BoneMarrowEDA

Patience Heath

2024-03-13

Information about the Dataset

The motivation of the study was to identify the most important factors influencing the success or failure of the transplantation procedure. In particular, the aim was to verify the hypothesis that increased dosage of CD34+ cells / kg extends overall survival time without simultaneous occurrence of undesirable events affecting patients' quality of life

Doing some quick analysis before I dive into the dataset more

```
new_data = complete_data
```

```
library(dlookr)
diagnose_paged_report(new_data)
```

```
##
##
## processing file: diagnosis_paged_temp.Rmd
```

```
##   |
```

We can also see from this that some variables can be removed (ex Recipient Age int, Recipientage10) as the Recipient age variable gives us enough information.

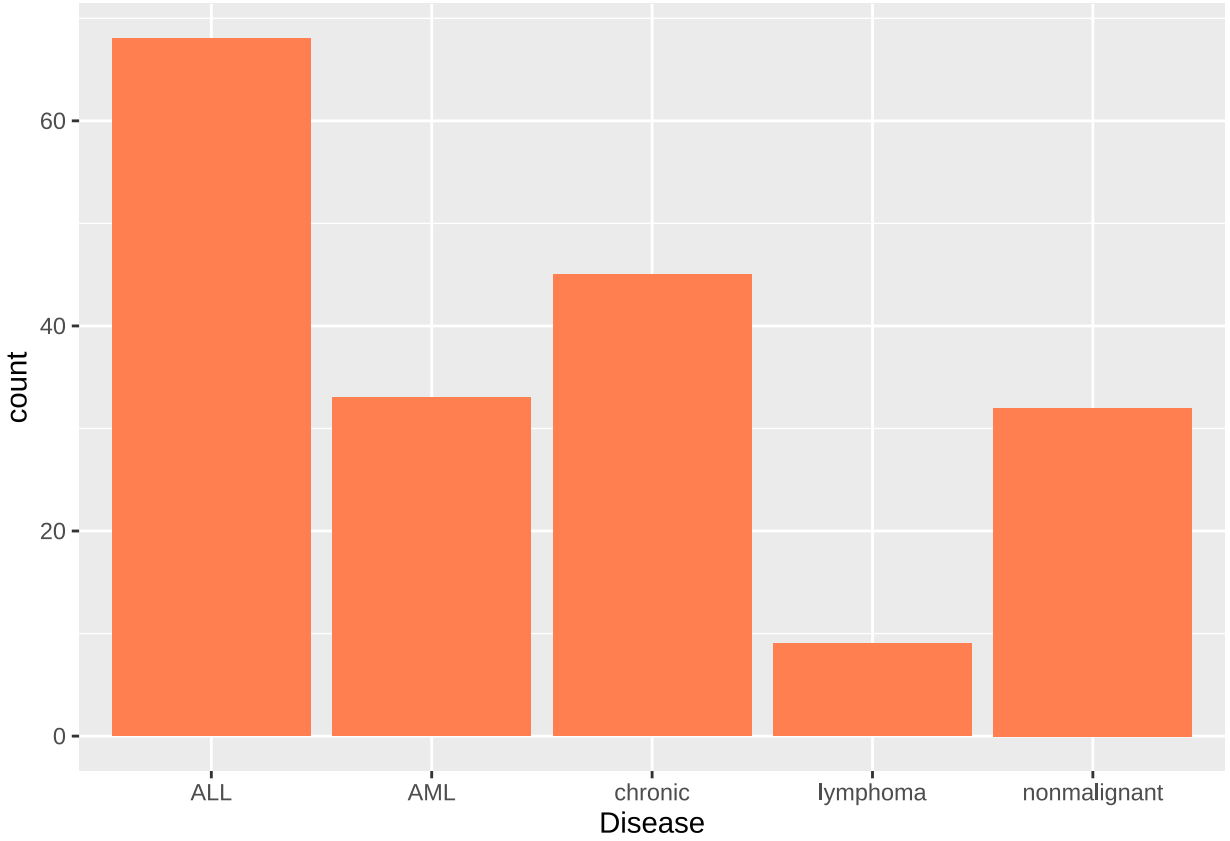
```
summary(new_data)
```

```
##  Recipientgender Stemcellsource    Donorage     Donorage35 IIIV    Gendermatch DonorABO RecipientABO
##  0: 75           0: 42          Min.   :18.65   0:104      0: 75   0:155       -1:28    -1:50
##  1:112           1:145          1st Qu.:27.04   1: 83      1:112   1: 32       0 :73    0 :48
##                                 Median :33.55                                  1 :71    1 :76
##                                 Mean   :33.47                                  2 :15    2 :13
##                                 3rd Qu.:40.12
##                                 Max.   :55.55
##
##  DonorCMV RecipientCMV         Disease   Riskgroup Txpostrelapse Diseasegroup HLAmatch HLAmismatch Ar
##  0:115    0: 82        ALL         :68   0:118     0:164         0: 32        0:94     0:159       -1
##  1: 72    1:105        AML         :33   1: 69     1: 23         1:155        1:65     1: 28       0
##                        chronic     :45                                       2:23                 1
##                        lymphoma    : 9                                       3: 5                 2
```

