CSCI 385 - Third Deliverable

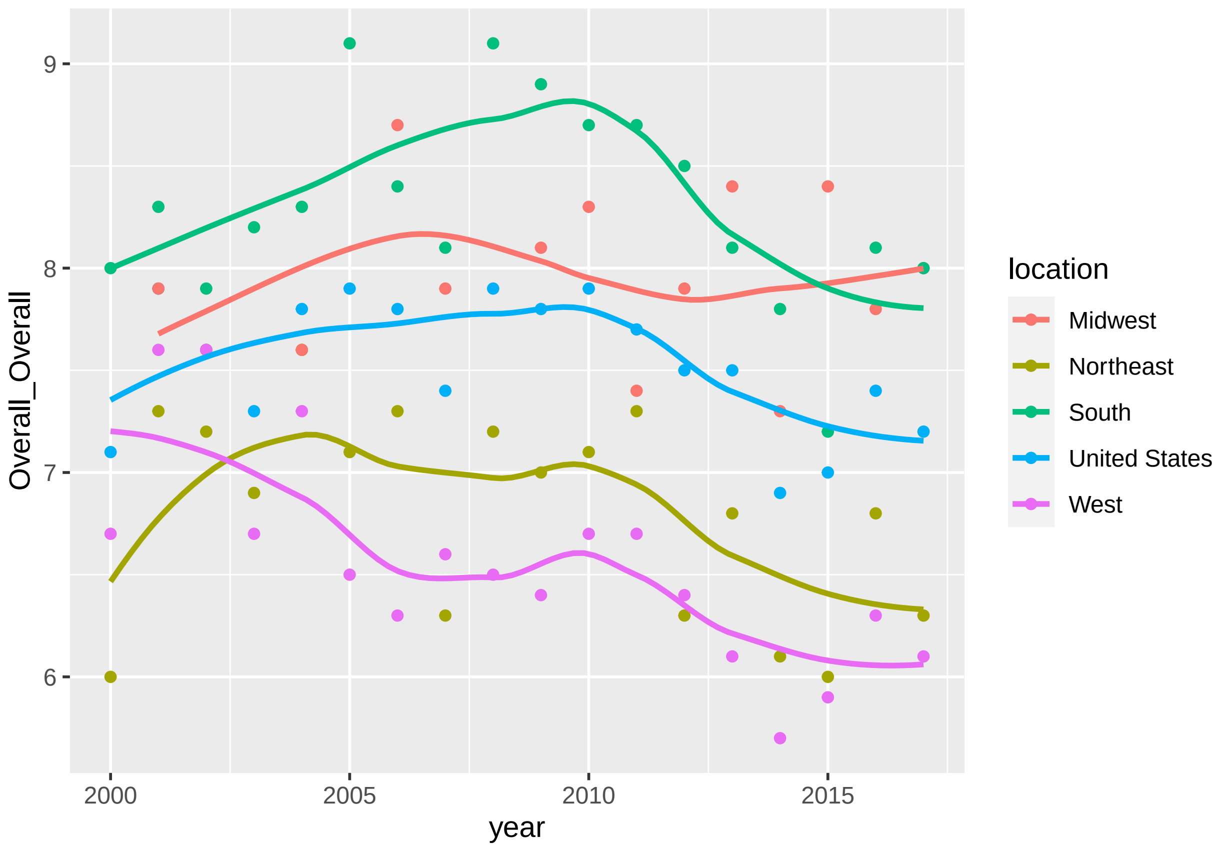
Garrett Welton

12/7/2020

Improvement

```
#API call
df <- read.socrata(
  "https://chronicdata.cdc.gov/resource/fwns-azgu.csv",
  app_token = "OJxxBUd1kEfoUDA5ACIVnD88u")

#Data Tidying
NHIS <- filter(df, !is.na(data_value), data_value_type == "Age-Standardized" | break_out_category == "Ag
  select(year, locationdesc, category, topic, data_value, break_out_category, break_out) %>%
  rename(location = locationdesc, percent_value = data_value)

#Spit into separate tables
cardioDiseases <- filter(NHIS, category == "Cardiovascular Diseases") %>%
  select(-category) %>%
  pivot_wider(names_from = c(break_out_category, break_out),
              values_from = percent_value) %>%
  select(-‘Age_25-44‘, -‘Age_18-24‘, -Race_Other, -‘Race_Non-Hispanic Asian‘) %>%
  drop_na()

cardioDiseases.majorCardiovascularDisease <- filter(cardioDiseases, topic == "Major Cardiovascular Disea
  select(-topic)
cardioDiseases.heartAttack <- filter(cardioDiseases, topic == "Acute Myocardial Infarction (Heart Attac
  select(-topic)
cardioDiseases.coronaryHeartDisease <- filter(cardioDiseases, topic == "Coronary Heart Disease") %>%
  select(-topic)
cardioDiseases.stroke <- filter(cardioDiseases, topic == "Stroke") %>%
  select(-topic)

#Data Exploration
ggplot(cardioDiseases.majorCardiovascularDisease, aes(x = year, y = Overall_Overall, color = location))+
  geom_point()+
  geom_smooth(se = FALSE)
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



This graph is a breakdown of the percent of the population of different regions of the US who have Major Cardiovascular Disease over time. It can be used to get an idea of the trends of Major Cardiovascular Disease in theses regions and how they compare.

Tables

