Flatiron Health Assessment

Karthik Mahanth Kattula

10/26/2020

#Load two datasets Diagnosis and Treatment as dataframes in R

```
diagnosis<-read.csv("DiagnosisSample01.csv")
treatment<-read.csv("TreatmentSample01.csv")
```

```
str(diagnosis)
```

```
## 'data.frame':    44 obs. of  5 variables:
##  $ PatientID        : int  2634 5657 7937 8615 4354 6922 7230 2038 2120 2407 ...
##  $ DiagnosisDate    : chr  "2/19/2011" "6/7/2012" "1/6/2013" "7/18/2013" ...
##  $ DiagnosisCode    : num  286 286 286 285 285 ...
##  $ Diagnosis        : chr  "Anemia" "Anemia" "Anemia" "Anemia" ...
##  $ IsCancerDiagnosis: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
str(treatment)
```

```
## 'data.frame':    714 obs. of  3 variables:
##  $ PatientID    : int  2038 2038 2038 2038 2038 2038 2038 2038 2120 2120 ...
##  $ TreatmentDate: chr  "2010-01-24" "2010-01-27" "2010-01-30" "2010-02-02" ...
##  $ DrugCode     : chr  "A" "A" "A" "A" ...
```

#We can see that diagnosis dataset consists of observations related to patients and their diagnosis whereas treatment data consists of observations related to treatment for each type of drug and the treatment dates for different patients

```
diag.fcols<-c("PatientID","DiagnosisCode","Diagnosis")
diagnosis[diag.fcols]<-lapply(diagnosis[diag.fcols],factor)
diagnosis$DiagnosisDate<-as.Date(diagnosis$DiagnosisDate,"%m/%d/%Y")
str(diagnosis)
```

```
## 'data.frame':    44 obs. of  5 variables:
##  $ PatientID        : Factor w/ 32 levels "2038","2120",..: 6 18 28 30 14 24 25 1 2 3 ...
##  $ DiagnosisDate    : Date, format: "2011-02-19" "2012-06-07" ...
##  $ DiagnosisCode    : Factor w/ 20 levels "153.3","153.4",..: 17 17 17 16 16 18 17 15 7 15 ...
##  $ Diagnosis        : Factor w/ 4 levels "Anemia","Breast Cancer",..: 1 1 1 1 1 1 1 2 2 2 ...
##  $ IsCancerDiagnosis: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

#From the above, we can conclude that there are 32 patients in the clinic taking diagnosis and the clinic is diagnosing for four different diseases

```
treat.fcols<-c("PatientID","DrugCode")
treatment[treat.fcols]<-lapply(treatment[treat.fcols],factor)
treatment$TreatmentDate<-as.Date(treatment$TreatmentDate,"%Y-%m-%d")
str(treatment)
```

```
## 'data.frame':    714 obs. of  3 variables:
##  $ PatientID    : Factor w/ 27 levels "2038","2120",..: 1 1 1 1 1 1 1 1 2 2 ...
##  $ TreatmentDate: Date, format: "2010-01-24" "2010-01-27" ...
##  $ DrugCode     : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 2 ...
```

#From the above, we can conclude that there are 27 patients who are taking treatment from atleast one type of drug A,B or C

#Question 1 First, the clinic would like to know for which diseases they are seeing patients for ?

```
unique(diagnosis$Diagnosis)
```

```
## [1] Anemia        Breast Cancer Colon Cancer  Hypertension
## Levels: Anemia Breast Cancer Colon Cancer Hypertension
```

#Overall, the clinic is seeing patients for four different types of diseases namely Anemia, Breast Cancer, Colon Cancer and Hypertension

#Question 1a Which types of cancer does the clinic see patients for ? First subset dataframe based on condition isCancerdiagnosis equals to True and then find different types of cancer in the resulting dataframe

```
cancer.diagnosis <- diagnosis[diagnosis$IsCancerDiagnosis==TRUE,]
unique(cancer.diagnosis$Diagnosis)
```

