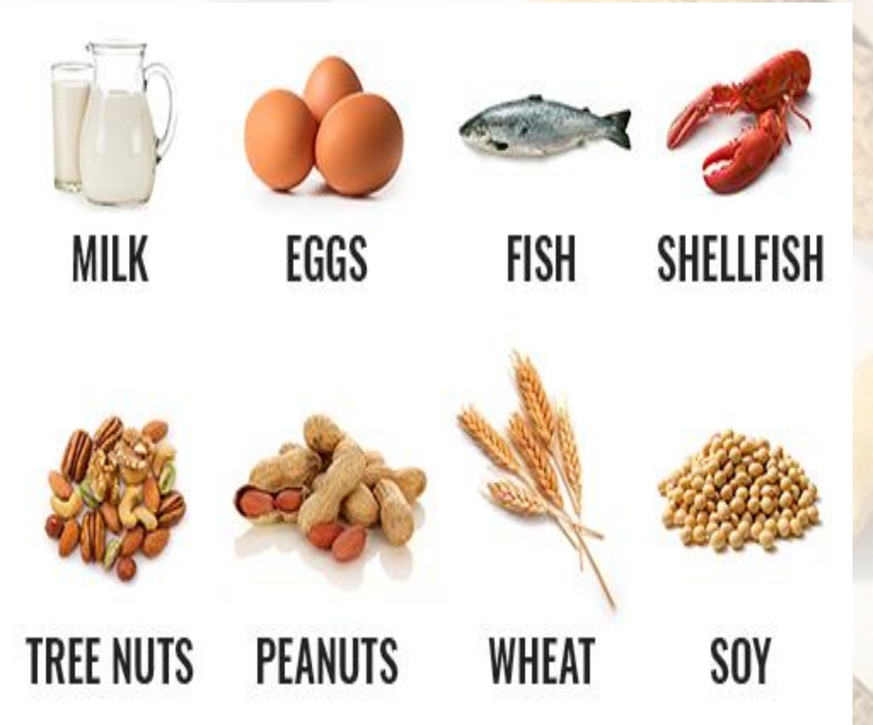
Childhood Allergy Prevalence

Alex Hart
Lucia Tablas
Sara Arasteh
Vivian Sun
Leah Krause

Introduction

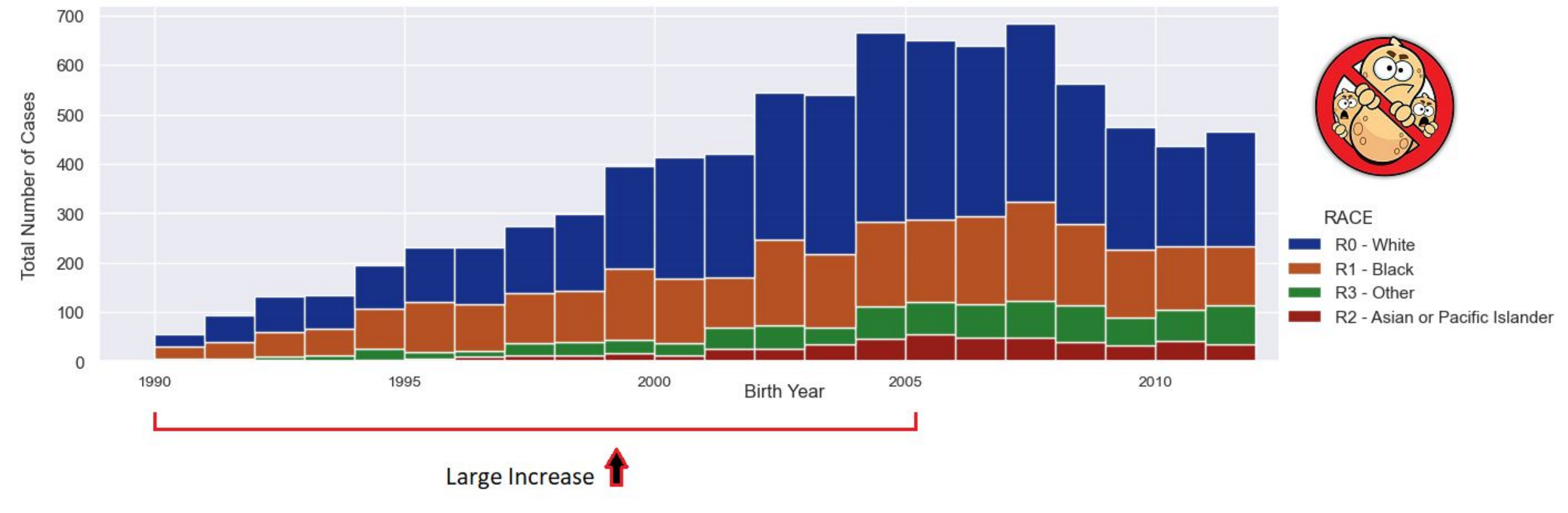
Food Allergy

- What..
  - The immune system reaction to a certain food
- How..
  - Can be moderate or strong



Importance

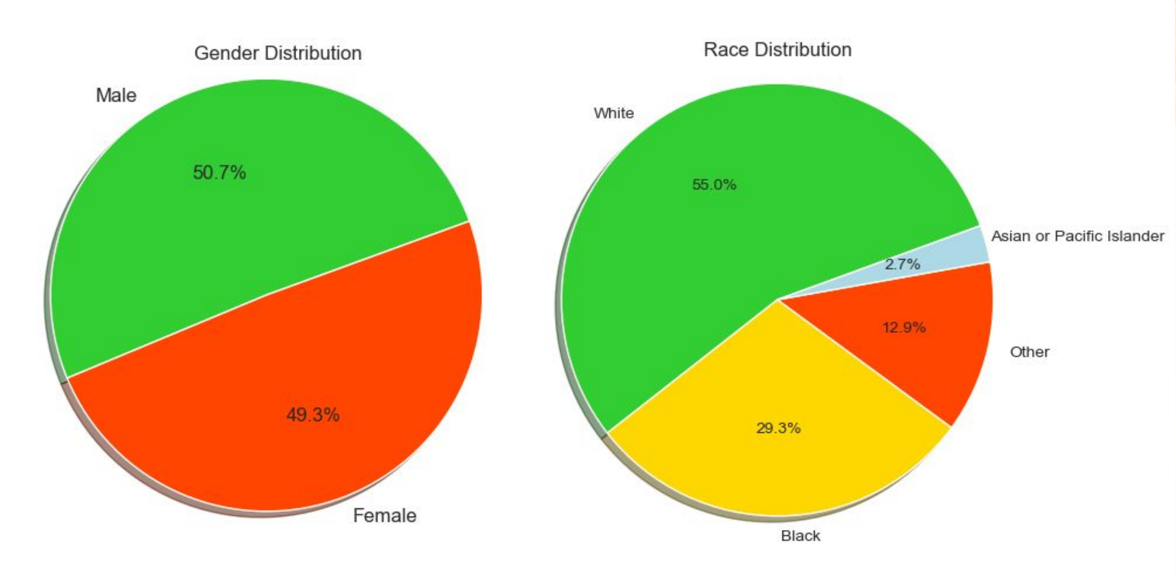
- The reported food allergy cases increased
- More cases for kids in the early years of their development



Large Increase

Data Set

Demographic Features



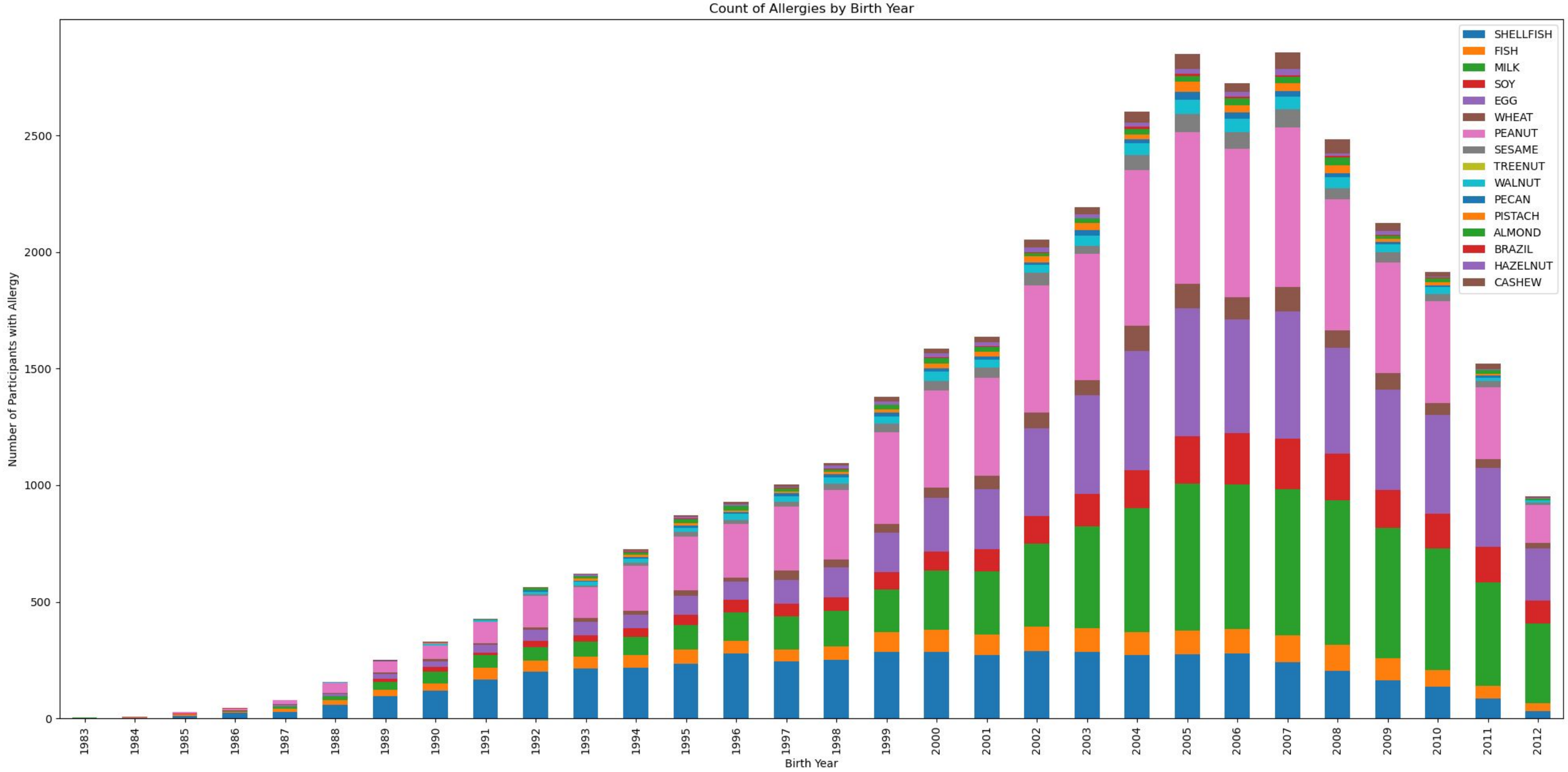
Questions Addressed:

1. Has the prevalence of food allergies **increased** over time?
2. What is the **likelihood** of allergens in each demographic?
3. Are certain allergies **correlated**?
4. Does population **density** define the **severity** level of reactions?