```
Heart <- read.csv("heart.csv")
head(Heart)
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1  52   1  0      125  212   0       1     168     0     1.0     2  2    3
## 2  53   1  0      140  203   1       0     155     1     3.1     0  0    3
## 3  70   1  0      145  174   0       1     125     1     2.6     0  0    3
## 4  61   1  0      148  203   0       1     161     0     0.0     2  1    3
## 5  62   0  0      138  294   1       1     106     0     1.9     1  3    2
## 6  58   0  0      100  248   0       0     122     0     1.0     1  0    2
##   target
## 1      0
## 2      0
## 3      0
## 4      0
## 5      0
## 6      1
```

```r
head(cardioDiseases.majorCardiovascularDisease)
```

```
## # A tibble: 6 x 12
##    year location Overall_Overall Gender_Male Gender_Female ‘Age_45-64‘ ‘Age_65+‘
##   <int> <chr>              <dbl>       <dbl>         <dbl>       <dbl>     <dbl>
## 1  2000 United ~             7.1         8.6           5.8         8        24.5
## 2  2000 Northea~             6           7.5           4.8         6.7      21.7
## 3  2000 South                8           9.9           6.4         9.6      26.4
## 4  2000 West                 6.7         7.9           5.6         7.1      23.6
## 5  2001 United ~             7.9         9.4           6.6         9.1      26.1
## 6  2001 Northea~             7.3         8.6           6.3         7.6      26.7
## # ... with 5 more variables: ‘Age_35+‘ <dbl>, ‘Age_75+‘ <dbl>,
## #   ‘Race_Non-Hispanic White‘ <dbl>, ‘Race_Non-Hispanic Black‘ <dbl>,
## #   Race_Hispanic <dbl>
```

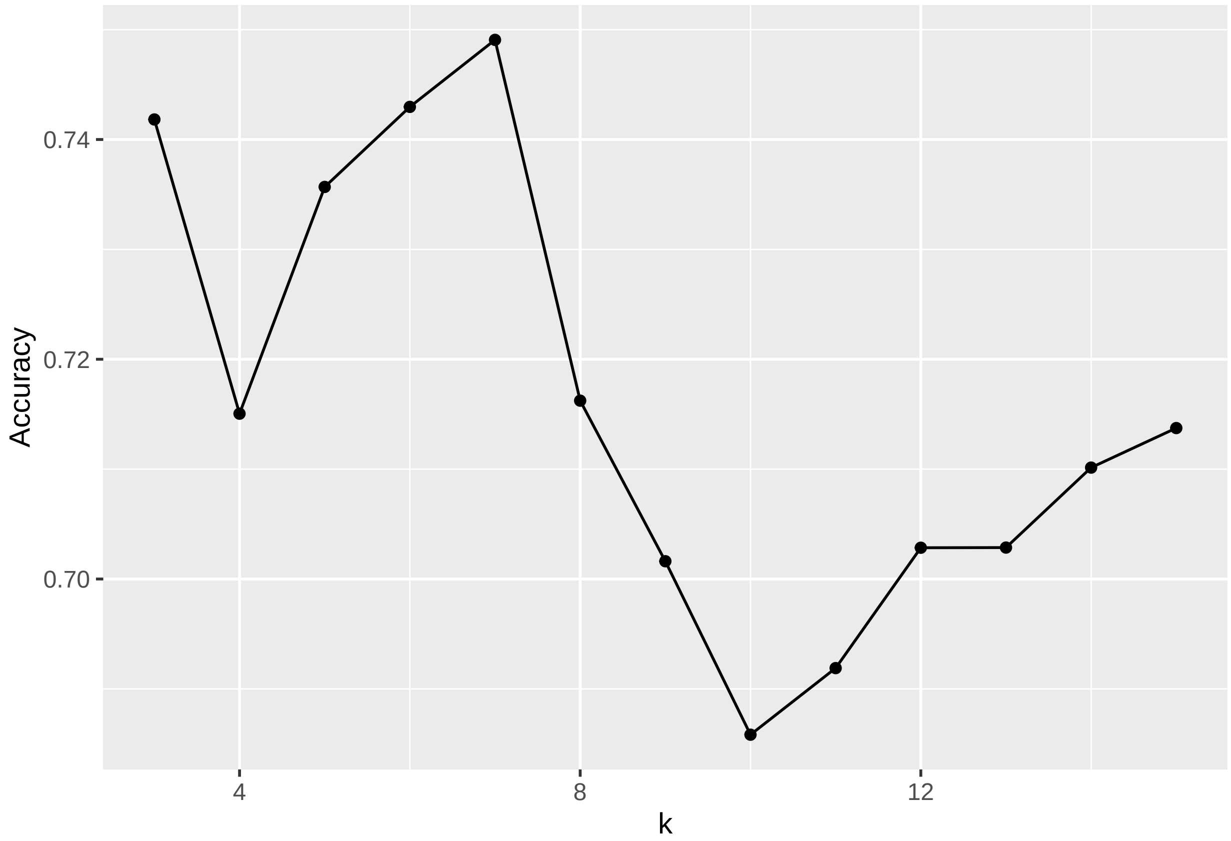
Validation

Five fold cross validation on knn model