```
## [1] Breast Cancer Colon Cancer
## Levels: Anemia Breast Cancer Colon Cancer Hypertension
```

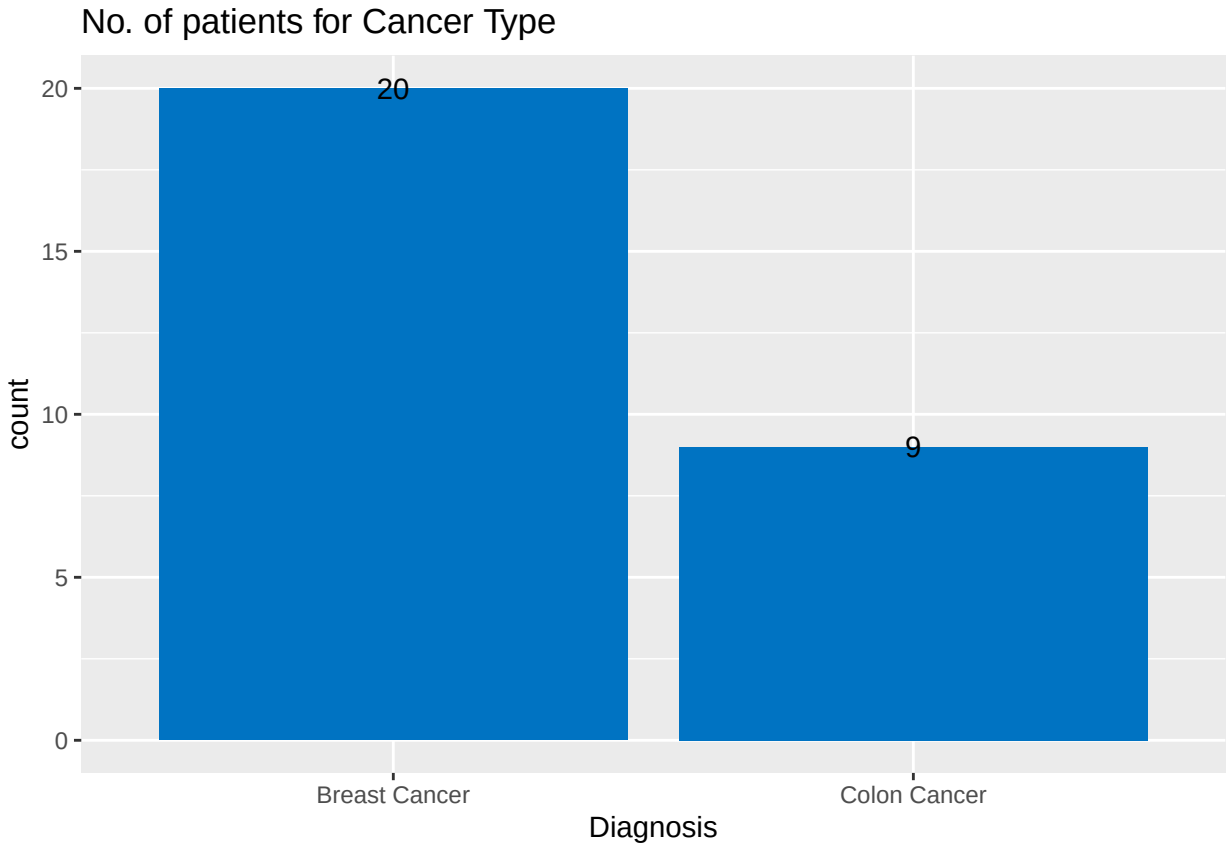
#We can conclude that from the sample given that clinic is seeing patients for two different types of cancer namely Breast Cancer and Colon Cancer

#Question 1b How many patients does the clinic see for each cancer type? In the resulting dataframe which is obtained by subsetting the iscancerdiagnosis=True group by diagnosis variable using dplyr package and find the count of unique patient id's for each type of cancer

```
df<-data.frame(cancer.diagnosis %>% group_by(Diagnosis) %>%
                 summarise(count=n_distinct(PatientID),.groups='drop'))
df
```

```
##       Diagnosis count
## 1 Breast Cancer    20
## 2  Colon Cancer     9
```

```
ggplot(df,aes(Diagnosis,count)) +geom_col(fill = "#0073C2FF") + ggtitle("No. of patients for Cancer Type
```



#We can conclude that clinic is seeing 20 patients suffering from Breast cancer type and 9 patients suffering from Colon Cancer Type

#Question 2 The clinic wants to know how long it takes for patients to start therapy after being diagnosed, which they consider to be helpful in understanding the quality of care for patients

#Question 2a How long after being diagnosed do cancer patients start treatment for each cancer type? For each of the cancer types Breast Cancer and Colon Cancer find the patients most recent diagnosed dates and earliest start date in treatment dataset for the same patients. Later find the difference between these two dates and then take the average of difference of the dates

