Project 1: Exploratory Data Analysis

Samuel Higgins

3/3/2020

Introduction

The datasets I have chosen for this project are DMepi and pr, which are both found within the epi package in R. Both datasets contain variables related to the onset of diabetes in a Danish population. Pr is specialized towards diabetes prevalence in Denmark and contains variables such as age, sex, population size, and number of diabetes patients. DMepi contains more of the same as pr, however it differs in that it has variables related to mortality and person-years, with and without the diabetes condition. It is important to note that DMepi does not contain data for females (even though it says it does when you print ?Dmepi), so the entirety of this analysis will be done in regards to the male population. I chose these datasets beacause I am in the process of pursuing a career in epidemiology and I thought that exploring the epi package would be good way to familiarize myself with some of the common funcitions and built-in datasets that I may encounter later on.

```
library(tidyverse)
library(Epi)
library(car)
library(mosaic)
library(reshape2)
library(forcats)
library(cluster)
library(GGally)

data("DMepi")
data("pr")
```

```
pr
```

```
##     A sex   X     N
## 1   0   F   1 30743
## 2   0   M   0 32435
## 3  10   F  84 32922
## 4  10   M  83 34294
## 5  11   F 106 32890
## 6  11   M  70 34644
## 7  12   F 104 33630
## 8  12   M 111 35439
## 9  13   F 133 33833
## 10 13   M 120 35681
## 11 14   F 144 34982
## 12 14   M 145 37062
## 13 15   F 149 35482
## 14 15   M 184 37064
## 15 16   F 189 34313
## 16 16   M 169 36169
## 17 17   F 166 34588
```

```
## 18 17   M 197 36576
## 19 18   F 191 33236
## 20 18   M 164 34915
## 21 19   F 191 33352
## 22 19   M 176 35233
## 23  1   F   3 31993
## 24  1   M   4 33984
## 25 20   F 217 32954
##  [ reached 'max' / getOption("max.print") -- omitted 175 rows ]
```

```
?pr
```

```
DMepi
```

```
##    sex A    P D.DM       Y.DM X D.nD     Y.nD
## 2    M 0 1996    0 0.48391513 1   28 35468.92
## 3    M 0 1997    0 0.63997262 2   19 35085.18
## 4    M 0 1998    0 1.64065708 4   20 34240.14
## 5    M 0 1999    0 0.55236140 4   11 34055.52
## 6    M 0 2000    0 2.50650240 4   21 34002.22
## 7    M 0 2001    0 0.11841205 1   16 34177.39
## 8    M 0 2002    0 0.01163587 1   21 33101.07
## 9    M 0 2003    0 0.69130732 3   15 33010.92
## 10   M 0 2004    0 1.69815195 4   16 33167.44
## 11   M 0 2005    0 0.29089665 1   16 33066.10
## 12   M 0 2006    0 0.99178645 3   11 33148.49
## 13   M 0 2007    0 0.58521561 1   20 33170.80
##  [ reached 'max' / getOption("max.print") -- omitted 4188 rows ]
```

```
?DMepi
```