```r
heart <- as_tibble(Heart) %>% select(-trestbps, -ca)
heart$target = as.factor(heart$target)

set.seed(132)

#Split training and test dataset
train_rows <- as.vector(createDataPartition(heart$target, p = 0.8, list = FALSE))

train <- heart[train_rows, ]
test <- heart[-train_rows, ]

train.def <- Heart[train_rows, 14]
test.def <- Heart[-train_rows, 14]

trControl <- trainControl(method  = "cv",
                          number  = 5)

fit <- train(target ~ .,
             method     = "knn",
             tuneGrid   = expand.grid(k = 3:15),
             trControl  = trControl,
             data       = train)
fit
```

```
## k-Nearest Neighbors
##
## 821 samples
##  11 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 657, 657, 657, 656, 657
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    3  0.7418256  0.4841923
##    4  0.7150480  0.4306468
##    5  0.7356837  0.4719789
##    6  0.7429712  0.4861451
##    7  0.7490687  0.4979775
##    8  0.7162306  0.4316379
##    9  0.7016186  0.4021549
##   10  0.6858315  0.3700342
##   11  0.6918921  0.3827283
##   12  0.7028455  0.4049560
##   13  0.7028603  0.4050660
##   14  0.7101404  0.4192869
##   15  0.7137398  0.4269098
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 7.
```

```
#Accuracy plot
ggplot(fit$results, aes(x=k, y=Accuracy)) +
  geom_point() +
  geom_line()
```



```
#Apply model
knn.7 <- knn(train, test, train.def, k = 7)

tab.7 <- table(knn.7, test.def)
confusionMatrix(tab.7)
```

```
## Confusion Matrix and Statistics
##
##      test.def
## knn.7  0  1
##     0 78 25
##     1 21 80
##
##                Accuracy : 0.7745
##                  95% CI : (0.7109, 0.8299)
##     No Information Rate : 0.5147
##     P-Value [Acc > NIR] : 1.793e-14
##
##                   Kappa : 0.5491
##
##  Mcnemar's Test P-Value : 0.6583
##
##             Sensitivity : 0.7879
##             Specificity : 0.7619
##          Pos Pred Value : 0.7573
##          Neg Pred Value : 0.7921
##              Prevalence : 0.4853
##          Detection Rate : 0.3824
##    Detection Prevalence : 0.5049
##       Balanced Accuracy : 0.7749
##
##        'Positive' Class : 0
##
```

There is not much of difference between the accuracy of our validation set and test set, suggesting that our model is not over fitting the data. This knn model has an accuracy of about 77% on the test set, which is over 70%, making it a successful model according to my definition of "Success" and "Failure". It is fairly accurate at predicting the presence of heart disease in patients based on their records.

Implications

The implications of these insights is that a doctor may be able to predict heart disease in patients earlier from patient records. Hopefully helping to reduce related deaths. One ethical issue that may arise is that the model has the potential to be used health insurance companies to increase the rate of or deny coverage to those who who it deems to have heart disease.