Exploratory Analysis

Section 1 – Cleaning Data:

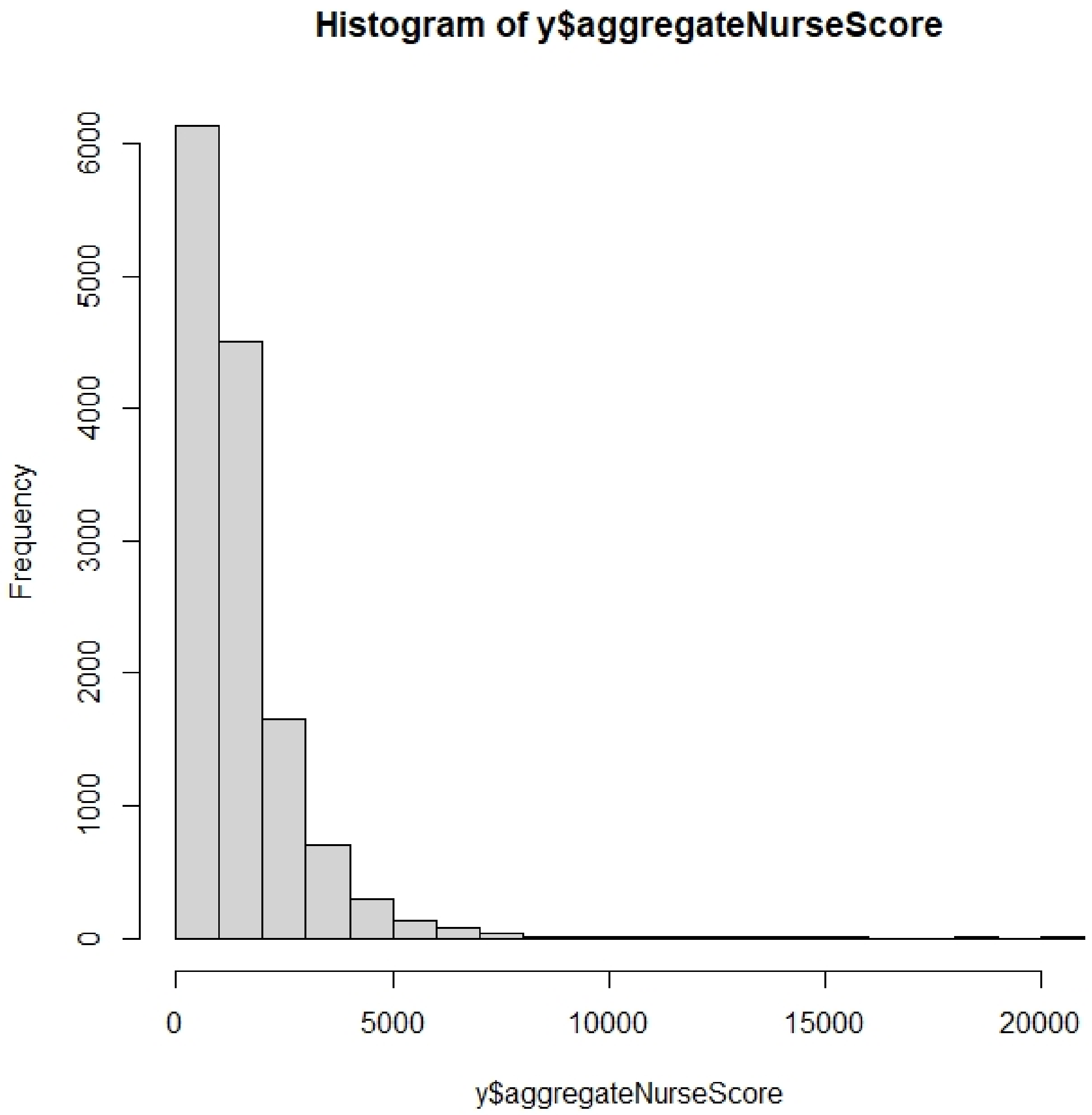
My data was originally in many parts, as ProPublica had collected the many aspects and provided the data they collected not the data after they processed it. Due to the data being in many parts, the main goal of the cleaning was to combine all the data into one R data frame object.