```
cancer.diagnosis.df<-cancer.diagnosis[,c("PatientID","DiagnosisDate","Diagnosis")]
```

```
duplicates<- cancer.diagnosis.df[(cancer.diagnosis.df$Diagnosis %in% cancer.diagnosis.df$Diagnosis[dupli
```

```
noduplicates<- cancer.diagnosis.df[!(cancer.diagnosis.df$Diagnosis %in% cancer.diagnosis.df$Diagnosis[du
```

#I seperated cancer diagnosis dataset into duplicates one and non duplicates one. My idea is to extract the most recent diagnosed date for patient id with duplicates and then r bind with no duplicates dataframe. So, then we will be having patient id's and their recent diagnosis date or last diagnosis date for different types of cancer namely Breast Cancer and Colon Cancer

```
duplicates
```

```
##    PatientID DiagnosisDate     Diagnosis
## 14      3757    2011-10-11 Breast Cancer
```

```
## 18      4374    2012-03-20 Breast Cancer
## 26      4374    2012-03-20 Breast Cancer
## 27      4374    2012-03-20 Breast Cancer
## 28      6877    2012-12-09 Breast Cancer
## 31      3095    2011-07-01  Colon Cancer
## 32      3095    2011-07-10  Colon Cancer
## 33      3449    2011-08-26  Colon Cancer
## 35      6877    2012-11-16  Colon Cancer
## 39      3449    2011-08-26  Colon Cancer
## 40      3757    2011-10-08  Colon Cancer
```

```
df4<-data.frame(duplicates%>%group_by(PatientID,Diagnosis)%>%summarise(date=max(DiagnosisDate),.groups=
df4
```

```
##   PatientID     Diagnosis       date
## 1      3095  Colon Cancer 2011-07-10
## 2      3449  Colon Cancer 2011-08-26
## 3      3757 Breast Cancer 2011-10-11
## 4      3757  Colon Cancer 2011-10-08
## 5      4374 Breast Cancer 2012-03-20
## 6      6877 Breast Cancer 2012-12-09
## 7      6877  Colon Cancer 2012-11-16
```

#R bind the above dataframe df4 with no duplicates dataframe. So, the resulting dataframe will be containing unique patientid and diagnosis and the most recent diagnosed date for the patient

```
df5<-data.frame(noduplicates%>% group_by(PatientID,Diagnosis)%>%summarise(date=max(DiagnosisDate),.grou
df5
```

```
##    PatientID     Diagnosis       date
## 1       2038 Breast Cancer 2010-01-21
## 2       2120 Breast Cancer 2010-01-09
## 3       2407 Breast Cancer 2010-06-13
## 4       2425 Breast Cancer 2010-12-15
## 5       2462 Breast Cancer 2011-01-07
## 6       2763 Breast Cancer 2011-04-19
## 7       2770  Colon Cancer 2011-04-06
## 8       3948 Breast Cancer 2011-12-18
## 9       4256 Breast Cancer 2011-11-07
## 10      4354 Breast Cancer 2012-02-04
## 11      4692 Breast Cancer 2012-04-27
## 12      5259 Breast Cancer 2012-05-13
## 13      6281 Breast Cancer 2012-08-12
## 14      6321 Breast Cancer 2012-09-06
## 15      6837  Colon Cancer 2012-10-08
## 16      6889 Breast Cancer 2012-11-17
## 17      6922  Colon Cancer 2012-11-07
## 18      7230  Colon Cancer 2013-01-02
## 19      7242  Colon Cancer 2013-01-11
## 20      7796 Breast Cancer 2013-01-16
## 21      7976 Breast Cancer 2013-03-06
## 22      9331 Breast Cancer 2013-08-23
```