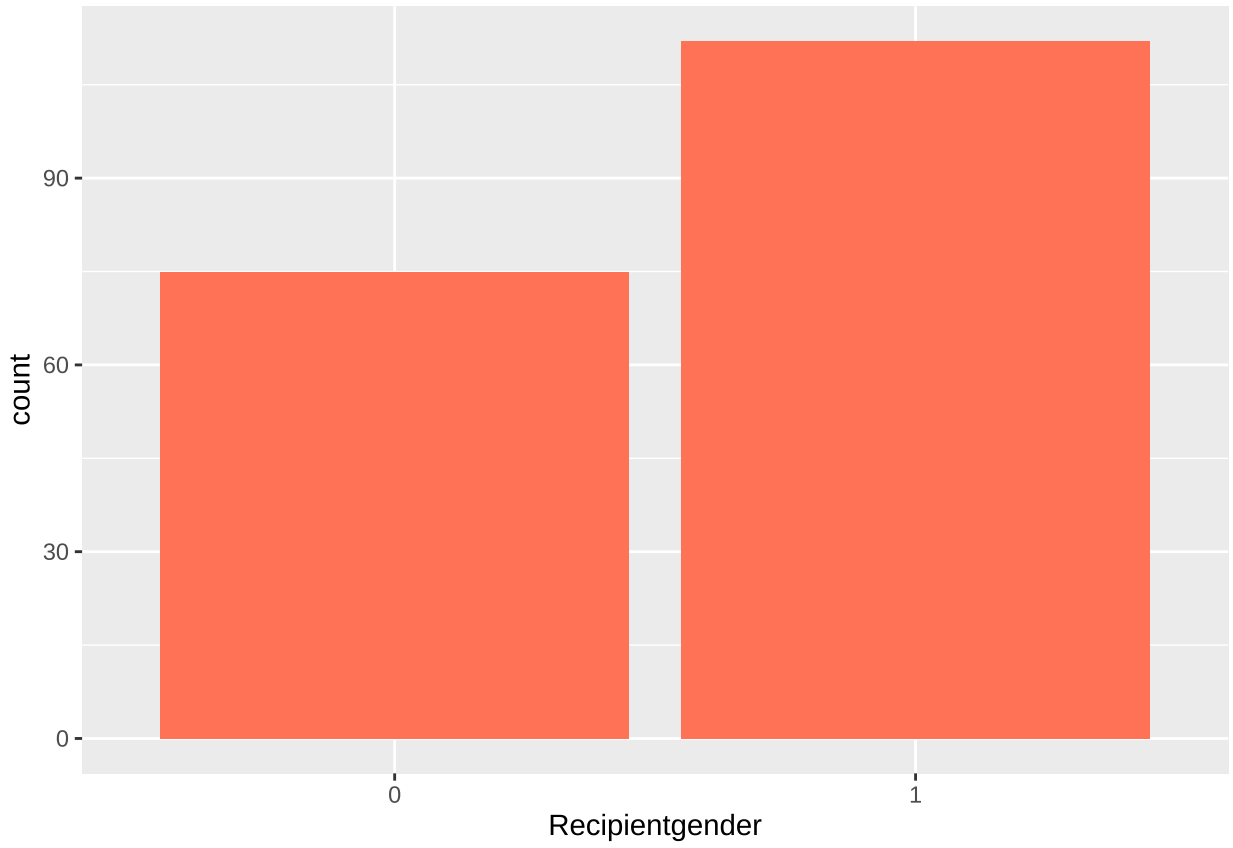
```
##                       nonmalignant:32
##
##
##   Recipientage    Recipientage10 Recipientageint Relapse aGvHDIIIIV extcGvHD  CD34kgx10d6       CD3d
##  Min.   : 0.600   0:99           0:47            0:159   0: 40      0: 34    Min.   : 0.79   Min.
##  1st Qu.: 5.050   1:88           1:51            1: 28   1:147      1:153    1st Qu.: 5.35   1st Qu.
##  Median : 9.600                  2:89                                        Median : 9.72   Median
##  Mean   : 9.932                                                              Mean   :11.89   Mean
##  3rd Qu.:14.050                                                              3rd Qu.:15.41   3rd Qu.
##  Max.   :20.200                                                              Max.   :57.78   Max.
##
##    Rbodymass      ANCrecovery     PLTrecovery     survival_time    survival_status
##  Min.   : 6.00   Min.   : 9.00   Min.   :  9.00   Min.   :   6.0   Min.   :0.0000
##  1st Qu.:19.00   1st Qu.:13.00   1st Qu.: 15.50   1st Qu.: 168.5   1st Qu.:0.0000
##  Median :33.00   Median :15.00   Median : 21.00   Median : 676.0   Median :0.0000
##  Mean   :35.78   Mean   :15.32   Mean   : 31.23   Mean   : 938.7   Mean   :0.4545
##  3rd Qu.:50.70   3rd Qu.:17.00   3rd Qu.: 29.50   3rd Qu.:1604.0   3rd Qu.:1.0000
##  Max.   :97.80   Max.   :26.00   Max.   :285.00   Max.   :3364.0   Max.   :1.0000
##
```