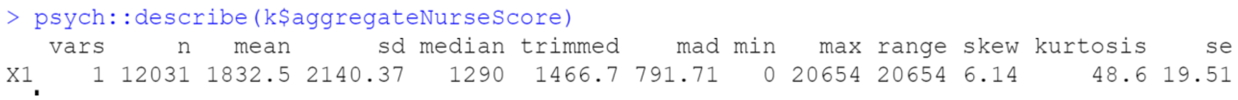
There were a number of values in the facilities data where the score was included as a string. An example being, above national average. To deal with these, I simply removed them. I chose this strategy because the final regression will use the facility scores on a continual basis to predict whether or not the nurses at the facility will be above or below average. Only knowing if it is above or below or at the national average would make the final prediction less accurate, as above average can mean many things. It would be impossible to tell from that description if it is in the 60th percentile or 99th. The nurse scores were either not included or an actual number. The NA values were removed as well, as there is no information about if they were above or below average.

Another aspect of this is the “creation” of data from the datasets. By creation I do not mean balancing, I mean aggregating the nurse scores for each facility. To accomplish this I used a nested loop, which while effective had a large run time. The number of nurses is 280970 and the number of facilities is 13572, making the number of loops run 3.81*10^9. Them I utilized the merge function to combine the R data frame object with the facility names and facility score and the R data frame object with the aggregate nurse score for each facility.

Section 2 – Summary Statistics:

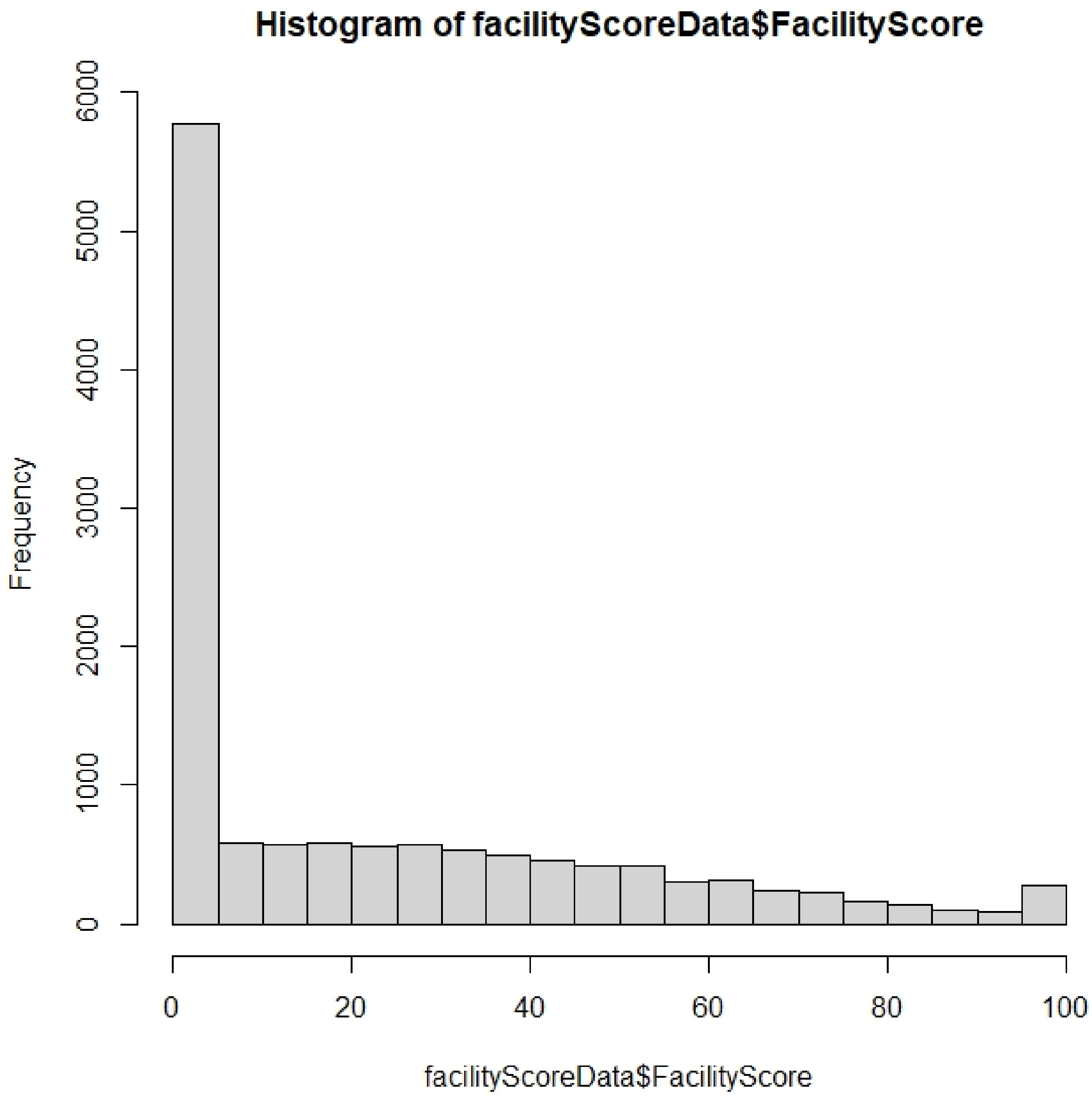
I started by looking at the descriptive statistics for the facility score and aggregate nurse score for each facility separately. To do this I utilized the R package psych, specifically the describe() function included in it. This returns the number of elements, the number of variables, the number of observations, the mean, the standard deviation, the median, trimmed mean, mean absolute deviation, min, max, range, skew, kurtosis and standard error. The values for both the facility score and aggregate nurse scores are included below. I am also including histograms of both datasets to use in analysis with the descriptive statistics.

