data_mining_using_R.R

Deepak

Sat Mar 10 20:44:08 2018

```
### Data Mining using dplyr
### Author: Deepak Agarwal

## Health care data
## Data Source - https://data.medicare.gov/data/physician-compare

# load the library
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
```

```
# read the data

data <- read.csv("Physician_Compare_National_Downloadable_File.csv")



# check first few rows of data

head(data)
```

```
##          NPI     PAC.ID Professional.Enrollment.ID Last.Name First.Name
## 1 1487927612 4880850486            I20120726000331      HALL     ESTHER
## 2 1235146762 2365435336            I20040406000367     WHITE    BARBARA
## 3 1346282258 5395768527            I20060113000139  DAVIDSON       JOHN
## 4 1932283124 5193762862            I20050415000143     CAGEN     STEVEN
## 5 1902950462 7416123666            I20120110000522      ESPY     LEISHA
## 6 1518981026 7719166586            I20110125001223  PETROSKY        DAN
##   Middle.Name Suffix Gender Credential
## 1           S             F
## 2           L             F        CSW
## 3           A             M        CSW
## 4           F             M         DC
## 5           H             F
## 6           M             M
##                         Medical.school.name Graduation.year
## 1         LIFE CHIROPRACTIC COLLEGE - WEST            2010
## 2                                    OTHER            1992
## 3                                    OTHER            1999
## 4 SHERMAN COLLEGE OF STRAIGHT CHIROPRACTIC            1997
## 5                LIFE CHIROPRACTIC COLLEGE            1985
## 6                                    OTHER            1976
##        Primary.specialty Secondary.specialty.1 Secondary.specialty.2
## 1           CHIROPRACTIC
## 2 CLINICAL SOCIAL WORKER
## 3 CLINICAL SOCIAL WORKER
## 4           CHIROPRACTIC
## 5           CHIROPRACTIC
## 6           CHIROPRACTIC
##   Secondary.specialty.3 Secondary.specialty.4 All.secondary.specialties
## 1
## 2
## 3
## 4
## 5
## 6
##          Organization.legal.name Group.Practice.PAC.ID
## 1                                                   NA
## 2                                                   NA
## 3                                                   NA
## 4 CAGEN FAMILY CHIROPRACTIC PLLC            7012954787
## 5                                                   NA
## 6                                                   NA
##   Number.of.Group.Practice.members Line.1.Street.Address
## 1                               NA    183 PLACERVILLE DR
## 2                               NA          163 ENGLE ST
## 3                               NA      9 KATTELVILLE RD
## 4                                2    1486 ASHEVILLE HWY
## 5                               NA      100 E GORDON AVE
```

```
## 6                               NA          125 N 6TH ST
##   Line.2.Street.Address Marker.of.address.line.2.suppression        City
## 1               SUITE A                                       PLACERVILLE
## 2                                                               ENGLEWOOD
## 3                                                              BINGHAMTON
## 4                                                                 BREVARD
## 5                                                               ROSSVILLE
## 6                                                                  ALPINE
##   State  Zip.Code Phone.Number Hospital.affiliation.CCN.1
## 1    CA 956673933   5306228041
## 2    NJ 076312530   2014102812
## 3    NY 139015821   6072456259
## 4    NC 287129524   8288857100
## 5    GA 307411348   7068667557
## 6    TX 798304607   4328371800
##   Hospital.affiliation.LBN.1 Hospital.affiliation.CCN.2
## 1
## 2
## 3
## 4
## 5
## 6
##   Hospital.affiliation.LBN.2 Hospital.affiliation.CCN.3
## 1
## 2
## 3
## 4
## 5
## 6
##   Hospital.affiliation.LBN.3 Hospital.affiliation.CCN.4
## 1
## 2
## 3
## 4
## 5
## 6
##   Hospital.affiliation.LBN.4 Hospital.affiliation.CCN.5
## 1                                                    NA
## 2                                                    NA
## 3                                                    NA
## 4                                                    NA
## 5                                                    NA
## 6                                                    NA
##   Hospital.affiliation.LBN.5 Professional.accepts.Medicare.Assignment
## 1                                                                   Y
## 2                                                                   Y
## 3                                                                   Y
## 4                                                                   Y
## 5                                                                   M
```

