Data Challegnge

Zihan Zhou

12/16/2019

Part1

Load Package and Data

```
library(feather)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v readr   1.3.1
## v tibble  2.1.3     v purrr   0.3.3
## v tidyr   1.0.0     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0
```

```
## -- Conflicts -------------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(qwraps2)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
options(qwraps2_markup = "markdown")
```

```
allergies = read_feather('../data/allergies.feather')
careplans = read_feather('../data/careplans.feather')
conditions = read_feather('../data/conditions.feather')
encounters = read_feather('../data/encounters.feather')
```

```
immunizations = read_feather('../data/immunizations.feather')
medications = read_feather('../data/medications.feather')
observations = read_feather('../data/observations.feather')
patients = read_feather('../data/patients.feather')
procedures = read_feather('../data/procedures.feather')
```

1. Data Inspection

We take a look of what information each dataframe contains one by one.

```
head(allergies)
```

```
## # A tibble: 6 x 6
##   START   STOP  PATIENT           ENCOUNTER            CODE DESCRIPTION
##   <chr>   <chr> <chr>             <chr>               <dbl> <chr>
## 1 1995-0~ <NA>  ab6d8296-d3c7-4f~ 9d87c22d-a777-426~ 4.26e8 Allergy to dai~
## 2 1995-0~ <NA>  ab6d8296-d3c7-4f~ 9d87c22d-a777-426~ 4.19e8 Allergy to tre~
## 3 1995-0~ <NA>  ab6d8296-d3c7-4f~ 9d87c22d-a777-426~ 4.19e8 Allergy to gra~
## 4 1995-0~ <NA>  ab6d8296-d3c7-4f~ 9d87c22d-a777-426~ 2.32e8 Dander (animal~
## 5 1995-0~ <NA>  ab6d8296-d3c7-4f~ 9d87c22d-a777-426~ 2.32e8 House dust mit~
## 6 1995-0~ <NA>  ab6d8296-d3c7-4f~ 9d87c22d-a777-426~ 4.19e8 Allergy to mou~
```

The `allergies` dataframe contains the information about one allergy description of a patient, with the ID of the patient, ID of the encounter and time periods of allergies.

```
head(careplans)
```

```
## # A tibble: 6 x 9
##   ID    START STOP  PATIENT ENCOUNTER    CODE DESCRIPTION REASONCODE
##   <chr> <chr> <chr> <chr>   <chr>       <dbl> <chr>            <dbl>
## 1 e031~ 2009~ 2009~ 719496~ 4d451e22~ 5.40e 7 Respirator~   10509002
## 2 e031~ 2009~ 2009~ 719496~ 4d451e22~ 3.05e 8 Recommenda~   10509002
## 3 e031~ 2009~ 2009~ 719496~ 4d451e22~ 3.72e 8 Deep breat~   10509002
## 4 26b8~ 2010~ 2010~ 719496~ bed7ecff~ 8.70e14 Urinary tr~   38822007
## 5 26b8~ 2010~ 2010~ 719496~ bed7ecff~ 2.23e 8 Discussion~   38822007
## 6 26b8~ 2010~ 2010~ 719496~ bed7ecff~ 1.71e 8 Urine scre~   38822007
## # ... with 1 more variable: REASONDESCRIPTION <chr>
```

The `careplans` dataframe contains treatments on a certain reason for an individual patient. The **CODE** column corresponds to treatments (e.g. 53950000 corresponds to 'Respiratory therapy' in the 0^{th} and 6^{th} sample). The **START** and **STOP** columns should be the dates that the patient starts and stops to receive the corresponding treatment.

```
head(conditions)
```

```
## # A tibble: 6 x 6
##   START   STOP    PATIENT          ENCOUNTER          CODE DESCRIPTION
##   <chr>   <chr>   <chr>            <chr>             <dbl> <chr>
## 1 2009-0~ 2009-0~ 71949668-1c2e-~ 4d451e22-a354-~ 1.05e7 Acute bronchitis ~
## 2 2010-1~ 2010-1~ 71949668-1c2e-~ bed7ecff-b41c-~ 3.88e7 Cystitis
## 3 2013-0~ 2013-0~ 71949668-1c2e-~ 6f2e3935-b203-~ 1.05e7 Acute bronchitis ~
## 4 2013-1~ 2014-0~ 71949668-1c2e-~ da4fd626-e74e-~ 7.29e7 Normal pregnancy
## 5 2014-0~ 2014-0~ 71949668-1c2e-~ b2e12445-b771-~ 1.96e8 Acute viral phary~
## 6 2017-0~ 2017-0~ 71949668-1c2e-~ 1c93058f-eeb2-~ 3.01e8 Escherichia coli ~
```