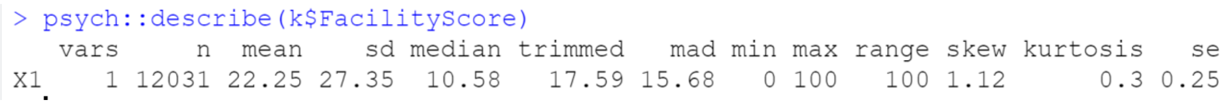




```
> psych::describe(k$aggregateNurseScore)
   vars     n   mean      sd median trimmed    mad min   max range skew kurtosis    se
X1    1 12031 1832.5 2140.37   1290  1466.7 791.71   0 20654 20654 6.14     48.6 19.51
```

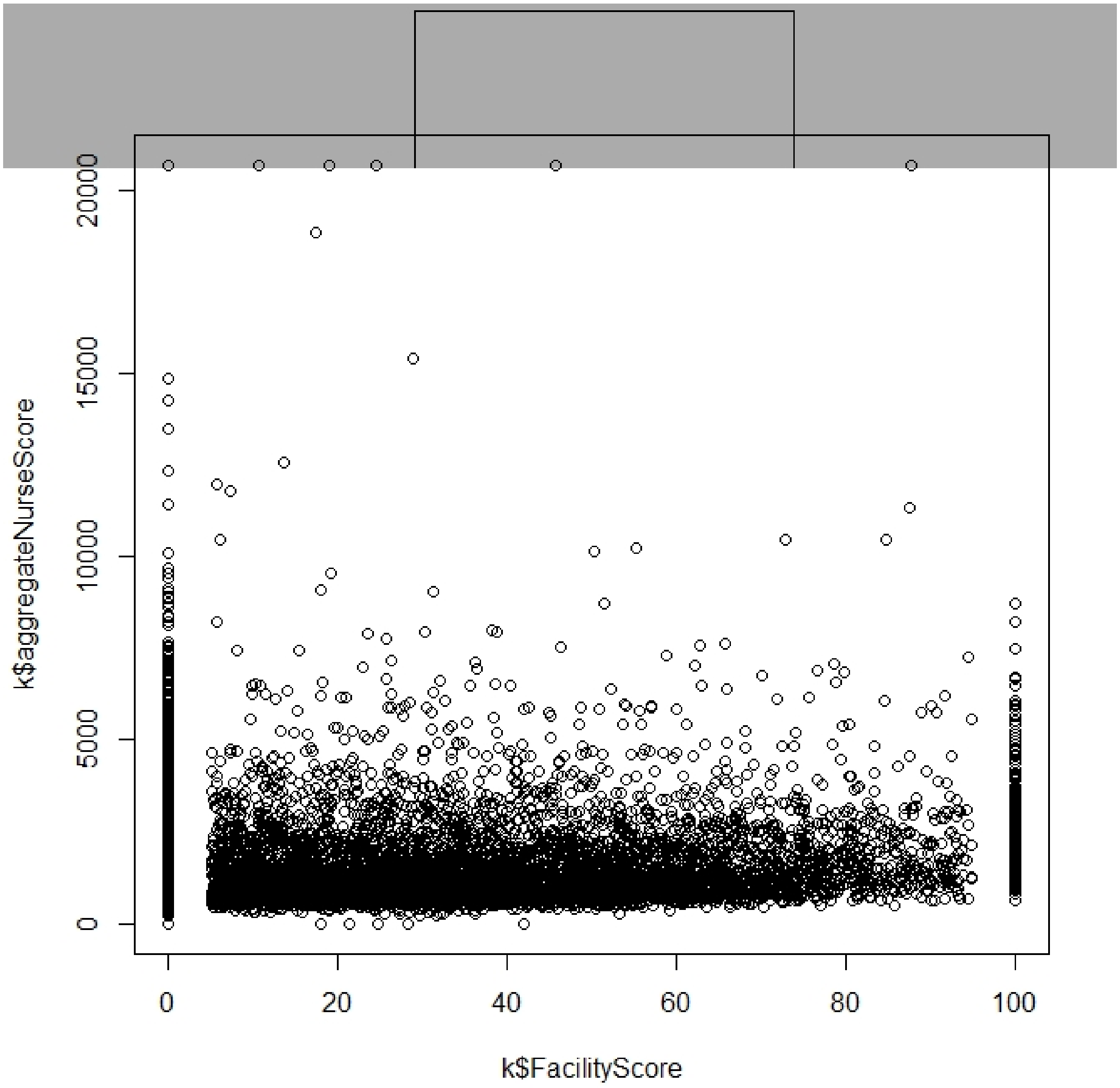
The above output is describing the aggregate nurse scores. By looking at the mean, median and histogram we can see that there are some much higher scores that are pulling the mean higher than the median. This shows how much better a few facilities are in terms of nurses than others. An interesting portion of this data is the min being 0. The facilities that have an aggregate nurse score of 0 have data on at least on nurse, and all nurses recorded are given a score of zero. This is interesting to me in that it seems odd that some nurses would be score at zero, but given that the binary logistic regression, which is our end goal, predicts binary classification it is only important that our facilities are rated either good or bad. To classify as either or good or bad I compared the nurse score to the median, and ranked the scores above or equal to good and the scores below as bad. This creates a balanced distribution and deals with the problem of outliers skewing the data.



