COGYES_Data

David Degnan

11/17/2018

DEMOGRAPHICS Data Set

Outlier concerns: None.

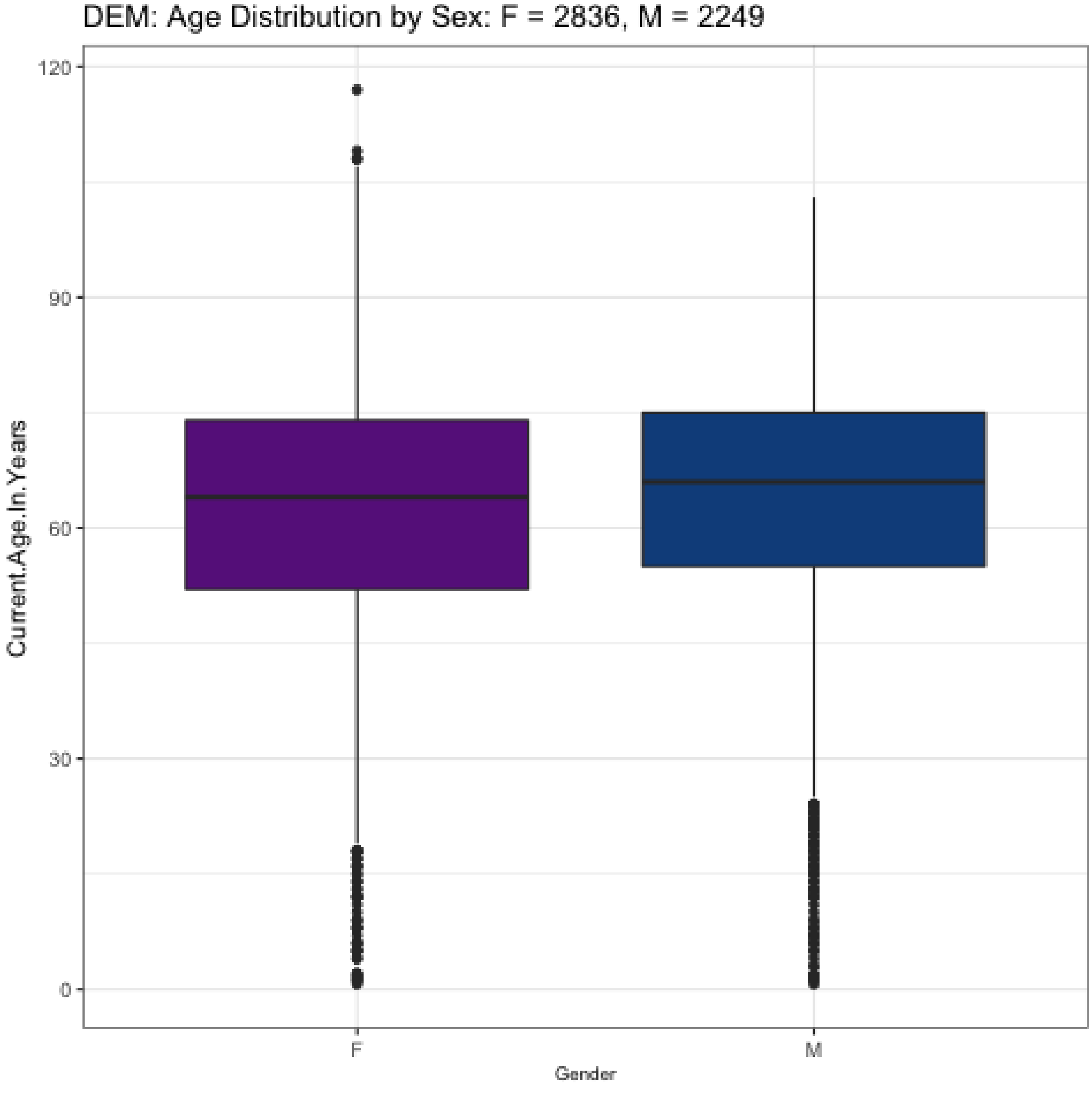

Figure 1: DEMOGRAPHICS: Age distribution split by sex. No outliers to be concerned about. Note more females than males.

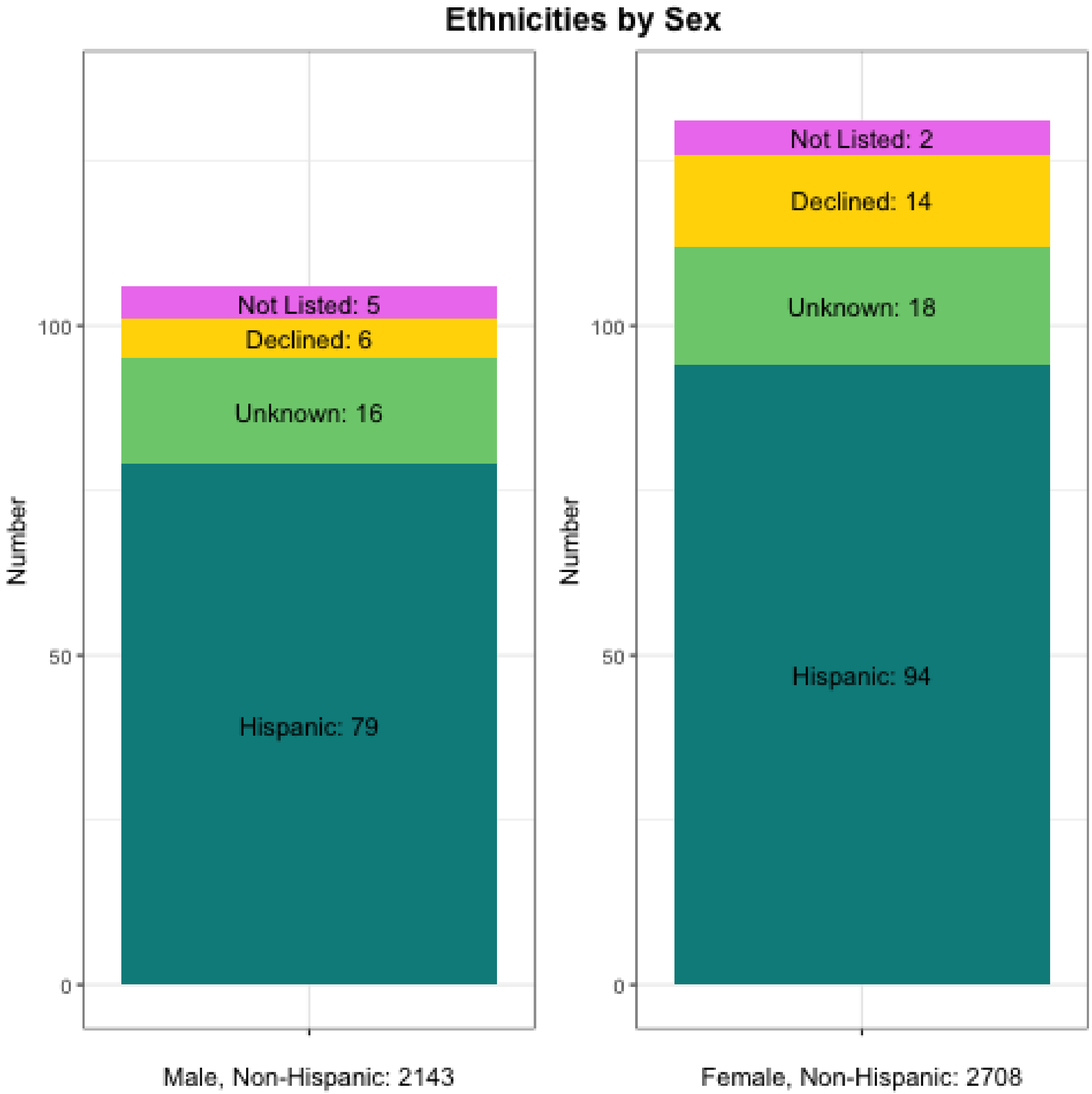

Figure 2: DEMOGRAPHICS: Ethnicity split by sex. Predominantly non-hispanic.

