Homework 3 Report Part 2

Noah Stafford

Data Diagonistics

Let's look at the two sex variables, which were provided in the complete_f2 and pheno_lipomics datasets. Here is a cross-tabluation of those two variables.

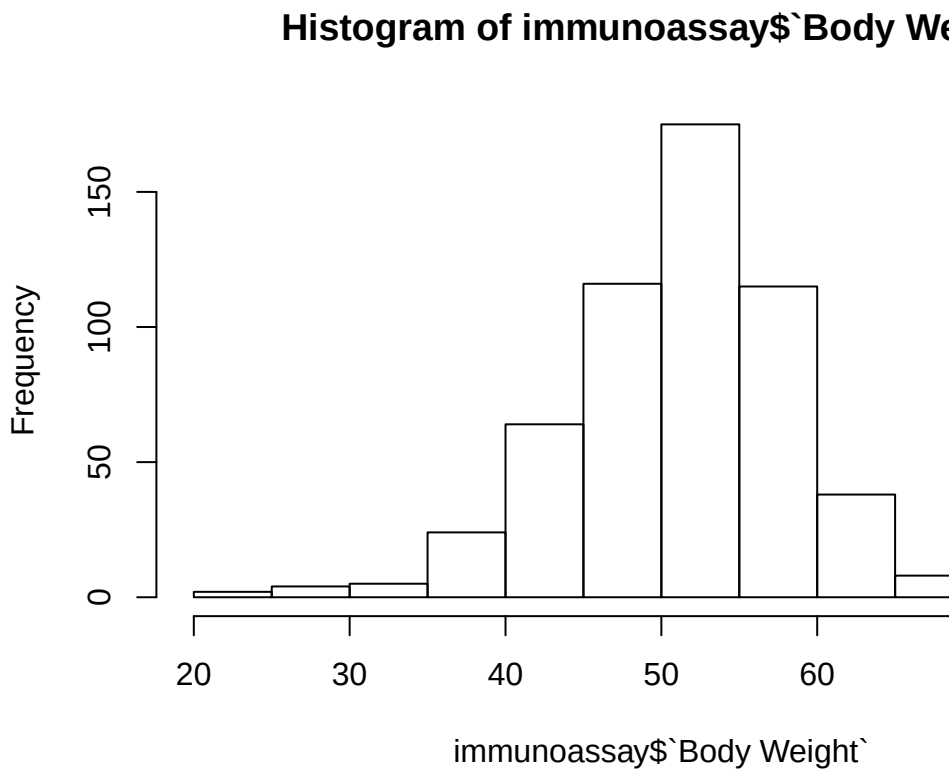
```
##
##       F   M
##   F 261   0
##   M   2 290
```

We see there are two mice that have inconsistent labels.