Tidying and Joining

```
untidy_pr <- pr %>%
  pivot_wider(names_from = sex, values_from = c("X", "N"))
untidy_pr
```

```
## # A tibble: 100 x 5
##        A   X_F   X_M   N_F   N_M
##    <dbl> <dbl> <dbl> <dbl> <dbl>
##  1     0     1     0 30743 32435
##  2    10    84    83 32922 34294
##  3    11   106    70 32890 34644
##  4    12   104   111 33630 35439
##  5    13   133   120 33833 35681
##  6    14   144   145 34982 37062
##  7    15   149   184 35482 37064
##  8    16   189   169 34313 36169
##  9    17   166   197 34588 36576
## 10    18   191   164 33236 34915
## # ... with 90 more rows
```

```
tidy_pr <- untidy_pr %>%
  pivot_longer(c("X_F", "X_M", "N_F", "N_M"), names_to = "sex", values_to = c("X", "N"))
tidy_pr
```

```
## # A tibble: 400 x 4
##        A sex       X     N
##    <dbl> <chr> <dbl> <dbl>
##  1     0 X_F       1    NA
##  2     0 X_M      NA     0
##  3     0 N_F   30743    NA
##  4     0 N_M      NA 32435
##  5    10 X_F      84    NA
##  6    10 X_M      NA    83
##  7    10 N_F   32922    NA
##  8    10 N_M      NA 34294
##  9    11 X_F     106    NA
## 10    11 X_M      NA    70
## # ... with 390 more rows
```

```
untidy_epi <- DMepi %>%
  pivot_wider(names_from = sex, values_from = X) %>% rename(M_Incidence = M)
untidy_epi
```

```
## # A tibble: 4,200 x 8
##        A     P  D.DM   Y.DM  D.nD   Y.nD M_Incidence     F
##    <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl>       <dbl> <dbl>
##  1     0  1996     0 0.484     28 35469.           1    NA
##  2     0  1997     0 0.640     19 35085.           2    NA
##  3     0  1998     0 1.64      20 34240.           4    NA
##  4     0  1999     0 0.552     11 34056.           4    NA
##  5     0  2000     0 2.51      21 34002.           4    NA
##  6     0  2001     0 0.118     16 34177.           1    NA
##  7     0  2002     0 0.0116    21 33101.           1    NA
##  8     0  2003     0 0.691     15 33011.           3    NA
##  9     0  2004     0 1.70      16 33167.           4    NA
## 10     0  2005     0 0.291     16 33066.           1    NA
## # ... with 4,190 more rows
```

```
tidy_epi <- untidy_epi %>%
  pivot_longer(M_Incidence, names_to = "sex", values_to = "X") %>%
  mutate(sex = recode(sex, "'M_Incidence'='M'")) %>%
  select(-"F")
tidy_epi
```

```
## # A tibble: 4,200 x 8
##        A     P  D.DM   Y.DM  D.nD   Y.nD sex       X
##    <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl> <chr> <dbl>
##  1     0  1996     0 0.484     28 35469. M         1
##  2     0  1997     0 0.640     19 35085. M         2
##  3     0  1998     0 1.64      20 34240. M         4
##  4     0  1999     0 0.552     11 34056. M         4
##  5     0  2000     0 2.51      21 34002. M         4
##  6     0  2001     0 0.118     16 34177. M         1
##  7     0  2002     0 0.0116    21 33101. M         1
##  8     0  2003     0 0.691     15 33011. M         3
##  9     0  2004     0 1.70      16 33167. M         4
## 10     0  2005     0 0.291     16 33066. M         1
## # ... with 4,190 more rows
```

```
full_epi <- full_join(DMepi, pr, by = c("sex", "A"))
full_epi
```

```
##    sex A    P D.DM       Y.DM X.x D.nD     Y.nD X.y     N
## 1    M 0 1996    0 0.48391513   1   28 35468.92   0 32435
## 2    M 0 1997    0 0.63997262   2   19 35085.18   0 32435
## 3    M 0 1998    0 1.64065708   4   20 34240.14   0 32435
## 4    M 0 1999    0 0.55236140   4   11 34055.52   0 32435
## 5    M 0 2000    0 2.50650240   4   21 34002.22   0 32435
## 6    M 0 2001    0 0.11841205   1   16 34177.39   0 32435
## 7    M 0 2002    0 0.01163587   1   21 33101.07   0 32435
## 8    M 0 2003    0 0.69130732   3   15 33010.92   0 32435
## 9    M 0 2004    0 1.69815195   4   16 33167.44   0 32435
## 10   M 0 2005    0 0.29089665   1   16 33066.10   0 32435
##  [ reached 'max' / getOption("max.print") -- omitted 4190 rows ]
```

Both datasets were relatively tidy, however I went ahead and untidyed them and tidyed them back together. I chose to do a full join because both datasets had columns I could not drop and would be invaluable later. I did a full join by sex and age but not "X" because both "X" variables were somewhat unrelated. "X" in DMepi equated to number of new diabetes diagnoses, while "X" in pr related to number of diabetes patients.