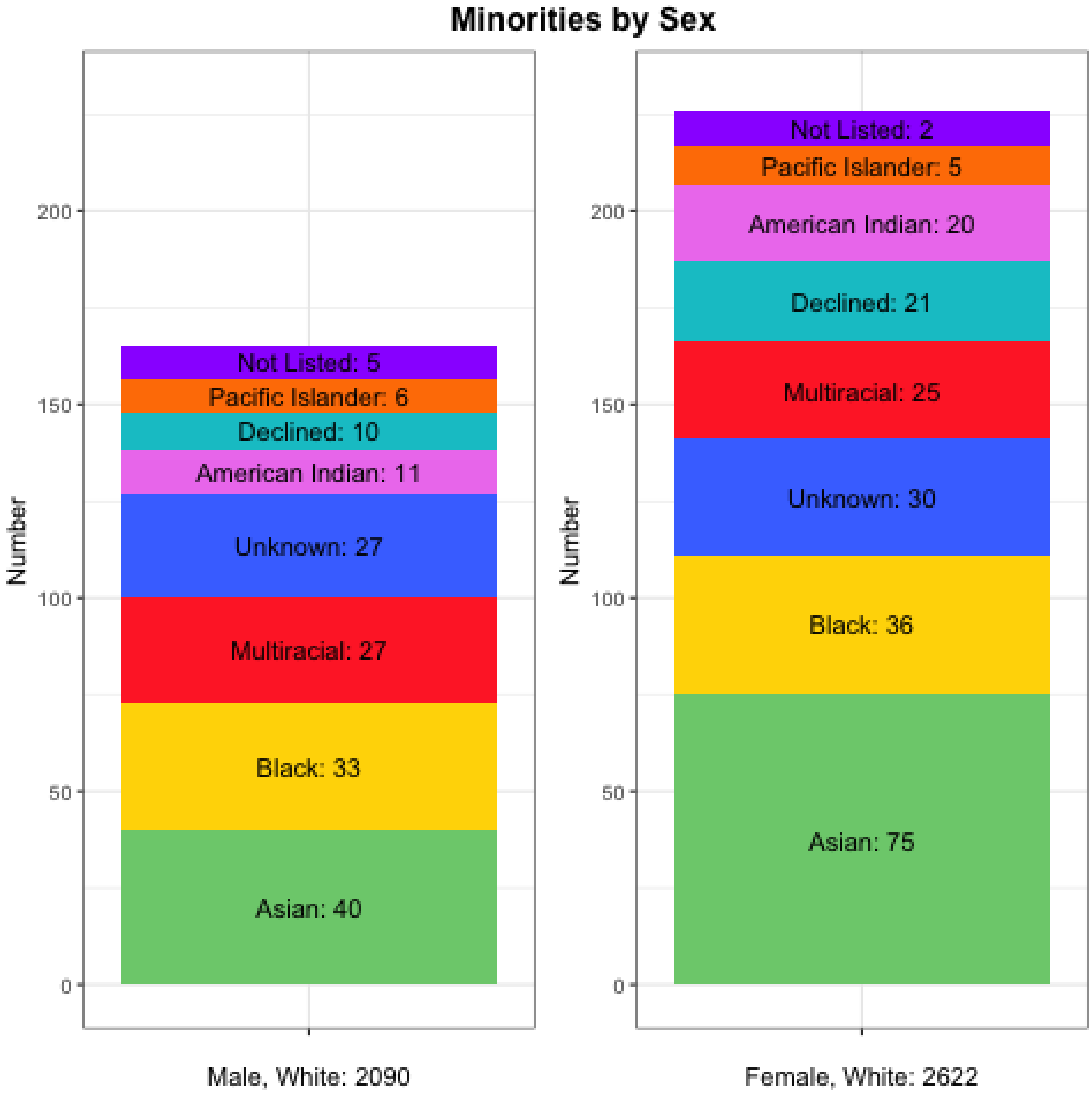

Figure 3: DEMOGRAPHICS: Race split by sex. Predominantly white.

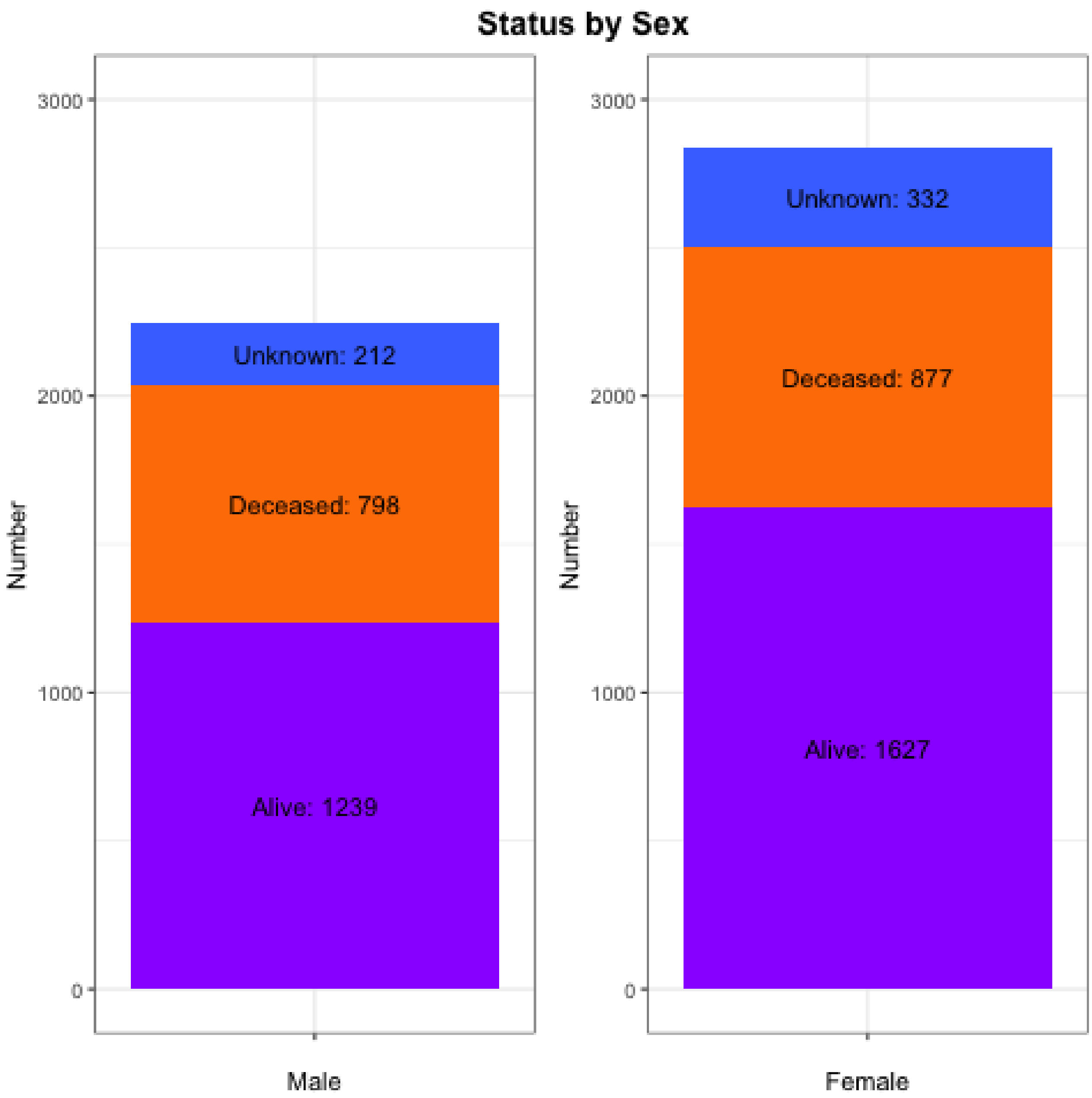

Figure 4: DEMOGRAPHICS: Status split by sex. Predominantly alive.

