```r
library(tidyverse)
```

```
## Warning: package ’tidyverse’ was built under R version 4.1.2

## Warning in as.POSIXlt.POSIXct(Sys.time()): unable to identify current timezone ’H’:
## please set environment variable ’TZ’

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## Warning: package ’ggplot2’ was built under R version 4.1.2

## Warning: package ’tibble’ was built under R version 4.1.2

## Warning: package ’tidyr’ was built under R version 4.1.2

## Warning: package ’readr’ was built under R version 4.1.2

## Warning: package ’purrr’ was built under R version 4.1.2

## Warning: package ’dplyr’ was built under R version 4.1.2

## Warning: package ’stringr’ was built under R version 4.1.2

## Warning: package ’forcats’ was built under R version 4.1.2

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

```r
library(ggplot2)
```

```r
cytof = read_csv("https://jfukuyama.github.io/teaching/stat670/notes/cytof_one_experiment.csv")
```

```
## Rows: 50000 Columns: 35

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (35): NKp30, KIR3DL1, NKp44, KIR2DL1, GranzymeB, CXCR6, CD161, KIR2DS4, ...

##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
head(cytof)
```

```
## # A tibble: 6 x 35
##    NKp30 KIR3DL1  NKp44 KIR2DL1 GranzymeB   CXCR6  CD161 KIR2DS4 NKp46  NKG2D
##    <dbl>   <dbl>  <dbl>   <dbl>     <dbl>   <dbl>  <dbl>   <dbl> <dbl>  <dbl>
## 1  0.188   3.62  -0.561  -0.294      2.48 -0.145  -0.315  1.94   4.08   2.62
## 2  1.03    1.70  -0.289  -0.480      3.26 -0.0339 -0.411  3.80   3.73  -0.483
## 3  3.00    6.14   1.90    0.482      4.28  1.95   -0.502 -0.320  4.56  -0.507
## 4  4.30   -0.221  0.243  -0.483      3.35  0.926   3.88  -0.170  4.48   1.93
## 5 -0.439  -0.504 -0.153   0.751      3.19 -0.0589  1.09  -0.0503 0.838 -0.458
## 6  2.09   -0.399  3.46   -0.520      4.35 -0.364  -0.571 -0.450  4.06   3.43
## # ... with 25 more variables: NKG2C <dbl>, X2B4 <dbl>, CD69 <dbl>,
## #   KIR3DL1.S1 <dbl>, CD2 <dbl>, KIR2DL5 <dbl>, DNAM.1 <dbl>, CD4 <dbl>,
## #   CD8 <dbl>, CD57 <dbl>, TRAIL <dbl>, KIR3DL2 <dbl>, MIP1b <dbl>,
## #   CD107a <dbl>, GM.CSF <dbl>, CD16 <dbl>, TNFa <dbl>, ILT2 <dbl>,
## #   Perforin <dbl>, KIR2DL2.L3.S2 <dbl>, KIR2DL3 <dbl>, NKG2A <dbl>,
## #   NTB.A <dbl>, CD56 <dbl>, INFg <dbl>
```

# Q.1)

## Answer

1. Converting the CyTOF dataset from wide form to a longer form
2. We first check the column names and number of col names

```
colnames(cytof)
```

```
##  [1] "NKp30"         "KIR3DL1"       "NKp44"         "KIR2DL1"
##  [5] "GranzymeB"     "CXCR6"         "CD161"         "KIR2DS4"
##  [9] "NKp46"         "NKG2D"         "NKG2C"         "X2B4"
## [13] "CD69"          "KIR3DL1.S1"    "CD2"           "KIR2DL5"
## [17] "DNAM.1"        "CD4"           "CD8"           "CD57"
## [21] "TRAIL"         "KIR3DL2"       "MIP1b"         "CD107a"
## [25] "GM.CSF"        "CD16"          "TNFa"          "ILT2"
## [29] "Perforin"      "KIR2DL2.L3.S2" "KIR2DL3"       "NKG2A"
## [33] "NTB.A"         "CD56"          "INFg"
```