QUESTION

1. Has the prevalence of food allergies **increased** over time?

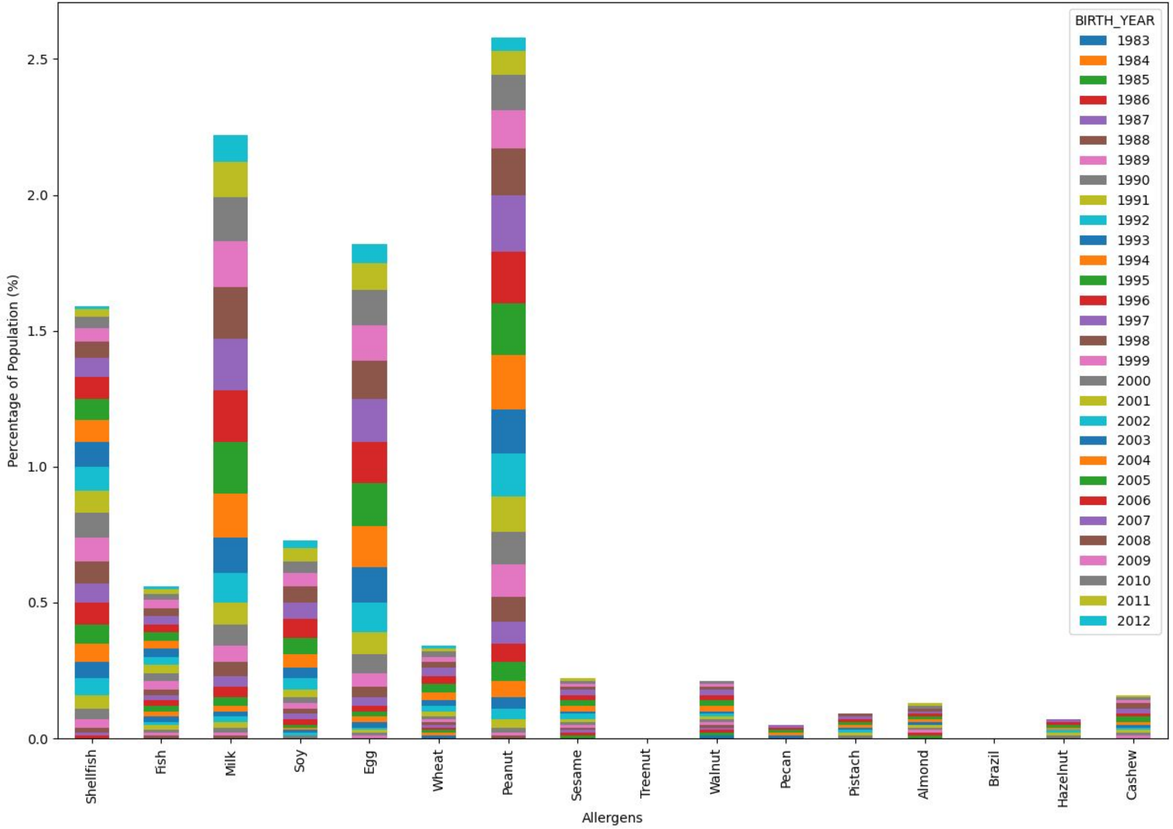
Allergies Through The Years



The Years Through Allergies

- ➔ We can see increases in peanut, milk, and egg allergies as birth years get later.

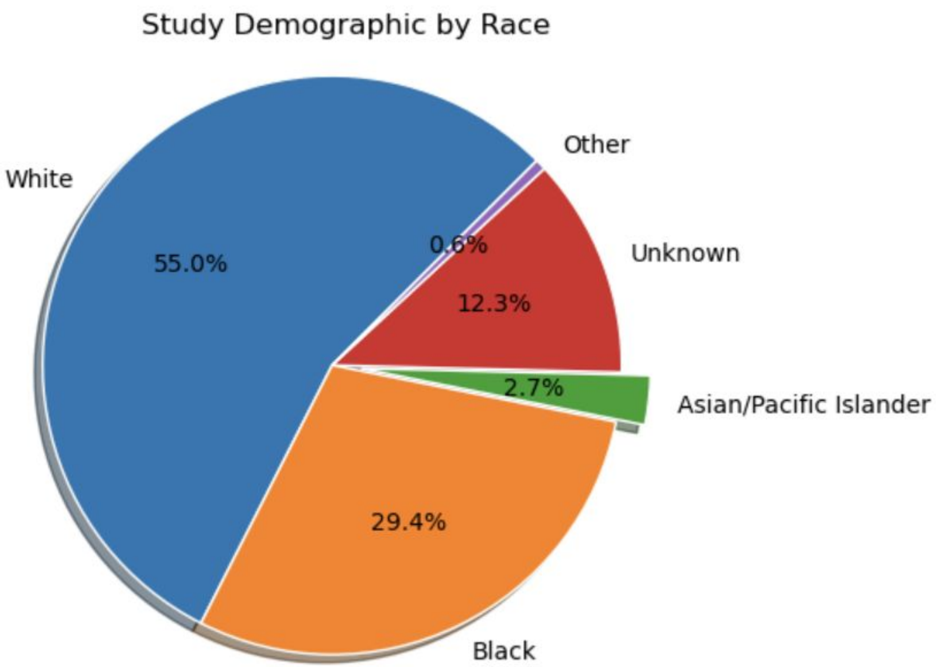
- ➔ The rest of the data is very squishy and consistent.

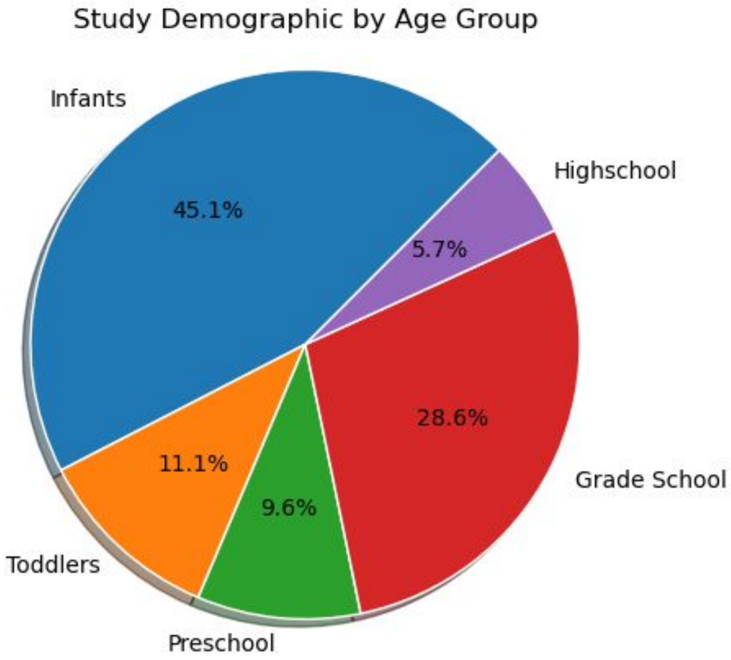


QUESTION

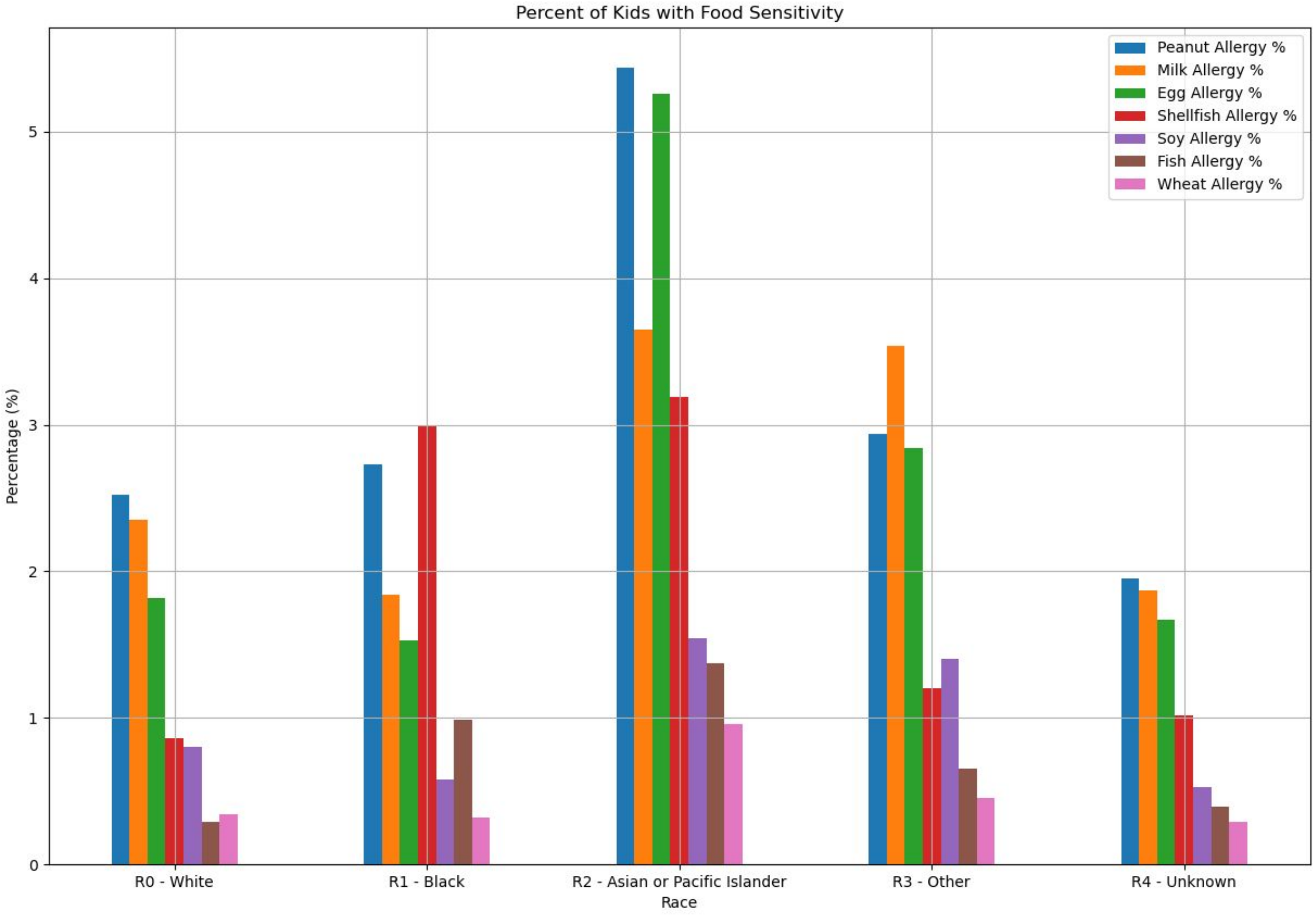
2. What is the **likelihood** of allergens in each demographic?