Wrangling

```
d_diabetes <- full_epi %>%
  rename(new_diabetes_diagnoses = X.x, num_diabetes_patients = X.y, Year = P,
                   Deaths_diabetes = D.DM, Deaths_wo.diabetes = D.nD, PY_diabetes = Y.DM, PY_w.diabetes

d_diabetes2 <- d_diabetes %>%
  filter(Year >= 2010) %>% select(A, Year, N, Deaths_diabetes, Deaths_wo.diabetes, new_diabetes_diagnose
  mutate(d_prevalence = num_diabetes_patients+new_diabetes_diagnoses/N, d_incidence = new_diabetes_diagn

d_diabetes2
```

```
##     A Year     N Deaths_diabetes Deaths_wo.diabetes new_diabetes_diagnoses
## 1   0 2010 32435               0                 10                      2
## 2   1 2010 33984               0                  9                      6
## 3   2 2010 33380               0                  7                      5
## 4   3 2010 34034               0                  4                      5
## 5   4 2010 33339               0                  2                      6
## 6   5 2010 33465               0                  2                     15
## 7   6 2010 33553               0                  1                      7
## 8   7 2010 33373               0                  3                      9
## 9   8 2010 33671               0                  4                     12
## 10  9 2010 34689               0                  2                     12
## 11 10 2010 34294               0                  0                     19
##    num_diabetes_patients d_prevalence  d_incidence
## 1                      0 6.166179e-05 6.166179e-05
## 2                      4 4.000177e+00 1.765537e-04
## 3                     12 1.200015e+01 1.497903e-04
## 4                     23 2.300015e+01 1.469119e-04
## 5                     17 1.700018e+01 1.799694e-04
## 6                     32 3.200045e+01 4.482295e-04
## 7                     47 4.700021e+01 2.086252e-04
## 8                     53 5.300027e+01 2.696791e-04
```

```
## 9                     62 6.200036e+01 3.563898e-04
## 10                    53 5.300035e+01 3.459310e-04
## 11                    83 8.300055e+01 5.540328e-04
##  [ reached 'max' / getOption("max.print") -- omitted 1389 rows ]
```

```
summ_diabetes <- d_diabetes2 %>%
  group_by(Year, A) %>% summarise_at(vars(Deaths_diabetes, Deaths_wo.diabetes, new_diabetes_diagnoses, n
summ_diabetes
```

```
## # A tibble: 700 x 34
## # Groups:   Year [7]
##     Year     A Deaths_diabetes~ Deaths_wo.diabe~ new_diabetes_di~
##    <dbl> <dbl>            <dbl>            <dbl>            <dbl>
##  1  2010     0                0                9                1
##  2  2010     1                0                7.5              5
##  3  2010     2                0                4.5              4
##  4  2010     3                0                3                6.5
##  5  2010     4                0                1.5              6.5
##  6  2010     5                0                1               11
##  7  2010     6                0                3                8.5
##  8  2010     7                0                2.5             10
##  9  2010     8                0                2.5             10.5
## 10  2010     9                0                1.5             15.5
## # ... with 690 more rows, and 29 more variables:
## #   num_diabetes_patients_mean <dbl>, Deaths_diabetes_sd <dbl>,
## #   Deaths_wo.diabetes_sd <dbl>, new_diabetes_diagnoses_sd <dbl>,
## #   num_diabetes_patients_sd <dbl>, Deaths_diabetes_var <dbl>,
## #   Deaths_wo.diabetes_var <dbl>, new_diabetes_diagnoses_var <dbl>,
## #   num_diabetes_patients_var <dbl>, Deaths_diabetes_IQR <dbl>,
## #   Deaths_wo.diabetes_IQR <dbl>, new_diabetes_diagnoses_IQR <dbl>,
## #   num_diabetes_patients_IQR <dbl>, Deaths_diabetes_min <dbl>,
## #   Deaths_wo.diabetes_min <dbl>, new_diabetes_diagnoses_min <dbl>,
## #   num_diabetes_patients_min <dbl>, Deaths_diabetes_max <dbl>,
## #   Deaths_wo.diabetes_max <dbl>, new_diabetes_diagnoses_max <dbl>,
## #   num_diabetes_patients_max <dbl>, Deaths_diabetes_q25 <dbl>,
## #   Deaths_wo.diabetes_q25 <dbl>, new_diabetes_diagnoses_q25 <dbl>,
## #   num_diabetes_patients_q25 <dbl>, Deaths_diabetes_q75 <dbl>,
## #   Deaths_wo.diabetes_q75 <dbl>, new_diabetes_diagnoses_q75 <dbl>,
## #   num_diabetes_patients_q75 <dbl>
```

I started off by renaming some of the column names so that they would be more coherent. I then filtered the data to where I would only get rows where the year was 2010-2016. I filtered this way because the data in pr was current for the year 2010. After arranging by year, I used mutate to calculate two new columns, one for diabetes prevalence and the other for diabetes incidence. Using this new dataset (d_diabetes2), I grouped by year and age and used summarise_at to select which numeric variables I wanted to calculate summary statistics for.

