

Name: Naquibuddin Shaik

Healthcare cost analysis

Project 7

DESCRIPTION

Background and Objective:

A nationwide survey of hospital costs conducted by the US Agency for Healthcare consists of hospital records of inpatient samples. The given data is restricted to the city of Wisconsin and relates to patients in the age group 0-17 years. The agency wants to analyse the data to research on healthcare costs and their utilization.

Domain: Healthcare

Dataset Description:

Here is a detailed description of the given dataset:

Attribute	Description
Age	Age of the patient discharged
Female	A binary variable that indicates if the patient is female
Los	Length of stay in days
Race	Race of the patient (specified numerically)
Totchg	Hospital discharge costs
Aprdrg	All Patient Refined Diagnosis Related Groups

Loading Libraries:

```
# loading libraries  
library(readxl) # for reading excel files  
library(dplyr) # for data manipulation  
library(corrplot) # for correlation-plot  
library(caTools) # for splitting dataset  
library(ggplot2) # for data visualization  
library(MLmetrics) # for machine learning metrics  
library(caret) # for feature importance  
library(lattice)
```

Data Loading, EDA, Data Preparation, Data Cleaning:

```
# removing objects in environment  
rm(list = ls())
```

```
# loading dataset  
mydf <- read_xlsx('hospitalcosts.xlsx')  
View(mydf)
```

```
# EDA  
dim(mydf)  
summary(mydf)  
str(mydf)
```

```
# checking and handling NA  
sapply(mydf, function(x) sum(is.na(x))) # count of NA for each feature
```

```
which(is.na(mydf$RACE)) # row number for NA in RACE
mydf <- na.omit(mydf)
```

```
# creating GENDER column
mydf$GENDER <- ifelse(mydf$FEMALE == 1, 'Female', 'Male')
mydf$GENDER <- as.factor(mydf$GENDER)
mydf$GENDER
```

```
# creating age class interval column i.e 0-5, 5-10, 10-15, 15-20
range(mydf$AGE)
bins <- seq(0,20,by=5)
bins
mydf$AGE_GROUP <- cut(mydf$AGE, bins, include.lowest = T)
summary(mydf)
View(mydf)
```

Questions:

1.To record the patient statistics, the agency wants to find the age category of people who frequently visit the hospital and has the maximum expenditure.

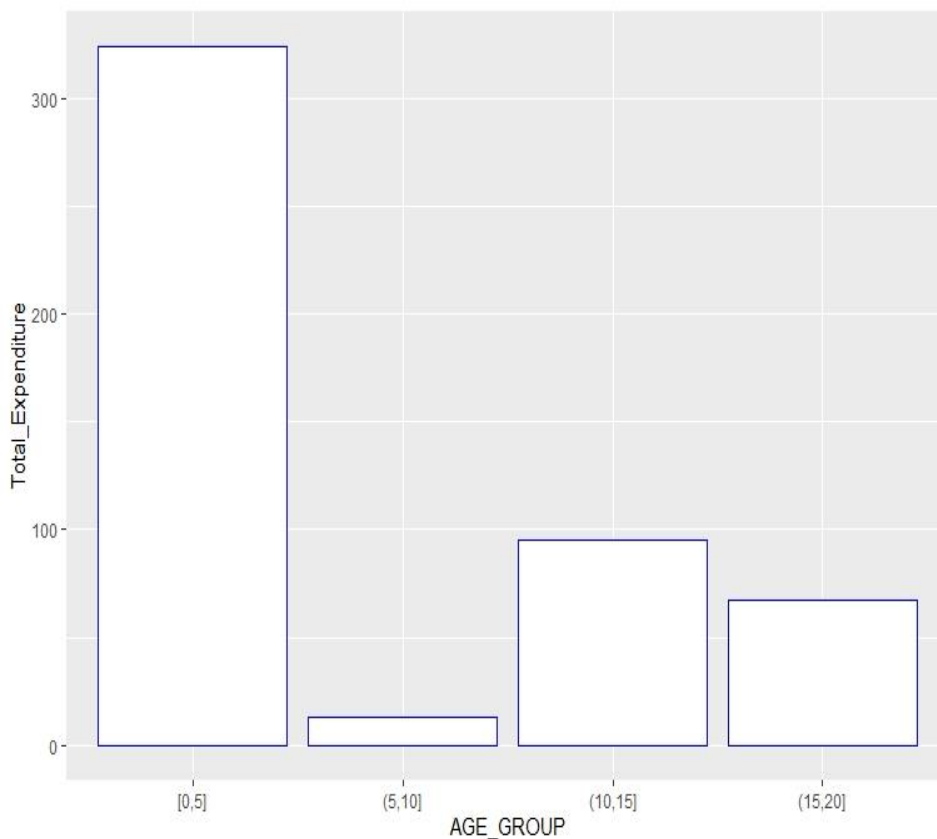
question-1

```
group_age <- group_by(mydf,AGE_GROUP)
age_res <- summarise(group_age, Total_Expenditure = n())
age_res[which.max(age_res$Total_Expenditure),]
ggplot(age_res, aes(x = AGE_GROUP, y = Total_Expenditure)) + geom_bar(stat =
'identity', color = 'blue', fill = 'white')
```

```

> # question-1
> group_age <- group_by(mydf, AGE_GROUP)
> age_res <- summarise(group_age, Total_Expenditure = n())
`summarise()` ungrouping output (override with `.groups` argument)
> age_res[which.max(age_res$Total_Expenditure),]
# A tibble: 1 x 2
  AGE_GROUP Total_Expenditure
  <fct>         <int>
1 [0,5]           324
> |

