

HW1

Jeffrey Liang, zl2974

Code available at https://github.com/zl2974/P8130_hw1_zl2974/

Problem 1

Descriptive statistics to the data

Descriptive statistics of the dataset was provided below, 1224 value missing occurred at Antibody level's columns. Also provided with categorical data's count table, no missing value in categorical data.

Table 1: Data summary

| | |
|------------------------|----------|
| Name | smell_df |
| Number of rows | 1491 |
| Number of columns | 5 |
| Column type frequency: | |
| factor | 3 |
| numeric | 2 |
| Group variables | None |

Variable type: factor

| skim_variable | n_missing | complete_rate | ordered | n_unique | top_counts |
|---------------|-----------|---------------|---------|----------|------------------------------|
| smell | 0 | 1 | FALSE | 3 | Alt: 1047, Nor: 410, Una: 34 |
| gender | 0 | 1 | FALSE | 2 | Fem: 981, Mal: 510 |
| age_category | 0 | 1 | FALSE | 3 | 31-: 810, 51+: 363, 18-: 318 |

Variable type: numeric

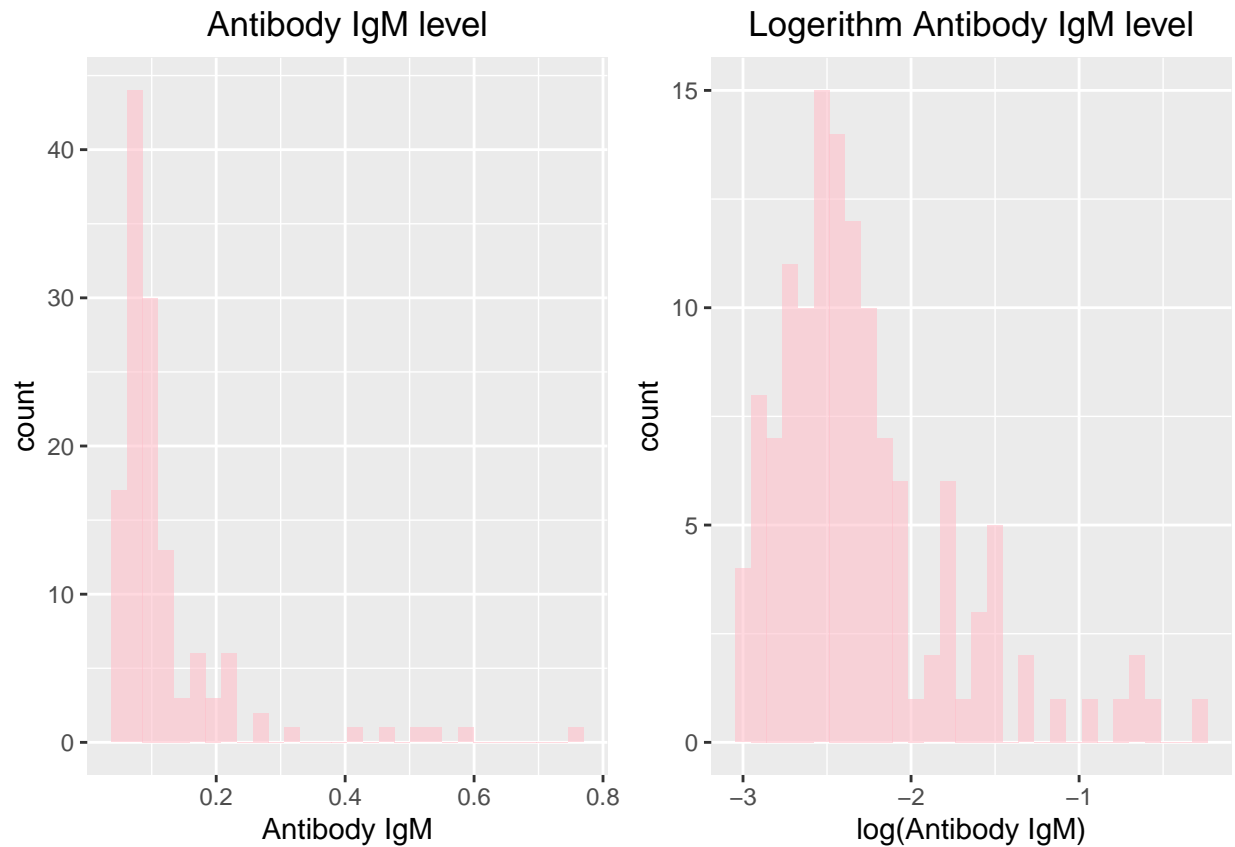
| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---------------|-----------|---------------|---------|--------|------|--------|---------|---------|---------|------|
| subject | 0 | 1.00 | 1413.75 | 881.07 | 1.00 | 632.50 | 1373.00 | 2190.50 | 2917.00 | |
| antibody_ig_m | 1224 | 0.18 | 0.12 | 0.11 | 0.05 | 0.07 | 0.09 | 0.13 | 1.05 | |

```
##
## Table: Categorical Variable Statistics
##
## | Overall (N=1491) |
## |:-----|:-----:|
```

| ## | age_category | | |
|----|---------------------|--------------|--|
| ## | - 18-30 | 318 (21.3%) | |
| ## | - 31-50 | 810 (54.3%) | |
| ## | - 51+ | 363 (24.3%) | |
| ## | - Missing | 0 | |
| ## | smell | | |
| ## | - Normal | 410 (27.5%) | |
| ## | - Altered | 1047 (70.2%) | |
| ## | - Unanswered/Others | 34 (2.3%) | |
| ## | - Missing | 0 | |
| ## | gender | | |
| ## | - Female | 981 (65.8%) | |
| ## | - Male | 510 (34.2%) | |
| ## | - Missing | 0 | |