```
## 6                                                                    M
##   Reported.Quality.Measures Used.electronic.health.records
## 1
## 2
## 3
## 4
## 5
## 6
##   Committed.to.heart.health.through.the.Million.HeartsÂ..initiative.
## 1
## 2
## 3
## 4
## 5
## 6
```

```
# check unique ids for NPI column

data %>% distinct(NPI) %>% tally()
```

```
##         n
## 1 1070395
```

```
# check unique ids for PAC.ID column

data %>% distinct(PAC.ID) %>% tally()
```

```
##         n
## 1 1070399
```

```
# check unique ids for both NPI, PAC.ID column combined

data %>% distinct(NPI,PAC.ID) %>% tally()
```

```
##         n
## 1 1070399
```

```
# Print the gender wise data (number of males,females)

data %>% group_by(Gender) %>% summarise(n=n())
```

```
## # A tibble: 3 x 2
##   Gender       n
##   <fctr>   <int>
## 1      F 1280809
## 2      M 1674452
## 3      U       1
```

```
# subset the data on the basis of unique PAC.ID to find male-female ratio

data_distinct <- data %>% distinct(PAC.ID,.keep_all = T)

# display the ratio (rounded upto four decimals) of males to females

data_distinct %>%
  group_by(Gender) %>%
      summarise(n=n()) %>%
            mutate(ratio=format(round(n/n[Gender=='F'],4),nsmall=4)) %>%
                                                    filter(ratio==max(ratio))
```

```
## # A tibble: 1 x 3
##   Gender      n  ratio
##   <fctr>  <int>  <chr>
## 1      M 578092 1.1743
```

```
# find the credential with highest male-female ratio

credential_female_ratio <- data_distinct %>%
                             group_by(Credential,Gender) %>%
                                          summarise(n=n()) %>%
                                               mutate(ratio=format(round((1/(n/n[Gen
der=='F'])),4),nsmall=4)) %>%
                                                                                filter
(Gender=='M',ratio==max(ratio)) %>%

              select(Credential,ratio)

# display the credential

credential_female_ratio[which.max(credential_female_ratio$ratio),]
```

```
## # A tibble: 1 x 2
## # Groups:   Credential [1]
##   Credential    ratio
##       <fctr>    <chr>
## 1        CNM 130.8571
```

```
# load the performance data for the physicians

performance_data <- read.csv("Physician_Compare_2015_Individual_EP_Public_Reporting___
Performance_Scores.csv")

# display the data

head(performance_data)
```

```
##          NPI   PAC.ID Last.Name First.Name Measure.Identifier
## 1 1508823618 42100117   GRIFFIN      DAVID      PQRS_EP_110_1
## 2 1508823618 42100117   GRIFFIN      DAVID      PQRS_EP_111_1
## 3 1508823618 42100117   GRIFFIN      DAVID      PQRS_EP_112_1
## 4 1508823618 42100117   GRIFFIN      DAVID      PQRS_EP_113_1
## 5 1508823618 42100117   GRIFFIN      DAVID      PQRS_EP_128_1
## 6 1508823618 42100117   GRIFFIN      DAVID      PQRS_EP_130_1
##                                                                    Measure.Title
## 1                           Preventive Care and Screening: Influenza Immunization
## 2                                   Pneumonia Vaccination Status for Older Adults
## 3                                                        Breast Cancer Screening
## 4                                                    Colorectal Cancer Screening
## 5 Preventive Care and Screening: Body Mass Index (BMI) Screening and Follow-Up Plan
## 6                       Documentation of Current Medications in the Medical Record
##   Inverse.Measure Measure.Performance.Rate Reporting.Mechanism
## 1               N                       21                 CLM
## 2               N                       28                 CLM
## 3               N                       37                 CLM
## 4               N                       22                 CLM
## 5               N                       42                 CLM
## 6               N                       92                 CLM
##   Reported.on.PC.Live.Site
## 1                        Y
## 2                        Y
## 3                        Y
## 4                        Y
## 5                        Y
## 6                        Y
```

```
# display the standard deviation where the average performance measure rate is greater
than 10

performance_data %>%
  group_by(PAC.ID) %>%
    summarise(avg_perf_measure=mean(Measure.Performance.Rate)) %>%
                                      filter(avg_perf_measure>=10) %>%
                                          summarise(sd=format(sd(avg_perf_measure),nsm
all = 10))
```

```
## # A tibble: 1 x 1
##              sd
##           <chr>
## 1 22.5100915537
```

```
# find the practitioners who have at least 10 performance measures

practitioner_atleast10_measures <- performance_data %>%
                                          group_by(PAC.ID) %>%
                                              summarise(perf_measures=n()) %>%
                                                        filter(perf_measures>=10)

head(practitioner_atleast10_measures)
```

```
## # A tibble: 6 x 2
##     PAC.ID perf_measures
##      <dbl>         <int>
## 1 42108672            10
## 2 42210056            10
## 3 42217044            10
## 4 42244816            10
## 5 42255168            10
## 6 42273351            13
```

```
# subset performance data for practitioners who have at least 10 performance measures

performance_data_atleast10measures <- performance_data %>%
                                        filter(PAC.ID %in% practitioner_atleast10_meas
ures$PAC.ID) %>%
                                                                        select(PAC.ID,Meas
ure.Performance.Rate)


# subset physician data for PAC.ID and credential

data_credentials <- data %>% select(PAC.ID,Credential)


# join the performance data and credential data to get credentials for practitioners w
ho have at least 10 performance measures

data_credentials_practitioner_atleast_10measure <- left_join(performance_data_atleast1
0measures,data_credentials,by="PAC.ID")


# calculate the absolute difference between average performance rate of practitioners
with credentials 'MD' and 'NP'

data_credentials_practitioner_atleast_10measure %>%
                                group_by(Credential) %>%
                                  summarise(avg_perf_measure=mean(Measure.Performance.