```

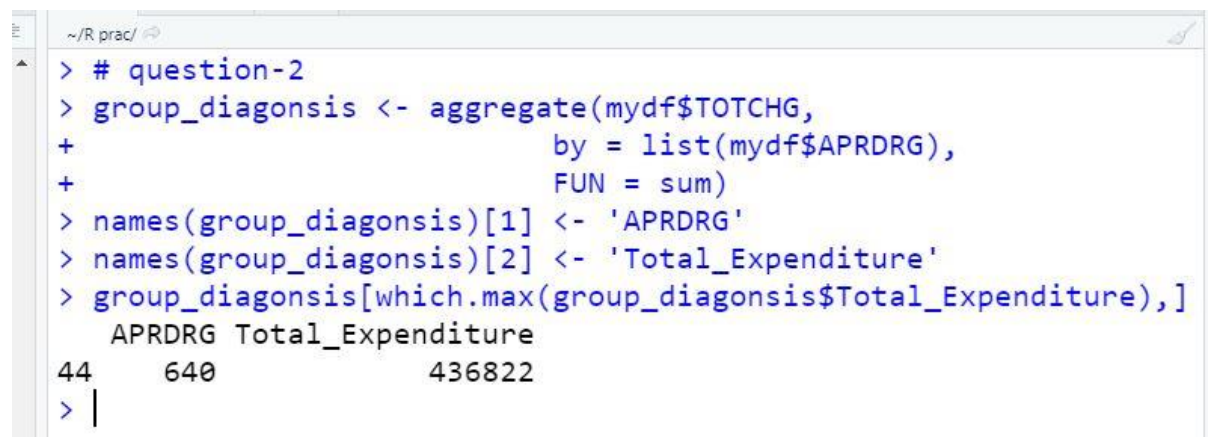


Hence from the bar plot we can say that the age category of people who frequently visit the hospital and has the maximum expenditure is between age of 0 to 5(both inclusive)

2. In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis-related group that has maximum hospitalization and expenditure.

question-2

```
group_diagnosis <- aggregate(mydf$TOTCHG,  
                             by = list(mydf$APRDRG),  
                             FUN = sum)  
names(group_diagnosis)[1] <- 'APRDRG'  
names(group_diagnosis)[2] <- 'Total_Expenditure'  
group_diagnosis[which.max(group_diagnosis$Total_Expenditure),]
```



```
> # question-2  
> group_diagnosis <- aggregate(mydf$TOTCHG,  
+                             by = list(mydf$APRDRG),  
+                             FUN = sum)  
> names(group_diagnosis)[1] <- 'APRDRG'  
> names(group_diagnosis)[2] <- 'Total_Expenditure'  
> group_diagnosis[which.max(group_diagnosis$Total_Expenditure),]  
  APRDRG Total_Expenditure  
44      640           436822  
> |
```

So Diagnosis-Related group which has the maximum expenditure is 640.