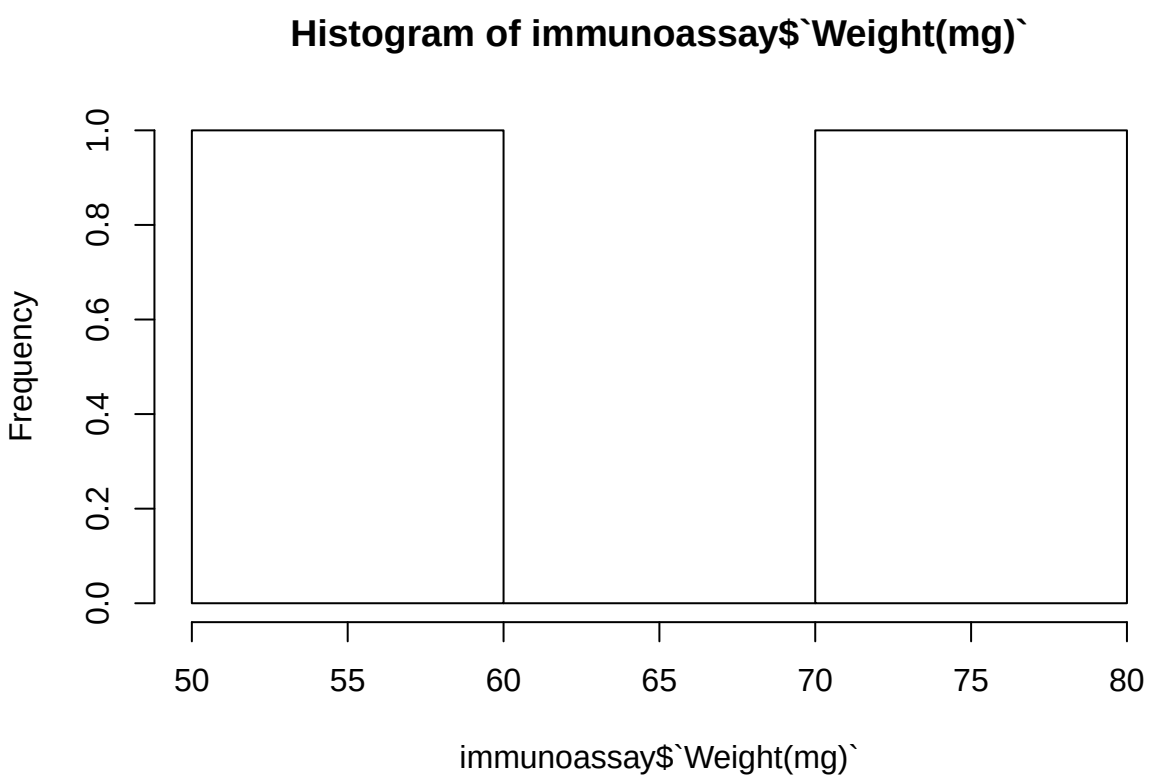
```
##   mouse_id sex_f2 sex_pl
## 1     3259      M      F
## 2     3260      M      F
```

From the data, we can see these mice are two mice with sequential mouse ids, pointing towards a data entry error.

I tried to compare the two general 'weight' variables, but it appears Weight(mg) is one of two values for all of the



mice, and is therefore mostly uninformative




```
## # A tibble: 3 x 4
##   `Weight(mg)` `mean(Body Weight)` `sd(Body Weight)`     n
##          <dbl>               <dbl>             <dbl> <int>
## 1         56.4                53.1            NaN        1
## 2         70.8                54              NaN        1
## 3         NA                  51.1              7.19   551
```