Rate)) %>%
                                                                  filter(Credential %in% c("M
D","NP")) %>%
                                                                    mutate(diff=format(abs(av
g_perf_measure[1]-avg_perf_measure[2]),
                                                                                       nsmal
l = 4))
```

```
## # A tibble: 2 x 3
##   Credential avg_perf_measure     diff
##       <fctr>            <dbl>    <chr>
## 1         MD         62.71128 6.971555
## 2         NP         55.73973 6.971555
```

```
# find if the average difference between performance measure of 'MD' and 'NP' is signi
ficant or not

practitioner_10measure_md <- data_credentials_practitioner_atleast_10measure %>%
                                                              filter(Credential=="MD") %
>%
                                                                      select(Measure.Per
formance.Rate)

practitioner_10measure_np <- data_credentials_practitioner_atleast_10measure %>%
                                                            filter(Credential=="NP")  %>%
                                                                select(Measure.Performanc
e.Rate)

# perform two sample t-test

ttest <- t.test(practitioner_10measure_md,practitioner_10measure_np,var.equal = T)

# dislay the results

ttest
```

```
##
##  Two Sample t-test
##
## data:  practitioner_10measure_md and practitioner_10measure_np
## t = 8.0384, df = 30341, p-value = 9.43e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.271647 8.671463
## sample estimates:
## mean of x mean of y
##  62.71128  55.73973
```

```
# display the actual p-value

format(ttest$p.value,scientific = F)
```

```
## [1] "0.000000000000000009429927"
```

```
## find the average performance measure of data practitioners whose graduation year is
between 1973-2003 both inclusive

# subset the practitoner data for the said years

data_practitioner_1973_2003 <- subset(data_distinct, Graduation.year>=1973 & Graduatio
n.year<=2003)


# join subsetted practitoner data with the performance data using PAC.ID

performance_data_atleast10measures_1973_2003 <- inner_join(performance_data_atleast10m
easures,
                                                           data_practitioner_1973_200
3,by="PAC.ID")

# find the average performance measure year-wise

performance_data_atleast10measures_1973_2003 %>%
                          group_by(Graduation.year) %>%
                                summarise(avg_perf_measure=mean(Measure.Performance.Ra
te))
```

```
## # A tibble: 31 x 2
##    Graduation.year avg_perf_measure
##              <int>            <dbl>
##  1            1973         67.17904
##  2            1974         64.97458
##  3            1975         68.04482
##  4            1976         68.25115
##  5            1977         58.11152
##  6            1978         63.22989
##  7            1979         65.08229
##  8            1980         64.46789
##  9            1981         65.59265
## 10            1982         65.86441
## # ... with 21 more rows
```

```r
## find the linear relationship between performance measure and graduation year for 19
73-2003

# get the target and predictor variables

y <- performance_data_atleast10measures_1973_2003$Measure.Performance.Rate
x <- performance_data_atleast10measures_1973_2003$Graduation.year

# fit the simple linear regresion model

fit <- lm(y~x)

# display the summary of model

summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.227 -24.261   6.903  32.479  37.009
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 345.80264   57.16022   6.050 1.48e-09 ***
## x            -0.14119    0.02872  -4.916 8.89e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.12 on 20922 degrees of freedom
## Multiple R-squared:  0.001154,   Adjusted R-squared:  0.001106
## F-statistic: 24.17 on 1 and 20922 DF,  p-value: 8.893e-07
```

```r
# find the p-value of the linear regression model

linregpval <- function (modelobject) {
  if (class(modelobject) != "lm") stop("Not an object of class 'lm' ")
  fstat <- summary(modelobject)$fstatistic
  pval <- pf(fstat[1],fstat[2],fstat[3],lower.tail=F)
  attributes(pval) <- NULL
  return(pval)
}

# display the p-value

format(linregpval(fit),scientific = F)
```

```
## [1] "0.000000889345"
```