Study Demographics

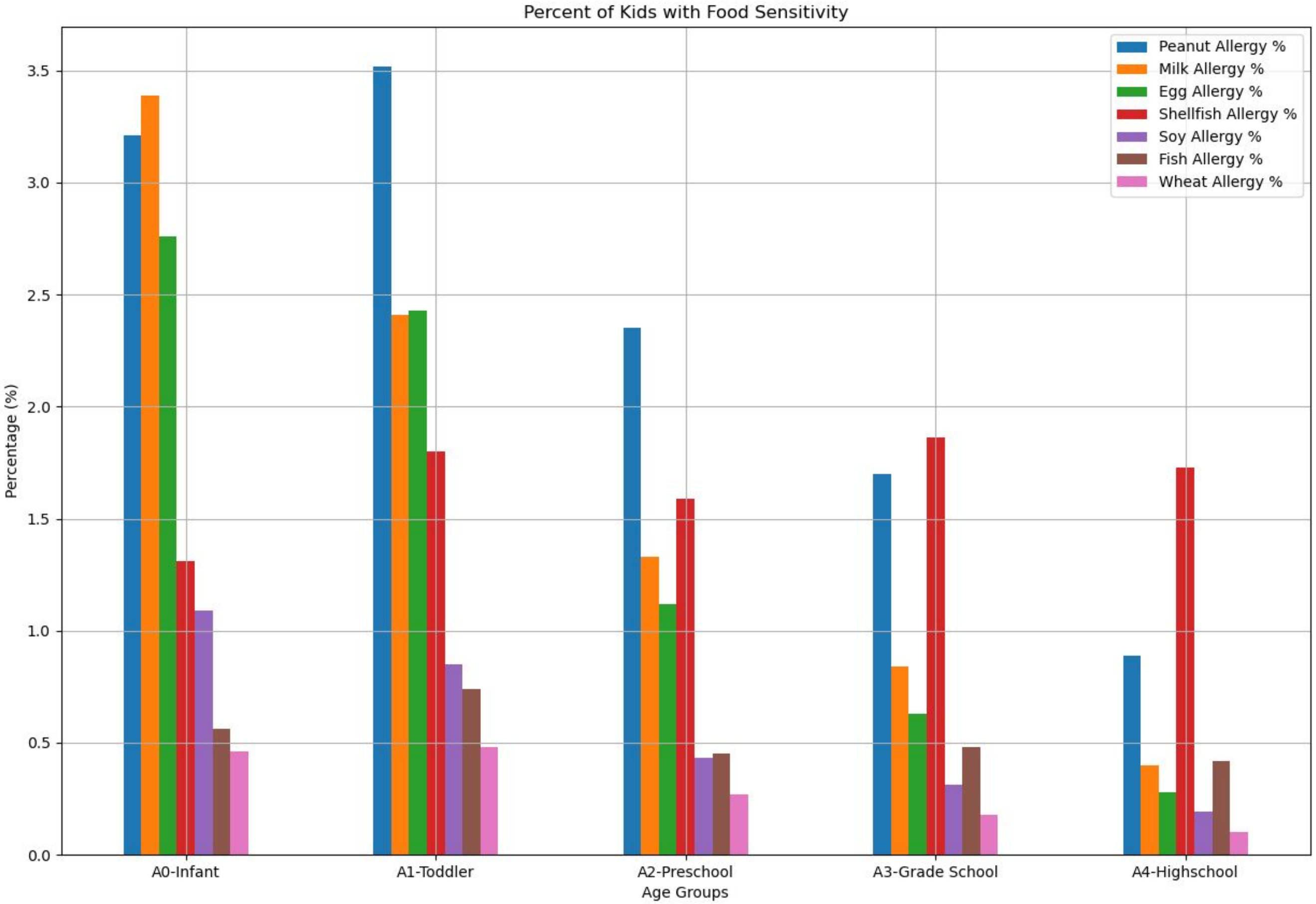




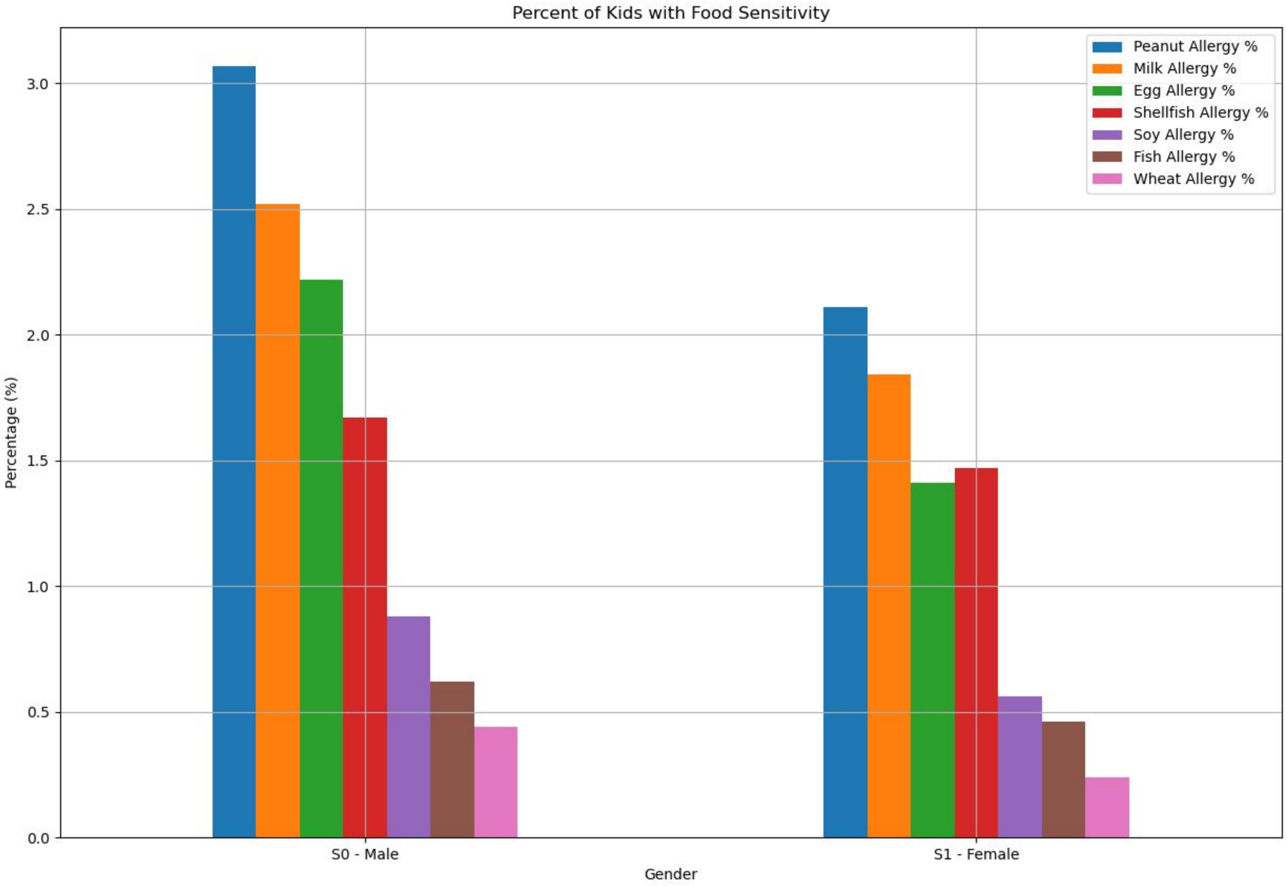
Normalized Race Distribution



Normalized Age Group Distribution



Gender Distribution



QUESTION

3. Are certain allergies **correlated**?

Allergen Indicators

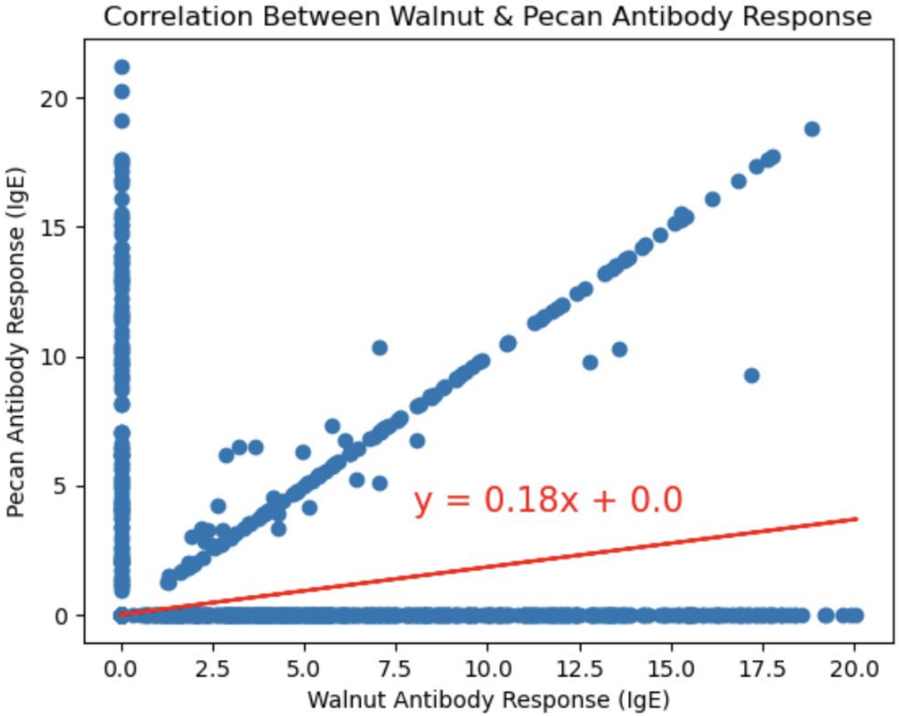
- ➔ Through correlation matrices using Pearson's R, we were able to determine which allergies (if any) are correlated

Pearson's Correlation Coefficient	Strength
$r < 0.3$	None or very weak
$0.3 \leq r < 0.5$	Weak
$0.5 \leq r < 0.7$	Moderate
$r \geq 0.7$	Strong

Overall Highest Correlated: Pecan and Walnut

Pearson's correlation coefficient of

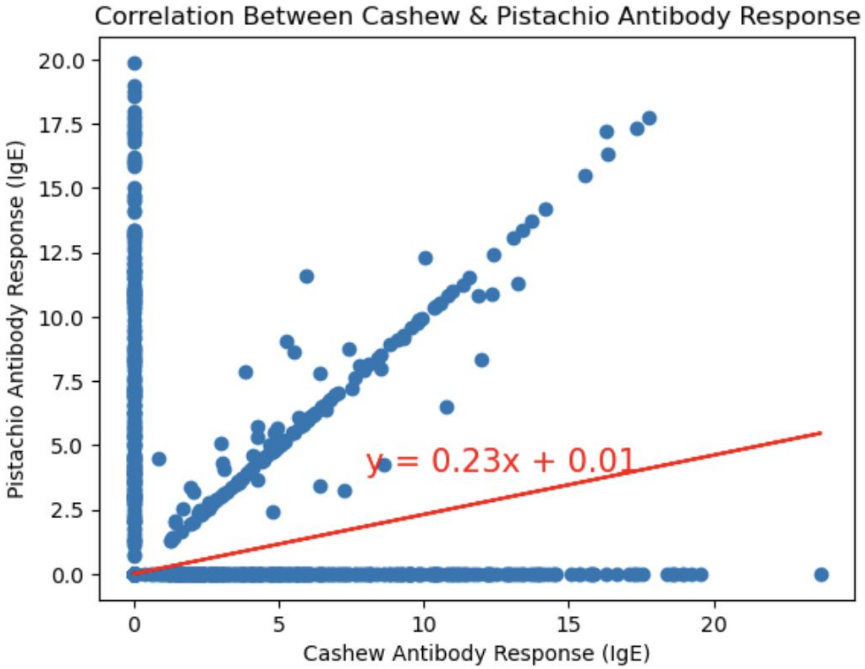
0.29 (Very weak)



Overall Second Highest Correlated: Cashew and Pistachio

Pearson's correlation coefficient of

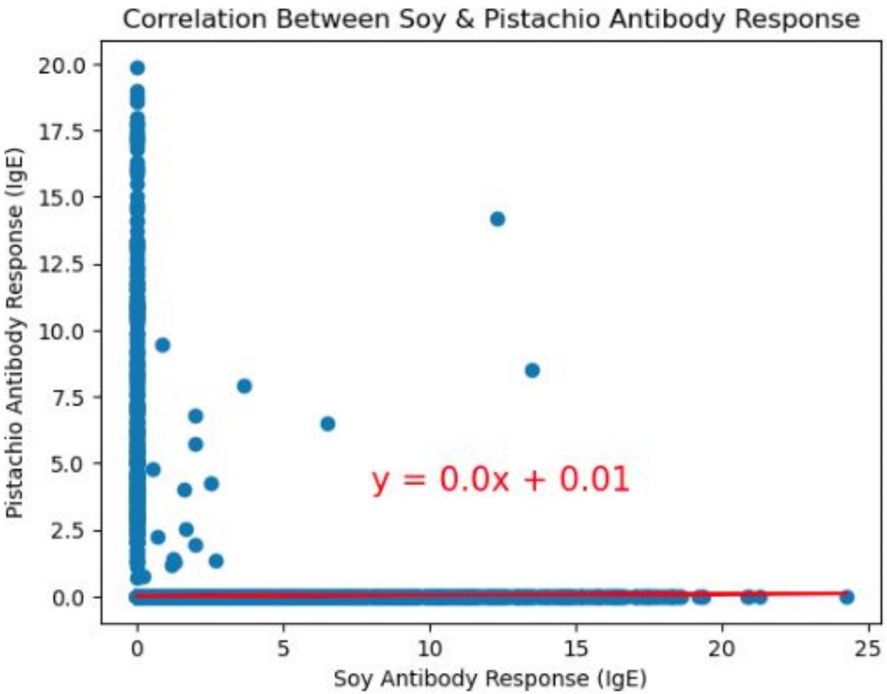
0.29 (Very weak)