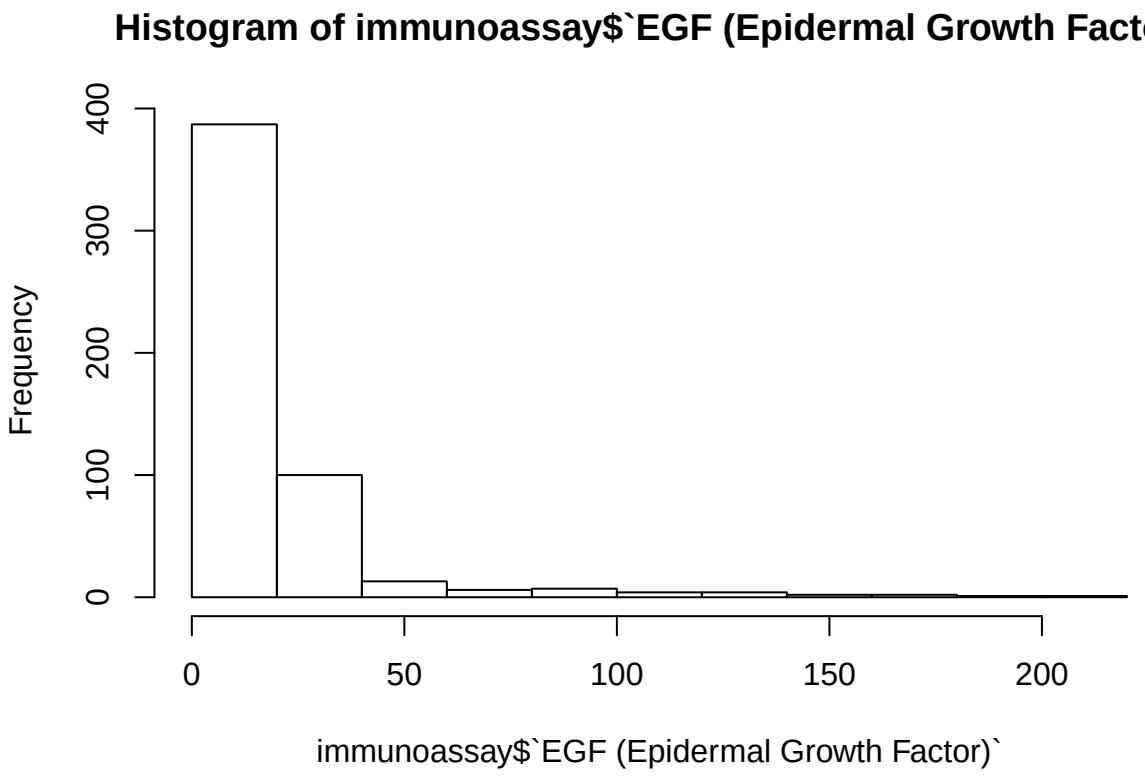
The truth is even more simple – there are only two non-NA observations for Weight(mg).

There are a couple of time series in the data. These can be looked at to see if their observations make sense. First, look at Body Length. From this table it appears that there was some change in measurements or units between week 8 and week 10. Mice should not be getting 80 cm shorter in two weeks.

```
## # A tibble: 6 x 4
##   mouse_id length_4_6_wk length_6_8_wk length_8_10_wk
##      <dbl>         <dbl>         <dbl>          <dbl>
## 1     3045            11             4         -80.2
## 2     3046            12             1         -80.5
## 3     3047            11             6         -81.3
## 4     3049            10             2          NA
## 5     3050            11             7         -88.3
## 6     3051            11             5         -86.3
```

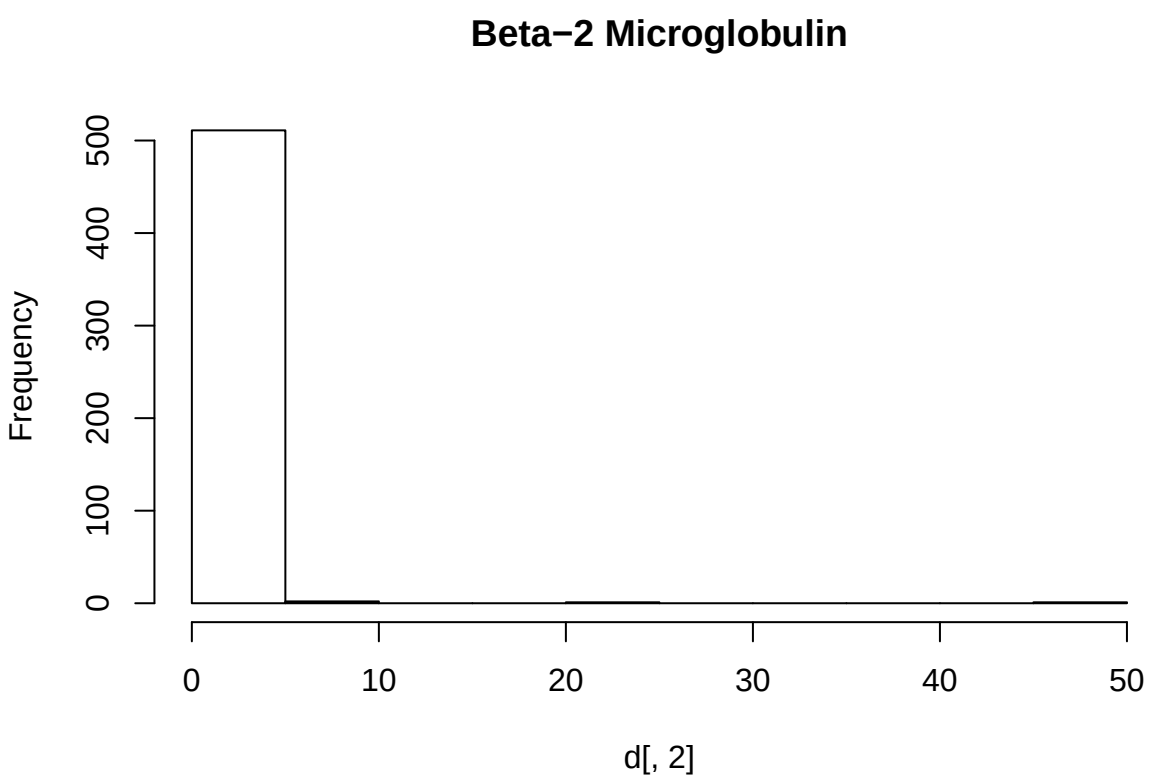
Now, let's check for extreme outliers.