data visualization

```
ggplot(data = new_data) +
  geom_bar(mapping = aes(x = Disease), fill = 'coral')
```

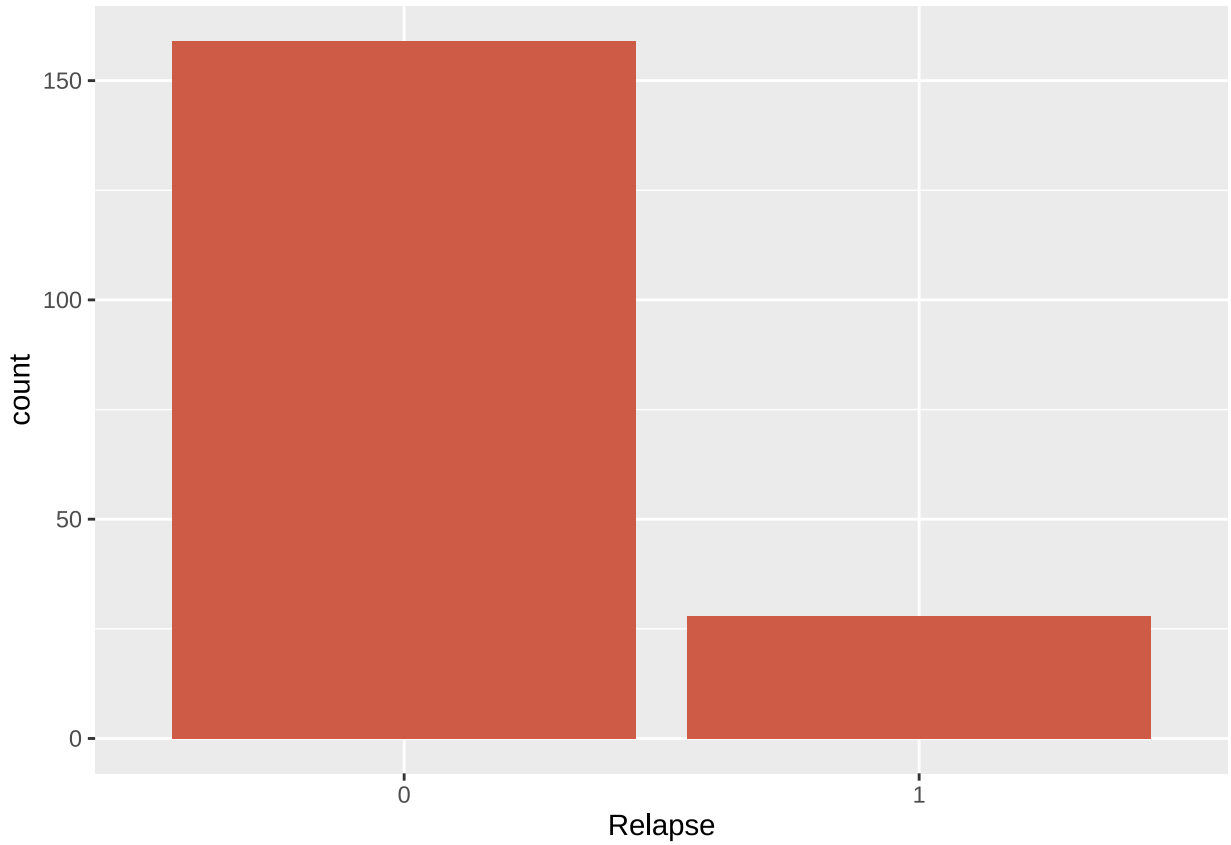



```
ggplot(data = new_data) +
  geom_bar(mapping = aes(x = Recipientgender), fill = 'coral1')
```