Overall Least Correlated: Soy and Pistachio

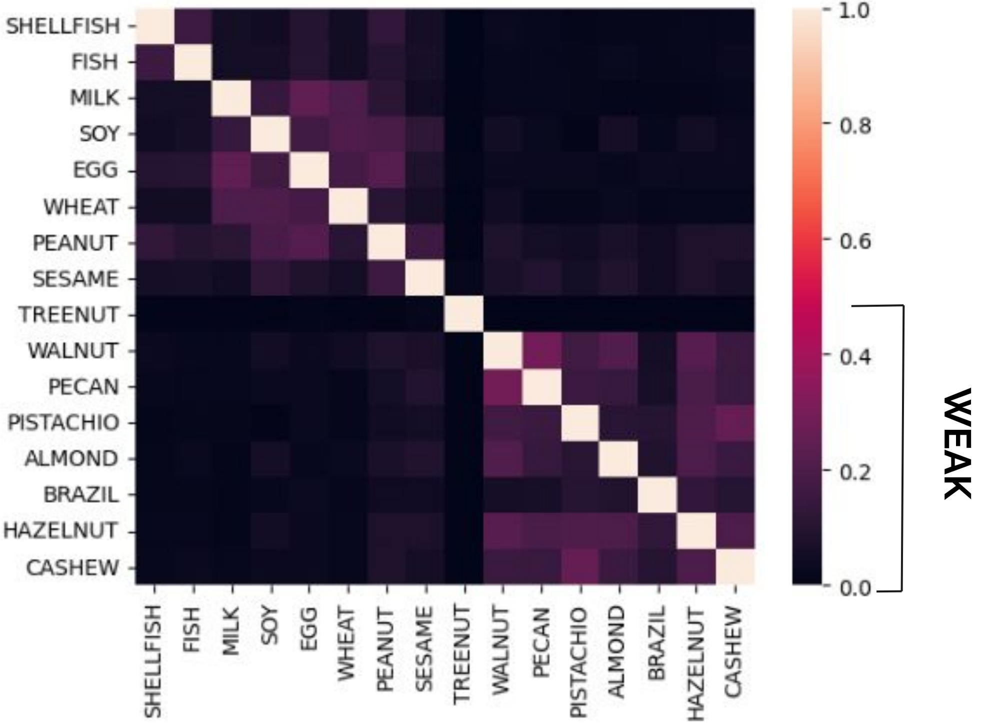
Pearson's correlation coefficient of

0.01 (None to very weak)



Correlation Throughout the Study

- ➔ None of the allergens ever indicate a correlation factor above **weak**
- ➔ We can infer that these allergies are most likely **independently developed**

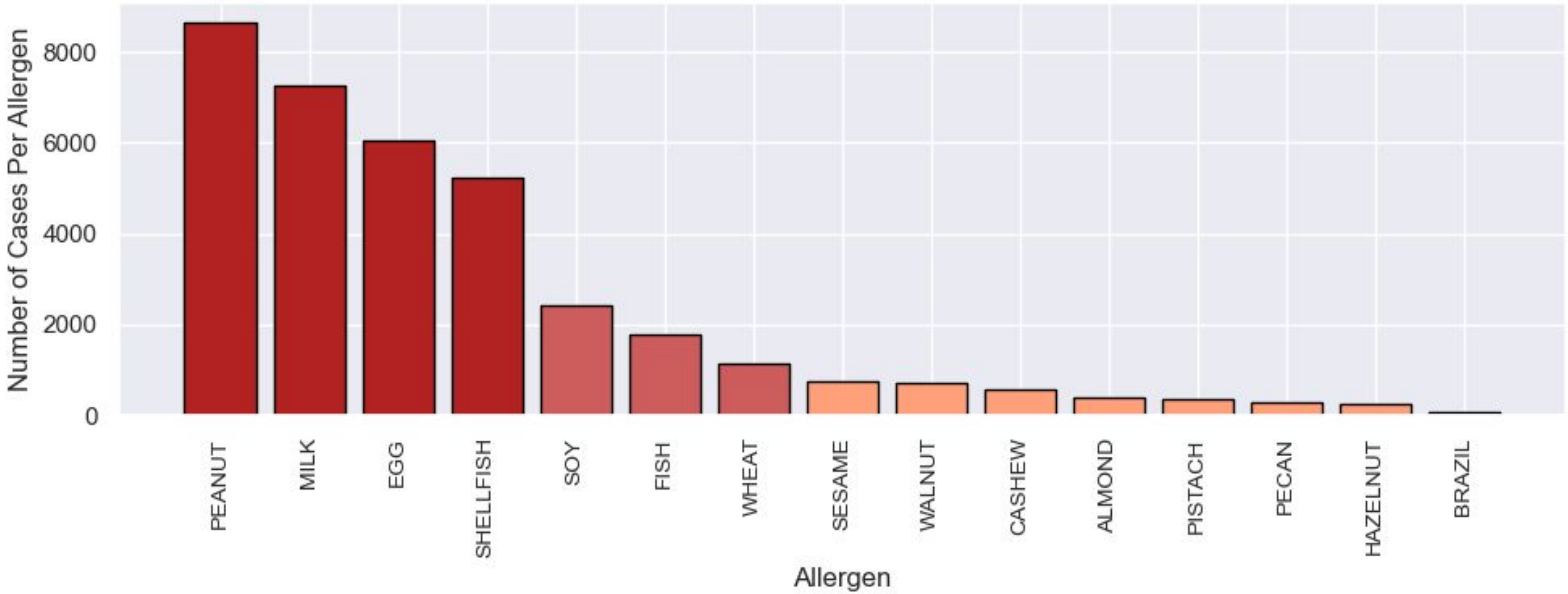


QUESTION

4. Does population **density** define the **severity** level of reactions?

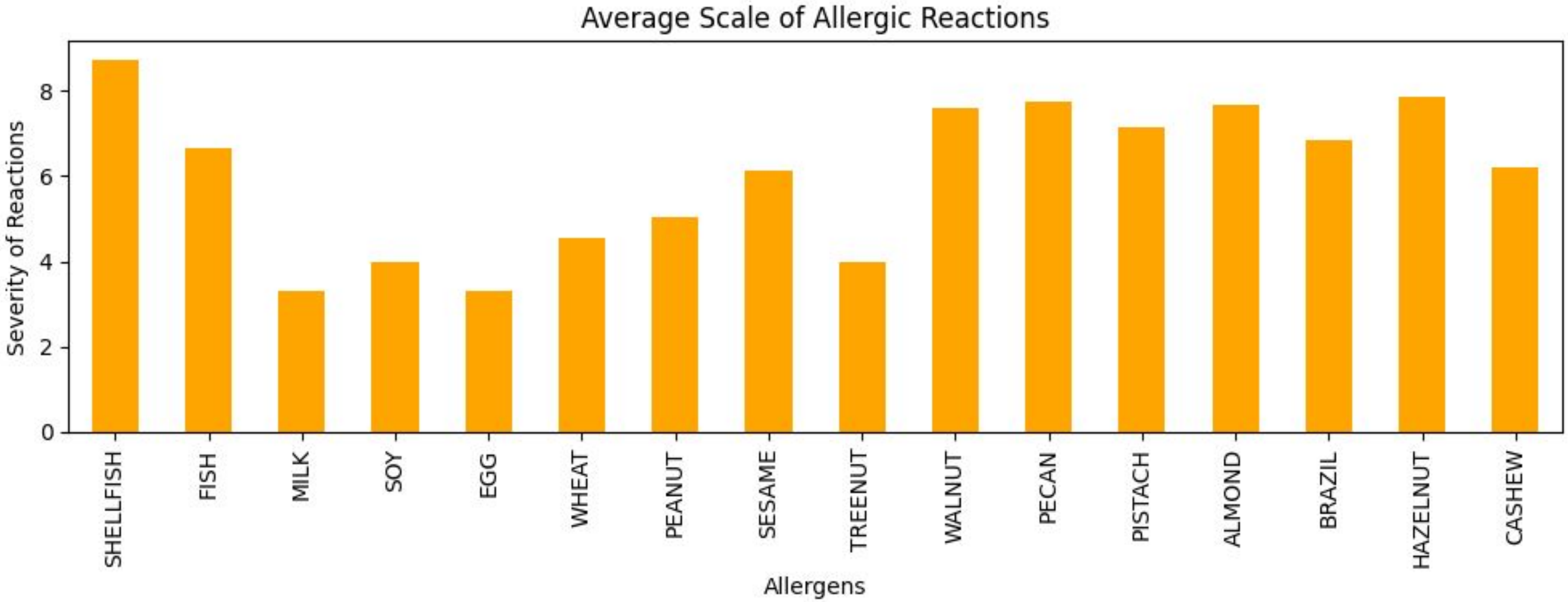
Allergen Distribution

- ➔ A quick visual analysis on the total number of each allergen recorded in this test
- ➔ All nut allergens appear to have a lower population in number of reactions recorded



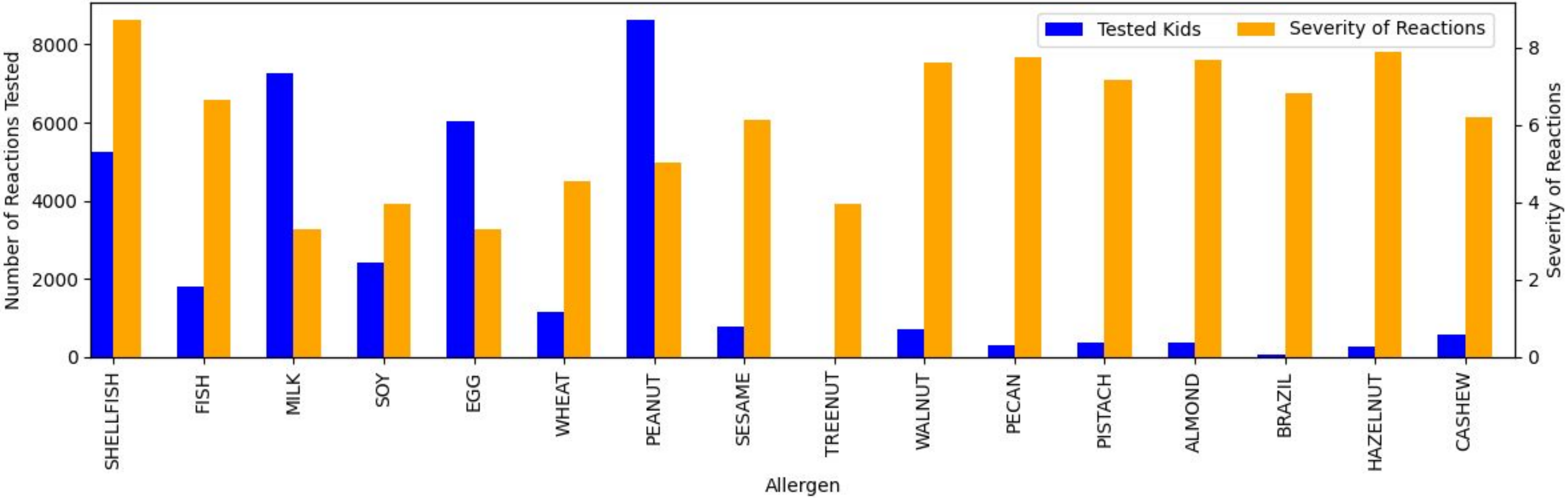
Allergen Reaction

- ➔ Nut allergens appear to be leading in terms of reaction severity levels
- ➔ However, Shellfish has the highest average severity level in reaction in this comparison



Final Comparison

- ➔ Although nuts have a low density in population and have severe allergic reactions, shellfish has both a high population as well as the highest average reaction within this test
- ➔ In this case severity of the reaction may not be directly related with the population density



Conclusions & Limitations

Limitations

- Sample not racially, ethnically, or geographically representative
- No environmental information available

Conclusions

- We can see a rise in allergic responses in kids born between 2004 and 2009
- Within the racial limitations of the study, Asian / Pacific Islanders demonstrated the highest likelihood to have a food sensitivity
- Allergens are most likely independently developed
- Population density in allergen groups does not define severity level