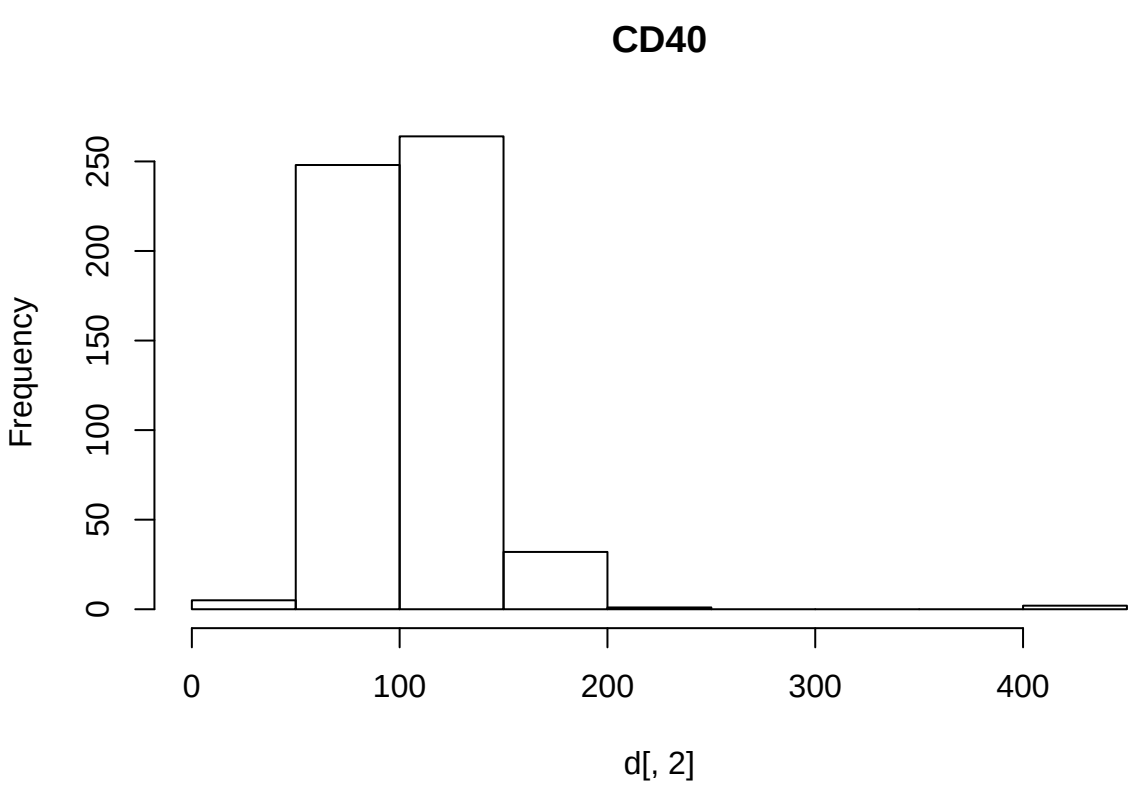
My first instinct was to apply a threshold, for example 2*IQR, and simply report observations that lay outside of that range. The issue I ran into, is that many of these variables have very skewed distributions. A variable such as EGF (Epidermal Growth Factor) will always report a lot of outliers at any threshold based on variance or inter-quartile range because it is so left skewed.



