HW1

Jeffrey Liang, zl2974

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

Problem 1

Descriptive statistics to the data

The data is a 1491 · 5 shape data with 1491 objects. There're 1224 missing values in the data. After cleaning missing values as well as the 'Unansered/Others' report from *Smell* variable, 259 objects remain. Descriptive statistics of the original dataset was provided below, 1224 value missing occurred at Antibody level's columns. Also provided with catagorical data's count table, no missing value in catagorcial 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
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: Catagorical Variable Statistics
##
```

1

```
## |                     | 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 logerithm of antibody level were provided. As shown below, even adjusted skewness with logerithm, antibody's level is highly concentrate on the left, and long tail on the right.
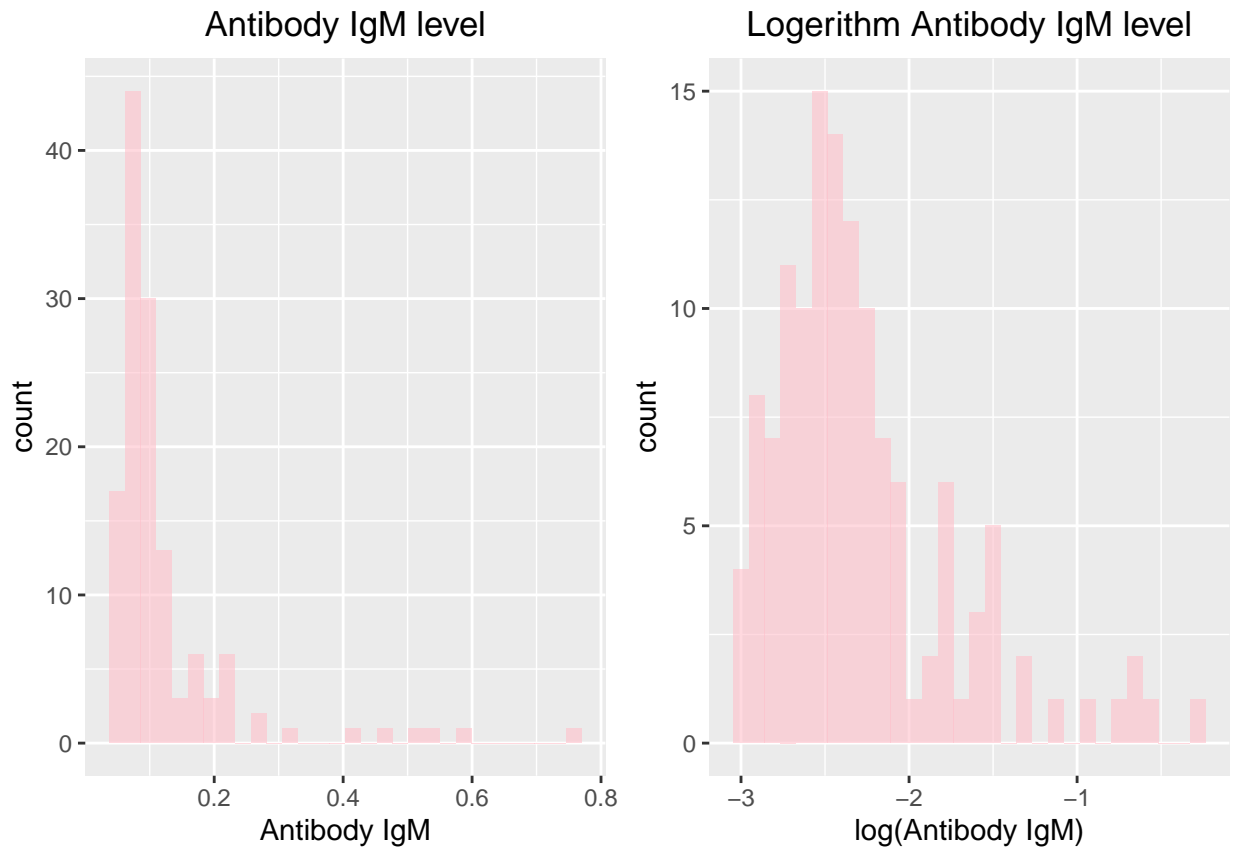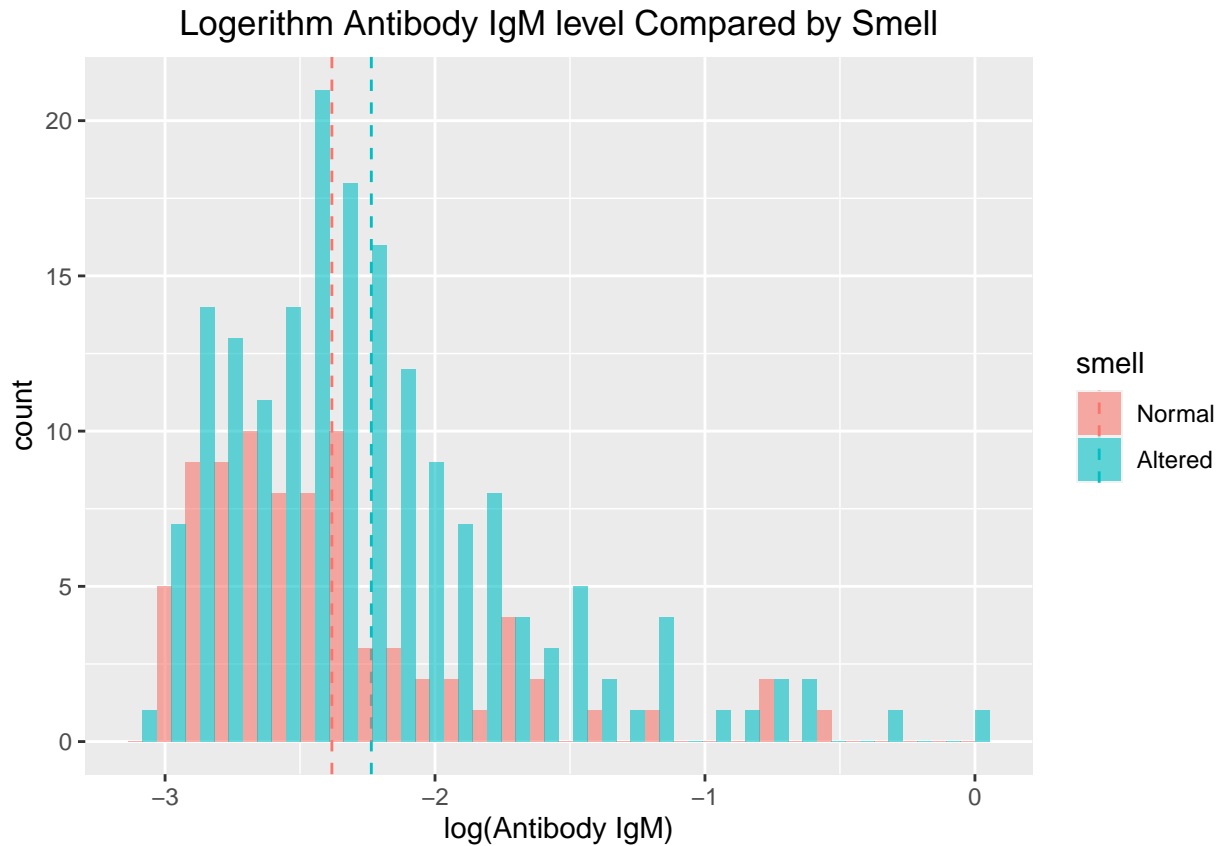
Smell *Alter* and *Normal* Comparison