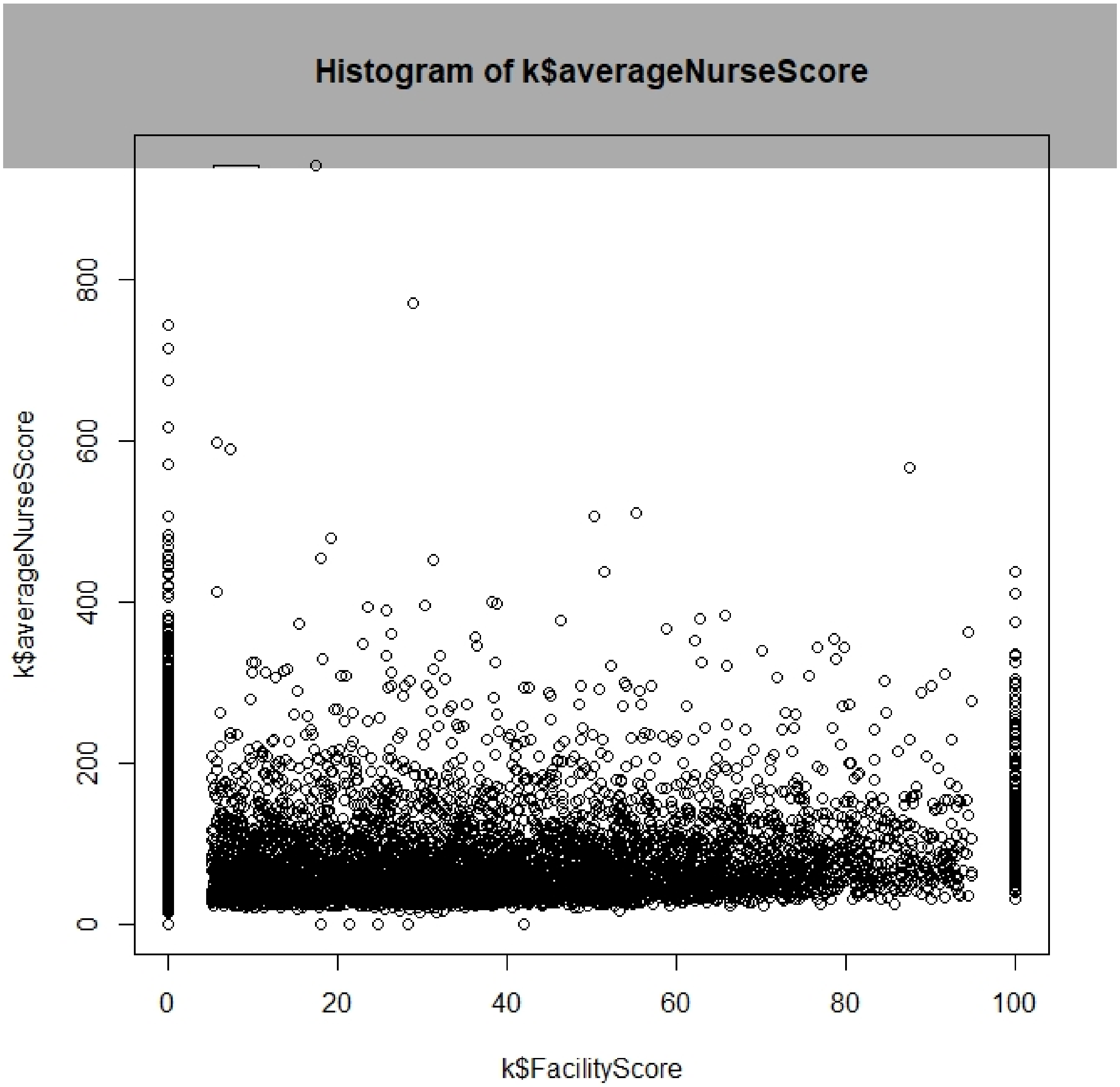


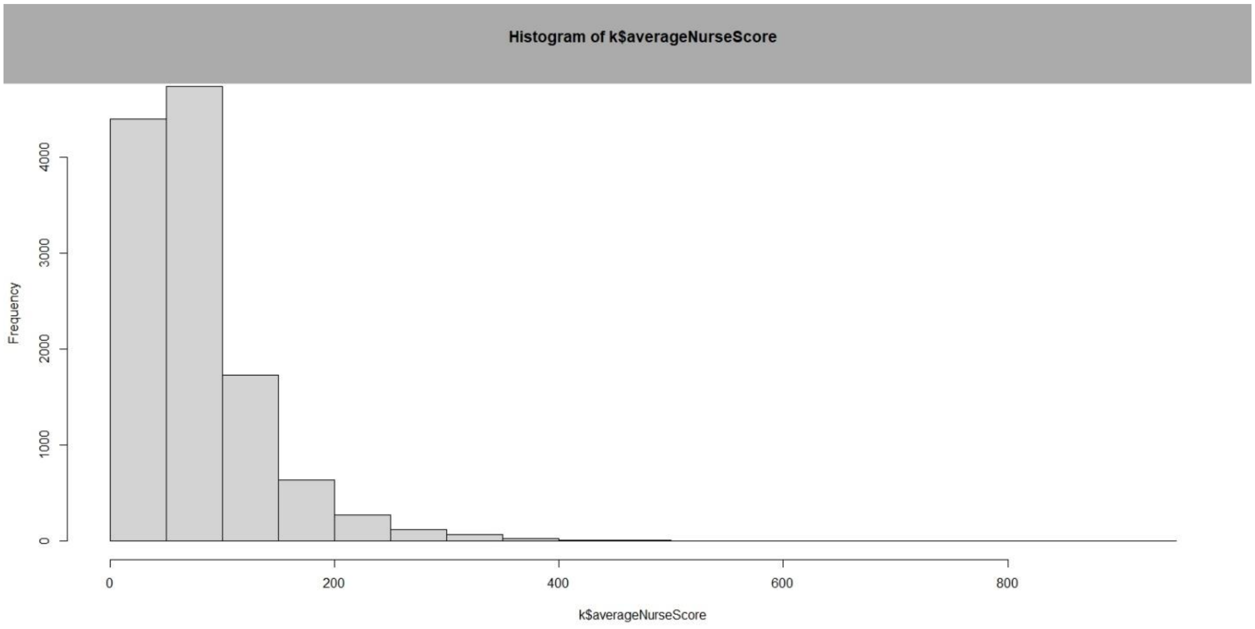
```
> psych::describe(k$FacilityScore)
   vars     n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 12031 22.25 27.35  10.58   17.59 15.68   0 100   100 1.12      0.3 0.25
```