```
res.df.nodup<-data.frame(rbind(df5,df4))
res.df.nodup
```

```
##    PatientID     Diagnosis       date
## 1       2038 Breast Cancer 2010-01-21
## 2       2120 Breast Cancer 2010-01-09
## 3       2407 Breast Cancer 2010-06-13
## 4       2425 Breast Cancer 2010-12-15
## 5       2462 Breast Cancer 2011-01-07
## 6       2763 Breast Cancer 2011-04-19
## 7       2770  Colon Cancer 2011-04-06
## 8       3948 Breast Cancer 2011-12-18
## 9       4256 Breast Cancer 2011-11-07
## 10      4354 Breast Cancer 2012-02-04
## 11      4692 Breast Cancer 2012-04-27
## 12      5259 Breast Cancer 2012-05-13
## 13      6281 Breast Cancer 2012-08-12
## 14      6321 Breast Cancer 2012-09-06
## 15      6837  Colon Cancer 2012-10-08
## 16      6889 Breast Cancer 2012-11-17
## 17      6922  Colon Cancer 2012-11-07
## 18      7230  Colon Cancer 2013-01-02
## 19      7242  Colon Cancer 2013-01-11
## 20      7796 Breast Cancer 2013-01-16
## 21      7976 Breast Cancer 2013-03-06
## 22      9331 Breast Cancer 2013-08-23
## 23      3095  Colon Cancer 2011-07-10
## 24      3449  Colon Cancer 2011-08-26
## 25      3757 Breast Cancer 2011-10-11
## 26      3757  Colon Cancer 2011-10-08
## 27      4374 Breast Cancer 2012-03-20
## 28      6877 Breast Cancer 2012-12-09
## 29      6877  Colon Cancer 2012-11-16
```

#Now from treatments dataset find unique patient id's and their starting dates of the therapy

```
treatment.q2<-treatment[,c("PatientID","TreatmentDate")]
df6<-treatment.q2%>%group_by(PatientID)%>%summarise(date=min(TreatmentDate),.groups='drop')
df6
```

```
## # A tibble: 27 x 2
##    PatientID date
##    <fct>     <date>
##  1 2038      2010-01-24
##  2 2120      2010-01-25
##  3 2407      2010-06-19
##  4 2425      2010-12-19
##  5 2462      2011-01-11
##  6 2763      2011-04-23
##  7 2770      2011-04-22
##  8 3095      2011-07-13
##  9 3449      2011-09-13
```

```
## 10 3757      2011-10-22
## # ... with 17 more rows
```

```
names(df6)<-c("PatientID","t.treatmentdate")
```

#Perform Left join operation on the above two dataframes df6 and res.df.nodup and then create a new variable and calculate the date difference between starting treatment date and most recent diagnosis date

```
res.df<- merge(x = res.df.nodup, y = df6, by = "PatientID", all.x = TRUE)
```

#Creating a new column and calculating the date difference in days between two columns most recent diagnosis date and earliest starting treatment date

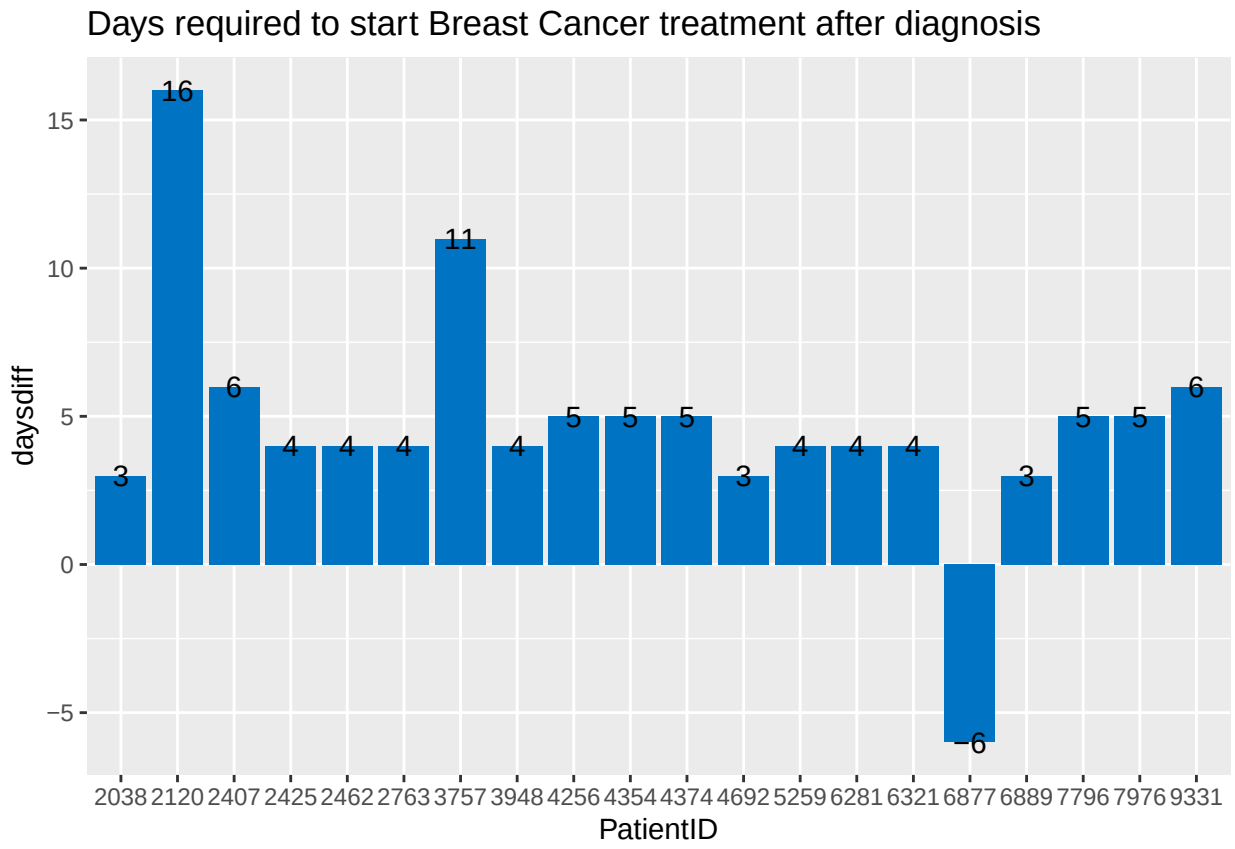
```
res.df$daysdiff <- res.df$t.treatmentdate - res.df$date
```

#Subset breast cancer days diff into another df and plotting the distribution to know about the distribution and to know the patients behavior, similarly for colon type of cancer doing the same