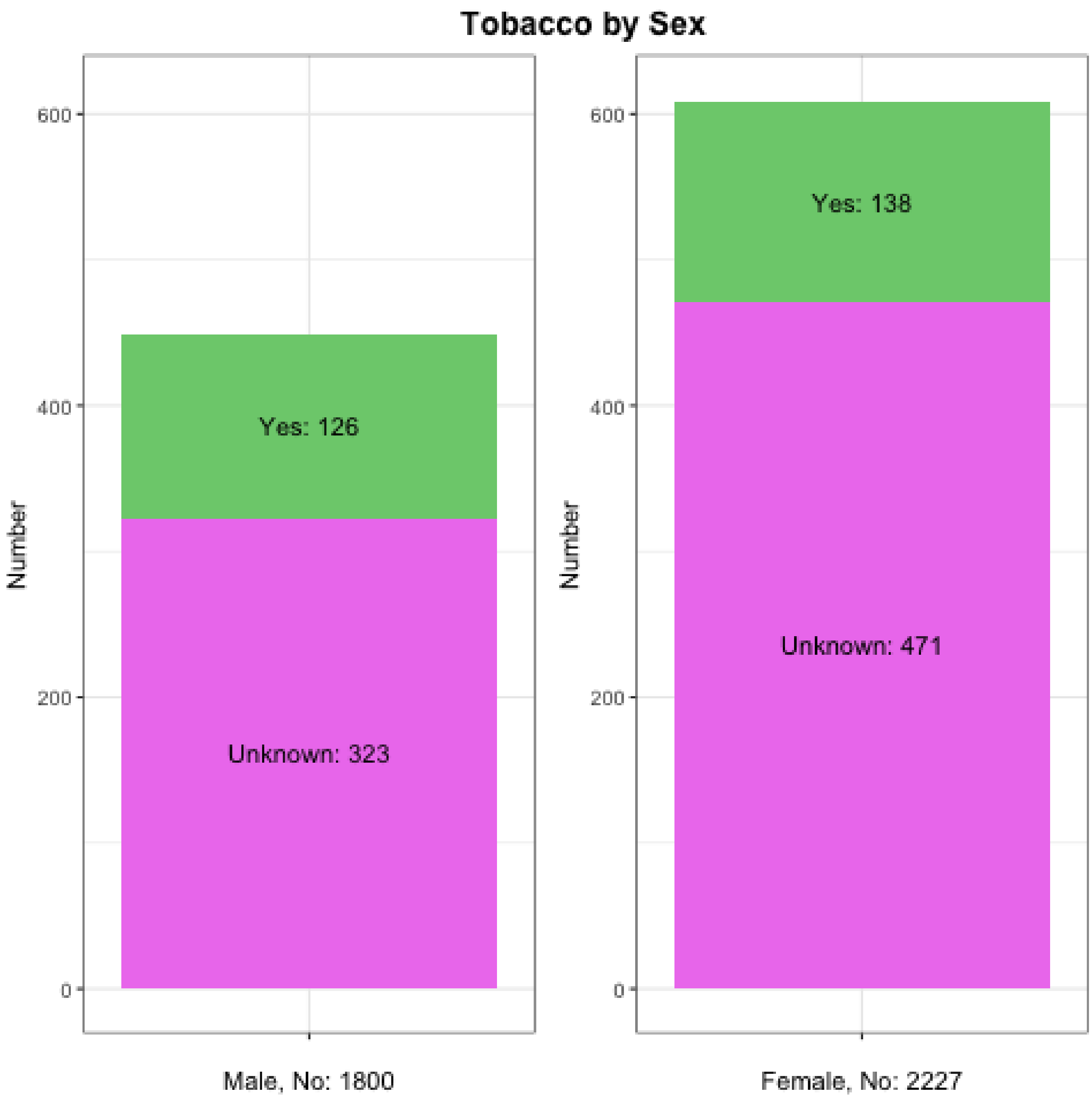

Figure 5: DEMOGRAPHICS: Tobacco usage split by sex. Predominantly “No”.

FLOW Data Set

Outlier concerns: None.