Grouped subjects with altered smell and normal, we have seen that subjects from group smell altered have higher antibody level compared with normal group. Also, spread of Altered group's antibody level is wider than the spread of normal smell group judging from both range and standard deviation.

```
##
## 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%)     |
```


Logerithm Antibody IgM level Compared by Smell

Problem 2

I've chosen **More Testing Needed: Most Coronavirus Cases Come From People Unaware They're Infected** *as my interest for problem 2's article.*

In the news of choice[1], the author introduced result from the article from Li,R et.al[2]. The author introduce the result of the paper via transcript and direct quotes for her article. For example, the article quoted *"Our findings also indicate that a radical increase in the identification and isolation of currently undocumented infections would be needed to fully control SARS-CoV2,"* directly to support her claim of *'Without enough testing to rapidly identify everyone who has the disease and separate them from others, the disease will continue to spread quickly.'*

In other news reports, Columbia's news use directed quotes to the paper[3]. In ScienceDaily, the report first summary the paper and following use transcript as well as quotes to introduce the contents of the paper.

There's a few relative news reports also referring to the same published article.

And here's **some** claims made in the article based on the Li,R's findings and are listed as followed:

- The ability of a respiratory disease to become a pandemic depends in part on how many infectious people are not aware that they're infectious
- That means if someone were able to completely isolate themselves immediately after they were exposed to the virus, it would tremendously cut down the total infection rate. Each person would only infect about one other person.
- Without enough testing to rapidly identify everyone who has the disease and separate them from others, the disease will continue to spread quickly.
- COVID-19 will be like H1N1 becoming a global endemic

The ScienceDaily report agree on bullet point 1 and 3, regarding asymtomatic and COVID-19 becoming a pandemic.

To validate these points in the report, we must first dig into the reference paper and make comparison on the claims in report and in the paper.

Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)

Li, R's paper tried to calculate the proportion of undocumented COVID-19 cases and their transmission. To unravel characteristic of undocumented COVID-19, the author conducted an observational study with reported infection case and 2.97 billion mobility data from tech company and applied bayesian inference model to interpret proportion of undocumented cases during 25 days before 10th, Jan. and 15 day after. And the paper found result:

- Basic reproductive number R_0 2.38(CI:2.03-2.77) during 10-23 Januray and only 14% total infections were reported (in another word: 86% of undocumented infections);
- These undocumented infections had halved transmission compared to reported cases;
- After implement of control measures(social-distancing and etc.), transmission rate of documented cases, β, dropped to 0.52 (CI:0.42–0.72) to 0.35 (0.28–0.45) as well as the R_0 (to 1.34(CI:1.10-1.67));
- Proportion of undocumented cases has dropped to 35%.
- fully identify undocumented disease is need to fully control COVID-19.
- COVID-19 can become the 5th global endemic

Summary

Comparing the news and paper, one have to admit that the Forbe's news has done good job putting the facts from the paper. All claims can be found from the *Introduction*, *Result* and *Conclusion* from the paper. But minor drawback of the news report is the definition of asymptomatic patients is not matching the define of undocumented infections of Li's paper. Asymptomatic patients are include in the undocumented infections, which stating the asymptomatic shares the characteristics of undocumented infections would over-estimated it.

Reference

1. *More Testing Needed: Most Coronavirus Cases Come From People Unaware They're Infected:* https://www.forbes.com/sites/tarahaelle/2020/03/20/more-testing-needed-most-coronavirus-cases-come-from-people-unaware-theyre-infected/#3ccf8ddc2988
2. *Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)*: https://science.sciencemag.org/content/sci/368/6490/489.full.pdf
3. *'Stealth Transmission' Fuels Fast Spread of Coronavirus Outbreak:* https://www.publichealth.columbia.edu/public-health-now/news/stealth-transmission-fuels-fast-spread-coronavirus-outbreak
4. *'Stealth transmission' fuels fast spread of coronavirus outbreak:* https://www.sciencedaily.com/releases/2020/03/200316141454.htm