3.To make sure that there is no malpractice, the agency needs to analyse if the race of the patient is related to the hospitalization costs.

question-3

```
group_race <- aggregate(mydf$TOTCHG,
                        by = list(mydf$RACE),
                        FUN = sum)

names(group_race)[1] <- 'RACE'
names(group_race)[2] <- 'Total_Expenditure'

group_race

anova_model <- aov(mydf$TOTCHG~mydf$RACE, data = mydf)

summary(anova_model)

barplot(group_race$Total_Expenditure,
        border = 'dark blue',
        main = 'Race - Total_Expenditure',
        xlab = 'Race',
        ylab= 'Total_Expenditure',
        names.arg = c(1:6))
```

```
> # question-3
> group_race <- aggregate(mydf$TOTCHG,
+                          by = list(mydf$RACE),
+                          FUN = sum)
> names(group_race)[1] <- 'RACE'
> names(group_race)[2] <- 'Total_Expenditure'
> group_race
  RACE Total_Expenditure
1    1           1341972
2    2             25213
3    3              3041
4    4               7034
5    5              6080
6    6               2698
> |
```

```

> anova_model <- aov(mydf$TOTCHG~mydf$RACE, data = mydf)
> summary(anova_model)

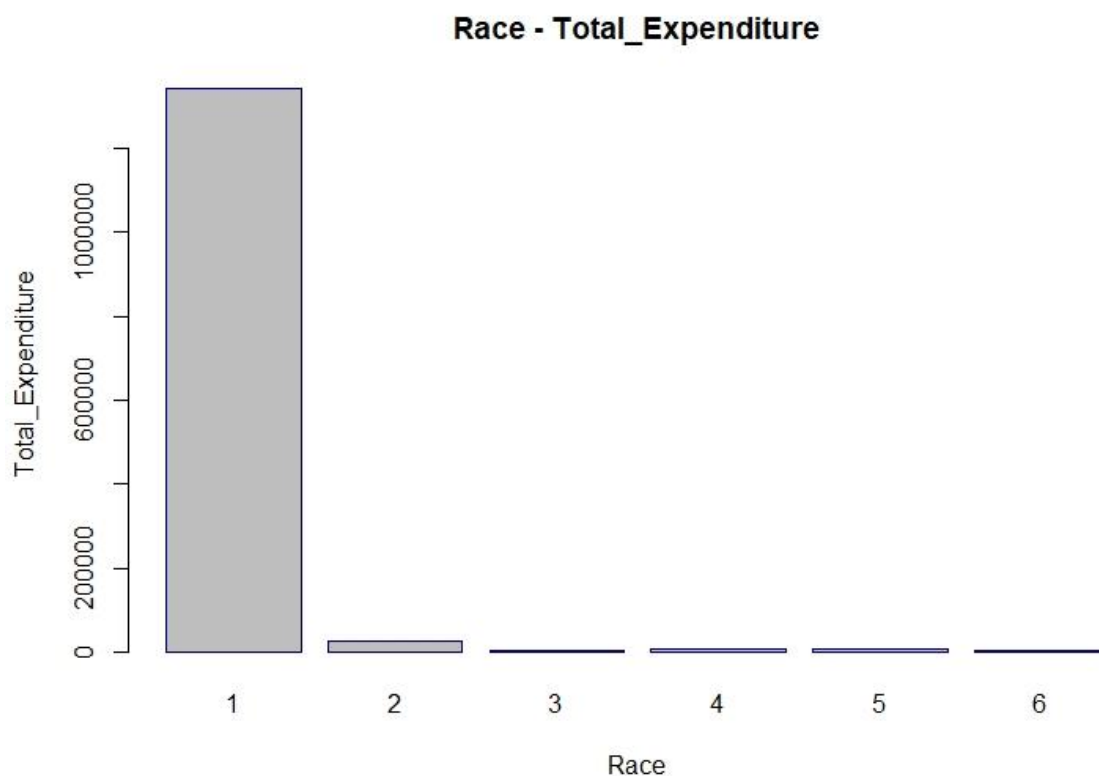
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mydf\$RACE	1	2.488e+06	2488459	0.164	0.686
Residuals	497	7.540e+09	15170268		

