EDA Homework 2 Solutions

Question 1

```
library(tidyverse)
```

```
## - Attaching packages ---------------------- tidyverse 1.2.1 -
```

```
##  tibble  1.4.2     purrr  0.2.5
##  tidyr  0.8.1     dplyr  0.7.6
##  readr  1.1.1     stringr 1.3.1
##  tibble  1.4.2     forcats 0.3.0
```

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

```
cytof = read_csv("~/GitHub/eda-fall-18/datasets/cytof_one_experiment.csv")
```

```
## Parsed with column specification:
## cols(
##  .default = col_double()
## )
```

```
## See spec(...)  for full column specifications.
```

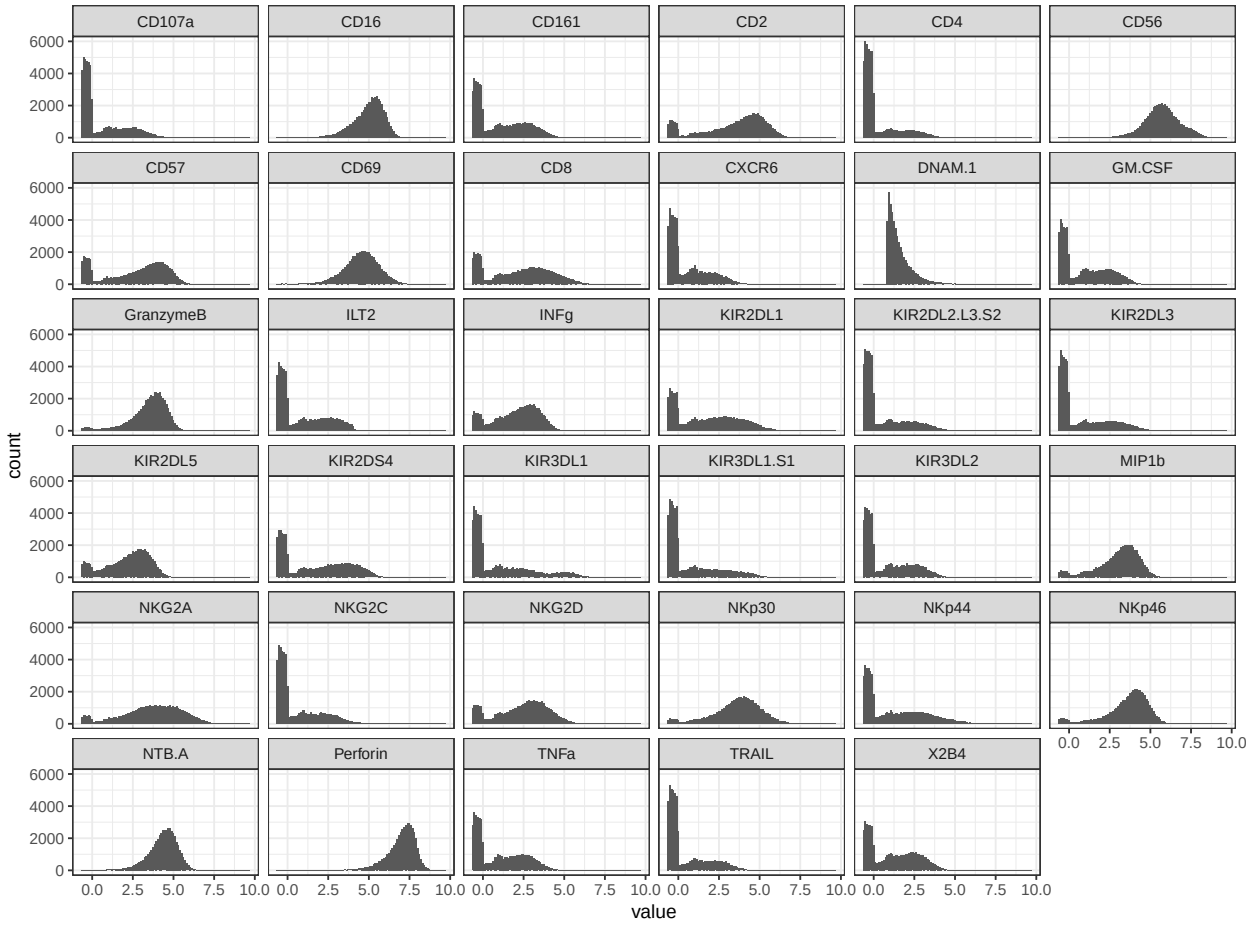
```
# melt is in dplyr
cytof_melted = melt(cytof)
```

```
## No id variables; using all as measure variables
```

```
## alternately
cytof_melted = gather(cytof, colnames(cytof), key = "marker", value = "value")
```

We can look at the distributions using histograms.