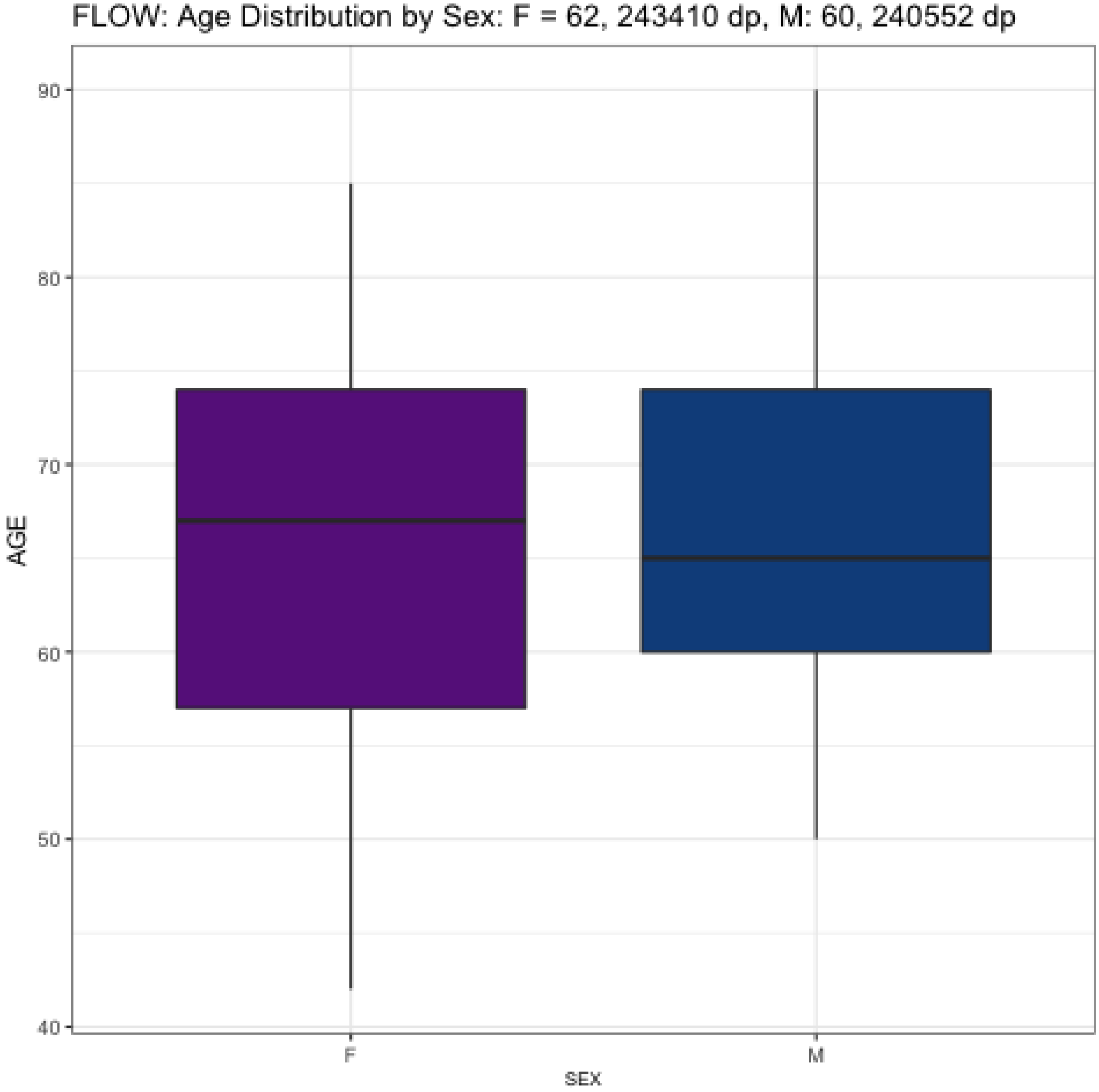

Figure 6: FLOW: Age split by sex. Note that there is significantly less than DEMOGRAPHIC data set.

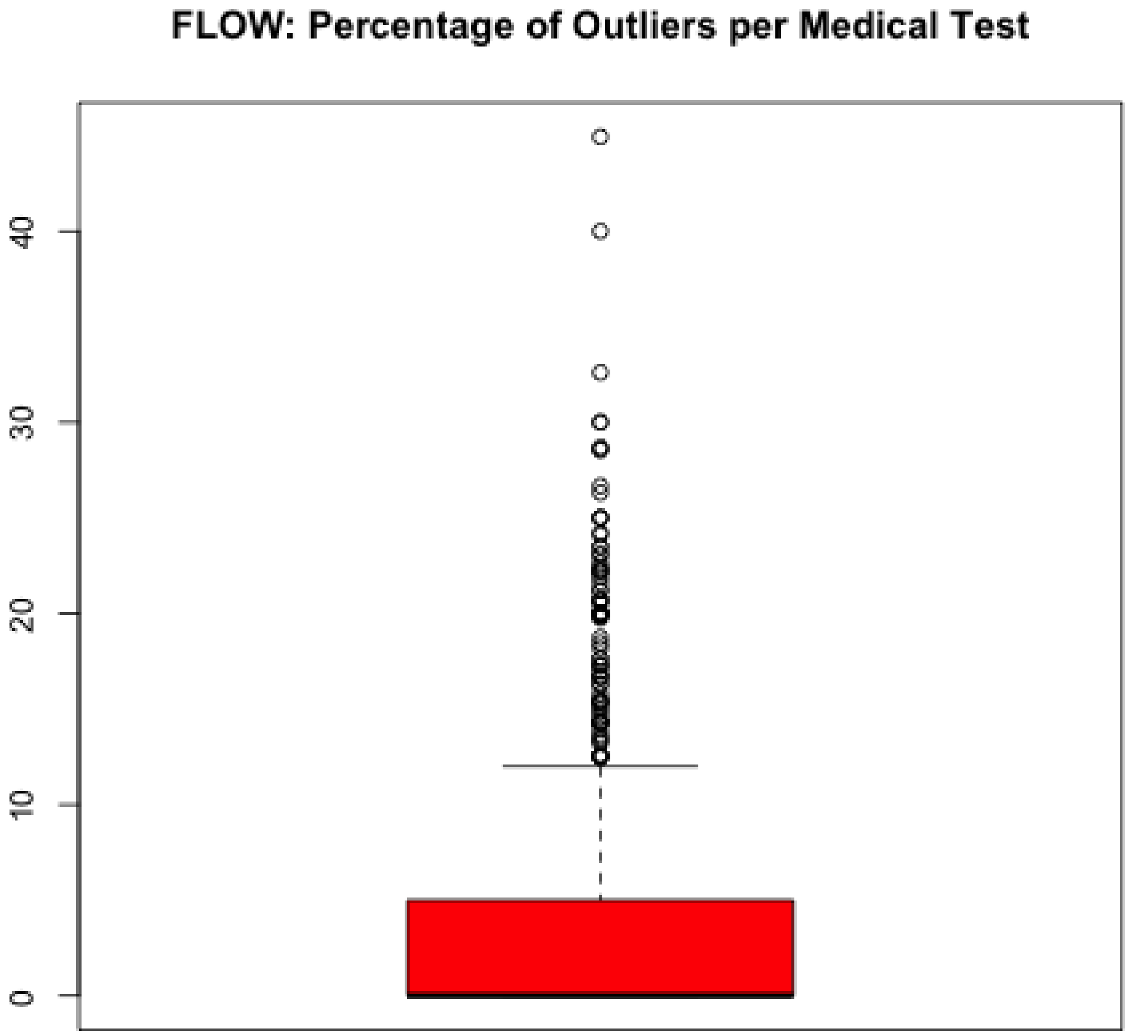

Figure 7: FLOW: Outlier percentage (number of outliers/total number of data points) for each test (every point is a medical test). Note that if we decided to remove outliers, we'd lose on average < 5% of the data. Also note that the largest outliers have a small number of data points.