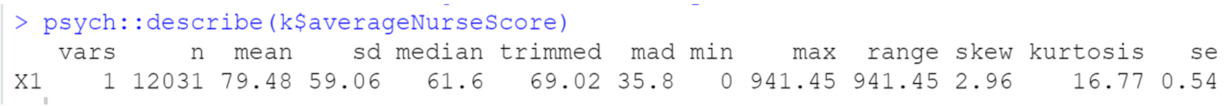
Looking at the descriptive statistics and the histogram we see an interesting distribution. There are a large number of facilities with ratings of zero, and this could potentially pose a problem. Unlike the nursing scores, these are not a binary classification and use their actual value in the prediction. Because of this, the reason for a score of zero is important. If the rating is zero because the facility is bad then the score is appropriate, however, if the score is zero because the report was filled out that way because it is easy, then the model will be less accurate because of it. Some descriptive statistics that embody this issue are the mean and median together. The median is substantially larger than the median because of the skewed distribution. However, this is not caused by large outliers, but rather many zeros.



The aggregate nurse score scatter plot above is showing that the points are very uniform and that making an accurate predictive model will be very difficult if not impossible from this data. To try and make the data more suited for this, I am going to look into average nurse score and then removing the zero values for the facility score. The average nurse score should remove the values that are really high because there are a lot of nurses and the removal of the zero values should help the future model by removing the questions that values with that facility score present.