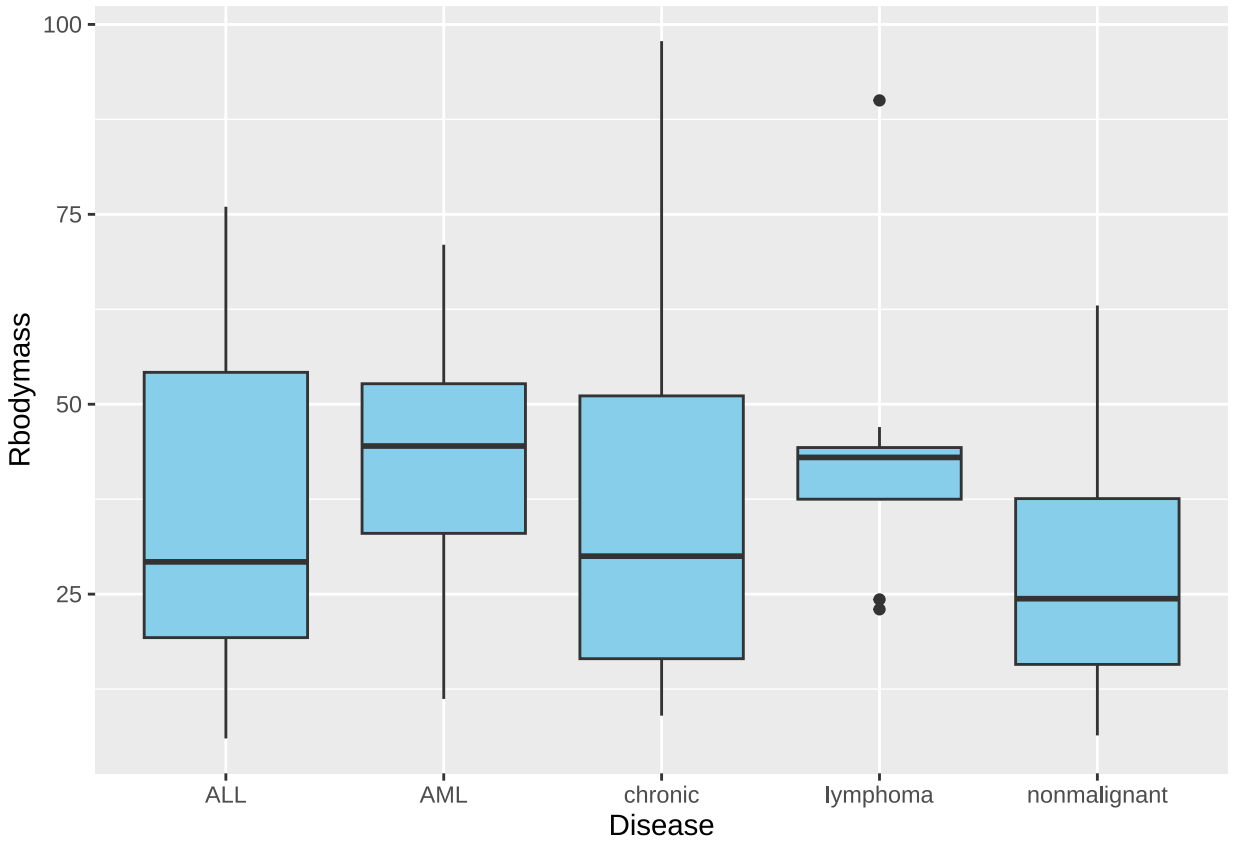
A lot of patients didn't relapse which is an adverse affect.

```
ggplot(data = new_data) +
  geom_bar(mapping = aes(x = Relapse), fill = 'coral3')
```