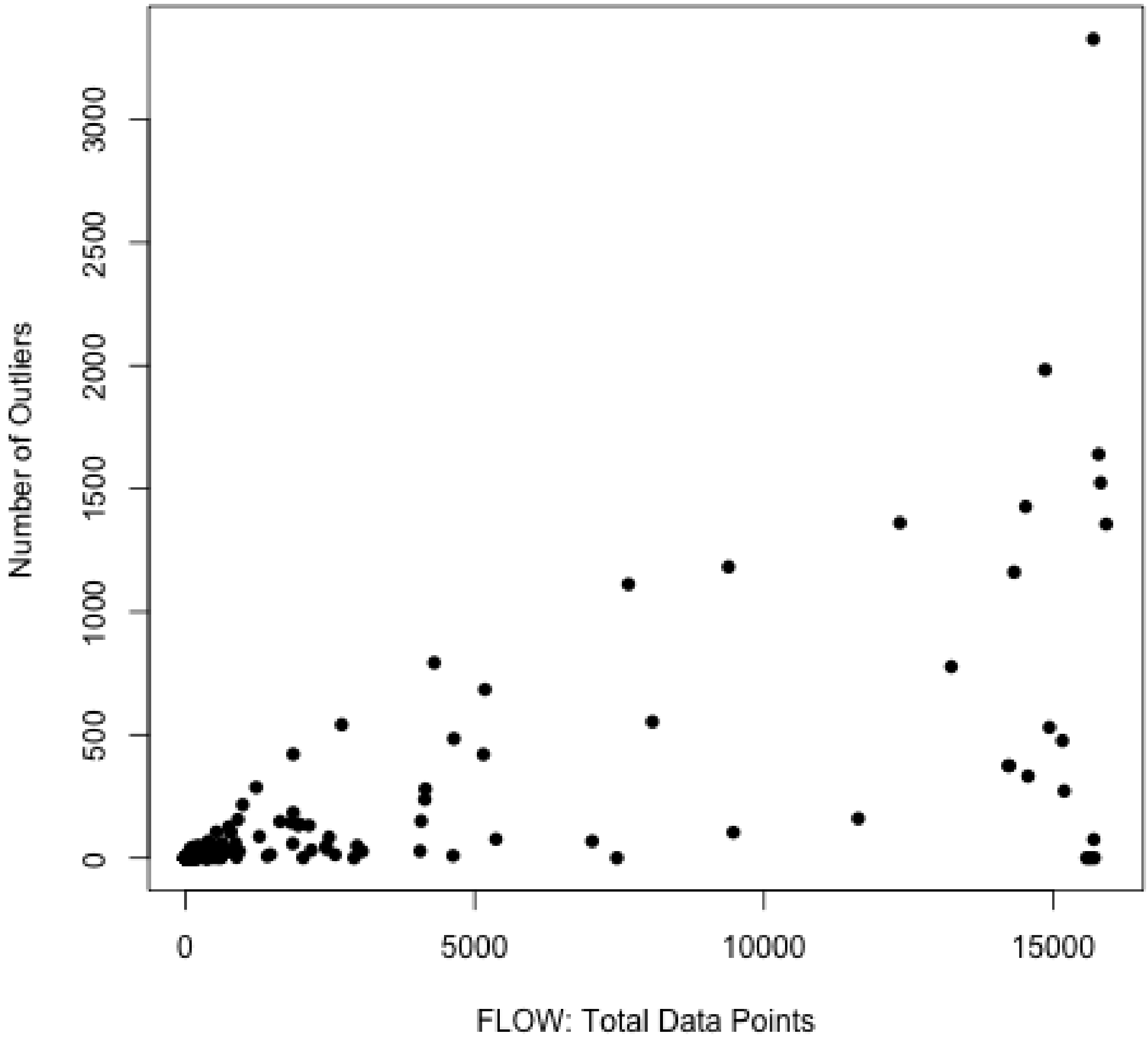

Figure 8: FLOW: As the number of data points increases, so does the number of outliers.

HOSPITALIZATIONS Data Set

Outlier concerns: Yes. Ages > 120.

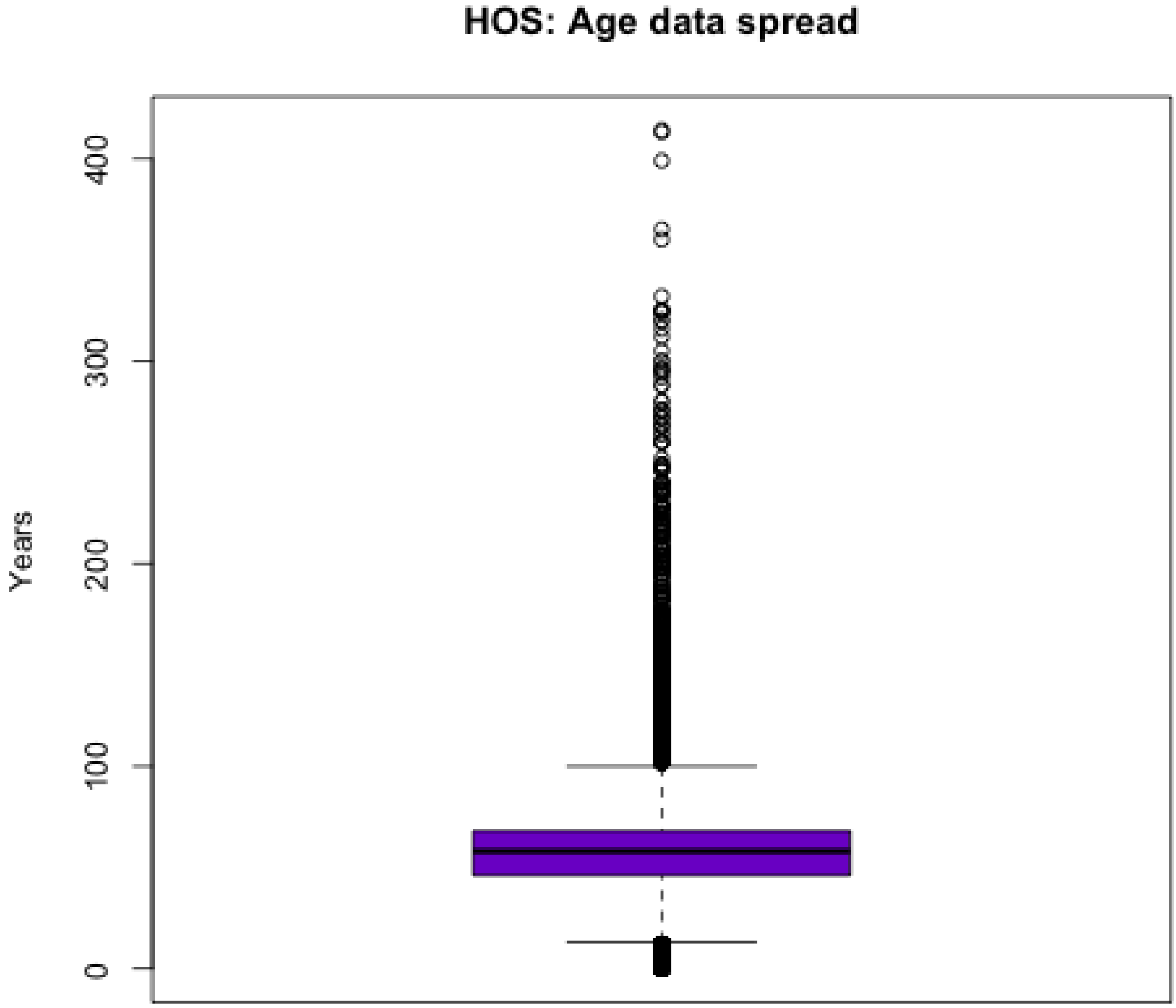

Figure 9: HOS: Note that there are quite a few ages over 120. These outliers will need to be accounted for, and a suggestion is to remove them from the data set. Age already has "NA" in its category.

RESULTS Data Set

Outlier concerns: No.