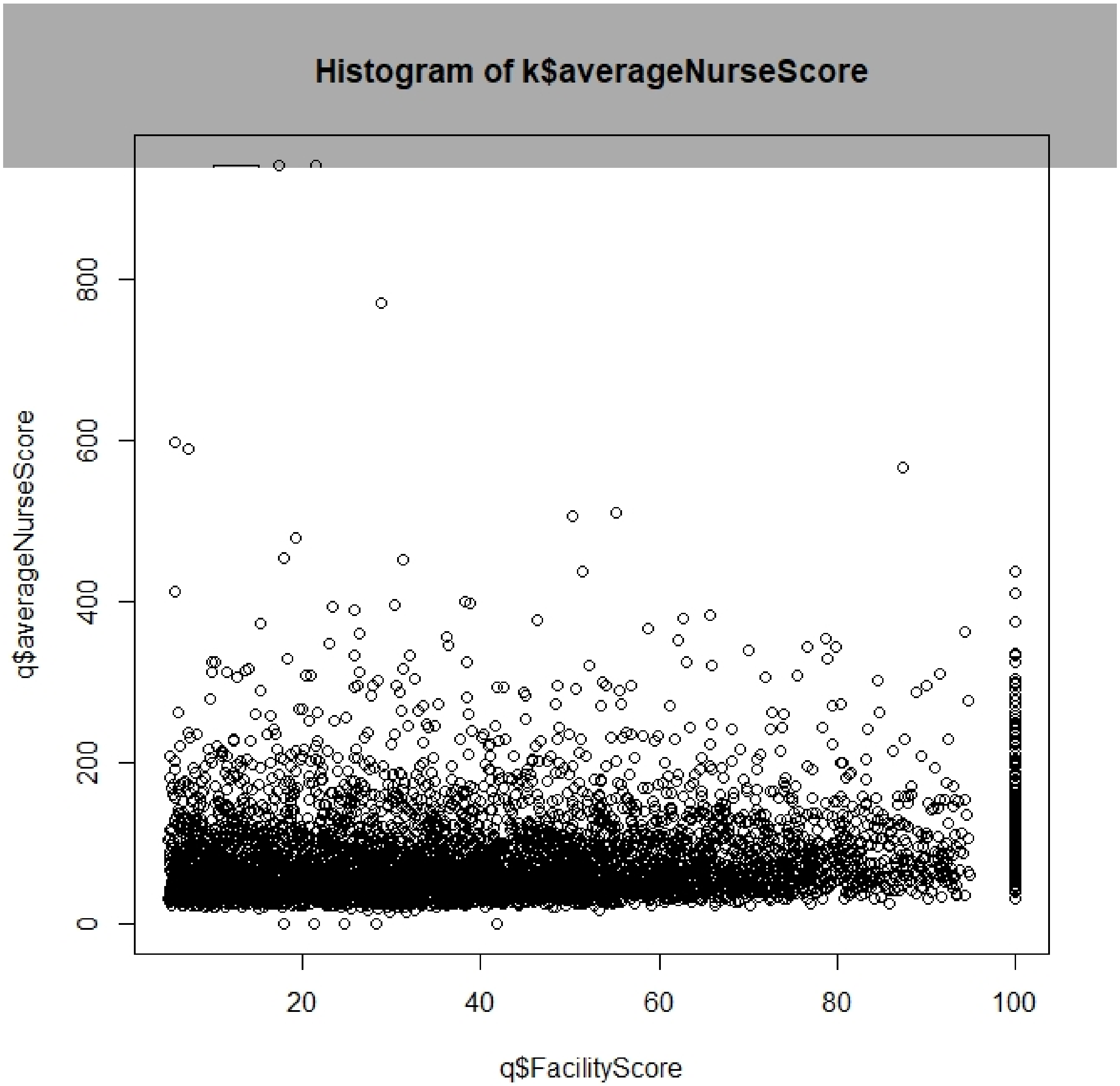


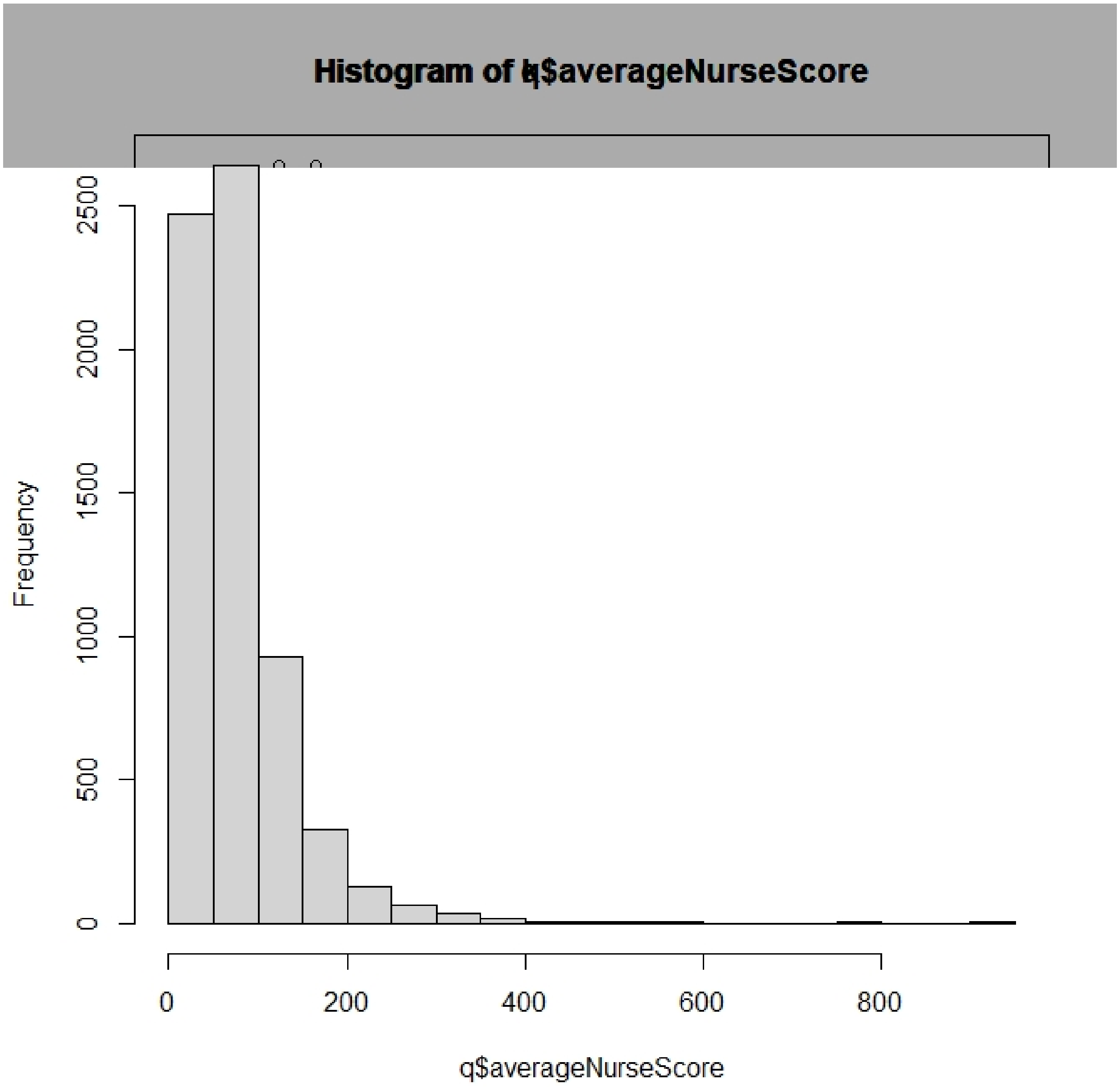


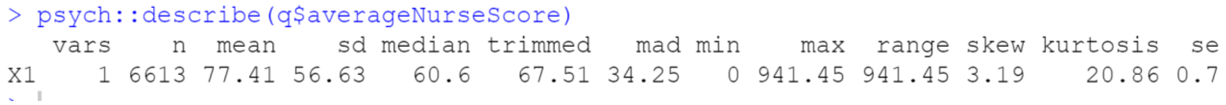
```
> psych::describe(k$averageNurseScore)
   vars     n  mean    sd median trimmed  mad min    max  range skew kurtosis   se
X1    1 12031 79.48 59.06   61.6   69.02 35.8   0 941.45 941.45 2.96    16.77 0.54
```

The three pictures above capture the data from average nurse scores for each facility with the facilities rated at zero included. This performance is better than the aggregate data sets performance in that the distribution of average scores is closer to a normal distribution than the aggregate scores were and the highest scores are closer to the average. However, even with the averages instead of the aggregate scores for the facilities, the scatter plot is still unclear about the trend. The average nurse scores for the facilities is trending flat, giving the idea that the score of the facilities is not an indicator of the quality of nurses. If one looks very closely at the bottom of the scatter plot, there appears a slightly positive sloped line, but given that there is a wide variation in scores there does not appear to be a correlation between the quality of a facility and the nurses there.

I looked at this with the average nurse scores again, but without the facility scores that were zero. The findings were not very different from the previous, but the descriptive statistics were different enough. The scatterplot shows essentially the same thing, as the points where the facility score is zero is removed but he rest is the exact same. The visuals are below.


Histogram of k$averageNurseScore
q$averageNurseScore
0
200
400
600
800
20
40
60
80
100
q$FacilityScore





```
> psych::describe(q$averageNurseScore)
   vars    n  mean    sd median trimmed   mad min    max  range skew kurtosis  se
X1    1 6613 77.41 56.63   60.6   67.51 34.25   0 941.45 941.45 3.19    20.86 0.7
```