The `conditions` is a subset of `careplans`, containing the diagnosis of a patient on an encounter and its time period.

```
head(encounters)
```

```
## # A tibble: 6 x 7
##   ID       DATE   PATIENT     CODE DESCRIPTION  REASONCODE REASONDESCRIPTI~
##   <chr>    <chr>  <chr>      <dbl> <chr>             <dbl> <chr>
## 1 5114a5b~ 2008-~ 71949668~ 1.85e8 Outpatient ~         NA <NA>
## 2 4d451e2~ 2009-~ 71949668~ 1.85e8 Encounter f~   10509002 Acute bronchiti~
## 3 bdb926b~ 2009-~ 71949668~ 1.85e8 Outpatient ~         NA <NA>
## 4 f45c623~ 2010-~ 71949668~ 6.98e8 Consultatio~         NA <NA>
## 5 bed7ecf~ 2010-~ 71949668~ 1.85e8 Encounter f~   38822007 Cystitis
## 6 3679652~ 2012-~ 71949668~ 1.85e8 Outpatient ~         NA <NA>
```

The `encounters` has the description of an encounter and its reason.

```
head(immunizations)
```

```
## # A tibble: 6 x 5
##   DATE    PATIENT           ENCOUNTER          CODE DESCRIPTION
##   <chr>   <chr>             <chr>             <dbl> <chr>
## 1 2008-0~ 71949668-1c2e-43~ 5114a5b4-64b8-47~   140 Influenza  seasonal  i~
## 2 2009-0~ 71949668-1c2e-43~ bdb926b8-5b6d-43~   140 Influenza  seasonal  i~
## 3 2012-0~ 71949668-1c2e-43~ 36796523-2672-46~   140 Influenza  seasonal  i~
## 4 2012-0~ 71949668-1c2e-43~ 36796523-2672-46~   113 Td (adult) preservativ~
## 5 2015-0~ 71949668-1c2e-43~ 323e1478-fdbf-49~   140 Influenza  seasonal  i~
## 6 2008-0~ 96b24072-e1fe-49~ 4e7beaee-50c2-46~   140 Influenza  seasonal  i~
```

The `immunizations` contains patients' immunization injection history.

```
head(medications)
```

```
## # A tibble: 6 x 8
##   START STOP  PATIENT ENCOUNTER   CODE DESCRIPTION REASONCODE
##   <chr> <chr> <chr>   <chr>      <dbl> <chr>            <dbl>
## 1 1988~ <NA>  719496~ 5114a5b4~ 8.34e5 Penicillin~   43878008
## 2 2007~ 2008~ 719496~ 5114a5b4~ 1.37e6 NuvaRing 0~         NA
## 3 2009~ 2009~ 719496~ 4d451e22~ 6.09e5 Acetaminop~   10509002
## 4 2010~ 2011~ 719496~ f45c623f~ 7.49e5 Levora 0.1~         NA
## 5 2010~ 2010~ 719496~ bed7ecff~ 5.69e5 Nitrofuran~   38822007
## 6 2010~ 2010~ 719496~ bed7ecff~ 1.09e6 Phenazopyr~   38822007
## # ... with 1 more variable: REASONDESCRIPTION <chr>
```