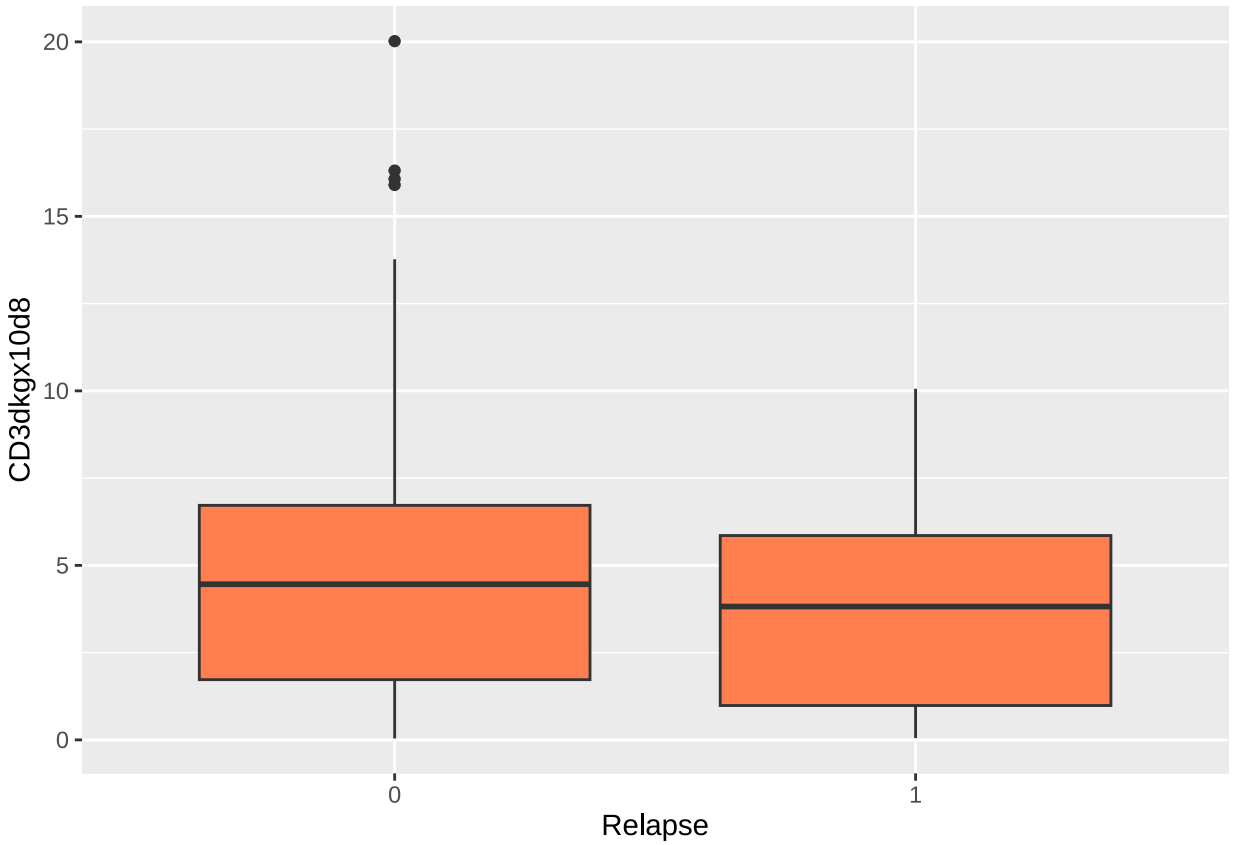
Checking the mean body weight of all the disease groups looks relativly similar besides lymphoma. Something to potentially check out with lymphoma

```
ggplot(data = new_data, mapping = aes(x = Disease, y = Rbodymass)) +
  geom_boxplot(fill = 'skyblue')
```