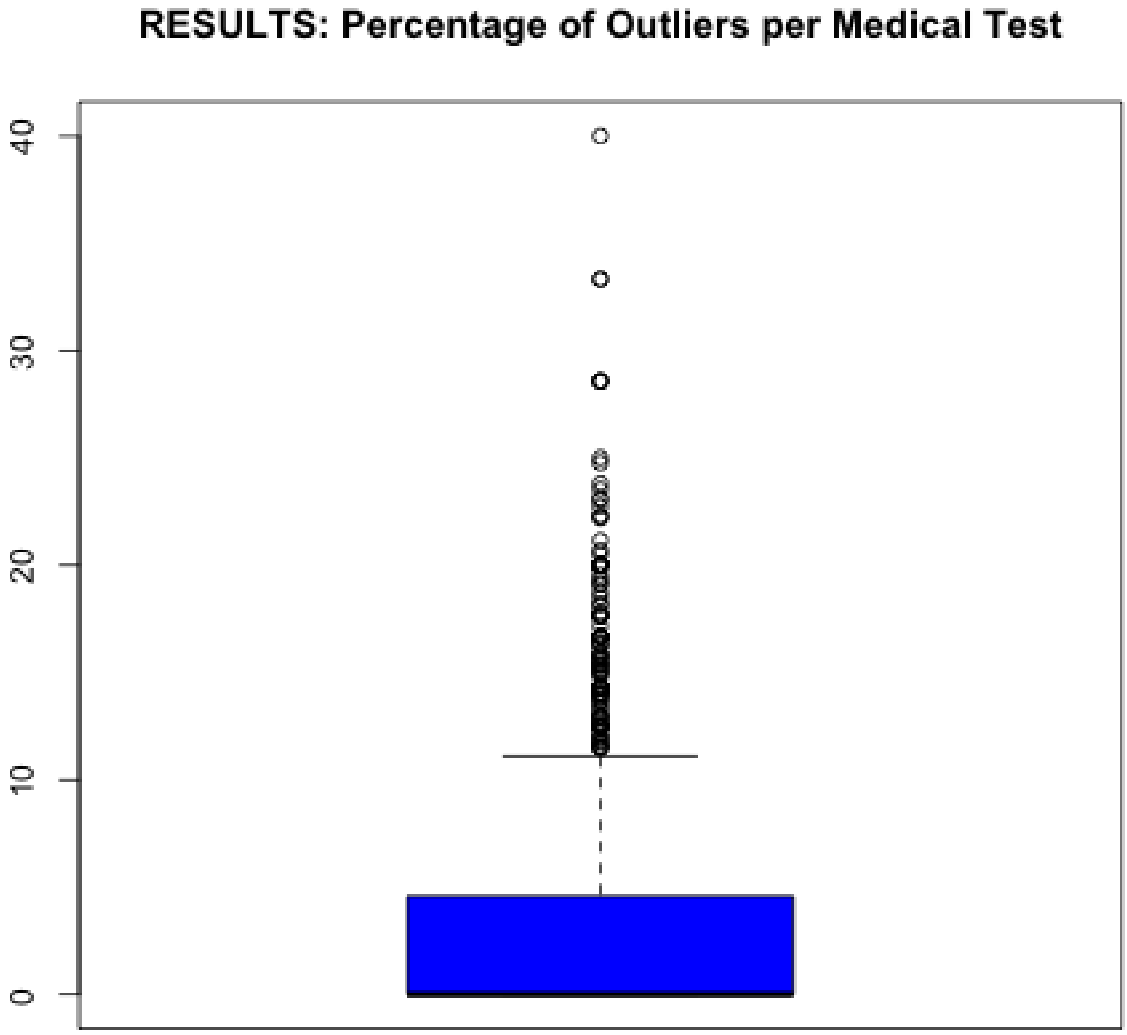

Figure 10: RESULTS: Outlier percentage (number of outliers/total number of data points) for each test (every point is a medical test). Note that if we decided to remove outliers, we'd lose on average < 5% of the data.

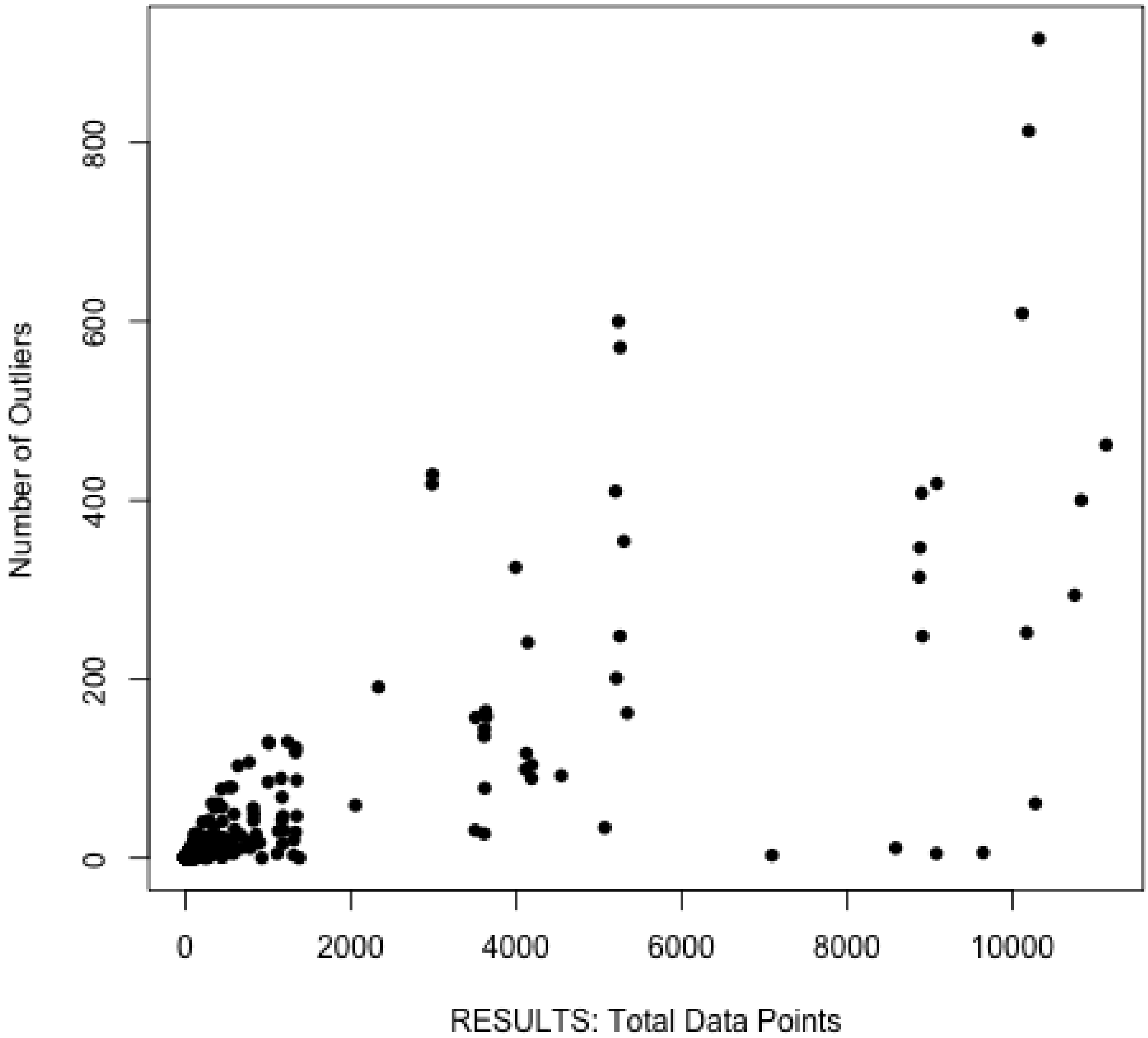

Figure 11: RESULTS: As the number of data points increases, so does the number of outliers.

SURGERY Data Set

Outlier concerns: No.