The `medications` has the similar structure with `careplans`, which substitutes treatments with medicines.

```
head(observations)
```

```
## # A tibble: 6 x 7
##   DATE    PATIENT         ENCOUNTER        CODE  DESCRIPTION    VALUE UNITS
##   <chr>   <chr>           <chr>            <chr> <chr>          <chr> <chr>
## 1 2008-0~ 71949668-1c2e-~ 5114a5b4-64b8-4~ 8302~ Body Height    166.~ cm
## 2 2008-0~ 71949668-1c2e-~ 5114a5b4-64b8-4~ 2946~ Body Weight    54.42 kg
## 3 2008-0~ 71949668-1c2e-~ 5114a5b4-64b8-4~ 3915~ Body Mass Ind~ 19.74 kg/m2
## 4 2008-0~ 71949668-1c2e-~ 5114a5b4-64b8-4~ 8480~ Systolic Bloo~ 139.0 mmHg
## 5 2008-0~ 71949668-1c2e-~ 5114a5b4-64b8-4~ 8462~ Diastolic Blo~ 89.0  mmHg
## 6 2009-0~ 71949668-1c2e-~ bdb926b8-5b6d-4~ 8302~ Body Height    166.~ cm
```

The `observations` contains basic body measure data of patients.

```r
head(patients)
```

```
## # A tibble: 6 x 17
##   ID    BIRTHDATE DEATHDATE SSN   DRIVERS PASSPORT PREFIX FIRST LAST
##   <chr> <chr>     <chr>     <chr> <chr>   <chr>    <chr>  <chr> <chr>
## 1 7194~ 1988-05-~ <NA>      999-~ S99965~ X222441~ Mrs.   Elly~ Koss~
## 2 c2ca~ 1936-09-~ 1987-11-~ 999-~ S99982~ X643523~ Mr.    Kim2~ Barr~
## 3 96b2~ 1939-08-~ <NA>      999-~ S99997~ X205134~ Ms.    Jacq~ Shan~
## 4 de43~ 1931-09-~ 1981-10-~ 999-~ S99920~ false    Mr.    Nich~ Lind~
## 5 7926~ 1968-08-~ <NA>      999-~ S99919~ X718721~ Mr.    Maxw~ Diet~
## 6 3141~ 1921-11-~ 2012-03-~ 999-~ S99945~ false    Mrs.   Marg~ Wuns~
## # ... with 8 more variables: SUFFIX <chr>, MAIDEN <chr>, MARITAL <chr>,
## #   RACE <chr>, GENDER <chr>, BIRTHPLACE <chr>, ADDRESS <chr>, AGE <dbl>
```

The `patients` contains personal information of patients, such as name, age, ID number, etc.

```r
head(procedures)
```

```
## # A tibble: 6 x 7
##   DATE   PATIENT  ENCOUNTER    CODE DESCRIPTION REASONCODE REASONDESCRIPTI~
##   <chr>  <chr>    <chr>       <dbl> <chr>           <dbl> <chr>
## 1 2013-~ 7194966~ 6f2e3935~ 2.34e 7 Measuremen~   10509002 Acute bronchiti~
## 2 2013-~ 7194966~ da4fd626~ 2.52e 8 Standard p~   72892002 Normal pregnancy
## 3 2014-~ 7194966~ 988f02a3~ 2.37e 8 Augmentati~   72892002 Normal pregnancy
## 4 2014-~ 7194966~ 988f02a3~ 1.15e 7 Cesarean s~   72892002 Normal pregnancy
## 5 2016-~ 7194966~ 8ae1f76d~ 1.70e 8 Insertion ~         NA <NA>
## 6 2009-~ 96b2407~ cac10621~ 4.28e14 Documentat~         NA <NA>
```

The `procedures` contains medical procedures patients receive in an encounter.

We have gone through all dataframes given. The dataset provides us with personal information of patients, along with reason for each encounter and specific treatments and medicines each patient receives.