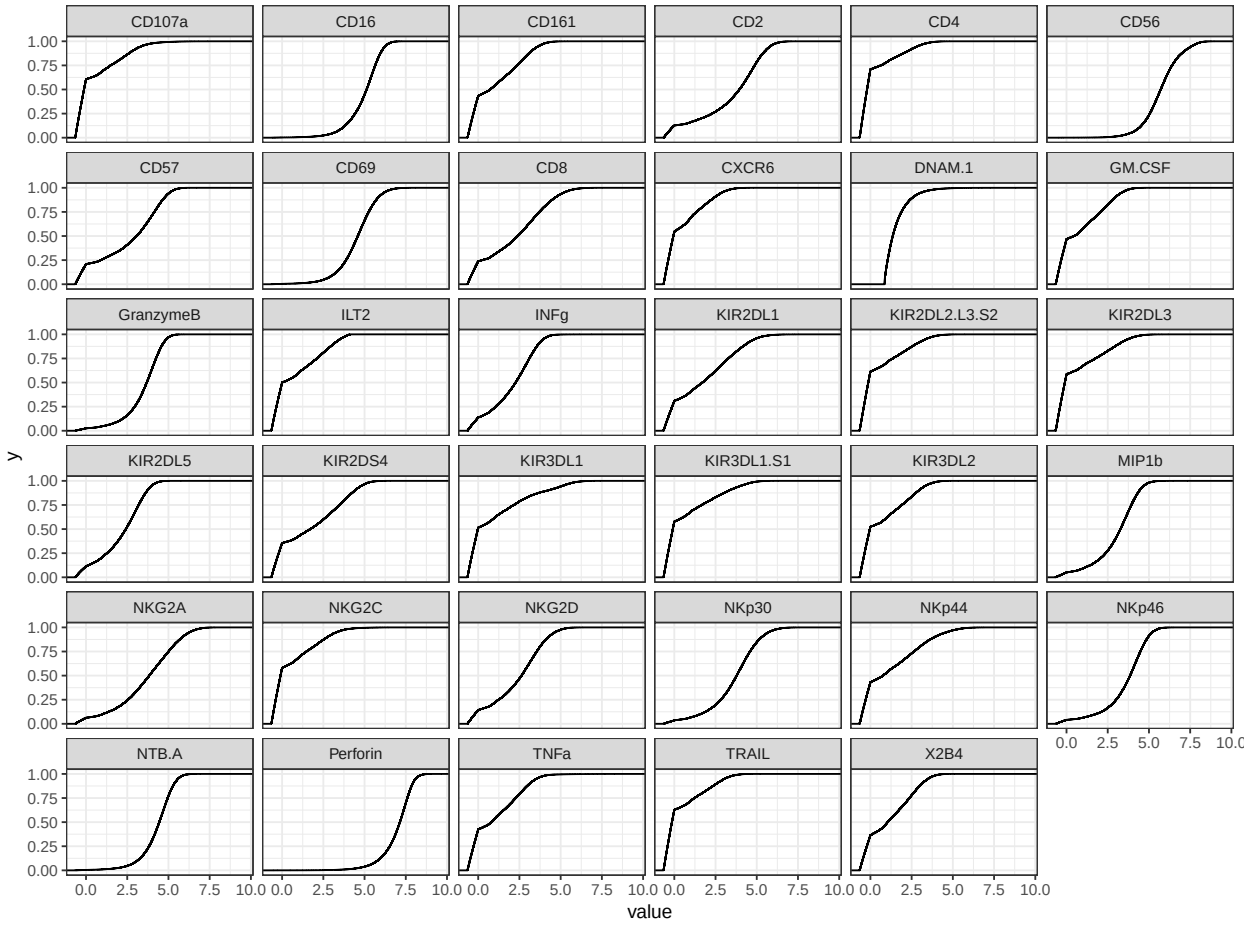
```
ggplot(cytof_melted) +
    geom_histogram(aes(x = value), bins = 100) +
    facet_wrap(~ marker)
```



From the histograms, we see that most of the distributions are strongly bimodal, although there are a few that have only one mode. The lower mode, when it exists, tends to be of the same, non-normal, shape. The upper mode varies in shape by marker, sometimes looking more normal and sometimes more skewed.