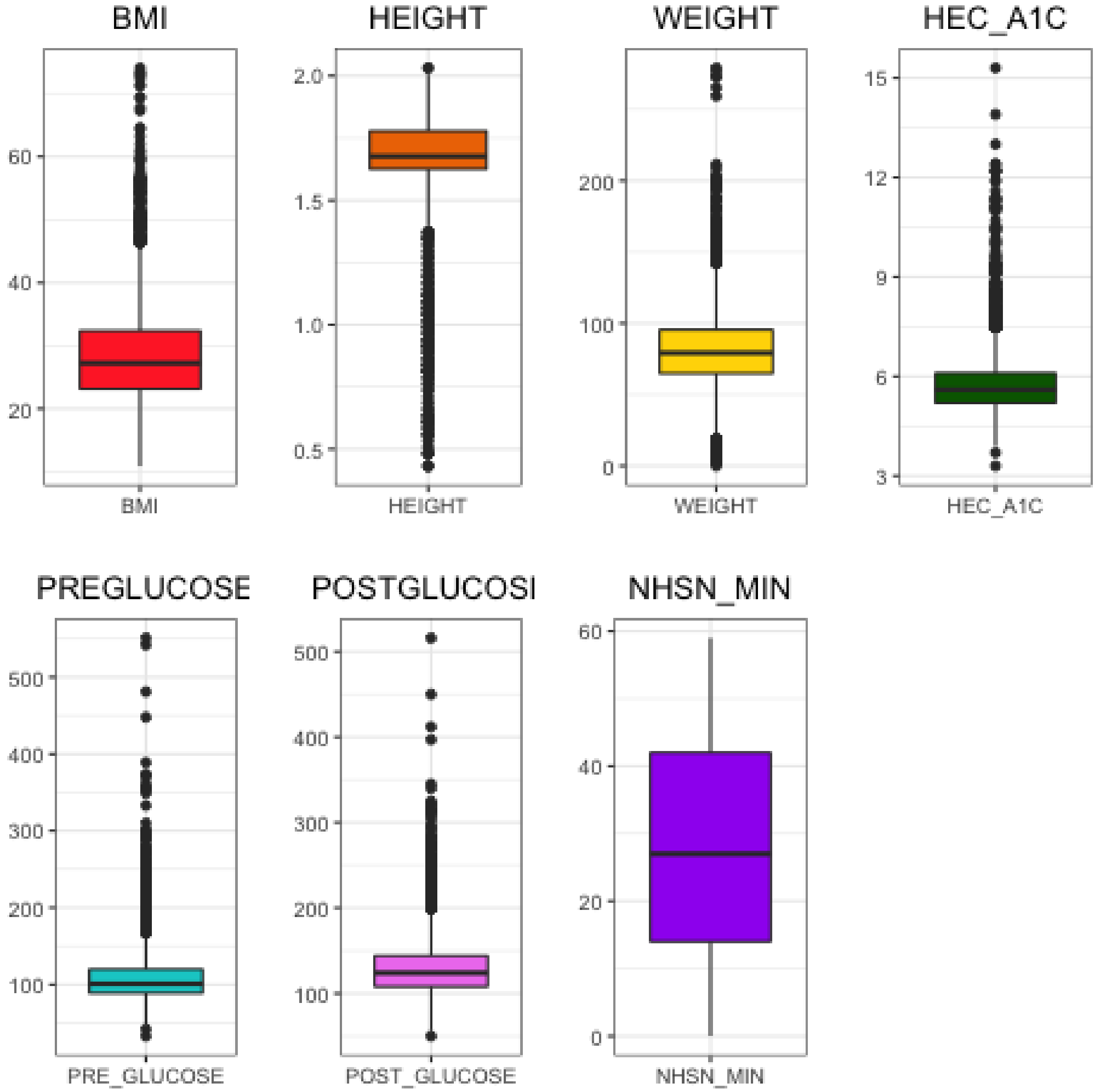

Figure 12: SURGERY: There are high values, but nothing seems unrealistic. Note that height is in meters, weight is in kilograms, and glucose levels are in mg/dL (average is around 100).

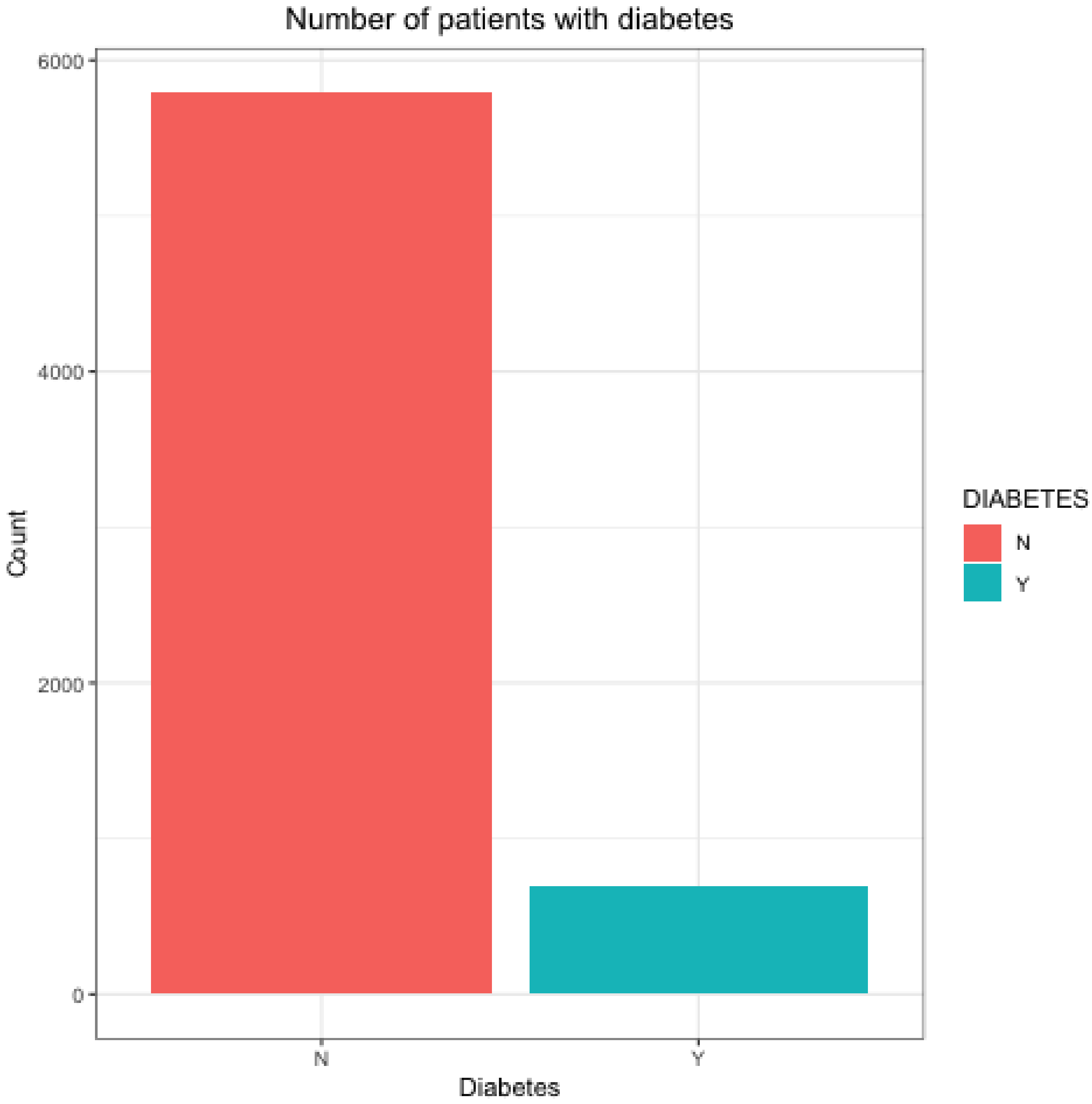

Figure 13: SURGERY: 352 patients in surgery had diabetes, while 2973 did not.

VISITS Data Set

Outlier concerns: Yes. BMI > 100. Height (in) > 118.