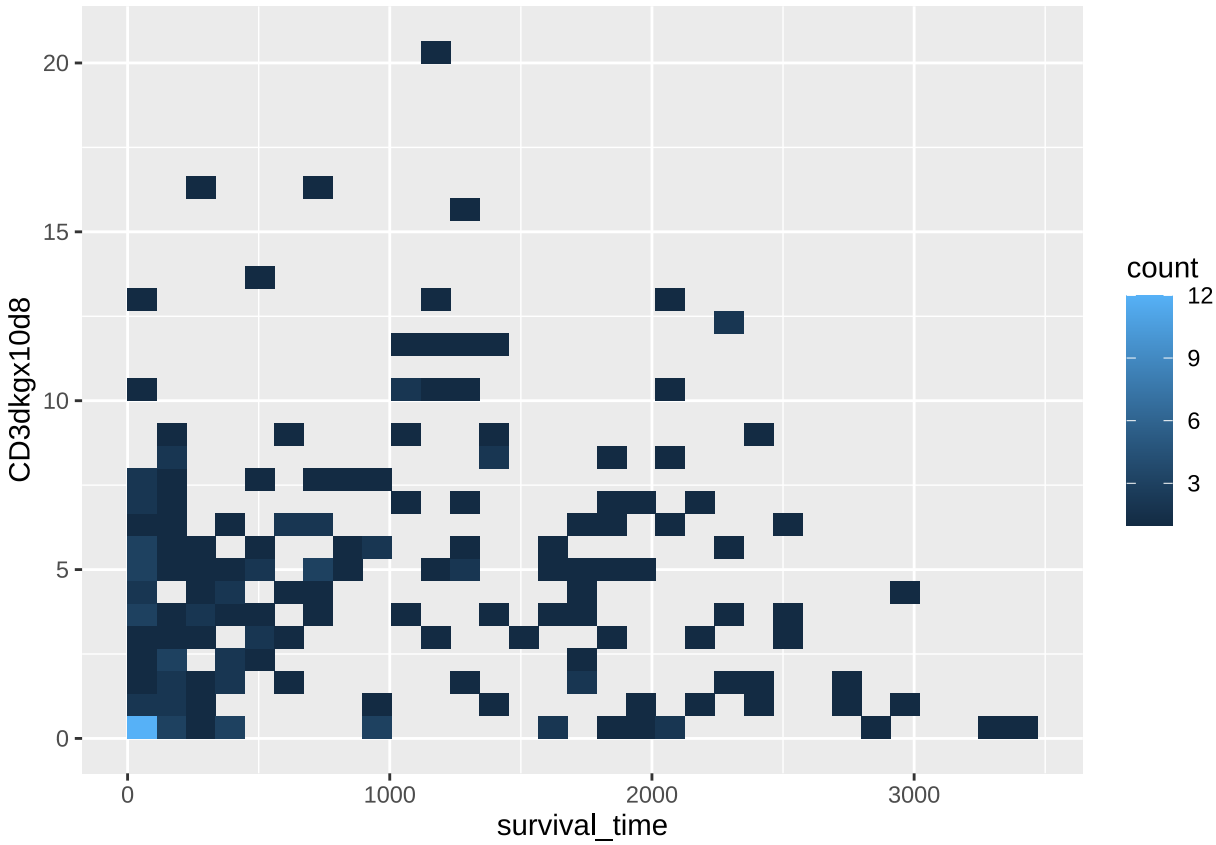
Seems like the mean CD3dkgx.. levels are even within the two groups. There are some data points that are higher in the non relapse group.

```
ggplot(data = new_data, mapping = aes(x = Relapse, y = CD3dkgx10d8)) +
  geom_boxplot(fill = 'coral')
```



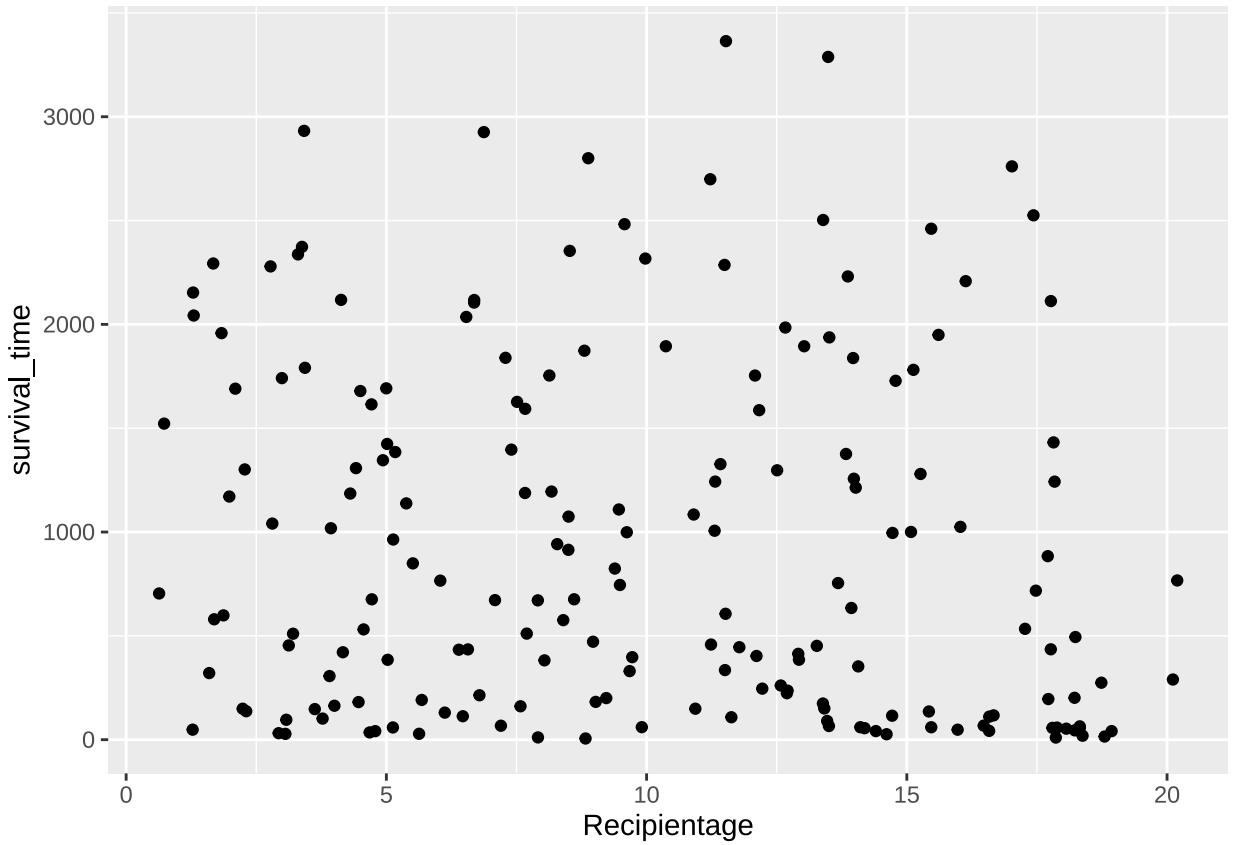
Based off this graph we wouldn't be able to immediately come to a conclusion about CD3dkgx... levels increasing survival time.

```
ggplot(data = new_data) +
  geom_bin2d(mapping = aes(x = survival_time, y = CD3dkgx10d8))
```



