Using data.table()

John Reddy Peasari

2/15/2021

Setting the directory path and loading the data

```r
dir <- getwd()
setwd(dir)
```

Loading required packages

Loadind the data files

```r
clinic <- fread("healthcare-data/Clinic.csv")
disease_map <- fread("healthcare-data/DiseaseMap.csv")
icd_codes <- fread("healthcare-data/ICDCodes.csv")
insurance_provider <- fread("healthcare-data/InsuranceProvider.csv")
mortality <- fread("healthcare-data/Mortality.csv")
outpatient_visit <- fread("healthcare-data/OutpatientVisit.csv")
patient <- fread("healthcare-data/Patient.csv")
patient_file <- fread("healthcare-data/PatientAnalyticFile.csv")
patient_insurance <- fread("healthcare-data/PatientInsurance.csv")
staff <- fread("healthcare-data/Staff.csv")
```

Question 1

Are men more likely to die than women in this group of patients? Assume people without a date of death in the mortality table are still alive.

```r
setkey(patient,PatientID)
setkey(mortality,PatientID)
merged <- mortality[patient]
no_males <- nrow(merged[!is.na(DateOfDeath) & Gender=="male" ])
no_females <- nrow(merged[!is.na(DateOfDeath) & Gender=="female" ])
print(paste(no_males, "-> Total no of men died"))
```

```
## [1] "3209 -> Total no of men died"
```

```r
print(paste(no_females, "-> Total no of women died"))
```

```
## [1] "3337 -> Total no of women died"
```

It was observed that both men and women died in almost equal numbers. But,the women are more likely to die than men.