Visualization

```
# cor prep
d_cor <- round(cor(d_diabetes2), 2)
d_cor
```

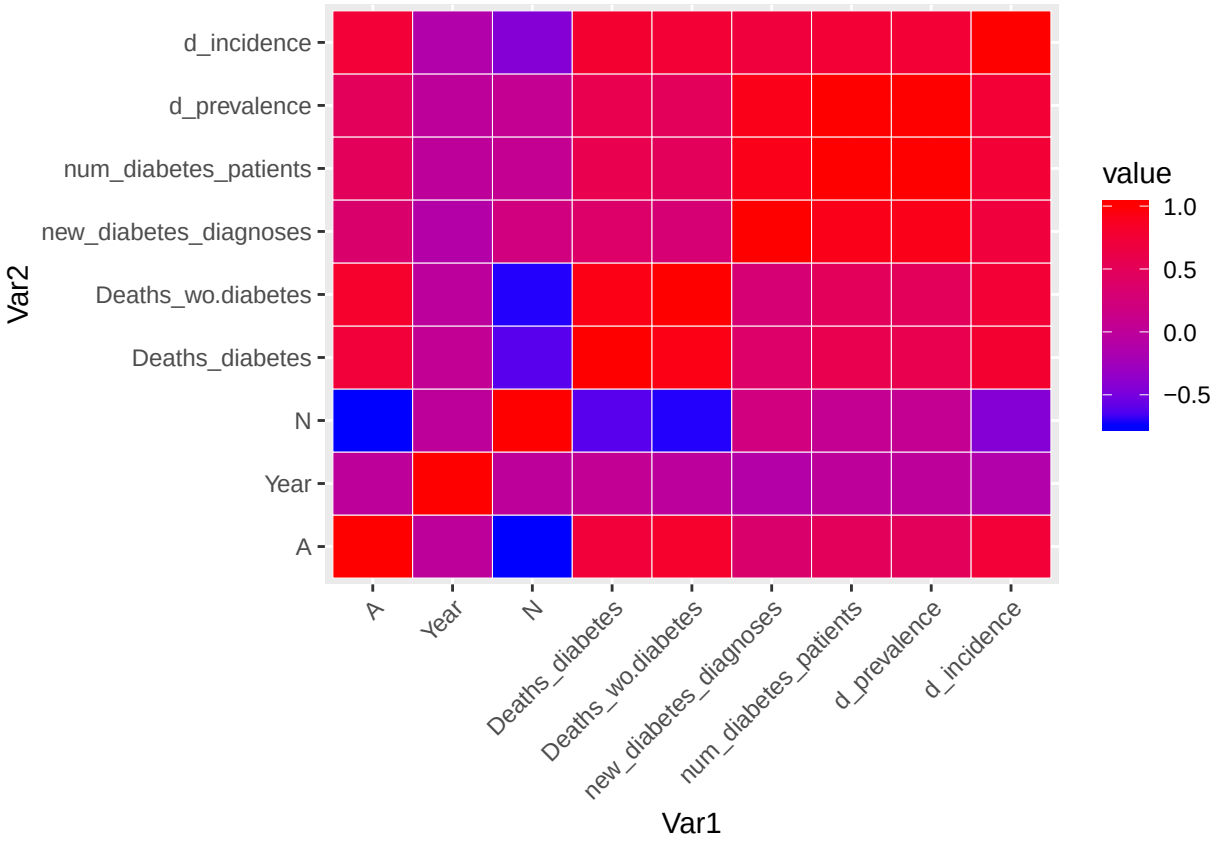
```
##                            A  Year     N Deaths_diabetes Deaths_wo.diabetes
## A                       1.00  0.00 -0.74            0.72                0.81
```

```
## Year                    0.00  1.00  0.00           0.04              -0.02
## N                      -0.74  0.00  1.00          -0.62              -0.72
## Deaths_diabetes         0.72  0.04 -0.62           1.00               0.93
## Deaths_wo.diabetes      0.81 -0.02 -0.72           0.93               1.00
## new_diabetes_diagnoses  0.35 -0.10  0.21           0.38               0.29
## num_diabetes_patients   0.48  0.00  0.06           0.58               0.48
## d_prevalence            0.48  0.00  0.06           0.58               0.48
## d_incidence             0.74 -0.12 -0.44           0.79               0.75
##                        new_diabetes_diagnoses num_diabetes_patients
## A                                        0.35                  0.48
## Year                                    -0.10                  0.00
## N                                        0.21                  0.06
## Deaths_diabetes                          0.38                  0.58
## Deaths_wo.diabetes                       0.29                  0.48
## new_diabetes_diagnoses                   1.00                  0.92
## num_diabetes_patients                    0.92                  1.00
## d_prevalence                             0.92                  1.00
## d_incidence                              0.71                  0.75
##                        d_prevalence d_incidence
## A                              0.48        0.74
## Year                           0.00       -0.12
## N                              0.06       -0.44
## Deaths_diabetes                0.58        0.79
## Deaths_wo.diabetes             0.48        0.75
## new_diabetes_diagnoses         0.92        0.71
## num_diabetes_patients          1.00        0.75
## d_prevalence                   1.00        0.75
## d_incidence                    0.75        1.00
```

```
d_melt <- melt(d_cor)
head(d_melt)
```