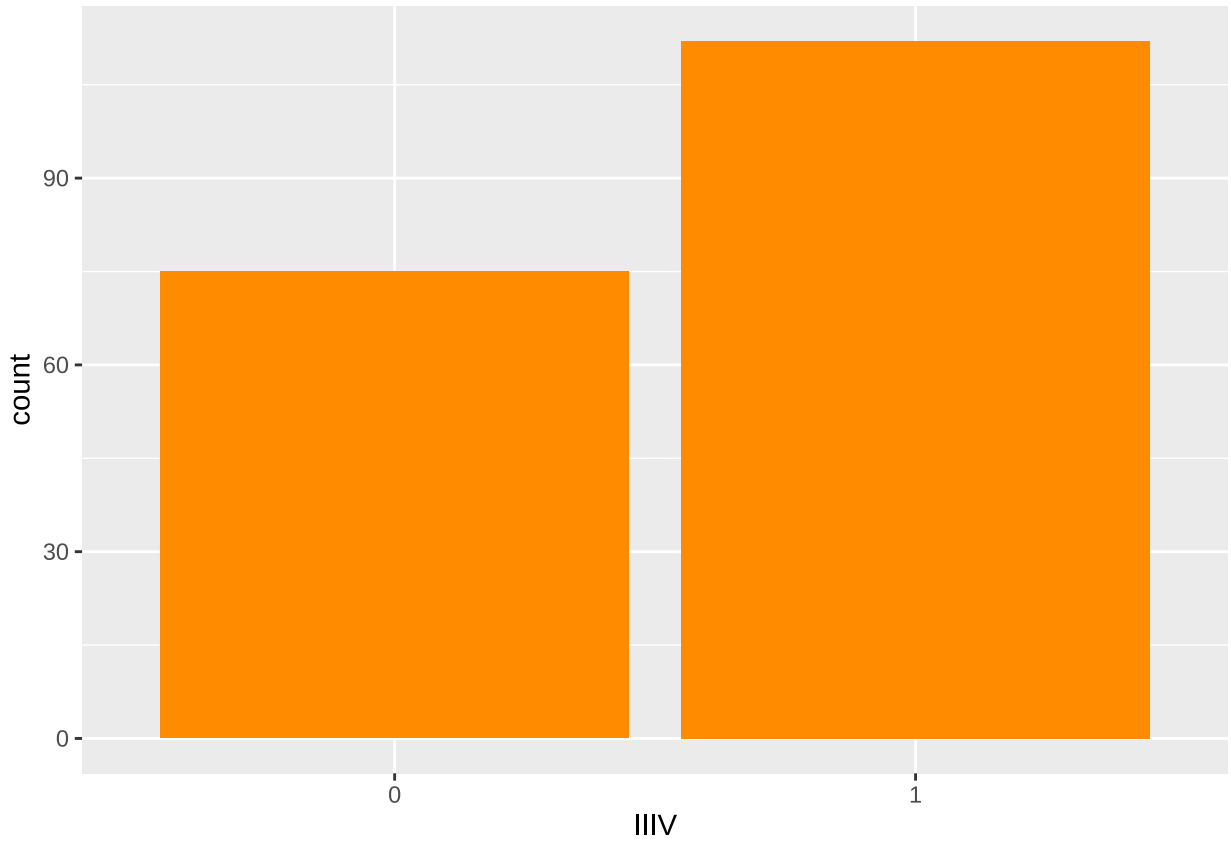
Checking if recipients age is positively or negatively correlated with survival time. As you can see there is no clear correlation between both variables.

```
ggplot(data = new_data, aes(x = Recipientage, y = survival_time)) +
  geom_jitter()
```