Question 2

Are patterns in the disease groups across gender. For every patient with at least one outpatient visit, identify if they have been diagnosed with any of the 22 conditions listed in the diseaseMap table at any time point. You will need to consider all three ICD columns in the outpatientVisit file (not just one). Create a table with the rate of disease for each condition for men, women, and all. It should look like this, where the XX% is the percent with the condition:

```
## Combining Patient and OutpatientVisit tables using setkey()
setkey(patient,PatientID)
setkey(outpatient_visit,PatientID)
merged1 <- patient[outpatient_visit]
merged1[1:10]
```

```
##     PatientID FirstName   LastName State ZipCode DateOfBirth Gender Race
##  1:         1     Diana Huddleston    WI   53186  1962-02-27 female
##  2:         1     Diana Huddleston    WI   53186  1962-02-27 female
##  3:         1     Diana Huddleston    WI   53186  1962-02-27 female
##  4:         1     Diana Huddleston    WI   53186  1962-02-27 female
##  5:         1     Diana Huddleston    WI   53186  1962-02-27 female
##  6:         1     Diana Huddleston    WI   53186  1962-02-27 female
##  7:         1     Diana Huddleston    WI   53186  1962-02-27 female
##  8:         1     Diana Huddleston    WI   53186  1962-02-27 female
##  9:         1     Diana Huddleston    WI   53186  1962-02-27 female
## 10:         1     Diana Huddleston    WI   53186  1962-02-27 female
##       Income VisitID StaffID  VisitDate ICD10_1 ICD10_2 ICD10_3 ClinicCode
##  1: 1076.168       1      46 2013-08-10  E10621    K269                  15
##  2: 1076.168       2      50 2013-12-02    K269  E10621                  55
##  3: 1076.168       3      13 2014-06-29  E10621    K269                   1
##  4: 1076.168       4      23 2014-09-19    K269  E10621                   3
##  5: 1076.168       5       9 2015-05-29    K269  E10621                   5
##  6: 1076.168       6      46 2016-05-07  E10621    K269                  15
##  7: 1076.168       7       7 2016-10-07  E10621    K269                  41
##  8: 1076.168       8      18 2016-11-07    K269  E10621                  31
##  9: 1076.168       9      23 2017-01-14    K269  E10621                   3
## 10: 1076.168      10       5 2017-01-29  E10621    K269                  14
```

```
## Getting PatientID and all the ICD10 columns (ICD10_1, ICD10_2, and ICD10_3)
all_ICD10 <- setDT(merged1)[, .(Freq = .N), by = .(PatientID, ICD10_1,ICD10_2,ICD10_3)]
all_ICD10[1:10]
```

```
##     PatientID ICD10_1 ICD10_2 ICD10_3 Freq
##  1:         1  E10621    K269            6
##  2:         1    K269  E10621            8
##  3:         2   C4650  O10019            4
##  4:         2  O10019   C4650            2
##  5:         3     B20   O1092           11
##  6:         3  O1092     B20            6
##  7:         4   J452    E131           11
##  8:         4   E131    J452            5
##  9:         5 O10013                   15
## 10:         6  Z0000                    4
```

```r
## Here, I merged all the ICD10 columns into a single column "ICD10" with theri codes
ID_ICD10 <- pivot_longer(all_ICD10, cols=2:4, names_to = "ICD10_1_2_3", values_to = "ICD10")
ID_ICD10 <- data.table(ID_ICD10)
ID_ICD10 <- ID_ICD10[,list(PatientID,ICD10)]
ID_ICD10[1:10]
```

```
##     PatientID  ICD10
##  1:         1 E10621
##  2:         1   K269
##  3:         1
##  4:         1   K269
##  5:         1 E10621
##  6:         1
##  7:         2  C4650
##  8:         2 O10019
##  9:         2
## 10:         2 O10019
```

```r
df.long <- ID_ICD10
ID_ICD10[ID_ICD10 == ''] <- NA ## Added NA to empty cells and dropped cells with NA values
new_ID_ICD10 <- ID_ICD10 %>% drop_na()
new_ID_ICD10[1:10]
```

```
##     PatientID  ICD10
##  1:         1 E10621
##  2:         1   K269
##  3:         1   K269
##  4:         1 E10621
##  5:         2  C4650
##  6:         2 O10019
##  7:         2 O10019
##  8:         2  C4650
##  9:         3    B20
## 10:         3  O1092
```

```r
## After removing NA values. I got only cells with unique codes (Removed repetative codes for particula
unique_ICD10 <- new_ID_ICD10 %>% distinct()
unique_ICD10 <- unique_ICD10[, ICD10:=as.character(ICD10)]
unique_ICD10[1:10]
```

```
##     PatientID  ICD10
##  1:         1 E10621
##  2:         1   K269
##  3:         2  C4650
##  4:         2 O10019
##  5:         3    B20
##  6:         3  O1092
##  7:         4   J452
##  8:         4   E131
##  9:         5 O10013
## 10:         6  Z0000
```

```r
## Combining previous unique_ICD10 with DiseaseMap tables to map ICD10 codes for each patient
setkey(unique_ICD10,ICD10)
setkey(disease_map,ICD10)
merged2 <- unique_ICD10[disease_map]
order_merged2 <- merged2[order(-PatientID,decreasing=TRUE)]
order_merged2[1:10]
```

```
##     PatientID  ICD10 DiseaseMapID                      Condition
##  1:         1 E10621         1506 Diabetes_without_complications
##  2:         1   K269         1429          Peptic_ulcer_disease
##  3:         2  C4650         2049                         Cancer
##  4:         2 O10019         3077                   Hypertension
##  5:         3    B20         3026                            HIV
##  6:         3  O1092         3084                   Hypertension
##  7:         4   E131         1550 Diabetes_without_complications
##  8:         4   J452          886                      Pulmonary
##  9:         5 O10013         3076                   Hypertension
## 10:         8    I10         3073                   Hypertension
```

```r
## Combining Patient table with previous table to map PatientID, ICD10, Condition, and Gender for each
setkey(patient,PatientID)
setkey(order_merged2,PatientID)
merged3 <- order_merged2[patient]
merged3[1:5]
```

```
##    PatientID  ICD10 DiseaseMapID                      Condition FirstName
## 1:         1 E10621         1506 Diabetes_without_complications     Diana
## 2:         1   K269         1429          Peptic_ulcer_disease     Diana
## 3:         2  C4650         2049                         Cancer    Marion
## 4:         2 O10019         3077                   Hypertension    Marion
## 5:         3    B20         3026                            HIV    Sandra
##      LastName State ZipCode DateOfBirth Gender  Race     Income
## 1: Huddleston    WI   53186  1962-02-27 female       1076.16798
## 2: Huddleston    WI   53186  1962-02-27 female       1076.16798
## 3:     Poston    IL   60527  1859-09-11   male white  475.78109
## 4:     Poston    IL   60527  1859-09-11   male white  475.78109
## 5:      Hamby    IL   60126  1946-02-15 female white   30.74799
```

```r
## Getting only Condition and the Gender for each Condition for each patientID
Only_Condition_Gender <- merged3[,list(Condition,Gender)]
Only_Condition_Gender[1:10]
```

```
##                           Condition Gender
##  1: Diabetes_without_complications female
##  2:           Peptic_ulcer_disease female
##  3:                         Cancer   male
##  4:                   Hypertension   male
##  5:                            HIV female
##  6:                   Hypertension female
##  7: Diabetes_without_complications female
##  8:                      Pulmonary female
##  9:                   Hypertension female
## 10:                           <NA>   male
```

```
## Here, I dropped rows that contain "MISSING" keyword in Gender and blank values in the Condition
## First added NA to blank columns and dropped rows that has NA values
nrow(Only_Condition_Gender)
```

```
## [1] 30737
```

```
Only_Condition_Gender <- Only_Condition_Gender[!grepl("MISSING",Only_Condition_Gender$Gender),]
nrow(Only_Condition_Gender)
```

```
## [1] 29190
```

```
Only_Condition_Gender[Only_Condition_Gender == ''] <- NA
refined_table <- Only_Condition_Gender %>% drop_na()
nrow(refined_table)
```

```
## [1] 23775
```

```
## Converting final table to a frequency table with proportions using prop.table() function
Final_output <- as.table(table(refined_table))
Final_output <- prop.table(Final_output,1)*100
Final_result <- as.data.frame.matrix(Final_output)
names(Final_result)[1] <- "Women" ## Refined output
names(Final_result)[2] <- "Men"
Final_table <- transform(Final_result, All = (Women + Men)) ## New column with combined proportions
Final_table <- Final_table[, c(2, 1, 3)]
Final_table
```

```
##                                     Men    Women All
## Alcohol                        49.17241 50.82759 100
## Cancer                         48.42562 51.57438 100
## Congestive_heart_failure       63.25758 36.74242 100
## Dementia                       46.74868 53.25132 100
## Depression                     38.85153 61.14847 100
## Diabetes_with_complications    46.42375 53.57625 100
## Diabetes_without_complications 47.32288 52.67712 100
## Drugs                          46.98630 53.01370 100
## HIV                            50.90909 49.09091 100
## Hypertension                   51.38042 48.61958 100
## LiverMild                      47.97688 52.02312 100
## LiverSevere                    50.32397 49.67603 100
## Metastatic_solid_tumour        49.75845 50.24155 100
## Myocardial_infarction          63.68039 36.31961 100
## Obesity                        41.67781 58.32219 100
## Paralysis                          42.79835 57.20165 100
## Peptic_ulcer_disease               45.19774 54.80226 100
## Peripheral_vascular_disease        46.62005 53.37995 100
## Pulmonary                          49.05231 50.94769 100
## Renal                              46.92308 53.07692 100
## Rheumatic                          44.14414 55.85586 100
## Stroke                             51.71756 48.28244 100
```