```

> |

```



From the summary of anova test , p-value > 0.05 so we reject the fact that there is relation between race and hospital costs i.e null hypothesis is rejected.

4.To properly utilize the costs, the agency has to analyse the severity of the hospital costs by age and gender for the proper allocation of resources.

question-4

```
group_age_gender <- aggregate(mydf$TOTCHG,
                              by = list(mydf$AGE, mydf$GENDER),
                              FUN = sum)

names(group_age_gender)[1] <- 'AGE'
names(group_age_gender)[2] <- 'GENDER'
names(group_age_gender)[3] <- 'Total_Expenditure'

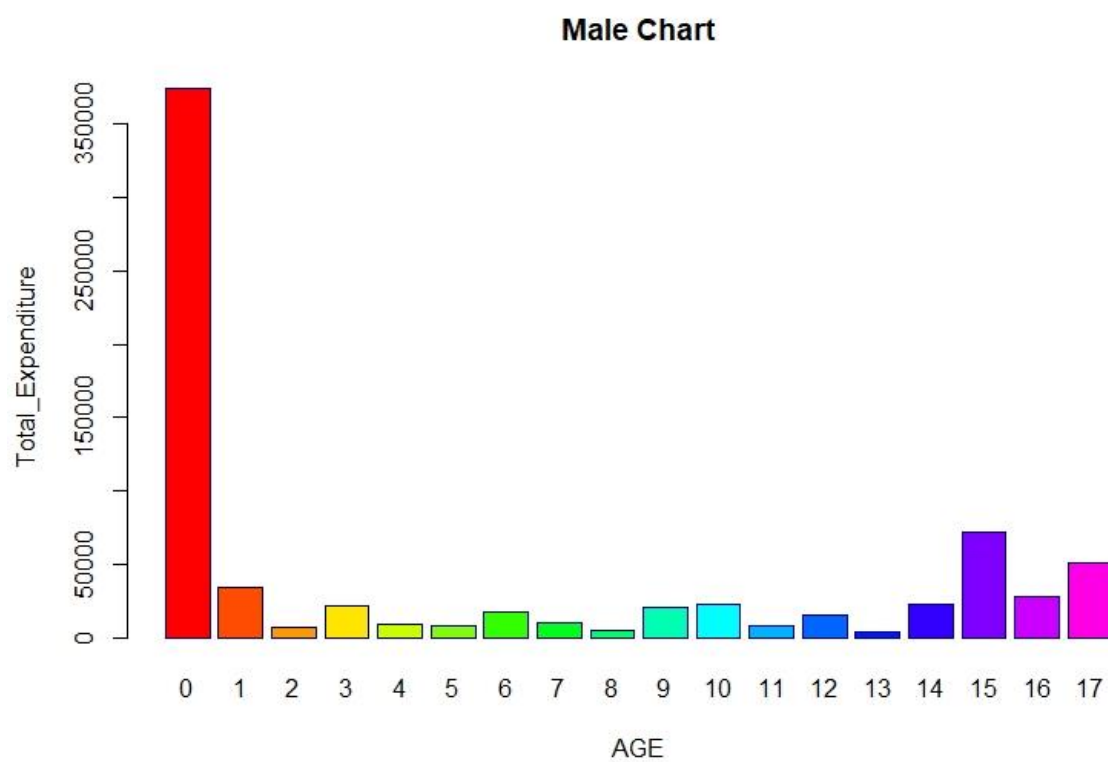
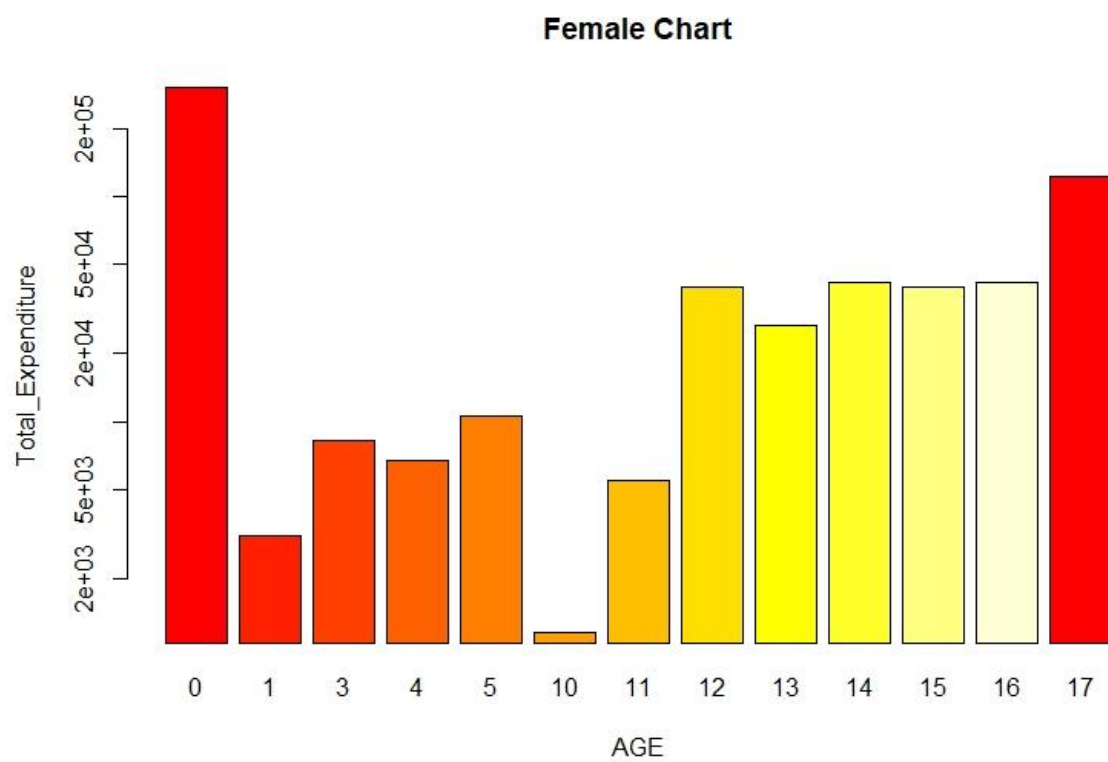
group_age_gender