```
ncol(cytof)
```

```
## [1] 35
```

```
cytof_long= cytof %>%
  pivot_longer(c(colnames(cytof)),names_to = "markers",values_to = "values")
```

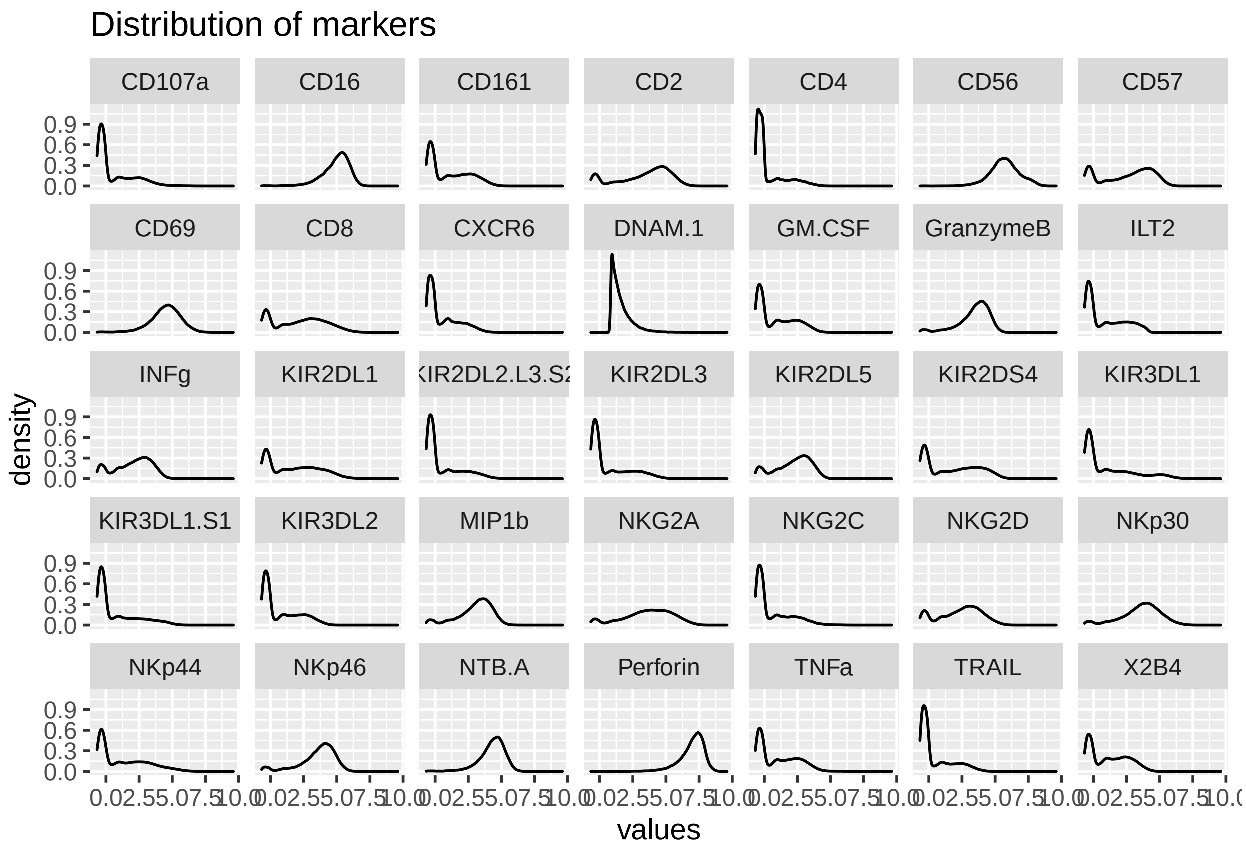
3. This is the data we get after converting to longer form