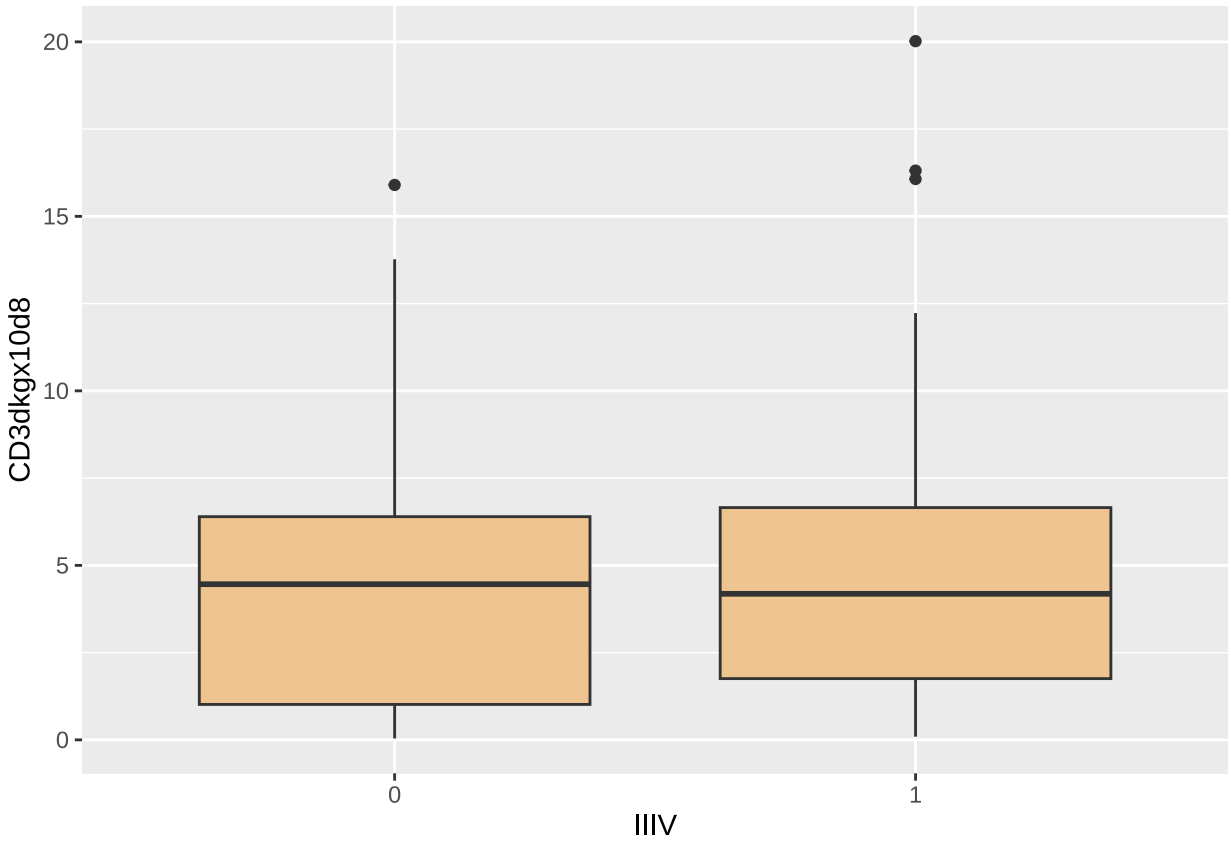
IIIV refers to the development of acute graft vs host disease stage which would be considered an adverse effect

```
ggplot(data = new_data) +
  geom_bar(mapping = aes(x = IIIV), fill = 'darkorange')
```



Checking if increased CD34kg.. levels have an effect on the patients developing acute graft. From the look at the boxplot it looks pretty even with some outliers.

```
ggplot(data = new_data, mapping = aes(x = IIIV, y = CD3dkgx10d8)) +
  geom_boxplot(fill = 'burlywood2')
```


20
15
10
5
0
CD3dkgx10d8
0
1
IIIV