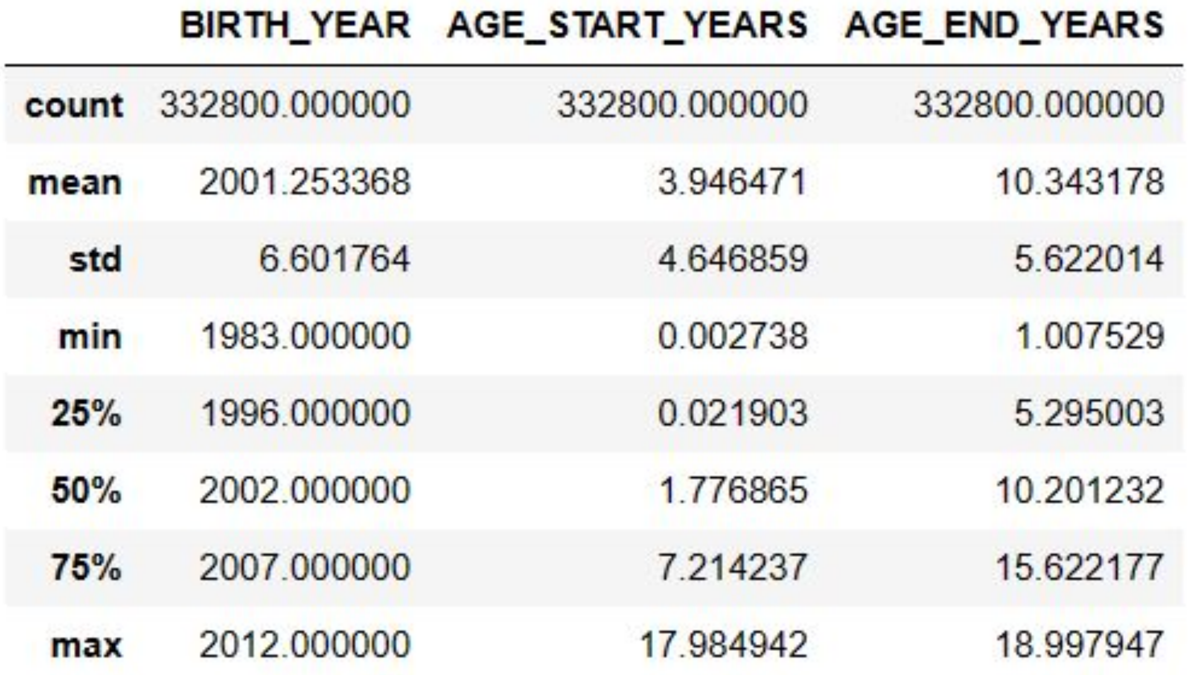
Next Steps

- Research into additives, pesticides in foods, and environmental factors
- Actual increase in allergies or perception due to improved scientific methods & information sharing or perceived
- Increased awareness for potential hazards in food sensitivities in parents and caregivers

Questions?

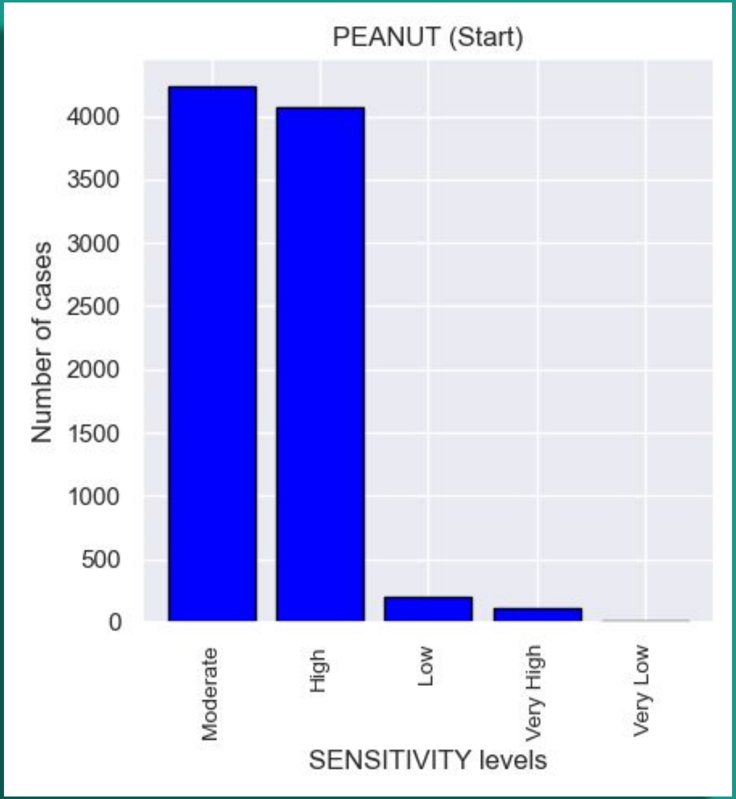
References

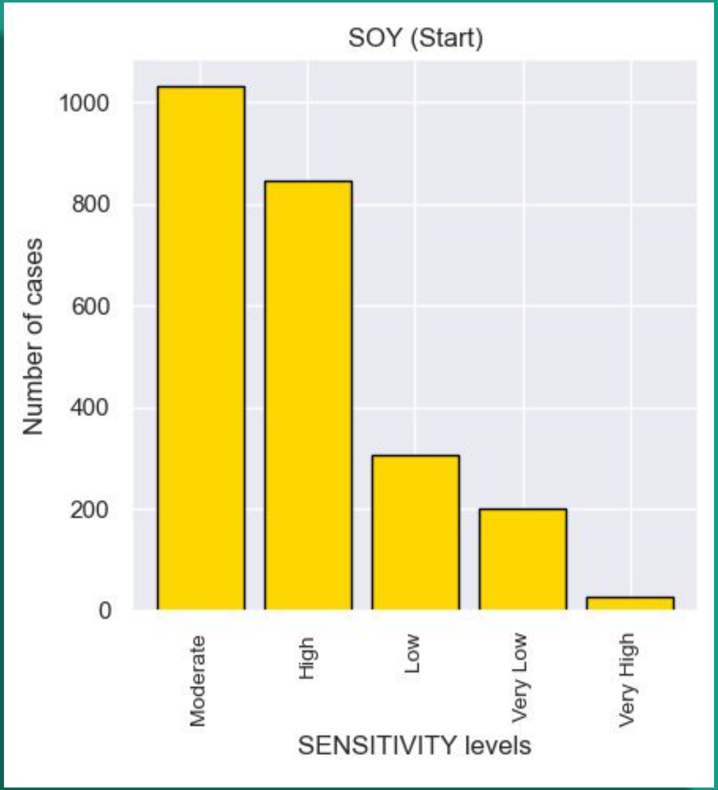
Source from Kaggle:
Childhood Allergies: Prevalence, Demographics



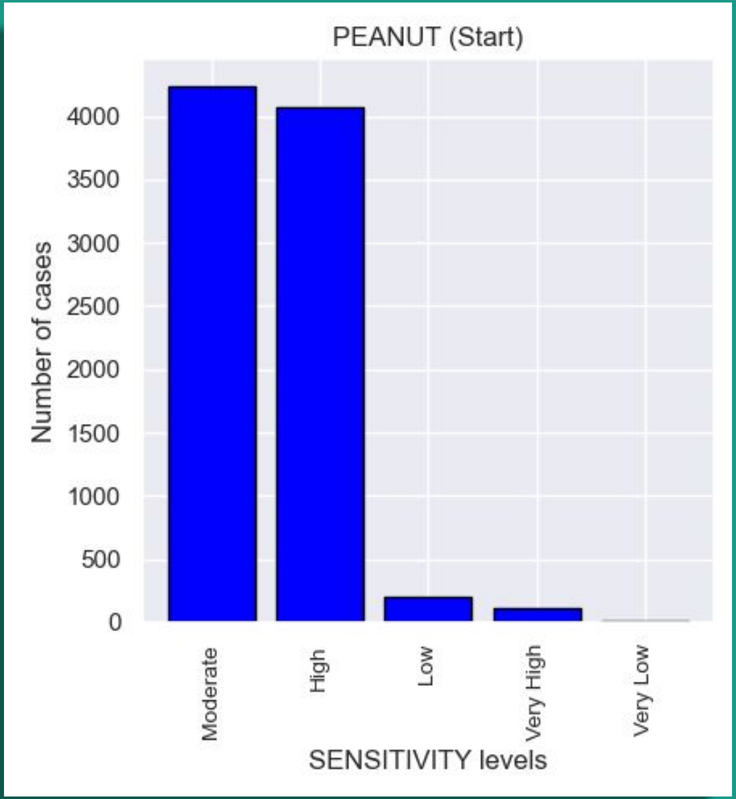
	BIRTH_YEAR	AGE_START_YEARS	AGE_END_YEARS
count	332800.000000	332800.000000	332800.000000
mean	2001.253368	3.946471	10.343178
std	6.601764	4.646859	5.622014
min	1983.000000	0.002738	1.007529
25%	1996.000000	0.021903	5.295003
50%	2002.000000	1.776865	10.201232
75%	2007.000000	7.214237	15.622177
max	2012.000000	17.984942	18.997947

Analysis

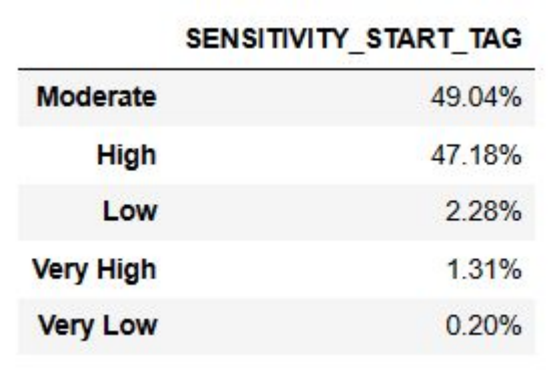

PEANUT (Start)
Number of cases
4000
3500
3000
2500
2000
1500
1000
500
0
Moderate
High
Low
Very High
Very Low
SENSITIVITY levels


SOY (Start)
Number of cases
1000
800
600
400
200
0
Moderate
High
Low
Very Low
Very High
SENSITIVITY levels

Analysis

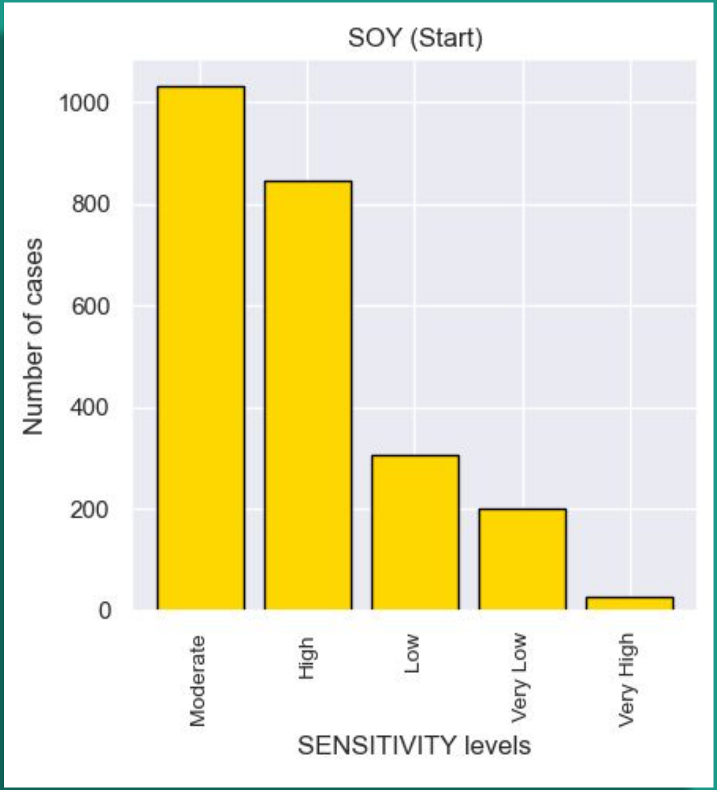


- Peanut allergies are mostly towards High and Moderate sensitivities.
- Low, Very Low and Very High are only happen in 4% of the sample.