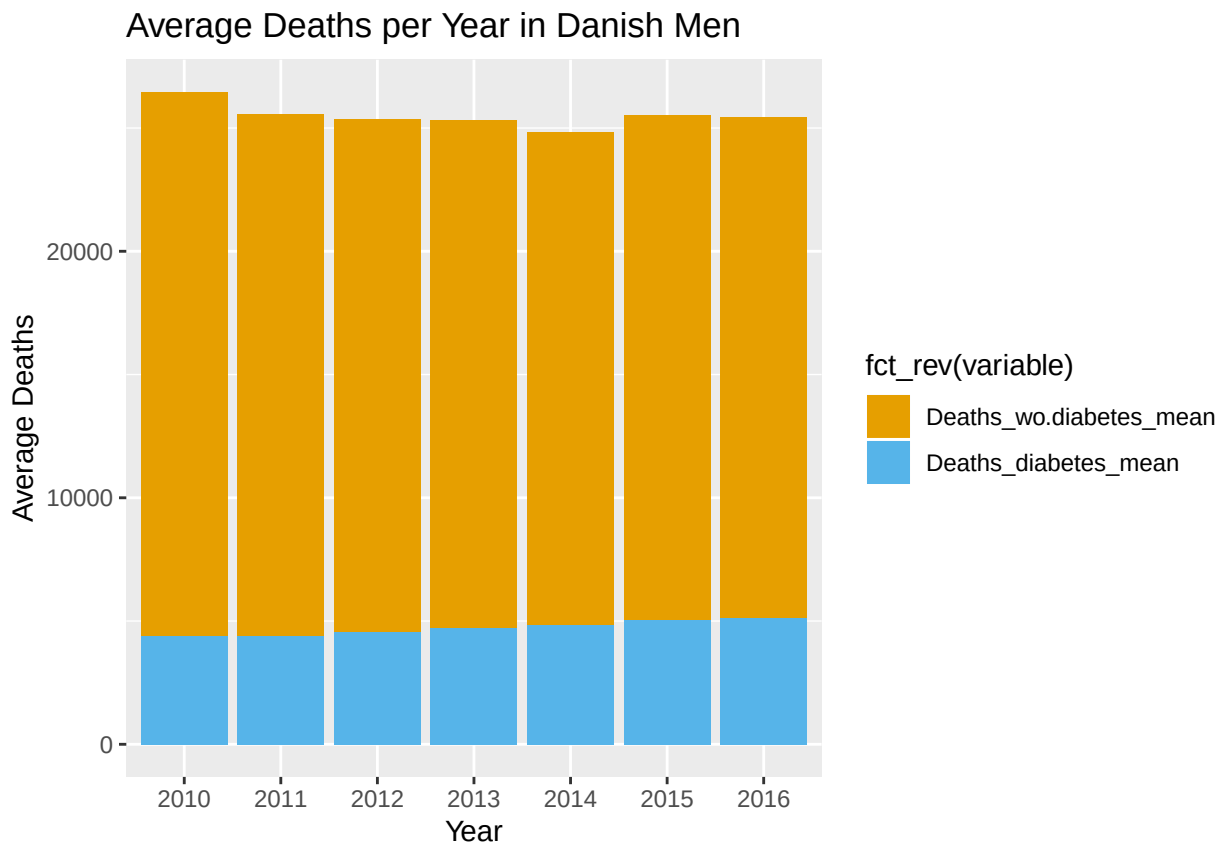
```
##                      Var1 Var2 value
## 1                       A    A  1.00
## 2                    Year    A  0.00
## 3                       N    A -0.74
## 4         Deaths_diabetes    A  0.72
## 5      Deaths_wo.diabetes    A  0.81
## 6 new_diabetes_diagnoses    A  0.35
```

```
d_melt %>%
  ggplot(aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "blue", high = "red") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



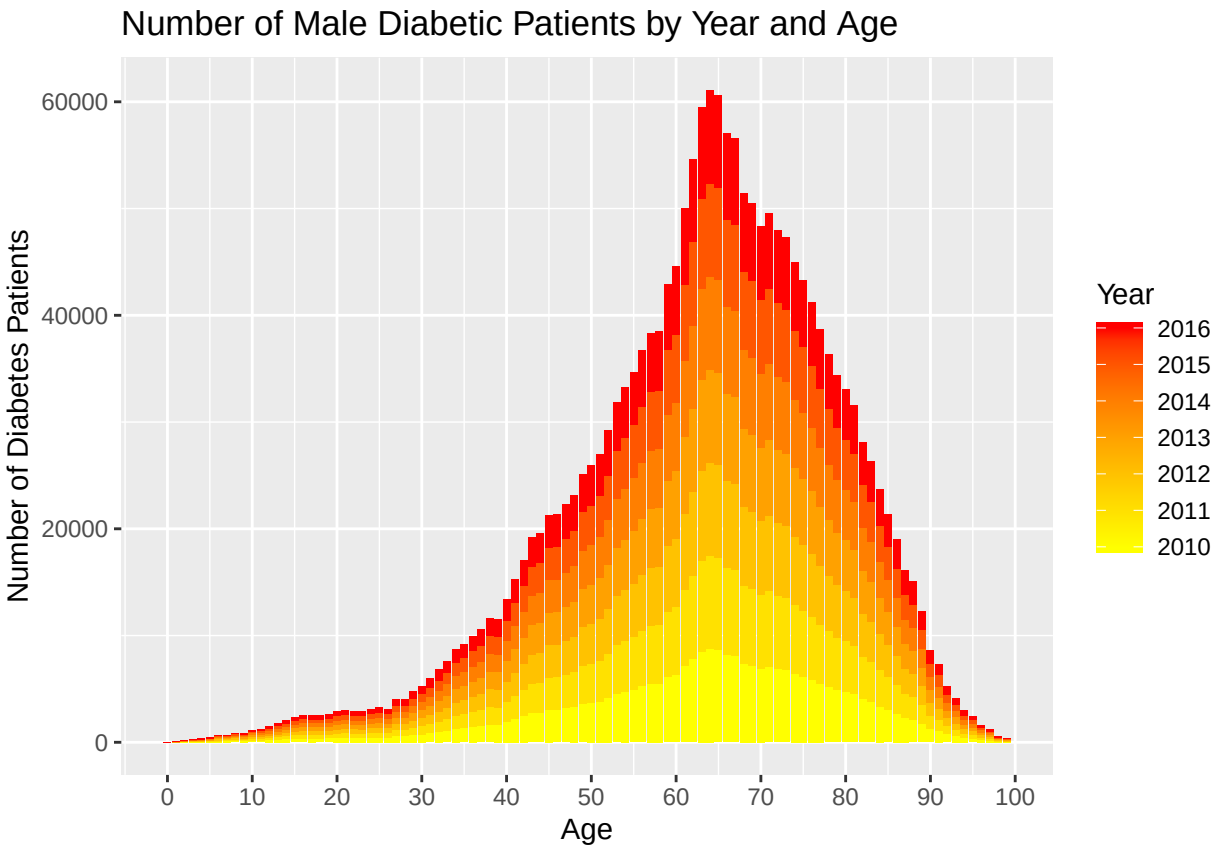
For the correlation heatmap, it's evident that the mortality variables (Deaths_diabetes and Deaths_wo.diabetes) had some positive correlation d_prevalence and d_incidence. Year and age had no correlation or negative correlation with most of our novel numeric variables(columns 4 through 6 on the heatmap) while age did have some correlation with our novel numeric variables.