2. Summary Statistics

First we filter out visits from 2008 to 2016 and summarize information from different dataframes into one.

```r
start_date <- as.Date('2008-01-01')
end_date <- as.Date('2016-12-31')
encounters$DATE <- as.Date(encounters$DATE)

# Construct the new dataframe based on encounters
new_df <- encounters %>%
  filter((DATE >= start_date) & (DATE <= end_date))

# Sort by DATE
#new_df <- new_df[with(new_df, order(DATE)),]

# Add patients' information
#new_df <- left_join(new_df, patients, by = c('PATIENT' = 'ID'))

# Add conditions
#new_df <- left_join(new_df, conditions,
 #                   by = c('PATIENT' = 'PATIENT', 'ID' = 'ENCOUNTER'))
```

```
build_table <- function(data, variable, n, col_name){
  #"""
  #Return the table with top n categories for variable in data
  #"""
  counts <- data %>%
    group_by_at(variable) %>%
    dplyr::count(sort = TRUE)

  tb_list = list()
  for(i in 1:n){
    name <- eval(parse(text = paste0('counts$',variable,'[',i,']')))
    temp <- paste0('tb_list$`',name,
                   '` <- ~ qwraps2::n_perc0(.data$',variable,' == "',name,'",na_rm = TRUE)')
    eval(parse(text = temp))
  }

  summary1 <- eval(parse(text=paste0('list("',col_name,'"= tb_list)')))

  tab1 <- summary_table(data, summary1)
  return(tab1)
}
```

```
#Create separate tables and bind them at last

table_reason <- build_table(new_df, 'DESCRIPTION', 6, 'Most common reason for visit')
cname <- 'Encounters (n = 18,110)'
colnames(table_reason) <- cname

table_diagnosis <- build_table(new_df %>% filter(!is.na(REASONDESCRIPTION)),
                               'REASONDESCRIPTION', 6, 'Most common medical conditions')
colnames(table_diagnosis) <- cname

table_race <- build_table(patients, 'RACE', length(unique(patients$RACE)),
                          'RACE')
colnames(table_race) <- cname

age_summary <- patients %>%
  dplyr::select(AGE) %>%
  qsummary(.,
           numeric_summaries = list('Minimum' = '~ min(%s)',
                                    'Maximum' = '~ max(%s)',
                                    'Median (interquartile range) age' = '~ median_iqr(%s)'))
table_age <- summary_table(patients, age_summary)
colnames(table_age) <- cname

no_summary <-
  list('No of patients' = list('N' = ~ length(.data$ID)))
table_no <- summary_table(patients, no_summary)
colnames(table_no) <- cname

table_sex <- build_table(patients, 'GENDER', 2, 'Gender')
colnames(table_sex) <- cname

final_table <- rbind(table_no, table_reason, table_diagnosis, table_age, table_race, table_sex)
```

3. Run a simple model

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
# Load in income data
income <- read_csv('../data/Income.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
income$zip <- substr(income$GEO_ID, 10, 14)
income <- tibble(zip = income$zip,
                 income = income$S1903_C02_001E)
income$zip <- as.character(income$zip)
income$income <- as.numeric(income$income)
```

```
## Warning: NAs introduced by coercion
```

```
# Add mortality
new_df <- left_join(new_df, patients %>%
                      select(ID, DEATHDATE), by = c('PATIENT' = 'ID'))

new_df <- new_df %>%
  mutate(DATE = lubridate::as_date(DATE), DEATHDATE = lubridate::as_date(DEATHDATE)) %>%
  mutate(MORTALITY = (DATE %--% DEATHDATE)/dyears(1))

# Add zipcode
patients <- patients %>%
  mutate(zip = substr(ADDRESS, nchar(ADDRESS)-7, nchar(ADDRESS)-3))

new_df <- left_join(new_df, patients %>%
                      select(ID, zip, RACE), by = c('PATIENT' = 'ID'))
new_df <- left_join(new_df, income, by = c('zip' = 'zip'))

