corrplot(cor(dfNumeric2, use="pairwise.complete.obs"), type = "upper", tl.col = "black", tl.srt = 45, tl.cex = 0.5, number.cex = 0.3, number.digits=2, method="color", addCoef.col="white")



#running vif  
fullModel <- lm(heart\_disease\_mortality\_per\_100k~., data=dfPartial)  
vif(fullModel)[,3]

## area\_\_rucc   
## 1.102268   
## econ\_\_economic\_typology   
## 1.114097   
## econ\_\_pct\_civilian\_labor   
## 1.951305   
## econ\_\_pct\_unemployment   
## 1.605450   
## econ\_\_pct\_uninsured\_adults   
## 2.455163   
## econ\_\_pct\_uninsured\_children   
## 1.962404   
## demo\_\_pct\_female   
## 1.367621   
## demo\_\_pct\_below\_18\_years\_of\_age   
## 1.885528   
## demo\_\_pct\_aged\_65\_years\_and\_older   
## 2.577154   
## demo\_\_pct\_hispanic   
## 10.398631   
## demo\_\_pct\_non\_hispanic\_african\_american   
## 12.766206   
## demo\_\_pct\_non\_hispanic\_white   
## 15.703254   
## demo\_\_pct\_american\_indian\_or\_alaskan\_native   
## 5.809900   
## demo\_\_pct\_asian   
## 3.171833   
## demo\_\_pct\_adults\_less\_than\_a\_high\_school\_diploma   
## 2.643435   
## demo\_\_pct\_adults\_with\_high\_school\_diploma   
## 2.321080   
## demo\_\_pct\_adults\_with\_some\_college   
## 1.663122   
## demo\_\_birth\_rate\_per\_1k   
## 1.798667   
## demo\_\_death\_rate\_per\_1k   
## 2.824588   
## health\_\_pct\_adult\_obesity   
## 2.290998   
## health\_\_pct\_adult\_smoking   
## 1.792162   
## health\_\_pct\_diabetes   
## 2.428579   
## health\_\_pct\_low\_birthweight   
## 2.068124   
## health\_\_pct\_excessive\_drinking   
## 1.456569   
## health\_\_pct\_physical\_inacticity   
## 2.523841   
## health\_\_air\_pollution\_particulate\_matter   
## 1.414969   
## health\_\_homicides\_per\_100k   
## 1.763969   
## health\_\_motor\_vehicle\_crash\_deaths\_per\_100k   
## 2.030200   
## health\_\_pop\_per\_dentist   
## 1.399649   
## health\_\_pop\_per\_primary\_care\_physician   
## 1.421994   
## yr   
## 1.094620

After finding what variables show linear dependence using the alias() function, we removed those variables from our data and proceed to using, the vif function from the car package. When using vif to address colinearity within the data, the rule of thumb is when a variable has a vif greater than 10, that variable can beremoved from data.Keeping this rul of thumb in mind, we find two variables that we can remove from our model: demo\_\_pct\_non\_hispanic\_african\_american (vif=12.77), and demo\_\_pct\_non\_hispanic\_white (vif=15.70).

#removing demo\_\_pct\_non\_hispanic\_african\_american & demo\_\_pct\_non\_hispanic\_white  
drop <- c('demo\_\_pct\_non\_hispanic\_african\_american',   
 'demo\_\_pct\_non\_hispanic\_white','demo\_\_pct\_adults\_bachelors\_or\_higher',   
 'area\_\_urban\_influence')  
dfPartial2 <- dfFull[, !(names(dfFull) %in% drop)]

#rerunning vif after removing demo\_\_pct\_non\_hispanic\_african\_american & demo\_\_pct\_non\_hispanic\_white  
fullModel2 <- lm(heart\_disease\_mortality\_per\_100k~., data=dfPartial2)  
vif(fullModel2)[,3]

## area\_\_rucc   
## 1.099099   
## econ\_\_economic\_typology   
## 1.111817   
## econ\_\_pct\_civilian\_labor   
## 1.945114   
## econ\_\_pct\_unemployment   
## 1.579173   
## econ\_\_pct\_uninsured\_adults   
## 2.451254   
## econ\_\_pct\_uninsured\_children   
## 1.949201   
## demo\_\_pct\_female   
## 1.364695   
## demo\_\_pct\_below\_18\_years\_of\_age   
## 1.866845   
## demo\_\_pct\_aged\_65\_years\_and\_older   
## 2.560031   
## demo\_\_pct\_hispanic   
## 2.029926   
## demo\_\_pct\_american\_indian\_or\_alaskan\_native   
## 1.319113   
## demo\_\_pct\_asian   
## 1.555094   
## demo\_\_pct\_adults\_less\_than\_a\_high\_school\_diploma   
## 2.600992   
## demo\_\_pct\_adults\_with\_high\_school\_diploma   
## 2.290180   
## demo\_\_pct\_adults\_with\_some\_college   
## 1.661161   
## demo\_\_birth\_rate\_per\_1k   
## 1.769539   
## demo\_\_death\_rate\_per\_1k   
## 2.808004   
## health\_\_pct\_adult\_obesity   
## 2.239012   
## health\_\_pct\_adult\_smoking   
## 1.697857   
## health\_\_pct\_diabetes   
## 2.391091   
## health\_\_pct\_low\_birthweight   
## 1.780874   
## health\_\_pct\_excessive\_drinking   
## 1.428387   
## health\_\_pct\_physical\_inacticity   
## 2.521789   
## health\_\_air\_pollution\_particulate\_matter   
## 1.393650   
## health\_\_homicides\_per\_100k   
## 1.623656   
## health\_\_motor\_vehicle\_crash\_deaths\_per\_100k   
## 2.008856   
## health\_\_pop\_per\_dentist   
## 1.398138   
## health\_\_pop\_per\_primary\_care\_physician   
## 1.407876   
## yr   
## 1.091360

After dropping a total of four variables: ’demo\_\_pct\_non\_hispanic\_african\_american’, ’demo\_\_pct\_non\_hispanic\_white’,’demo\_\_pct\_adults\_bachelors\_or\_higher’, and ’area\_\_urban\_influence’ the vif factors of the remaining variables are less than 10. We are now ready to proceed with the regression analysis portion of this project.

# Regression Analysis

## Regression Models: Introduction

“Restatement of problem and overall approach.”

### Model Selection

### Checking Assumptions

### Comparing Competing Models

## Parameter Interpretation

## Regression Models: Conclusion

Lorem ipsum dolor sit amet, te his wisi voluptatum. Legimus mentitum senserit ad per. Te cum iudico everti concludaturque, id eam congue primis dolores. Nec at quas augue feugiat, dolor perpetua id sed, ei paulo tritani nec.

# Secondary Analysis

## 2-Way ANOVA: Introduction

## Main Analysis Content

## 2-Way ANOVA / Time Series: Conclusion

# Appendix