```
head(cytof_long)
```

```
## # A tibble: 6 x 2
##   markers   values
##   <chr>      <dbl>
## 1 NKp30      0.188
## 2 KIR3DL1    3.62
## 3 NKp44     -0.561
## 4 KIR2DL1   -0.294
## 5 GranzymeB  2.48
## 6 CXCR6     -0.145
```

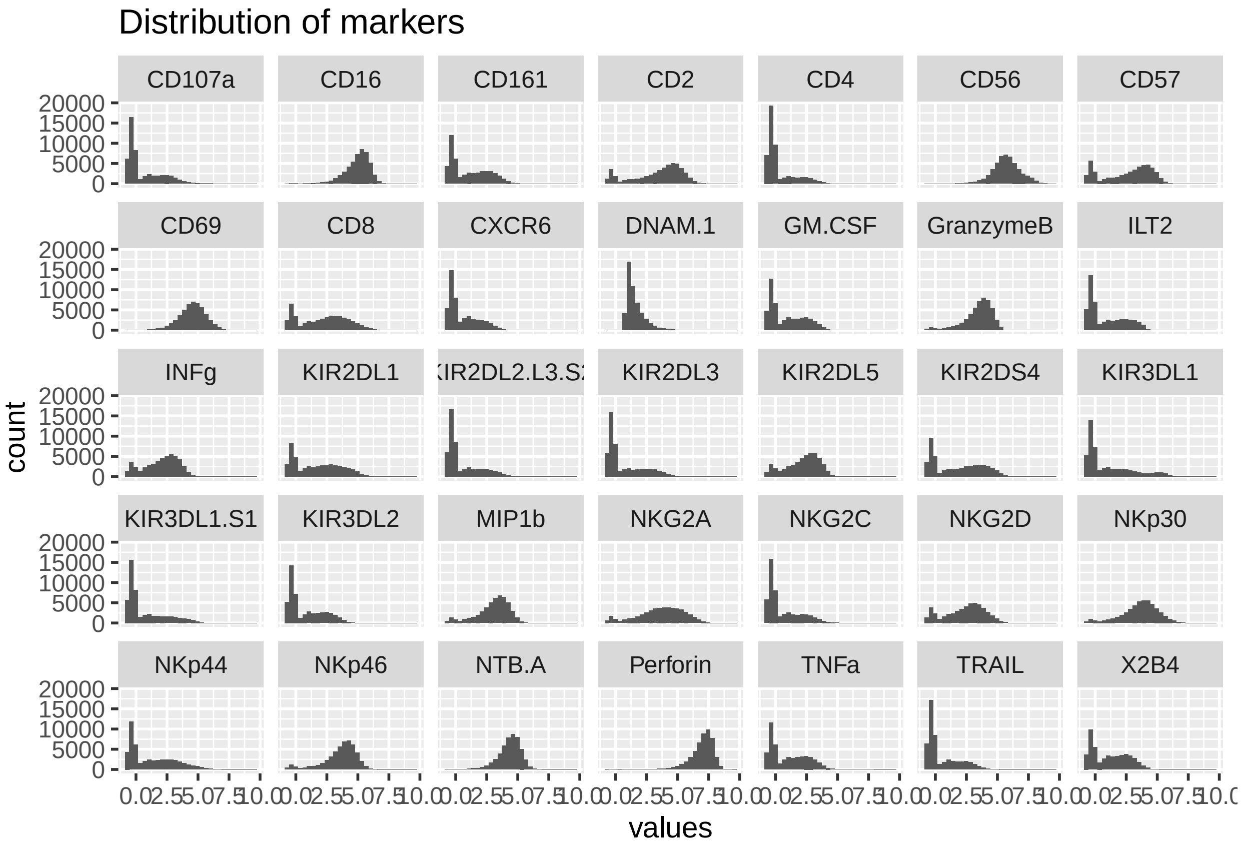
4. Now we plot the faceted plots for distributions of all markers using density plot and also we created a histogram plot for the same

```
marker_gg = ggplot(cytof_long,aes(x=values)) + geom_density()
marker_gg + facet_wrap(~ markers , nrow = 5) + ggtitle("Distribution of markers")
```



```
marker_gg = ggplot(cytof_long,aes(x=values)) + geom_histogram()
marker_gg + facet_wrap(~ markers , nrow = 5) + ggtitle("Distribution of markers")
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



5. From the above faceted plots, we see the folloing:

a. Markers like CD56 and CD69 show similar distribution.
b. When we check the mean and median values for all the distribution using summary method(refer Q.3),we see that there are very few markers who follow a normal distribution.
c. Mostly the markers are right-skewed and left-skewed based on the values of mean and median.
d. The standard deviation for most of the markers is around 1 or in most cases greater than 1. So the distribution is quite spread out out.
e. Also we see that markers KRDL1 and CD8 are also similar in distribution.

## Q.2)

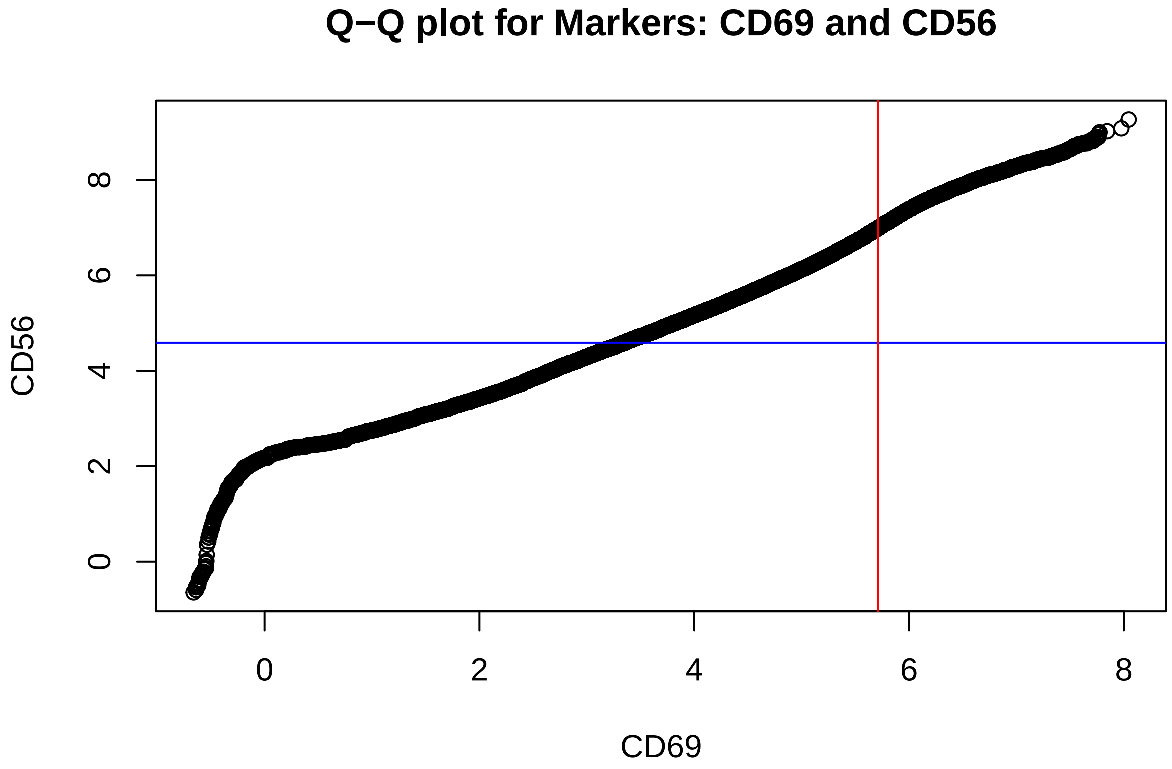
## Answer:

1. Here from looking at the graph, we took two sets of markers
2. For the first set, we took the markers CD69 and CD56 because they almost are similar in distribution.

```
CD69 <- filter(cytof_long, markers == "CD69") %>%  pull(values)
```

```
CD56 <- filter(cytof_long, markers == "CD56") %>%  pull(values)
```

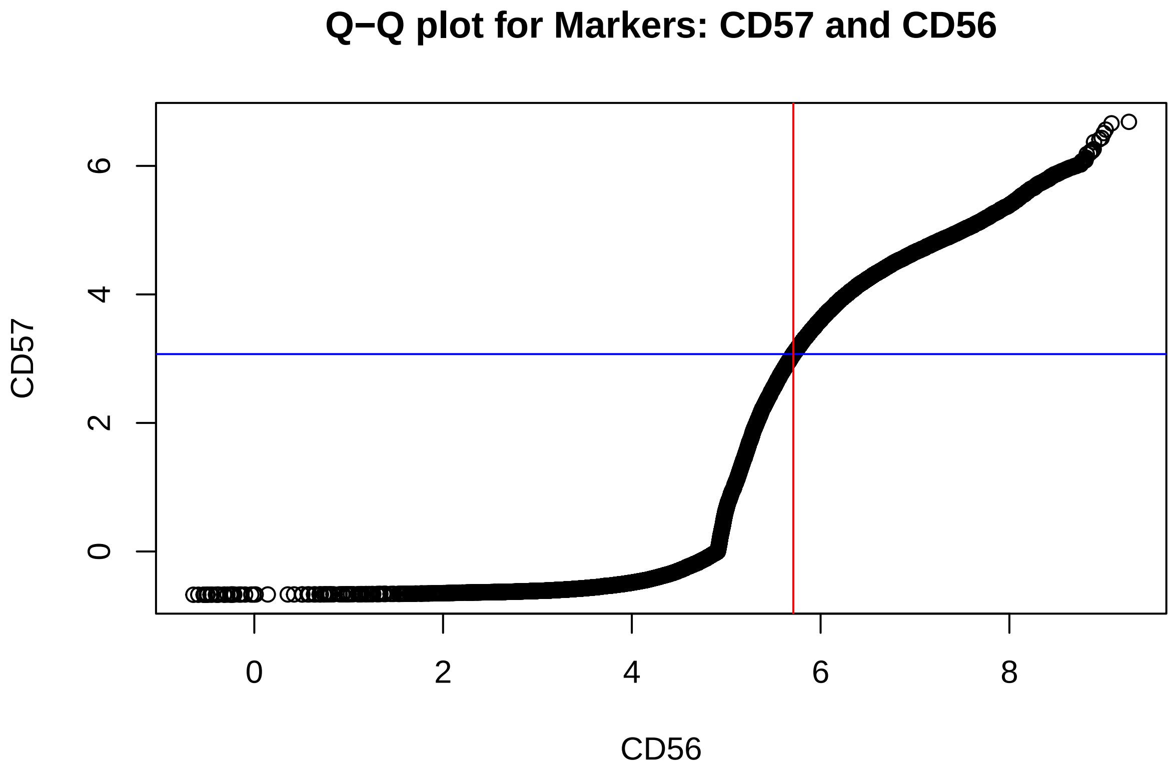
```r
qqplot(x=CD69,y=CD56)
abline(h=median(CD69),col="blue")
abline(v=median(CD56),col="red")
title("Q-Q plot for Markers: CD69 and CD56")
```