We can also use ecdf plots:

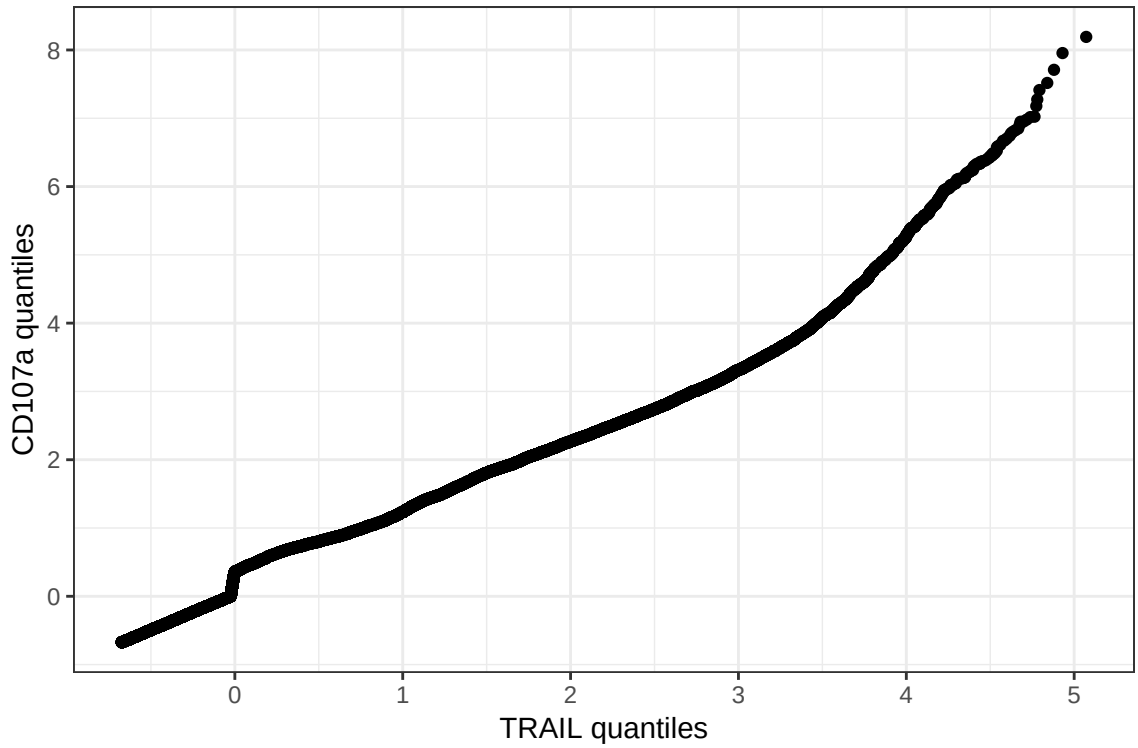
```
ggplot(cytof_melted) +
    stat_ecdf(aes(x = value)) +
    facet_wrap(~ marker)
```



From the ecdfs, we can easily read off the fraction of points in the two modes: these range from a high of nearly .75 for CD4 down to zero for markers like perforin.

Question 2

```
ggplot(cytof) +
    stat_qq(aes(sample = CD107a), distribution = function(p) quantile(cytof$TRAIL, probs = p)) +
    xlab("TRAIL quantiles") + ylab("CD107a quantiles")
```



The QQ plot has two regions: the first where the quantiles fall exactly along a straight line (for values less than zero for both of the markers), and the second where they are shifted, but still roughly fall along a straight line.

This indicates that below zero, the two distributions have exactly the same shape, while above zero the distributions are offset a bit and are of similar but not identical shape. The increasing slope of the quantile plot near at the upper values indicates that CD107a has heavier tails than TRAIL.