female <- group_age_gender[group_age_gender$GENDER == 'Female', ]
female

barplot(female$Total_Expenditure,
        col = heat.colors(12),
        log = "y",
        main = 'Female Chart',
        names.arg = female$AGE)

male <- group_age_gender[group_age_gender$GENDER == 'Male',]
male

barplot(male$Total_Expenditure,
        border = 'dark blue',
        main = 'Male Chart',
        col = rainbow(20),
        names.arg = male$AGE)
```

From the summary of Multiple Regression below, we can observe the p-value(level of significance) of age as well as gender which is very low. Hence we can say that age and gender has impact on hospital costs.

5 Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race.

```
# question - 5
```

```
finaldf <- mydf[,c(1:4)] # selecting features required to perform multiple regression
```

```
View(finaldf)
```

```
categcol <- finaldf[,c(1,2,4)] # selecting categorical features to create dummy variables
```

```
fact<- data.frame(apply(categcol, function(x) factor(x)))
```

```
str(fact)
```

```
names(fact)
```

```
# creating dummy variables for factor attributes
```

```
dummies<- data.frame(apply(fact,
                           function(x) data.frame(model.matrix(~x,data = fact))[, -1]))
```

```
View(dummies)
```

```
names(dummies)
```

```
finaldf <- cbind(finaldf[,3], dummies)
```

```
View(finaldf)
```

```
regressor <- lm(LOS~.,data = finaldf)
```

```
summary(regressor)
```

```

> # question - 5
> finaldf <- mydf[,c(1:4)] # selecting features required to perform multiple regression
> View(finaldf)
> catecol <- finaldf[,c(1,2,4)] # selecting categorical features to create dummy variables
>
> fact<- data.frame(sapply(catecol, function(x) factor(x)))
> str(fact)
'data.frame': 499 obs. of 3 variables:
 $ AGE : chr "17" "17" "17" "17" ...
 $ FEMALE: chr "1" "0" "1" "1" ...
 $ RACE : chr "1" "1" "1" "1" ...
> names(fact)
[1] "AGE" "FEMALE" "RACE"
> # creating dummy variables for factor attributes
> dummies<- data.frame(sapply(fact,
+ function(x) data.frame(model.matrix(x(~x,data = fact))[, -1])))
> View(dummies)
> names(dummies)
 [1] "AGE.x1" "AGE.x10" "AGE.x11" "AGE.x12" "AGE.x13"
 [6] "AGE.x14" "AGE.x15" "AGE.x16" "AGE.x17" "AGE.x2"
[11] "AGE.x3" "AGE.x4" "AGE.x5" "AGE.x6" "AGE.x7"
[16] "AGE.x8" "AGE.x9" "FEMALE" "RACE.x2" "RACE.x3"
[21] "RACE.x4" "RACE.x5" "RACE.x6"
>

```

```

~/R prac/
> regressor <- lm(LOS~.,data = finaldf)
> summary(regressor)

Call:
lm(formula = LOS ~ ., data = finaldf)

Residuals:
    Min       1Q   Median       3Q      Max