3. Here since the distributions are almost similar, the points for the quartiles lie almost in a straight line except for the region near the origin. 4. Near this region, the quartiles seem to increase exponentially 5. We then plot the median for both the marker and we see that the median value for marker CD56 is around 4.5 and the median for marker CD69 is around 5.7 6. Since median values of both the markers are different and shape of distribution is almost the same, this is a simple shift. 7. For the parts around -0.6 to 0 the quartiles are increasing and then after 0 the quartiles fall in a line.

```r
CD57 <- filter(cytof_long, markers == "CD57") %>%  pull(values)
```

```r
qqplot(x=CD56,y=CD57)
abline(h=median(CD57),col='blue')
abline(v=median(CD56),col='red')
title("Q-Q plot for Markers: CD57 and CD56")
```



8. For the second set, we took the markers CD57 and CD56 because they are dissimilar in distribution. 9. Here since the distributions are not similar, the points for the quartiles do not lie in a straight line. 10. We check the median values for the markers CD57 and CD56. 11. The median value for CD57 is around 3 and median value for CD56 is around 5.7 12. Since the values of medians for both the markers is different and the shape of distribution is also different the shift for this distribution seems to be complex 13. For the parts around -0.6 to 0, there seems to be outliers in that region, after around 0.2 till 4, the quartiles are steady and after 4.5 onwards the quartiles are exponentially increaing.