This makes any one-size fits all analysis for finding outliers difficult. Ultimately, just looking at a bunch of histograms might not be a horrible option. The main issue as I see it is a lack of domain knowledge about these measurements, and a lack of a collaborator to work with who can validate whether measurements could be as extreme as they are observed to be. Below I plot histograms for all available numeric variables, the totality of which I will spare you, but I'll put up 2 just for posterity.

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(column)` instead of `column` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```


Beta−2 Microglobulin
Frequency
0
100
200
300
400
500
0
10
20
30
40
50
d[, 2]



Visual inspection reveals a large amount of variables with potential outliers. I am unsure if this is characteristic of the domain, but I am seeing a lot of variables with a mode of some low value, with a smattering of measuremens an order of magnitude higher. I would say in general that if I were tasked with this project, I would not feel comfortable moving forward with the analysis of this data without a lot more input from a collaborator with a knowledge of these measurements.

From visual inspection of these histograms, variables with clear possible outliers include: Beta-2 Microglobulin, CD40, CD40 Ligand, GST-alpha, IL-11 (Interleukin-11), IL-6 (Interleukin-6), KC/GROalpha, MIP-1beta, MIP-2, vWF, Clusterin, Fibrinogen, Growth Hormone, IP-10, Osteopontin, SCF, TIMP-1, VEGF, and 10wk Insulin.

Some variables don't have very many observations. Let's make sure to note which they are.

```
##                                                                col_name
## 1                                                          Weight(mg)
## 2                                                 IL-3 (Interleukin-3)
## 3                                                       Growth Hormone
## 4                                               IL-17 (Interleukin-17)
## 5  GM-CSF (Granulocyte Macrophage-Colony Stimulating Factor)
## 6                                                  NGAL (Lipocalin-2)
## 7                                          IFN-gamma (Interferon-gamma)
## 8                          GST-alpha (Glutathione S-Transferase alpha)
## 9                                                 IL-7 (Interleukin-7)
## 10                                              IL-11 (Interleukin-11)
## 11                          TNF-alpha (Tumor Necrosis Factor-alpha)
## 12                                      IL-12p70 (Interleukin-12p70)
## 13                                              IL-2 (Interleukin-2)
## 14                                  FGF-9 (Fibroblast Growth Factor-9)
```

```
## 15                                                    IL-4 (Interleukin-4)
## 16                                                    IL-6 (Interleukin-6)
## 17 RANTES (Regulation Upon Activation, Normal T-Cell Expressed and Secreted)
## 18                                              vWF (von Willebrand Factor)
## 19                                                          Endothelin-1
## 20                        MIP-1beta (Macrophage Inflammatory Protein-1beta)
##    obs_rows
## 1         2
## 2        16
## 3        38
## 4        56
## 5        78
## 6        91
## 7        98
## 8       106
## 9       120
## 10      131
## 11      131
## 12      136
## 13      151
## 14      161
## 15      202
## 16      210
## 17      280
## 18      480
## 19      486
## 20      486
```

This table has the amount of observed values for each column. Sorting by these counts, notice that 18 variables have observed values for less than half of the observations. This is not necessarily bad – we can also see that the 19th lowest variable has 480 out of 553 rows observed, meaning out of the 123 variables, only 18 have large amounts of missing data.