-3.262 -1.224 -0.892  0.045 37.776

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.95535     0.24457  12.084 <2e-16 ***
AGE.x1       -1.20910     1.09842  -1.101  0.2716
AGE.x10      -0.27254     1.71648  -0.159  0.8739
AGE.x11      -1.65823     1.23557  -1.342  0.1802
AGE.x12      -0.71661     0.90295  -0.794  0.4278
AGE.x13      -0.86106     0.84041  -1.025  0.3061
AGE.x14      -0.16271     0.72444  -0.225  0.8224
AGE.x15       0.03803     0.66785   0.057  0.9546
AGE.x16      -1.33221     0.68452  -1.946  0.0522 .
AGE.x17      -0.50059     0.59066  -0.848  0.3971
AGE.x2       -0.95535     3.41674  -0.280  0.7799
AGE.x3       0.28840     1.97773   0.146  0.8841
AGE.x4      -1.08973     2.41786  -0.451  0.6524
AGE.x5      -0.58973     2.41786  -0.244  0.8074
AGE.x6      -0.45535     2.42218  -0.188  0.8510
AGE.x7      -2.62201     1.98274  -1.322  0.1867
AGE.x8      -1.49810     2.53185  -0.592  0.5543
AGE.x9      -0.95535     2.42218  -0.394  0.6935
FEMALE       0.26877     0.32509   0.827  0.4088
RACE.x2      0.08552     1.49616   0.057  0.9544
RACE.x3      0.77589     3.41835   0.227  0.8205
RACE.x4      0.54007     2.00086   0.270  0.7873
RACE.x5     -0.95535     1.98274  -0.482  0.6301
RACE.x6     -0.42362     2.43389  -0.174  0.8619
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.408 on 475 degrees of freedom
Multiple R-squared:  0.02263,    Adjusted R-squared:  -0.0247
F-statistic: 0.4781 on 23 and 475 DF,  p-value: 0.982

```

Since the p-value(level of significance) is very high so we can say that there is no relation between length of stay and age, gender, race. We can't predict length of stay with only these features.

6.To perform a complete analysis, the agency wants to find the variable that mainly affects hospital costs.

```
# question-6
imp <- varImp(regressor)
barplot(imp$Overall, names.arg = row.names(imp))
cor <- cor(mydf[,1:6])
corrplot(cor,
  method = "color",
  outline = T,
  cl.pos = 'n',
  rect.col = "black",
  tl.col = "indianred4",
  addCoef.col = "black",
  number.digits = 2,
  number.cex = 0.60,
  tl.cex = 0.7,
  cl.cex = 1,
  col = colorRampPalette(c("green4","white","red"))(100))
```

From the below figures of Bar plot of VarImp and Heatmap of dataset, we can say that gender and age(specifically 0 and 17) are of more important features.

	AGE	FEMALE	LOS	RACE	TOTCHK	APDRG
AGE	1	0.24	-0.07	0.02	0.13	0.15
FEMALE	0.24	1	0.04	-0.04	-0.06	0.25
LOS	-0.07	0.04	1	-0.02	0.62	0.01
RACE	0.02	-0.04	-0.02	1	-0.02	-0.04
TOTCHK	0.13	-0.06	0.62	-0.02	1	-0.33
APDRG	0.15	0.25	0.01	-0.04	-0.33	1