# Create income table
income_summary <- new_df %>%
  dplyr::select(income) %>%
  qsummary(.,
           numeric_summaries = list('Minimum' = '~ min(%s)',
                                    'Maximum' = '~ max(%s)',
                                    'Median (interquartile range) age' = '~ median_iqr(%s)'))
table_income <- summary_table(new_df, income_summary)
colnames(table_income) <- cname

final_table <- rbind(final_table, table_income)
print(final_table)
```

	Encounters (n = 18,110)
No of patients	
N	1462
Most common reason for visit	
Outpatient Encounter	7,758 (43)
Encounter for symptom	2,600 (14)
Patient encounter procedure	1,363 (8)
Prenatal visit	1,211 (7)
Outpatient procedure	1,161 (6)
Consultation for treatment	806 (4)
Most common medical conditions	
Normal pregnancy	1,657 (32)
Viral sinusitis (disorder)	956 (19)
Acute viral pharyngitis (disorder)	540 (10)
Acute bronchitis (disorder)	469 (9)
Child attention deficit disorder	177 (3)
Otitis media	172 (3)
AGE	
Minimum	0
Maximum	101
Median (interquartile range) age	51.00 (24.00, 77.00)
RACE	
white	1,085 (74)
hispanic	155 (11)
asian	93 (6)
black	90 (6)
black or african american	39 (3)
Gender	
M	741 (51)
F	721 (49)
income	
Minimum	18843
Maximum	191744
Median (interquartile range) age	78,114.00 (62,157.00, 95,592.00)
Unknown	4,647/18,110 (26)

```
new_df_2 <- new_df %>%
  filter(!is.na(MORTALITY))

fit <- lm(MORTALITY ~ income + RACE, data = new_df_2)
summary(fit)
```

```
##
## Call:
## lm(formula = MORTALITY ~ income + RACE, data = new_df_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2503 -1.9843 -0.4489  1.6740  7.2203
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.444e+00  2.601e-01  13.239  < 2e-16 ***
```

```
## income                         -1.493e-06  2.259e-06  -0.661   0.5089
## RACEblack                      -2.361e-02  4.047e-01  -0.058   0.9535
## RACEblack or african american   7.428e-01  3.292e-01   2.256   0.0242 *
## RACEhispanic                   -1.383e+00  3.268e-01  -4.231 2.49e-05 ***
## RACEwhite                      -5.377e-01  2.209e-01  -2.434   0.0151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 1258 degrees of freedom
##   (328 observations deleted due to missingness)
## Multiple R-squared:  0.03367,    Adjusted R-squared:  0.02983
## F-statistic: 8.767 on 5 and 1258 DF,  p-value: 3.465e-08
```

It seems like mortality has little relevance to income, but rather higher relevance to race.

Part2

1.

I might consider using time-series models for the prediction, since we are given a large amount of data, with a time-series nature.

2.

Let A denote the event that the chosen individual has the condition. Then $P(A) = \frac{1}{12}$. We want $P(A|\text{test positive})$.

$$
\begin{aligned}
P(A|\text{test positive}) &= \frac{P(\text{test positive}|A)P(A)}{P(\text{test positive})} \\
&= \frac{P(\text{test} = 1, +|A)P(A) + P(\text{test} = 2, +|A)P(A)}{\sum_{i=1}^{2}\left[P(\text{test} = i, +, A) + P(\text{test} = i, +, A^c)\right]} \\
&= \frac{P(+|\text{test} = 1, A)P(\text{test} = 1|A)P(A) + P(+|\text{test} = 2, A)P(\text{test} = 2|A)P(A)}{\sum_{i=1}^{2}\left[P(\text{test} = i, +, A) + P(\text{test} = i, +, A^c)\right]} \\
&= \frac{1 \times \frac{1}{3} \times \frac{1}{12} + \frac{5}{6} \times \frac{2}{3} \times \frac{1}{12}}{1 \times \frac{1}{3} \times \frac{1}{12} + \frac{5}{6} \times \frac{2}{3} \times \frac{1}{12} + \sum_{i=1}^{2} P(\text{test} = i, +, A^c)} \\
&= \frac{\frac{2}{27}}{\frac{2}{27} + P(+|\text{test} = 1, A^c)P(\text{test} = 1|A^c)P(A^c) + P(+|\text{test} = 2, A^c)P(\text{test} = 2|A^c)P(A^c)} \\
&= \frac{\frac{2}{27}}{\frac{2}{27} + \frac{1}{2} \times \frac{1}{3} \times \frac{11}{12} + \frac{1}{4} \times \frac{2}{3} \times \frac{11}{12}} = \frac{8}{41}
\end{aligned}
$$