## Q.3)

## Answer:

1. We first computed the statistics on the basis of all markers
2. From this, we get the value of mean,median,1st quartile,3rd quartile,min and max

```
by(cytof_long,cytof_long$markers,summary)
```

```
## cytof_long$markers: CD107a
##    markers              values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.:-0.3994
##  Mode  :character   Median :-0.1223
##                     Mean   : 0.6076
##                     3rd Qu.: 1.4912
##                      Max.   : 8.1919
## ------------------------------------------------------------
## cytof_long$markers: CD16
##    markers              values
##  Length:50000       Min.   :-0.6416
##  Class :character   1st Qu.: 4.4233
##  Mode  :character   Median : 5.1230
##                     Mean   : 4.9491
##                     3rd Qu.: 5.6465
##                     Max.   : 7.5307
## ------------------------------------------------------------
## cytof_long$markers: CD161
##    markers              values
##  Length:50000       Min.   :-0.6738
##  Class :character   1st Qu.:-0.2929
##  Mode  :character   Median : 0.7257
##                     Mean   : 1.0596
##                     3rd Qu.: 2.3461
##                     Max.   : 5.6971
## ------------------------------------------------------------
## cytof_long$markers: CD2
##    markers              values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.: 2.2902
##  Mode  :character   Median : 3.9454
##                     Mean   : 3.4134
##                     3rd Qu.: 4.8724
##                     Max.   : 7.7211
## ------------------------------------------------------------
## cytof_long$markers: CD4
##    markers              values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.:-0.4379
##  Mode  :character   Median :-0.2036
##                     Mean   : 0.2869
##                     3rd Qu.: 0.6275
##                     Max.   : 4.9322
## ------------------------------------------------------------
## cytof_long$markers: CD56
##    markers              values
##  Length:50000       Min.   :-0.645
##  Class :character   1st Qu.: 5.051
##  Mode  :character   Median : 5.711
##                     Mean   : 5.715
##                     3rd Qu.: 6.401
##                     Max.   : 9.267
## ------------------------------------------------------------
## cytof_long$markers: CD57
##    markers              values
##  Length:50000       Min.   :-0.6735
##  Class :character   1st Qu.: 0.8916
##  Mode  :character   Median : 3.0709
##                     Mean   : 2.5780
##                     3rd Qu.: 4.1450
##                      Max.   : 6.6852
## -----------------------------------------------------------
## cytof_long$markers: CD69
##    markers              values
##  Length:50000       Min.   :-0.6611
##  Class :character   1st Qu.: 3.8747
##  Mode  :character   Median : 4.5885
##                     Mean   : 4.5058
##                     3rd Qu.: 5.2523
##                     Max.   : 8.0455
## -----------------------------------------------------------
## cytof_long$markers: CD8
##    markers              values
##  Length:50000       Min.   :-0.6738
##  Class :character   1st Qu.: 0.2508
##  Mode  :character   Median : 2.4007
##                     Mean   : 2.2130
##                     3rd Qu.: 3.7001
##                     Max.   : 7.4813
## -----------------------------------------------------------
## cytof_long$markers: CXCR6
##    markers              values
##  Length:50000       Min.   :-0.67376
##  Class :character   1st Qu.:-0.36745
##  Mode  :character   Median :-0.05814
##                     Mean   : 0.54687
##                     3rd Qu.: 1.35477
##                     Max.   : 4.76702
## -----------------------------------------------------------
## cytof_long$markers: DNAM.1
##    markers              values
##  Length:50000       Min.   :0.8563
##  Class :character   1st Qu.:1.0443
##  Mode  :character   Median :1.3535
##                     Mean   :1.5843
##                     3rd Qu.:1.8796
##                     Max.   :8.7841
## -----------------------------------------------------------
## cytof_long$markers: GM.CSF
##    markers              values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.:-0.3228
##  Mode  :character   Median : 0.4404
##                     Mean   : 0.8877
##                     3rd Qu.: 2.0638
##                     Max.   : 5.0265
## -----------------------------------------------------------
## cytof_long$markers: GranzymeB
##    markers              values
##  Length:50000       Min.   :-0.672
##  Class :character   1st Qu.: 2.951
##  Mode  :character   Median : 3.683
##                     Mean   : 3.457
##                     3rd Qu.: 4.242
##                      Max.   : 6.185
## ------------------------------------------------------------
## cytof_long$markers: ILT2
##    markers              values
##  Length:50000       Min.   :-0.673893
##  Class :character   1st Qu.:-0.344650
##  Mode  :character   Median : 0.004518
##                     Mean   : 0.868112
##                     3rd Qu.: 2.102205
##                     Max.   : 4.129952
## ------------------------------------------------------------
## cytof_long$markers: INFg
##    markers              values
##  Length:50000       Min.   :-0.6735
##  Class :character   1st Qu.: 1.0488
##  Mode  :character   Median : 2.2654
##                     Mean   : 2.0348
##                     3rd Qu.: 3.1019
##                     Max.   : 7.6256
## ------------------------------------------------------------
## cytof_long$markers: KIR2DL1
##    markers              values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.:-0.1282
##  Mode  :character   Median : 1.7049
##                     Mean   : 1.7765
##                     3rd Qu.: 3.2782
##                     Max.   : 7.9213
## ------------------------------------------------------------
## cytof_long$markers: KIR2DL2.L3.S2
##    markers              values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.:-0.3990
##  Mode  :character   Median :-0.1301
##                     Mean   : 0.5516
##                     3rd Qu.: 1.3537
##                     Max.   : 5.8729
## ------------------------------------------------------------
## cytof_long$markers: KIR2DL3
##    markers              values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.:-0.3910
##  Mode  :character   Median :-0.1020
##                     Mean   : 0.7100
##                     3rd Qu.: 1.7530
##                     Max.   : 6.4319
## ------------------------------------------------------------
## cytof_long$markers: KIR2DL5
##    markers              values
##  Length:50000       Min.   :-0.6738
##  Class :character   1st Qu.: 1.2697
##  Mode  :character   Median : 2.4158
##                     Mean   : 2.1593
##                     3rd Qu.: 3.1928
##                     Max.   : 5.7212
## ------------------------------------------------------------
## cytof_long$markers: KIR2DS4
##    markers              values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.:-0.2043
##  Mode  :character   Median : 1.7103
##                     Mean   : 1.7629
##                     3rd Qu.: 3.4708
##                     Max.   : 6.5207
## ------------------------------------------------------------
## cytof_long$markers: KIR3DL1
##    markers              values
##  Length:50000       Min.   :-0.67380
##  Class :character   1st Qu.:-0.35515
##  Mode  :character   Median :-0.02122
##                     Mean   : 1.05674
##                     3rd Qu.: 2.15491
##                     Max.   : 7.36360
## ------------------------------------------------------------
## cytof_long$markers: KIR3DL1.S1
##    markers              values
##  Length:50000       Min.   :-0.67390
##  Class :character   1st Qu.:-0.38403
##  Mode  :character   Median :-0.09276
##                     Mean   : 0.73871
##                     3rd Qu.: 1.65773
##                     Max.   : 6.26132
## ------------------------------------------------------------
## cytof_long$markers: KIR3DL2
##    markers              values
##  Length:50000       Min.   :-0.67390
##  Class :character   1st Qu.:-0.35674
##  Mode  :character   Median :-0.03381
##                     Mean   : 0.77490
##                     3rd Qu.: 1.90662
##                     Max.   : 5.28953
## ------------------------------------------------------------
## cytof_long$markers: MIP1b
##    markers              values
##  Length:50000       Min.   :-0.6737
##  Class :character   1st Qu.: 2.3758
##  Mode  :character   Median : 3.2699
##                     Mean   : 3.0124
##                     3rd Qu.: 3.9292
##                     Max.   : 7.3079
## ------------------------------------------------------------
## cytof_long$markers: NKG2A
##    markers              values
##  Length:50000       Min.   :-0.6733
##  Class :character   1st Qu.: 2.5479
##  Mode  :character   Median : 3.8345
##                     Mean   : 3.6444
##                     3rd Qu.: 5.0034
##                      Max.   : 8.1448
## ------------------------------------------------------------
## cytof_long$markers: NKG2C
##    markers               values
##  Length:50000       Min.   :-0.67387
##  Class :character   1st Qu.:-0.38872
##  Mode  :character   Median :-0.09717
##                     Mean   : 0.60702
##                     3rd Qu.: 1.47588
##                     Max.   : 6.78993
## ------------------------------------------------------------
## cytof_long$markers: NKG2D
##    markers               values
##  Length:50000       Min.   :-0.6736
##  Class :character   1st Qu.: 1.2433
##  Mode  :character   Median : 2.6266
##                     Mean   : 2.3635
##                     3rd Qu.: 3.5503
##                     Max.   : 6.8310
## ------------------------------------------------------------
## cytof_long$markers: NKp30
##    markers               values
##  Length:50000       Min.   :-0.6733
##  Class :character   1st Qu.: 2.8238
##  Mode  :character   Median : 3.7796
##                     Mean   : 3.5948
##                     3rd Qu.: 4.5907
##                     Max.   : 7.8212
## ------------------------------------------------------------
## cytof_long$markers: NKp44
##    markers               values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.:-0.2904
##  Mode  :character   Median : 0.7593
##                     Mean   : 1.2652
##                     3rd Qu.: 2.6436
##                     Max.   : 7.2905
## ------------------------------------------------------------
## cytof_long$markers: NKp46
##    markers               values
##  Length:50000       Min.   :-0.6721
##  Class :character   1st Qu.: 3.0094
##  Mode  :character   Median : 3.8535
##                     Mean   : 3.5701
##                     3rd Qu.: 4.4796
##                     Max.   : 6.6703
## ------------------------------------------------------------
## cytof_long$markers: NTB.A
##    markers               values
##  Length:50000       Min.   :-0.6737
##  Class :character   1st Qu.: 3.8259
##  Mode  :character   Median : 4.4428
##                     Mean   : 4.3019
##                     3rd Qu.: 4.9508
##                        Max.   : 6.8866
## ---------------------------------------------------------------
## cytof_long$markers: Perforin
##    markers              values
##  Length:50000       Min.   :-0.6189
##  Class :character   1st Qu.: 6.5303
##  Mode  :character   Median : 7.1411
##                     Mean   : 6.9619
##                     3rd Qu.: 7.5950
##                     Max.   : 9.6139
## ---------------------------------------------------------------
## cytof_long$markers: TNFa
##    markers              values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.:-0.2809
##  Mode  :character   Median : 0.7920
##                     Mean   : 1.0509
##                     3rd Qu.: 2.2773
##                     Max.   : 8.2290
## ---------------------------------------------------------------
## cytof_long$markers: TRAIL
##    markers              values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.:-0.4084
##  Mode  :character   Median :-0.1442
##                     Mean   : 0.4736
##                     3rd Qu.: 1.2266
##                     Max.   : 5.1661
## ---------------------------------------------------------------
## cytof_long$markers: X2B4
##    markers              values
##  Length:50000       Min.   :-0.6739
##  Class :character   1st Qu.:-0.2154
##  Mode  :character   Median : 1.0444
##                     Mean   : 1.1576
##                     3rd Qu.: 2.3652
##                     Max.   : 5.0679
```