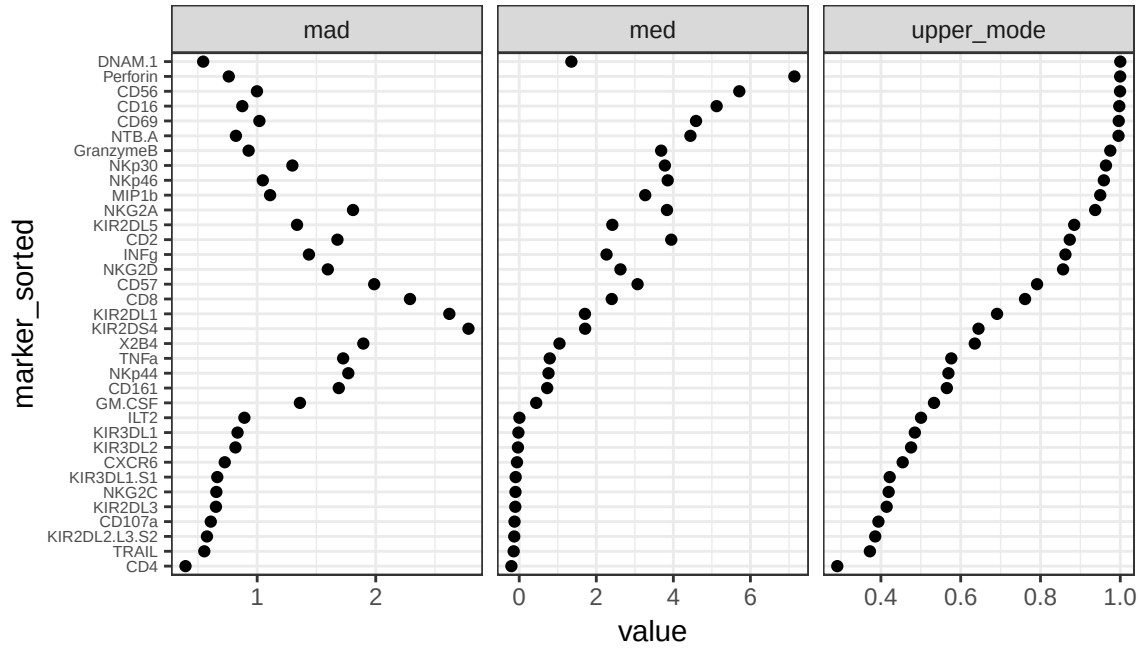
Side note: You can check that the lower modes have exactly the same shape for any pair of markers that have a substantial number of points in the lower mode. I believe it is because the variation in the lower mode is artificial: the number that came off of the machine was 0 (it detected zero of whatever the marker was), but the software used to process the data afterwards adds a bit of noise. This seems silly, but it isn't a big deal if we know about it.

Question 3

```
## The simplest way to create a couple of summaries
## of the data is using group_by and summarise
cytof_summaries = cytof_melted %>%
    group_by(marker) %>%
    summarise(med = median(value),
              upper_mode = mean(value >= 0),
              mad = mad(value))
```

```
## the following code adds a bit to the code above so that
## we can plot more nicely. It's not strictly necessary for
## the problem, but it makes for a nicer figure.
cytof_summaries = cytof_melted %>%
    group_by(marker) %>%
    summarise(med = median(value),
              upper_mode = mean(value >= 0),
              mad = mad(value)) %>%
    # sort the data frame by the value of upper_mode
    arrange(upper_mode) %>%
    # marker_sorted has its levels in order of the values
    # of upper_mode so that they will plot in that order
    mutate(marker_sorted = factor(marker, levels = marker)) %>%
    # we gather all the summaries into one variable so we
    # can plot them in different facets
    gather(med, upper_mode, mad, key = type, value = value)
```

```
ggplot(cytof_summaries) +
    geom_point(aes(y = marker_sorted, x = value)) +
    facet_wrap(~ type, scales = "free_x") +
    theme(axis.text.y = element_text(size=6))
```



We have plotted median absolute deviations (MADs), medians, and an approximation of the fraction of points that are in the upper mode for each marker. In some ways these are more useful than the full histograms or density estimates: we can identify quickly which markers are always on, or have fraction of points in the upper mode equal to 1. These are DNAM1, perforin,

CD56, CD16, CD69, and NTBA.

We see that the fraction of cells in the upper mode goes a long way toward determining both the center, as measured by the median, and the spread, as determined by the MAD, of the distributions. Notice the shape of the MAD distribution in particular: the MAD is highest for those points with an intermediate fraction of points in the upper mode. There is a good reason for this: the spread can be re-expressed as the between-mode spread plus the within-mode spread, and the between-mode spread will be highest when the points are equally spread between the two modes.

The disadantage of these plots compared with those of the full distributions is that you can't identify things like bimodality. Using these on their own, you're likely to miss things about the data, so they are best used in conjunction with the plots of the full distribution.