```
mdf <- summ_diabetes %>% select(Year, Deaths_diabetes_mean, Deaths_wo.diabetes_mean)
mdf_melt <- melt(mdf, id.vars = "Year")

mdf_melt %>%
  ggplot(aes(x = factor(Year), y = value, fill = fct_rev(variable))) +
  geom_bar(stat = "identity", position = "stack") +
  scale_x_discrete(labels = c("2010", "2011", "2012", "2013", "2014", "2015", "2016")) +
  labs(x = "Year", y = "Average Deaths") + ggtitle("Average Deaths per Year in Danish Men") +
  scale_fill_discrete(name = NULL, labels = c("Non-Diabetic Deaths (avg)", "Diabetic Deaths (avg)")) +
  scale_fill_manual(values = c("#E69F00", "#56B4E9"))
```



For the second plot, I created a stacked barplot showing the average amount of deaths for each year with the cause of mortality as a result of diabetes or without diabetes. It might be hard to see, but average diabetic deaths gradually increases each year while the average non-diabetic deaths varied for each year.

```
d_diabetes2 %>%
  ggplot(aes(y = num_diabetes_patients, x = A, fill = Year)) +
  geom_bar(stat = "identity", aes(fill = Year)) +
  scale_x_continuous(breaks = seq(0,100, 10)) +
  labs(x = "Age", y= "Number of Diabetes Patients") + ggtitle("Number of Male Diabetic Patients by Year
  scale_fill_gradient(low = "yellow", high = "red")
```



For my third plot, I created a barplot showing the number of male diabetic patients by age per year. For each year, it is evident that the number of diabetic patients gradually increases by age with more diabetic patients being within the 60 to 70 age group. Also for each year, the total number of diabetic patients increases. The number of diabetic patients seems to drop off after 70, which is most likely attributed to mortality.

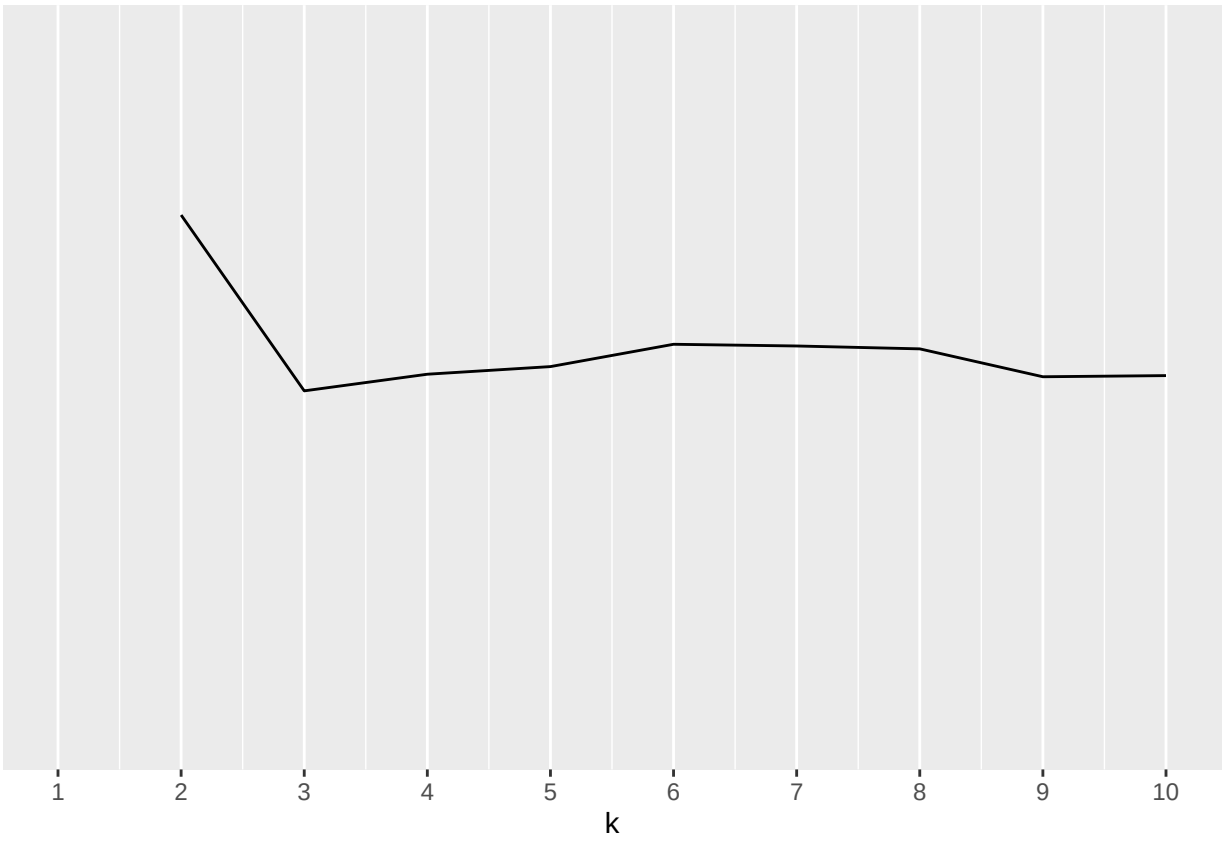
Dimenisonality Reduction