Question 3

Calculate the mortality rate for every year between 2005 and 2018. Is it generally increasing, or decreasing? Assume patients are only at risk of death as of their first visit (in the outpatient Visit file). Once they have died, they are no longer at risk in subsequent year.

```
## Loading data
mortality <- fread("healthcare-data/Mortality.csv")
outpatient_visit <- fread("healthcare-data/OutpatientVisit.csv")
```

```
## Getting total no of deaths in a year from 2005 to 2018
order_mortality <- mortality[order(-DateOfDeath,decreasing=TRUE)]
t1 <- format(order_mortality$DateOfDeath, format = "%Y")
deaths <- as.data.frame(table(t1))
names(deaths)[1] <- "Year"
names(deaths)[2] <- "Deaths"
deaths
```

```
##    Year Deaths
## 1  2005     79
## 2  2006    235
## 3  2007    356
## 4  2008    423
## 5  2009    479
## 6  2010    567
## 7  2011    605
## 8  2012    689
## 9  2013    715
## 10 2014    710
## 11 2015    702
## 12 2016    710
## 13 2017    601
## 14 2018    223
```

```
## Getting total popolation in an year from 2005 to 2018
outpatient_visit <- fread("healthcare-data/OutpatientVisit.csv")

outpatient_visit <- outpatient_visit[,list(PatientID,VisitDate)]
outpatient_visit$VisitDate <- format(as.Date(outpatient_visit$VisitDate, format="%y/%m/%d"),"%Y")
cc <- setDT(outpatient_visit)[, .(Freq = .N), by = .(PatientID, VisitDate)]
bb <- cc[order(-VisitDate,decreasing=TRUE)]
total_population <- as.data.frame(table(cc$VisitDate))
names(total_population)[1] <- "Year"
names(total_population)[2] <- "TotalPopulation"
total_population
```

```
##    Year TotalPopulation
## 1  2005             859
## 2  2006            2106
## 3  2007            3234
## 4  2008            4165
## 5  2009            5116
## 6  2010            5823
## 7  2011            6456
## 8  2012            7065
## 9  2013            7406
## 10 2014            7885
## 11 2015            8326
## 12 2016            8324
## 13 2017            7316
## 14 2018            4308
```

```
## Combining dataframes
TotalPopulation <- total_population$TotalPopulation
final <- cbind(deaths,TotalPopulation)
final
```

```
##    Year Deaths TotalPopulation
## 1  2005     79             859
## 2  2006    235            2106
## 3  2007    356            3234
## 4  2008    423            4165
## 5  2009    479            5116
## 6  2010    567            5823
## 7  2011    605            6456
## 8  2012    689            7065
## 9  2013    715            7406
## 10 2014    710            7885
## 11 2015    702            8326
## 12 2016    710            8324
## 13 2017    601            7316
## 14 2018    223            4308
```

```
Mortality_rate <- transform(final, MortalityRate = (Deaths / TotalPopulation)*100)
Mortality_rate
```

```
##    Year Deaths TotalPopulation MortalityRate
## 1  2005     79             859      9.196740
## 2  2006    235            2106     11.158594
## 3  2007    356            3234     11.008040
## 4  2008    423            4165     10.156062
## 5  2009    479            5116      9.362783
## 6  2010    567            5823      9.737249
## 7  2011    605            6456      9.371128
## 8  2012    689            7065      9.752300
## 9  2013    715            7406      9.654334
## 10 2014    710            7885      9.004439
## 11 2015    702            8326      8.431420
## 12 2016    710            8324      8.529553
## 13 2017    601            7316      8.214872
## 14 2018    223            4308      5.176416
```

It was observed that the mortality rate suddenly increased from 2005 to 2007. But, from the year 2009 it has decreased to the lowest value during 2005 to 2018 years.