Distribution of IgM

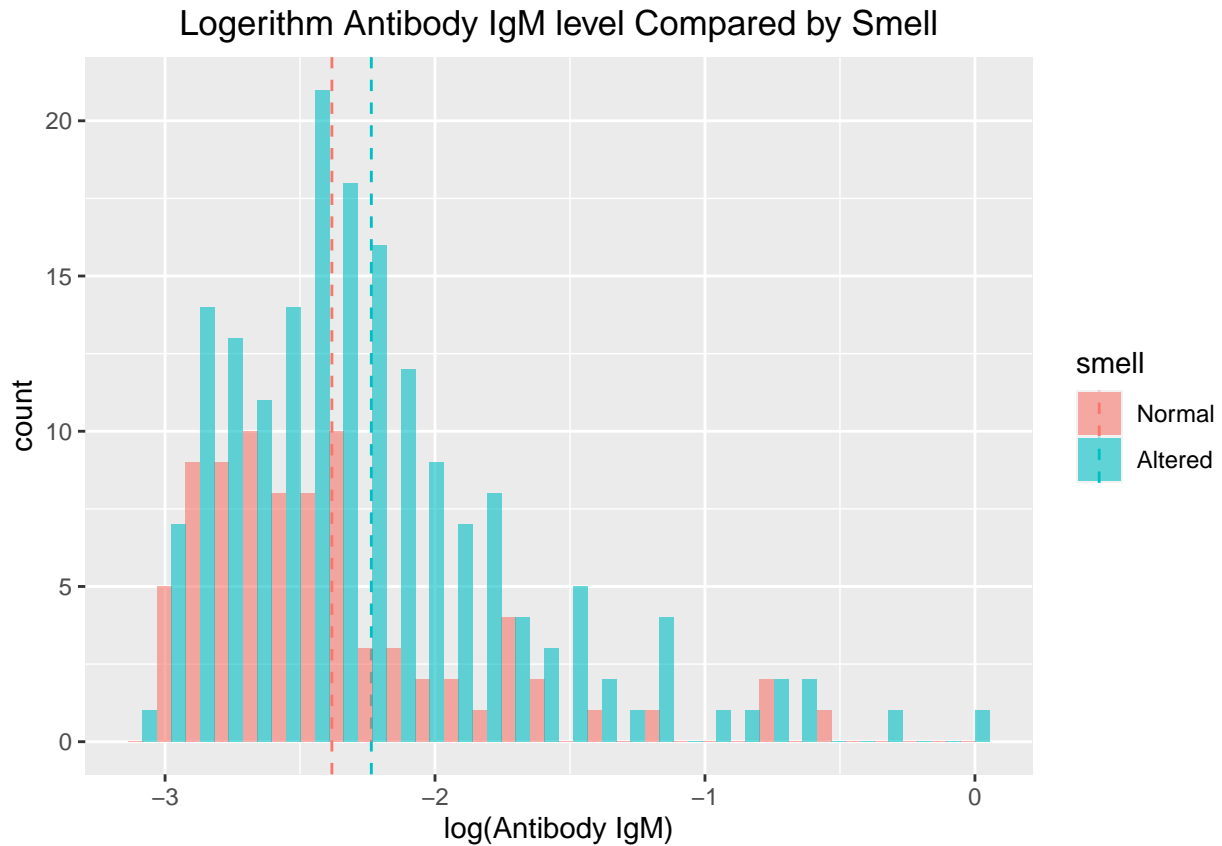
One can first notice from descriptive statistics above that antibody IgM's level is extreme left skewed. To detail, Both histograms of antibody level and the logarithm of antibody level were provided. As shown below, even adjusted skewness with logarithm, antibody's level is highly concentrate on the left, and long tail on the right.



Smell Alter and Normal comparison

Grouped subjects with altered smell and normal, we have seen that subjects from group smell altered have higher antibody level compared with normal group.

```
##
## Table: Smell
##
## |               | Normal (N=81) | Altered (N=178) |
## |-----|-----|-----|
## |antibody_ig_m |               |               |
## |  Mean (SD)   | 0.110 (0.092) | 0.130 (0.119)   |
## | Median (Q1, Q3) | 0.081 (0.065, 0.107) | 0.097 (0.072, 0.137) |
## | Min - Max    | 0.049 - 0.589 | 0.048 - 1.048   |
## |gender        |               |               |
## |  Female      | 47 (58.0%)   | 120 (67.4%)     |
## |  Male        | 34 (42.0%)   | 58 (32.6%)      |
## |age_category  |               |               |
## |  18-30       | 13 (16.0%)   | 36 (20.2%)      |
## |  31-50       | 43 (53.1%)   | 101 (56.7%)     |
## |  51+         | 25 (30.9%)   | 41 (23.0%)      |
```



Problem 2