```
pam1 <- d_diabetes2 %>% pam(2)
pam1
```

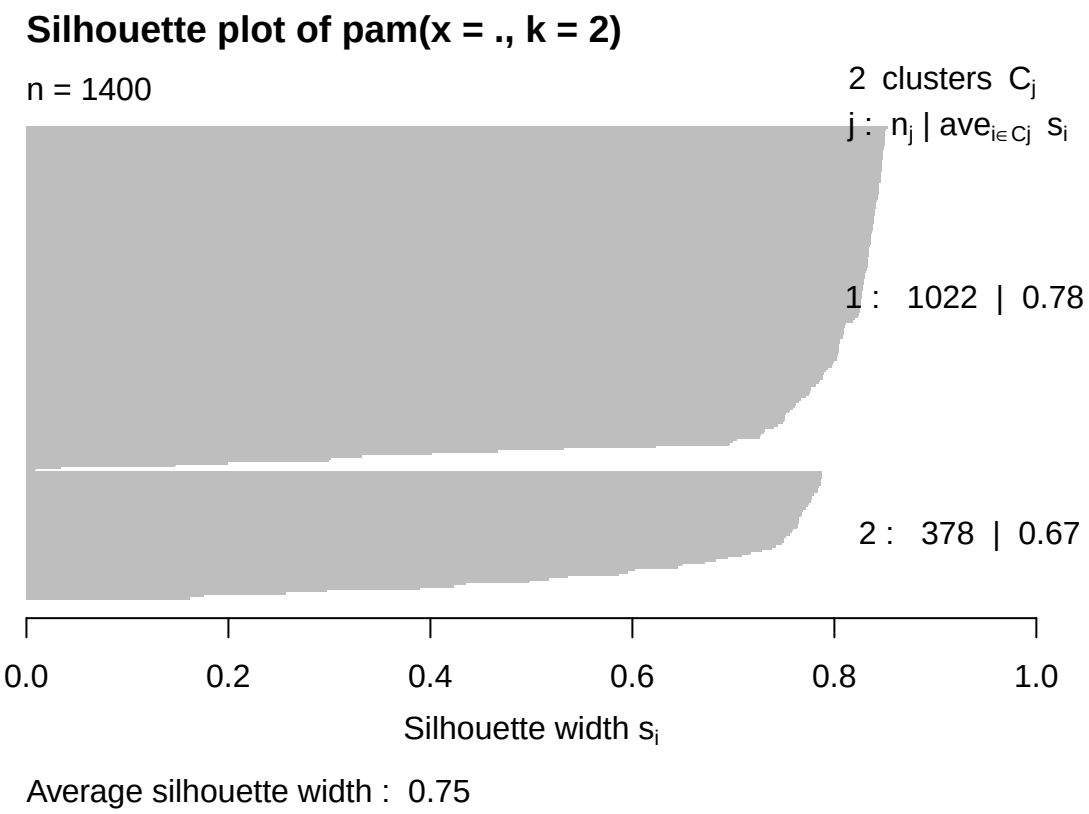
```
## Medoids:
##        ID  A Year     N Deaths_diabetes Deaths_wo.diabetes
## [1,]  533 32 2012 34208               1                 17
## [2,] 1084 83 2015  8549             160                652
##      new_diabetes_diagnoses num_diabetes_patients d_prevalence  d_incidence
## [1,]                     33                   584      584.001 0.0009646866
## [2,]                     77                  1579     1579.009 0.0090069014
## Clustering vector:
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2
##  [75] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [ reached getOption("max.print") -- omitted 1300 entries ]
## Objective function:
##    build     swap
## 4690.390 4447.071
##
```