```
breast.cancer.df <- res.df[res.df$Diagnosis=='Breast Cancer',]
colon.cancer.df <- res.df[res.df$Diagnosis=='Colon Cancer',]

ggplot(breast.cancer.df,aes(PatientID,daysdiff)) + geom_col(fill = "#0073C2FF") + ggtitle("Days require
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```



#From the above distribution we can see that the patient with id 6877 has started the treatment even before diagnosis and and patients with id 2120 and 3757 is taking bit longer to start their treatment when compared with other patients. In order to know reason why, we need to further investigate if there are any specific reasons for this type of delay.On an average we can say that breast cancer patients are taking approximately 3-5 days after their diagnosis. For calculation purposes exclude PatientID 2120,3757 and 6877 and then calculate average of daysdiff which will give approximately the days required to start treatment after diagnosis for breast cancer

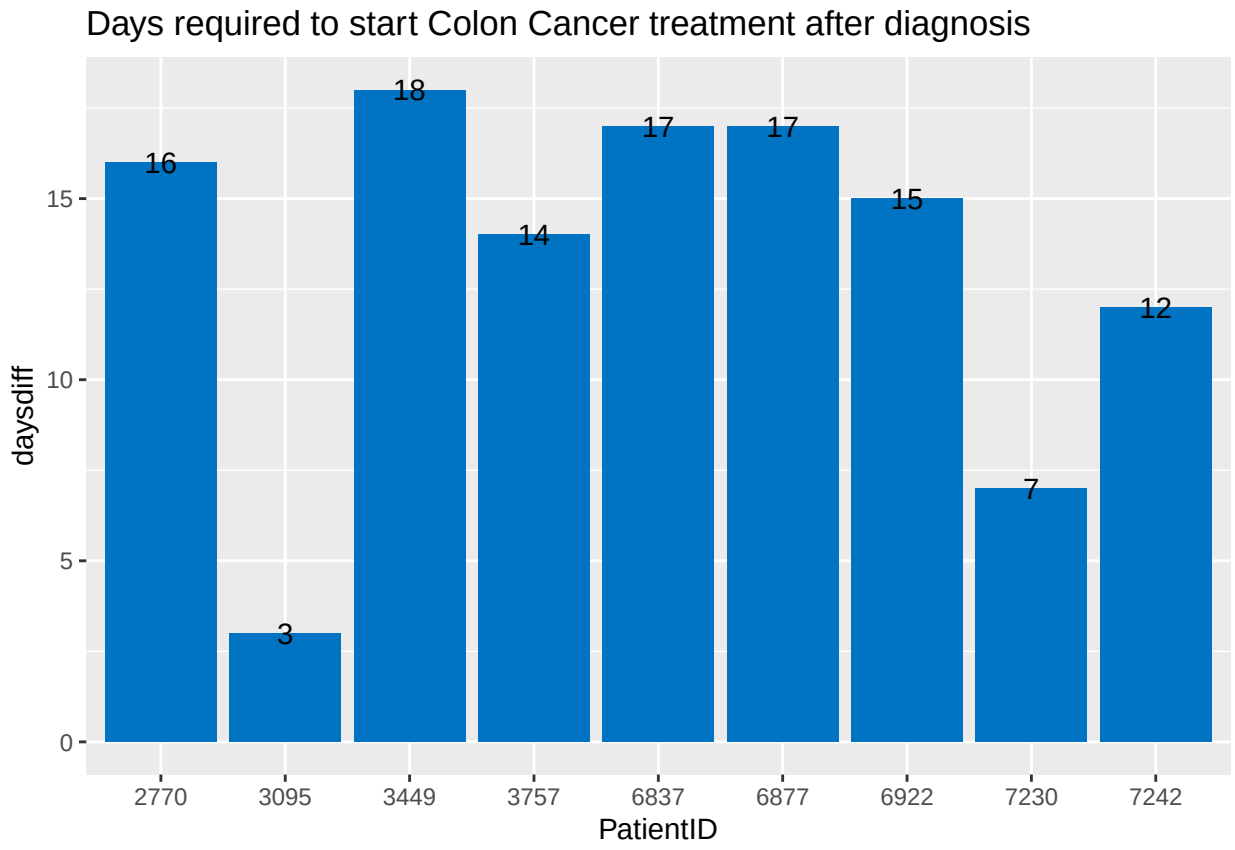
```
breast.cancer.df <- breast.cancer.df[ !(breast.cancer.df$PatientID %in% c(2120,3757,6877)), ]
mean(breast.cancer.df$daysdiff)
```

```
## Time difference of 4.352941 days
```