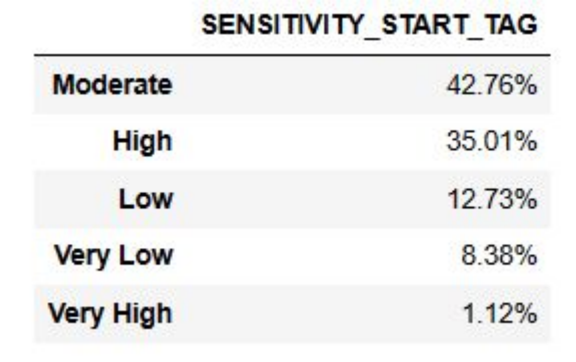


	SENSITIVITY_START_TAG
Moderate	49.04%
High	47.18%
Low	2.28%
Very High	1.31%
Very Low	0.20%

Analysis



- Soy allergy is less populated compared to Peanuts.
- More cases with Low, Very Low and Very High (22%)



	SENSITIVITY_START_TAG
Moderate	42.76%
High	35.01%
Low	12.73%
Very Low	8.38%
Very High	1.12%

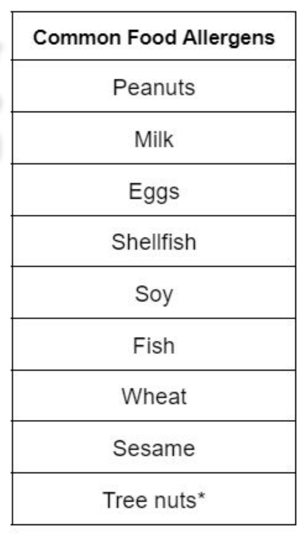
Introduction

Motivation:
We were interested in seeing if there was a perceived or actual increase in food allergies in children in recent years

Analyze the prevalence of allergies/sensitivities of the most common food allergens for the population of kids. This analysis is based on the need for the food industry due to food recalls and implementation of allergen control strategies. The population selected from different:

- Birth Years (1983-2012)
- Ethnicities (Hispanic, non_Hispanic)
- Races (Black,White,Asian and Pacific Islander)

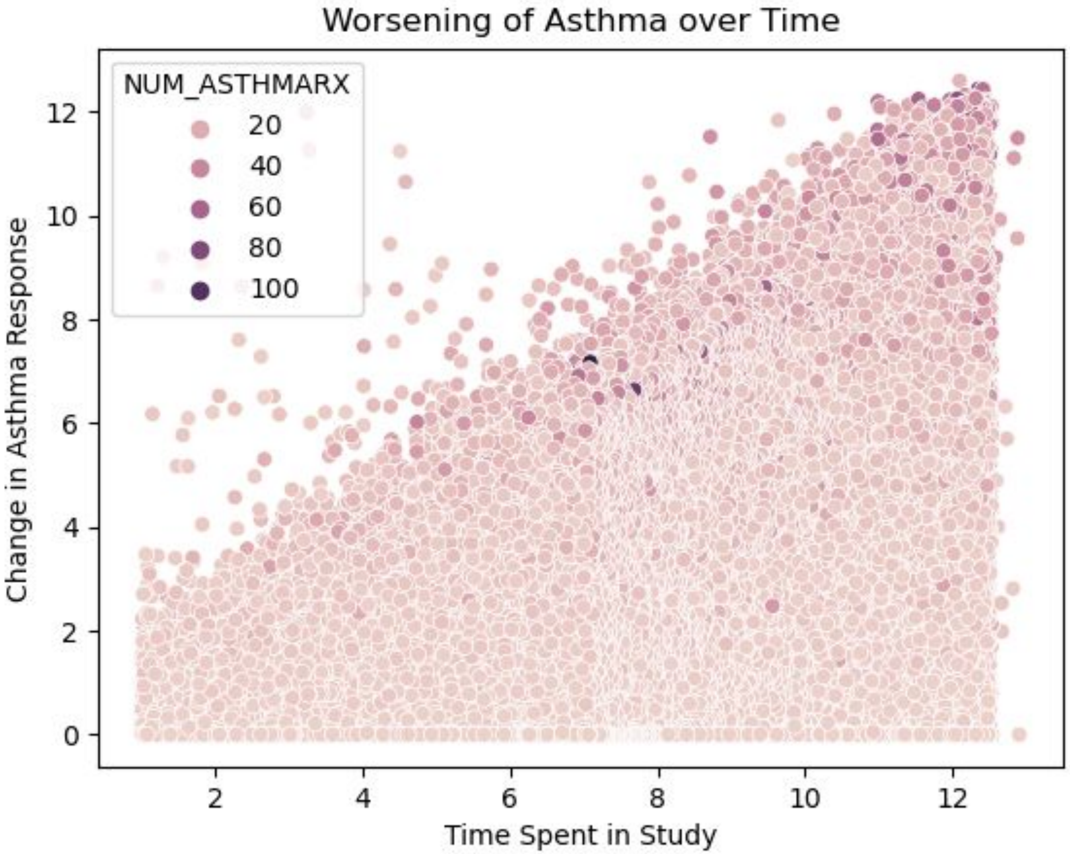
Motivation



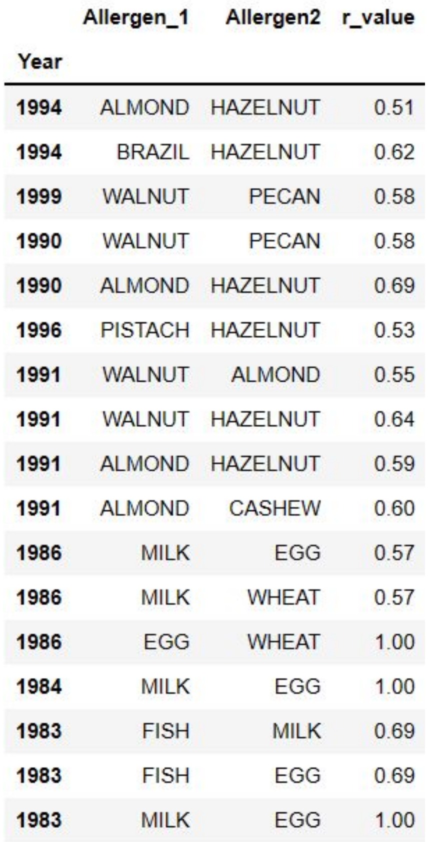
Common Food Allergens
Peanuts
Milk
Eggs
Shellfish
Soy
Fish
Wheat
Sesame
Tree nuts*

Tree nuts* includes: *walnut, almond, hazelnut, pecan ,brazil nuts, cashew and pistachio.*

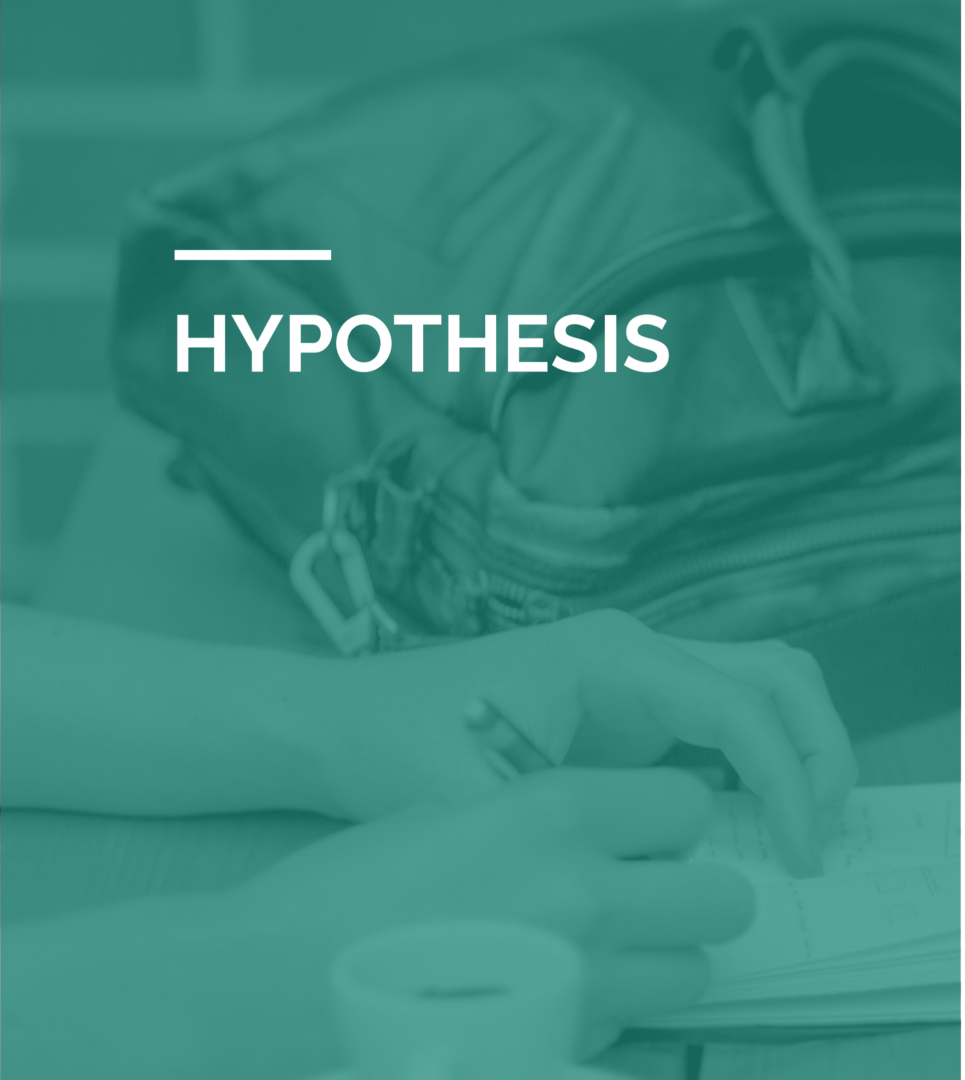
- This very messy graph tells us a few things:
  - Kids that spent longer in the study tended to have worsening asthma throughout time.
  - The color of the dot is the number of asthma prescriptions the kids had, which tends to be on the lower side.
- My inference from this data is that parents kept their kids with worsening conditions in the study in a hope of finding their asthma triggers



Highest Allergy Correlations Indexed by Birth Year



Year	Allergen_1	Allergen2	r_value
1994	ALMOND	HAZELNUT	0.51
1994	BRAZIL	HAZELNUT	0.62
1999	WALNUT	PECAN	0.58
1990	WALNUT	PECAN	0.58
1990	ALMOND	HAZELNUT	0.69
1996	PISTACH	HAZELNUT	0.53
1991	WALNUT	ALMOND	0.55
1991	WALNUT	HAZELNUT	0.64
1991	ALMOND	HAZELNUT	0.59
1991	ALMOND	CASHEW	0.60
1986	MILK	EGG	0.57
1986	MILK	WHEAT	0.57
1986	EGG	WHEAT	1.00
1984	MILK	EGG	1.00
1983	FISH	MILK	0.69
1983	FISH	EGG	0.69
1983	MILK	EGG	1.00



HYPOTHESIS

H_n: **Null Hypothesis:** If your antibody response to one allergen increases over the span of 6 years, then your response to other allergens will be unchanged.

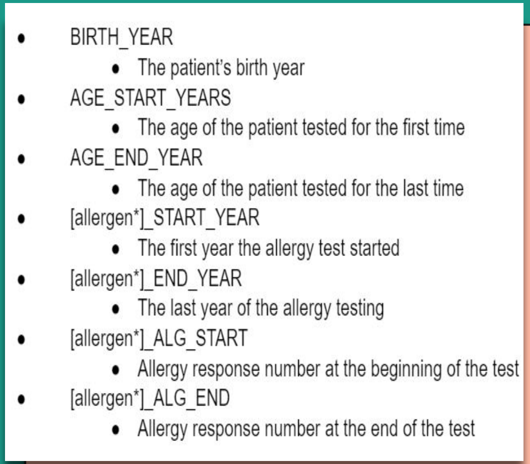
H_a: **Alternate Hypothesis:** If your antibody response to one allergen increases over the span of 6 years, then your response to other allergens will also increase.