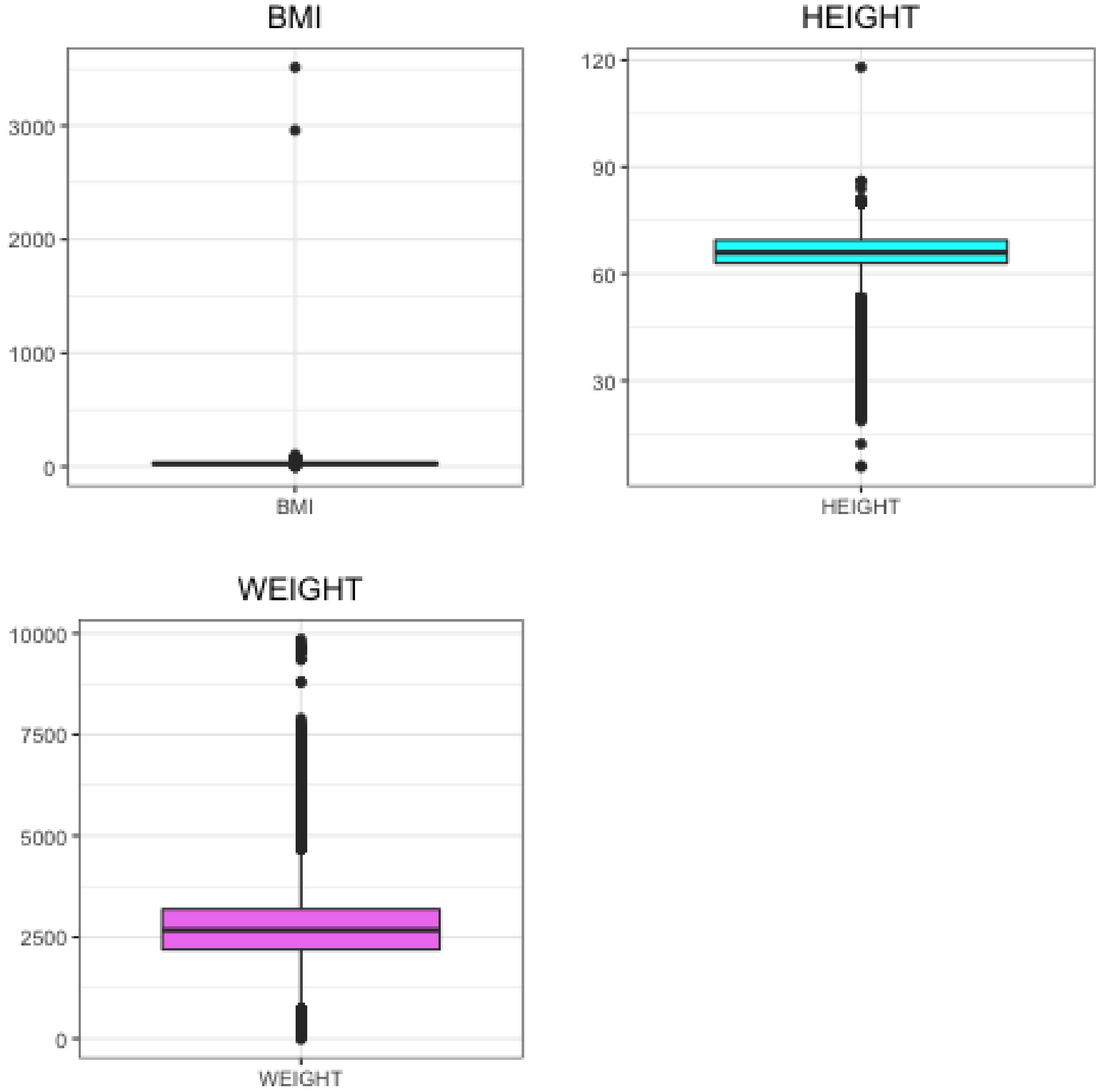

FIGURE 14: VISITS: Two high BMI datapoints that should be removed. Height is likely in inches, as 120cm = 3.3 feet. This means that the 120 datapoint should be removed. Weight is in ounces, as 2500 oz is 156 lbs and 10000 ounces is 625 pounds. Conclusion: BMI and HEIGHT have a few outliers (likely typos) that can be quickly removed.

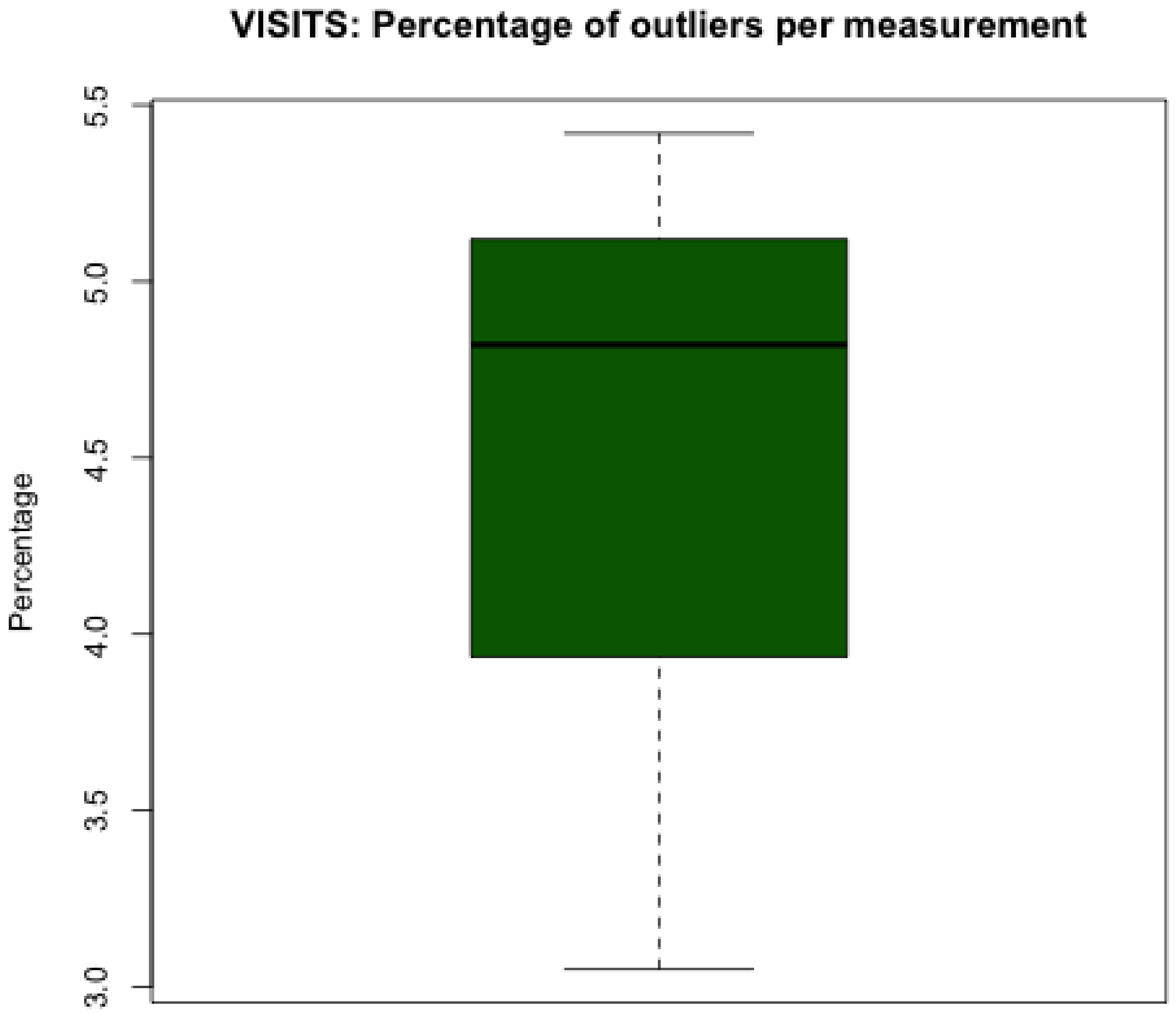

FIGURE 15: VISITS: Outlier percentage (number of outliers/total number of data points) for each test (every point is a medical test). If we decide to remove all outliers, which is not suggested, we will lose less than 6% of all data points.