#On an average approximately its taking 4 days to start the treatment for breast cancer after diagnosis

```
ggplot(colon.cancer.df,aes(PatientID,daysdiff)) + geom_col(fill = "#0073C2FF") + ggtitle("Days required
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```



#From the above plot, we can see that only patients with ID's 3095 and 7230 are starting their treatments early whereas other patients are taking on an average approximately 13-15 days to start their treatment after diagnosis. This suggests that patients are taking some time to start the treatment if they are diagnosed with Colon Cancer whereas patients with Breast Cancer are starting their treatments within 3-4 days after their diagnosis

```
mean(colon.cancer.df$daysdiff)
```

```
## Time difference of 13.22222 days
```

#Lets examine the Patient ID 6837

```
res.df.6877 <- res.df[res.df$PatientID %in% ("6877"),]
res.df.6877
```

```
##    PatientID     Diagnosis       date t.treatmentdate daysdiff
## 21      6877 Breast Cancer 2012-12-09      2012-12-03  -6 days
## 22      6877  Colon Cancer 2012-11-16      2012-12-03  17 days
```

#From the above table we can see that the patient is first diagnosed with Colon Cancer on 2012-11-16 and he started treatment from 2012-12-03 whereas when he diagnosed again on 2012-12-09 the diagnosis is of Breast Cancer type. This is why we see negative value of days diff

#Question2 b Are there any patients which are diagnosed but not treated at practice ? To check patients who are diagnosed but are not treated at practice, first find patients in diagnosis dataset and then find unique patients in treatment dataset. Later, perform set difference operation which will give patients who are diagnosed but not treated at practice

```
df1<-unique(diagnosis$PatientID)
df2<-unique(treatment$PatientID)
df3<-setdiff(df1,df2)
df3
```

```
## [1] "2634" "5657" "7937" "8615" "8827"
```

#There are 5 patients who are diagnosed but not yet treated at practice and the respective Patient ID's for them is 2634,5657,7937,8615 and 8827

#Question 3 After being treated with first line of treatment(a drug or combination of drugs), what proportion of all cancer patients go on to be treated with a second line of treatment Based on my understanding, first line of treatment means a patients is taking therapy from any one drug or combination of drugs. So my approach is to find the number of patients who are taking treatment from one drug for few days and then taking another drug. When we examine the dataset we can see there are few patients who are taking multiple drugs on same day. It also comes under first line of treatment. We need to know the patients who took either one drug or combination of drugs on same day and then after few days they are taking again new drugs. For example one patient is taking drug A or B or C in his entire treatment. Then we can exclude him whereas on other hand if a patient is taking treatment with drug A for few days and then same patient started taking drug B (or) when one patient is taking Drug A and B on same day for few days and then after few days if he starts taking Drug C then all these scenarios will come under second line of treatment