3. We also computed other statistical data which was not previously computed using the summary method
4. Here we computed the measures of spread of distributions for all the markers by first finding the standard deviation and then the IQR
5. We see that for most of the markers, the standard deviation is greater than 1 and very close to 1 so we can say that the data is more spread out for the markers.

```
cytof_long %>% group_by(markers) %>% summarise(sd(values))
```

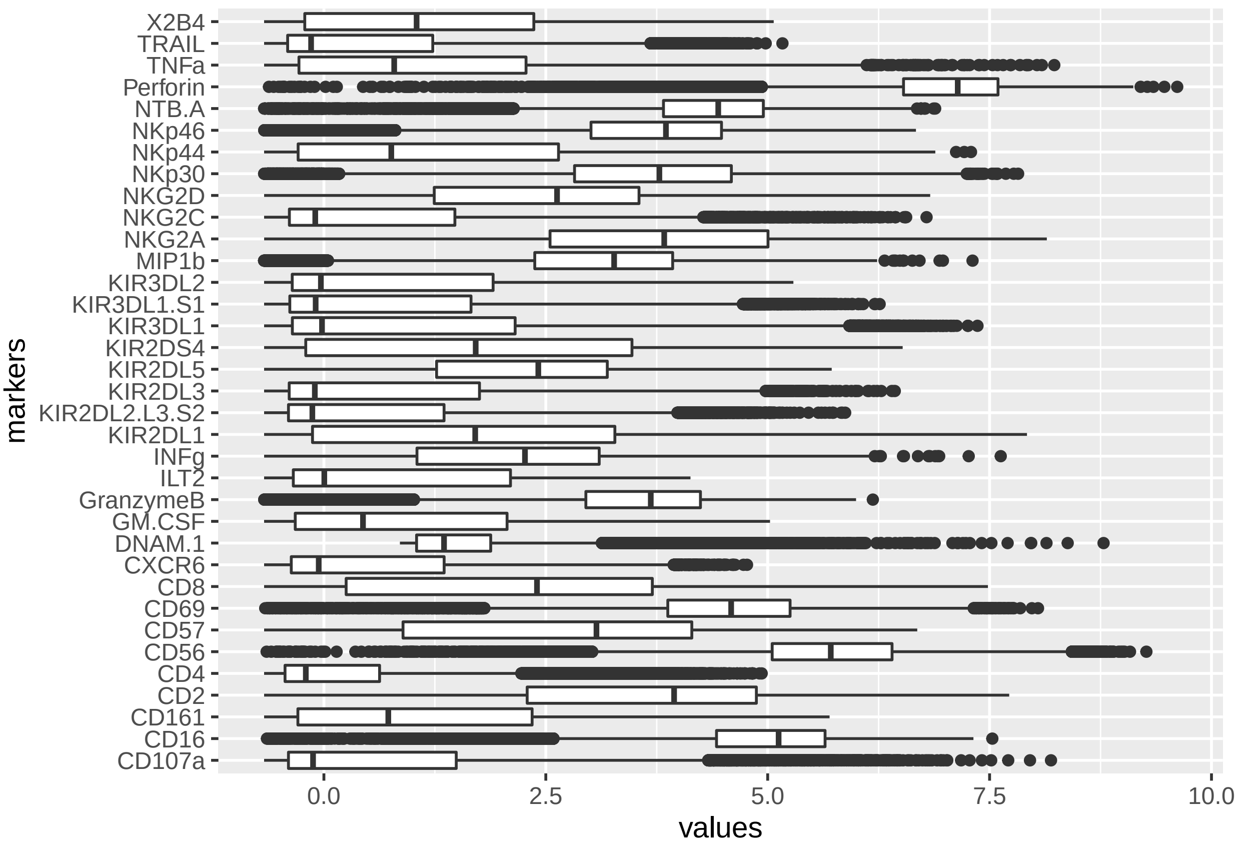
```
## # A tibble: 35 x 2
##    markers `sd(values)`
##    <chr>          <dbl>
##  1 CD107a          1.39
##  2 CD16            1.01
##  3 CD161           1.48
##  4 CD2             1.94
##  5 CD4              1.13
##  6 CD56             1.11
##  7 CD57             1.89
##  8 CD69             1.13
##  9 CD8              1.89
## 10 CXCR6            1.17
## # ... with 25 more rows
```

```
cytof_long %>% group_by(markers) %>% summarise(IQR(values))
```

```
## # A tibble: 35 x 2
##    markers `IQR(values)`
##    <chr>            <dbl>
##  1 CD107a            1.89
##  2 CD16              1.22
##  3 CD161             2.64
##  4 CD2               2.58
##  5 CD4               1.07
##  6 CD56              1.35
##  7 CD57              3.25
##  8 CD69              1.38
##  9 CD8               3.45
## 10 CXCR6             1.72
## # ... with 25 more rows
```

6. Now for plotting the statistical data, we have chosen box plot.
7. From box plot we get to see the data for the median, the 1st and 3rd quartile, the min and the max values for all the markers
8. As compared to the full distribution we plotted before, this gives a fairly better statistical measure for the data as opposed to the histogram where we have to estimate the values of mean and spread.

```
ggplot(cytof_long) +
  aes(x = markers, y =values) +
  geom_boxplot() + coord_flip()
```


X2B4
TRAIL
TNFa
Perforin
NTB.A
NKp46
NKp44
NKp30
NKG2D
NKG2C
NKG2A
MIP1b
KIR3DL2
KIR3DL1.S1
KIR3DL1
KIR2DS4
KIR2DL5
KIR2DL3
KIR2DL2.L3.S2
KIR2DL1
INFg
ILT2
GranzymeB
GM.CSF
DNAM.1
CXCR6
CD8
CD69
CD57
CD56
CD4
CD2
CD161
CD16
CD107a
markers
0.0
2.5
5.0
7.5
10.0
values