Data Source

Source from Kaggle:
Childhood Allergies: Prevalence, Demographics

Input source includes related columns for our analysis:

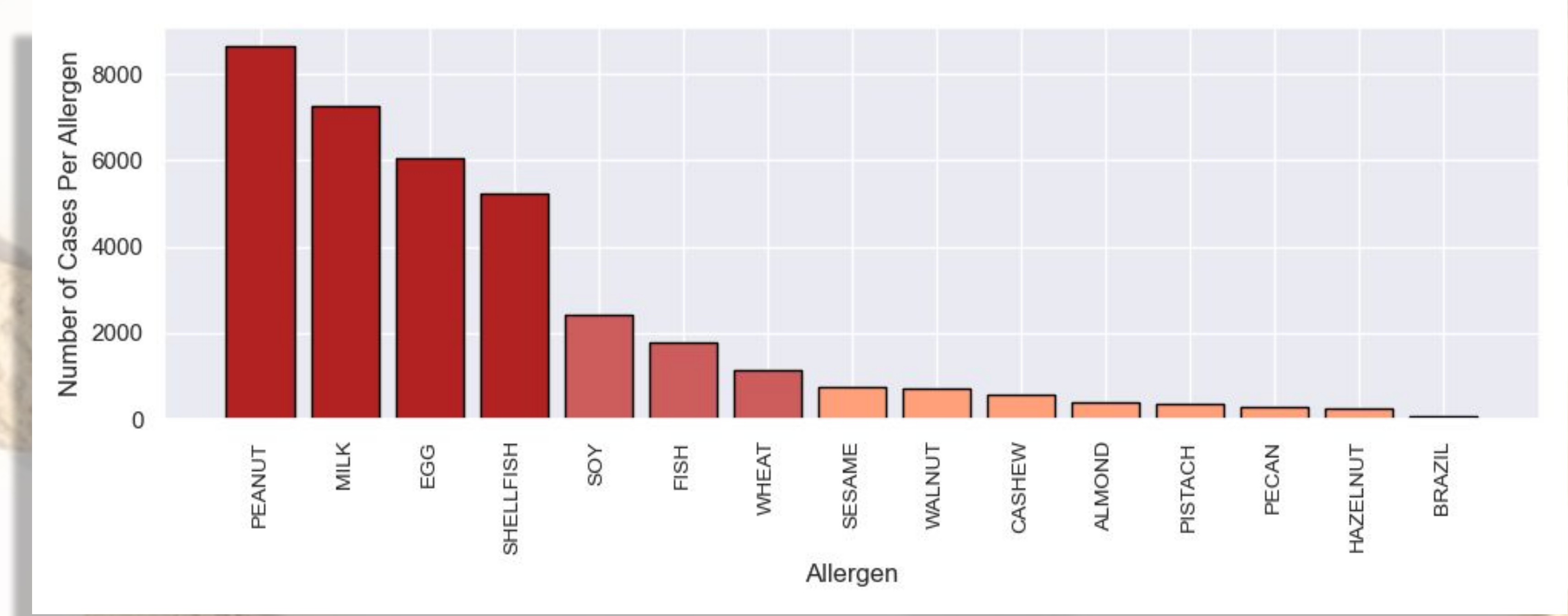
- Differences across different demographics
- Focus on one particular type of allergy
- Correlation of different allergens



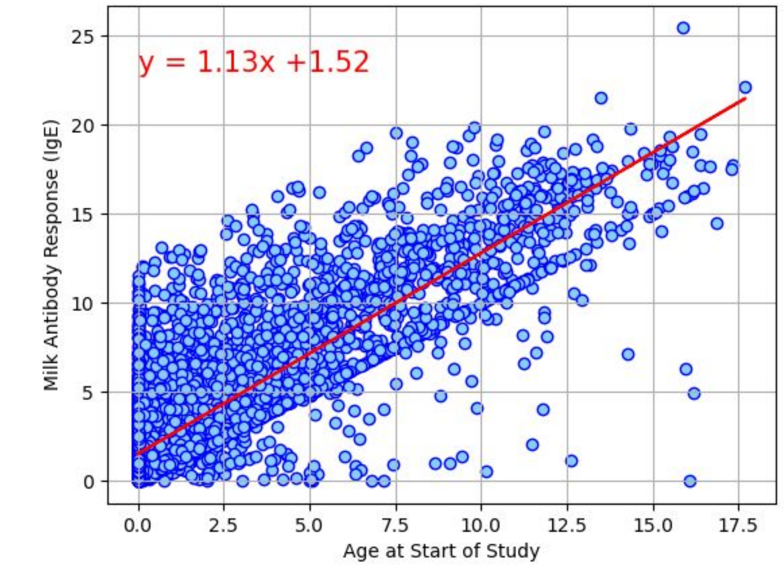
- BIRTH_YEAR
  - The patient's birth year
- AGE_START_YEARS
  - The age of the patient tested for the first time
- AGE_END_YEAR
  - The age of the patient tested for the last time
- [allergen*]_START_YEAR
  - The first year the allergy test started
- [allergen*]_END_YEAR
  - The last year of the allergy testing
- [allergen*]_ALG_START
  - Allergy response number at the beginning of the test
- [allergen*]_ALG_END
  - Allergy response number at the end of the test

*allergen includes:[SHELLFISH, FISH, MILK, SOY, EGG, WHEAT, PEANUT, SESAME, WALNUT, PECAN, PISTACH, ALMOND, BRAZIL, HAZELNUT, CASHEW]

Introduction

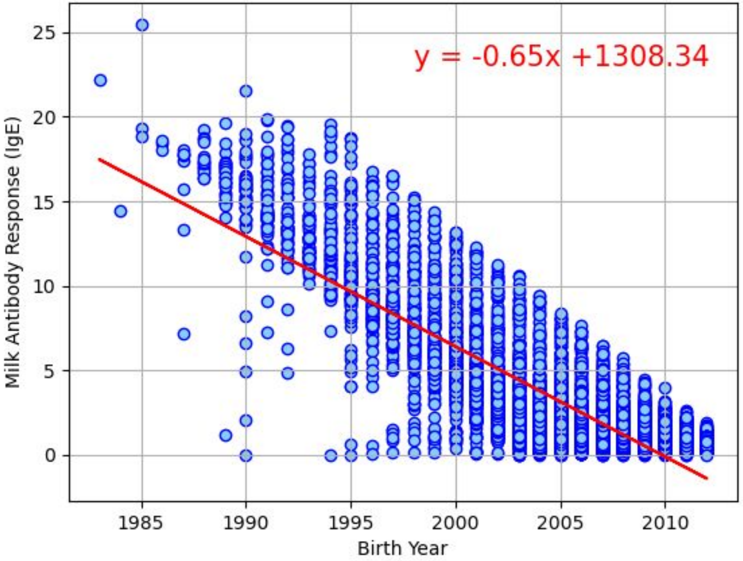


Correlation between Age and Antibody Response



`r-value is: 0.8453348158418111`

strong positive correlation



`r-value is: -0.8306979666422377`

strong negative correlation

Cleaning Data

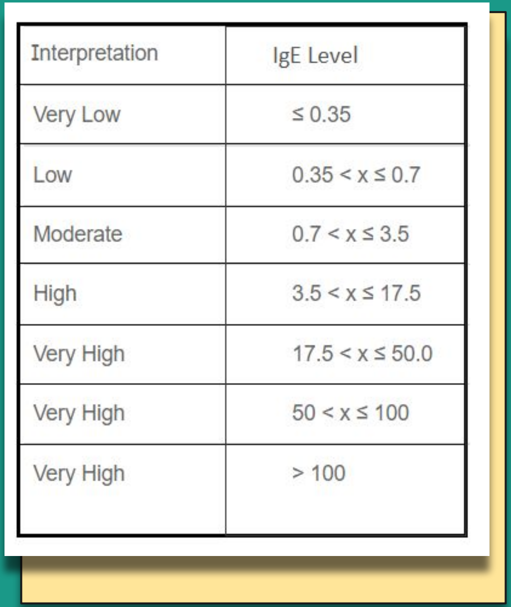
[allergen*]_ALG_START and **[allergen*]_ALG_END** Columns:

- Allergy response number (**IgE reactivity level**)
- Level of antibodies produced after being exposed to the allergen
- How strong your immune system reacts to the allergen

The dataset includes a large number of 'NA' entries:

- Each patient is not necessarily allergic to all allergens
- Not applicable allergen shows with the NA entry
- Considered as 'no allergy'.

There are wrong entries (negative values) for BIRTH_YEAR, AGE_START_YEAR, AGE_END_YEAR,.. which were removed from the table.



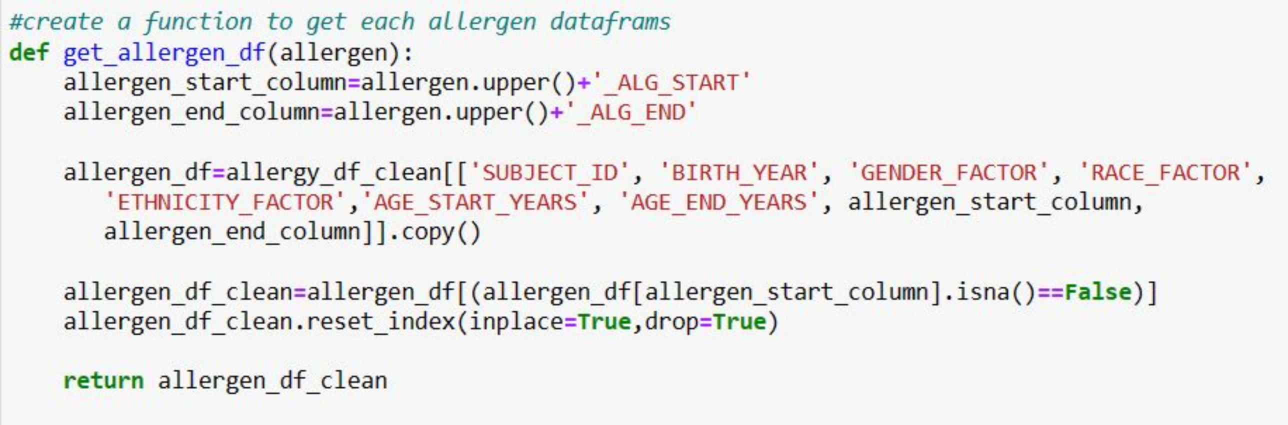
Interpretation	IgE Level
Very Low	≤ 0.35
Low	0.35 < x ≤ 0.7
Moderate	0.7 < x ≤ 3.5
High	3.5 < x ≤ 17.5
Very High	17.5 < x ≤ 50.0
Very High	50 < x ≤ 100
Very High	> 100

Initial Processes

Allergens Processing

"get_allergen (allergen_name)" function

- Create the allergen dataframe
- Analyze each specific allergen
- Focus on the patients with this particular allergen

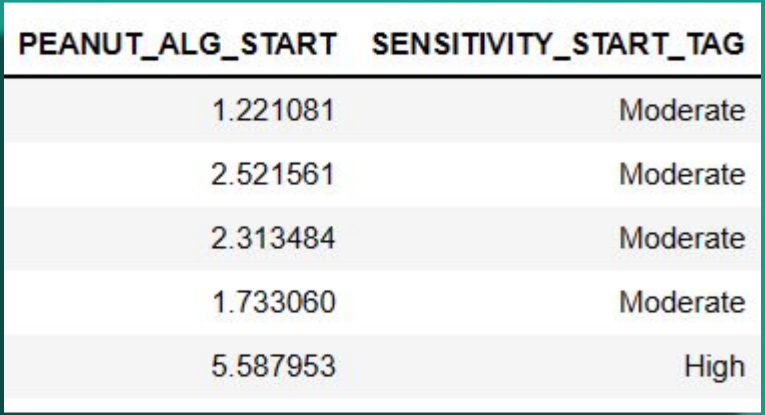


```
#create a function to get each allergen dataframs
def get_allergen_df(allergen):
    allergen_start_column=allergen.upper()+'_ALG_START'
    allergen_end_column=allergen.upper()+'_ALG_END'

    allergen_df=allergy_df_clean[['SUBJECT_ID', 'BIRTH_YEAR', 'GENDER_FACTOR', 'RACE_FACTOR',
       'ETHNICITY_FACTOR','AGE_START_YEARS', 'AGE_END_YEARS', allergen_start_column,
       allergen_end_column]].copy()

    allergen_df_clean=allergen_df[(allergen_df[allergen_start_column].isna()==False)]
    allergen_df_clean.reset_index(inplace=True,drop=True)

    return allergen_df_clean
```