```
df8<-treatment%>%group_by(PatientID)%>%count(drg=DrugCode)
df8<-df8%>%group_by(PatientID)%>%summarise(d1=min(n),d2=max(n),.groups='drop')
df8$diff = df8$d2 - df8$d1
df8<-df8[df8$diff>0,]
df8
```

```
## # A tibble: 7 x 4
##   PatientID    d1    d2  diff
```

```
##   <fct>     <int> <int> <int>
## 1 3948          9    16     7
## 2 4692         15    20     5
## 3 5259         16    21     5
## 4 6281         18    20     2
## 5 6321         21    24     3
## 6 6837         16    17     1
## 7 7242         20    41    21
```

#There are 7 patients who are taking second line of treatment of the total 27 Patients. So the proprotion of patients who go on to be treated with second line of treatment is $7/27 = 0.25925$. In the clinic we can say that approx 26% of patients are undergoing second line of treatment from all the patients in the given sample

#Question 4 How does each drug used at the clinic compare in terms of its duration of therapy? Duration of therapy is different for different patients, One way is for each drug find the min and max dates of drug types A, B and C. But, this will not give accurate one, instead my approach is to find the min and max dates for each patient and drug combination which will give the duration of therapy for a specific patient belonging to different drug types A, B and C and then calculate average of the difference between dates which in general gives the effectiveness of drug and then compare the averages for different drug types

```
drug.effect <- treatment%>%group_by(PatientID,DrugCode)%>%summarise(start_date=min(TreatmentDate),end_da
drug.effect
```

```
## # A tibble: 44 x 4
##    PatientID DrugCode start_date end_date
##    <fct>     <fct>    <date>     <date>
##  1 2038      A        2010-01-24 2010-02-20
##  2 2120      A        2010-01-25 2010-03-02
##  3 2120      B        2010-01-25 2010-03-02
##  4 2407      A        2010-06-19 2010-08-03
##  5 2407      B        2010-06-19 2010-08-03
##  6 2425      A        2010-12-19 2011-02-08
##  7 2425      B        2010-12-19 2011-02-08
##  8 2462      A        2011-01-11 2011-03-04
##  9 2462      B        2011-01-11 2011-03-04
## 10 2763      A        2011-04-23 2011-06-22
## # ... with 34 more rows
```

#From the above table, we see that few patients are taking two drugs and three drugs. Lets examine the Patient ID's 2120,2407,5259 and 6321

```
df7<-drug.effect[drug.effect$PatientID %in% c("2120","2407","5259","6321"),]
df7
```

```
## # A tibble: 10 x 4
##    PatientID DrugCode start_date end_date
##    <fct>     <fct>    <date>     <date>
##  1 2120      A        2010-01-25 2010-03-02
##  2 2120      B        2010-01-25 2010-03-02
##  3 2407      A        2010-06-19 2010-08-03
##  4 2407      B        2010-06-19 2010-08-03
##  5 5259      A        2012-05-17 2012-07-19
```

```
##  6 5259      B         2012-05-17 2012-07-19
##  7 5259      C         2012-07-24 2012-10-30
##  8 6321      A         2012-09-10 2012-11-19
##  9 6321      B         2012-09-10 2012-11-19
## 10 6321      C         2012-11-24 2013-03-08
```

#From the above table, we see that Patient ID's 2120 and 2407 are taking therapy from two different drugs A and B from starting to end whereas Patient ID's 5259 and 6321 started their therapy with two different drugs A and B and after few days then they are using Drug C. To be precise when we examine this sample we see that it takes 5 days to start the therapy from Drug Type C after having done therapy with drugs A and B

```
drug.effect$daysdiff <- drug.effect$end_date - drug.effect$start_date
drug.effect<-drug.effect%>% group_by(DrugCode)%>%summarise(average=mean(daysdiff),.groups='drop')
drug.effect
```

```
## # A tibble: 3 x 2
##   DrugCode average
##   <fct>    <drtn>
## 1 A         56.13333 days
## 2 B         63.46154 days
## 3 C        112.06250 days
```

#From the above summary we conclude that Duration of therapy with Drug Code C requires more number of days followed by Drug Code B and A