```
## Available components:
##  [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
##  [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

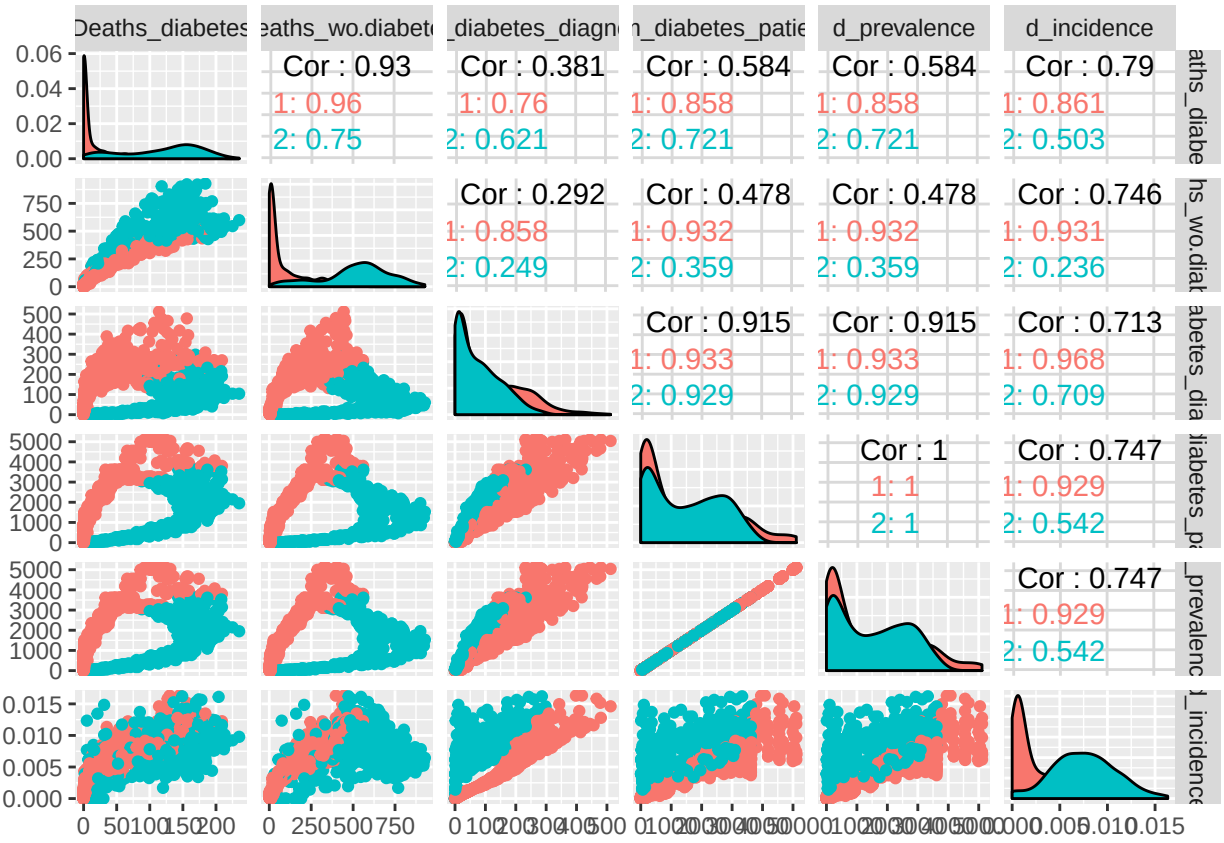
```
sil_width <- vector()
for(i in 2:10){
  pam_fit <- d_diabetes2 %>% pam(i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}

ggplot() +
  geom_line(aes(x = 1:10), y = sil_width) +
  scale_x_continuous(name = "k", breaks = 1:10)
```




```
plot(pam1, which = 2)
```




```
ggpairs(d_diabetes2, columns = 4:9, aes(color=as.factor(pam1$clustering)))
```



For the dimensionality portion, I performed PAM clustering on the d_diabetes2 dataset. Originally, I had PAM set to 3 clusters, but eventually chose 2 clusters. The average silhouette width was 0.75, indicating a strong structure was found. I then visualized all pairwise combinations on my novel numeric variables. A majority of the clusters correlated strongly with each other. New diabetes diagnoses correlated poorly with our mortality variables (Diabetic Deaths, Non-diabetic Deaths).