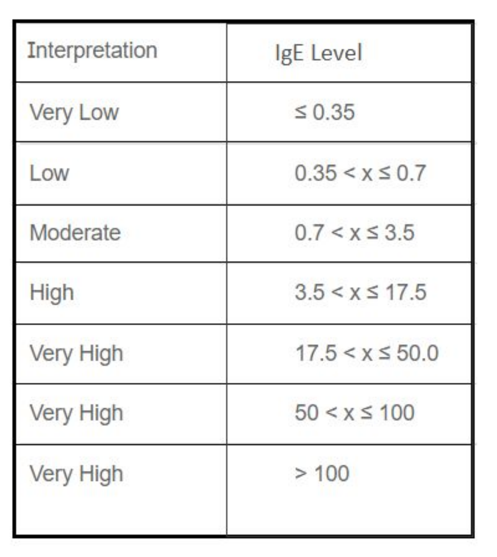
Initial Processes

Add SENSITIVITY_TAG columns

- Add new columns based on the IgE range to each allergen dataframe
- Analyze the sensitivity range distribution



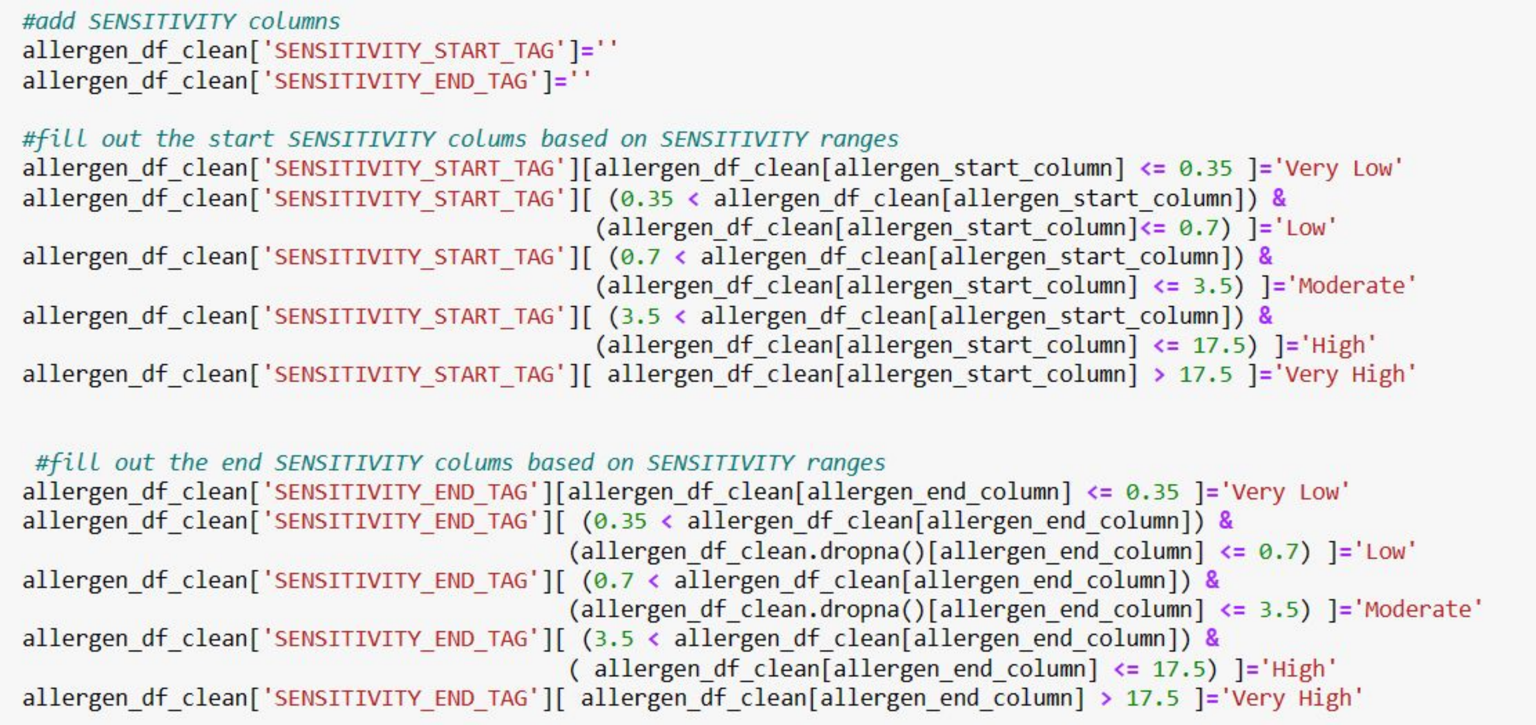
PEANUT_ALG_START	SENSITIVITY_START_TAG
1.221081	Moderate
2.521561	Moderate
2.313484	Moderate
1.733060	Moderate
5.587953	High



Interpretation	IgE Level
Very Low	≤ 0.35
Low	0.35 < x ≤ 0.7
Moderate	0.7 < x ≤ 3.5
High	3.5 < x ≤ 17.5
Very High	17.5 < x ≤ 50.0
Very High	50 < x ≤ 100
Very High	> 100

Initial Processes

SENSITIVITY_TAG columns code

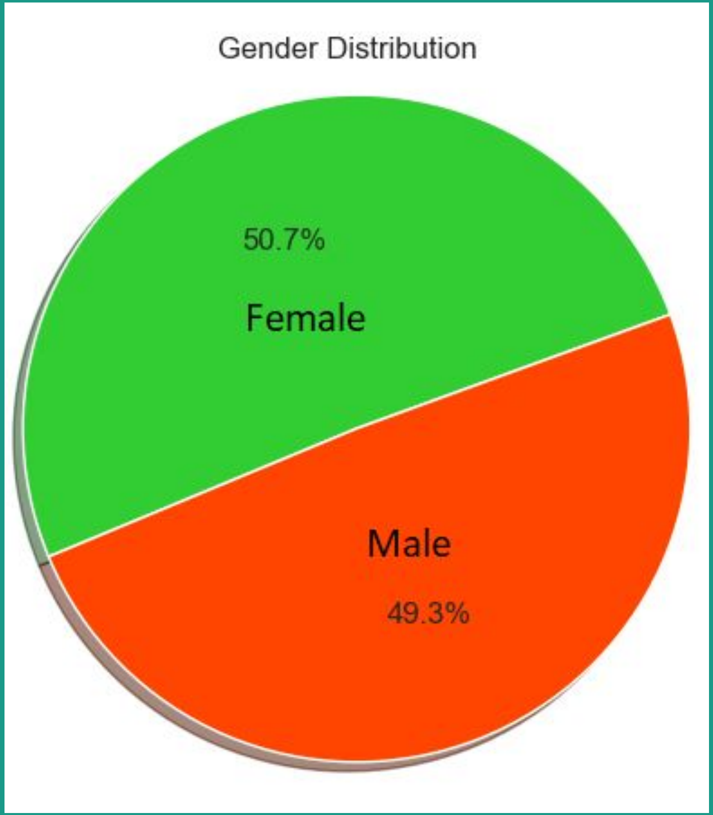


```
#add SENSITIVITY columns
allergen_df_clean['SENSITIVITY_START_TAG']=''
allergen_df_clean['SENSITIVITY_END_TAG']=''

#fill out the start SENSITIVITY colums based on SENSITIVITY ranges
allergen_df_clean['SENSITIVITY_START_TAG'][allergen_df_clean[allergen_start_column] <= 0.35 ]='Very Low'
allergen_df_clean['SENSITIVITY_START_TAG'][ (0.35 < allergen_df_clean[allergen_start_column]) &
                                           (allergen_df_clean[allergen_start_column]<= 0.7) ]='Low'
allergen_df_clean['SENSITIVITY_START_TAG'][ (0.7 < allergen_df_clean[allergen_start_column]) &
                                           (allergen_df_clean[allergen_start_column] <= 3.5) ]='Moderate'
allergen_df_clean['SENSITIVITY_START_TAG'][ (3.5 < allergen_df_clean[allergen_start_column]) &
                                           (allergen_df_clean[allergen_start_column] <= 17.5) ]='High'
allergen_df_clean['SENSITIVITY_START_TAG'][ allergen_df_clean[allergen_start_column] > 17.5 ]='Very High'


 #fill out the end SENSITIVITY colums based on SENSITIVITY ranges
allergen_df_clean['SENSITIVITY_END_TAG'][allergen_df_clean[allergen_end_column] <= 0.35 ]='Very Low'
allergen_df_clean['SENSITIVITY_END_TAG'][ (0.35 < allergen_df_clean[allergen_end_column]) &
                                         (allergen_df_clean.dropna()[allergen_end_column] <= 0.7) ]='Low'
allergen_df_clean['SENSITIVITY_END_TAG'][ (0.7 < allergen_df_clean[allergen_end_column]) &
                                         (allergen_df_clean.dropna()[allergen_end_column] <= 3.5) ]='Moderate'
allergen_df_clean['SENSITIVITY_END_TAG'][ (3.5 < allergen_df_clean[allergen_end_column]) &
                                         ( allergen_df_clean[allergen_end_column] <= 17.5) ]='High'
allergen_df_clean['SENSITIVITY_END_TAG'][ allergen_df_clean[allergen_end_column] > 17.5 ]='Very High'
```

Initial Processes

Demographic Overview

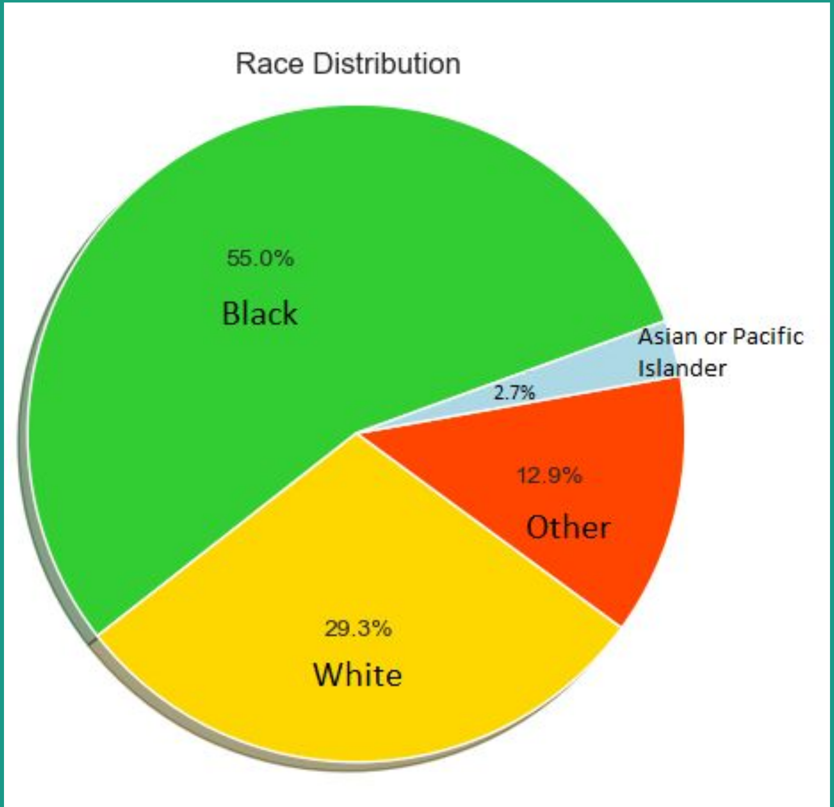
- Gender
  - Populations are almost equally distributed in terms of gender



Initial Processes

Demographic Overview

- Race
  - Low percentage in Asian/Pacific Islander(2.7%)
  - Undefined races(12.9%)
  - Black and White races are the most represented in the study.


Race Distribution
Black
Asian or Pacific Islander
Other
White