Trends with Food Allergies, Asthma, and Allergic Rhinitis in Children

Mateo Biggs

Introduction

Food allergies are a growing public health concern, particularly among children, where their prevalence has been steadily increasing in recent decades. This report analyzes a large data set of over 330,000 pediatric medical records to uncover trends in food allergies, asthma, and allergic rhinitis. By identifying common allergens and tracking their prevalence over time, the analysis aims to shed light on potential causes, demographic patterns, and the broader impact of food allergies on children's health. These insights can inform future research, healthcare interventions, and policy decisions.

Methodology

- Data Source: Data set from Zenodo containing +330,000 children medical records from The Children's Hospital of Philadelphia.
- Tools: R Studio, `tidyverse`, `lubridate`, `ggplot2`, `corrplot`.
- Steps: Data cleaning, visualization, statistical analysis.

Run once to have required packages:

```r
# Load required libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(lubridate)
library(ggplot2)
```

Here is the .csv file that holds the data for our allergen analysis. The file comes from Zenodo, and it analyzes about 333,000 children by birth year (starting/end age), gender, race/ethnicity, payer (if medicaid or not), allergen (start/end), and associated illnesses (eczema, asthma, and rhinitis). Removes starting age that are in the negatives and the tree nut allergen as it is an outlier.

```r
# Read and examine data
allergy_data <- read.csv("food-allergy-analysis-Zenodo.csv", stringsAsFactors = FALSE)

# Define the expected levels for each factor
```

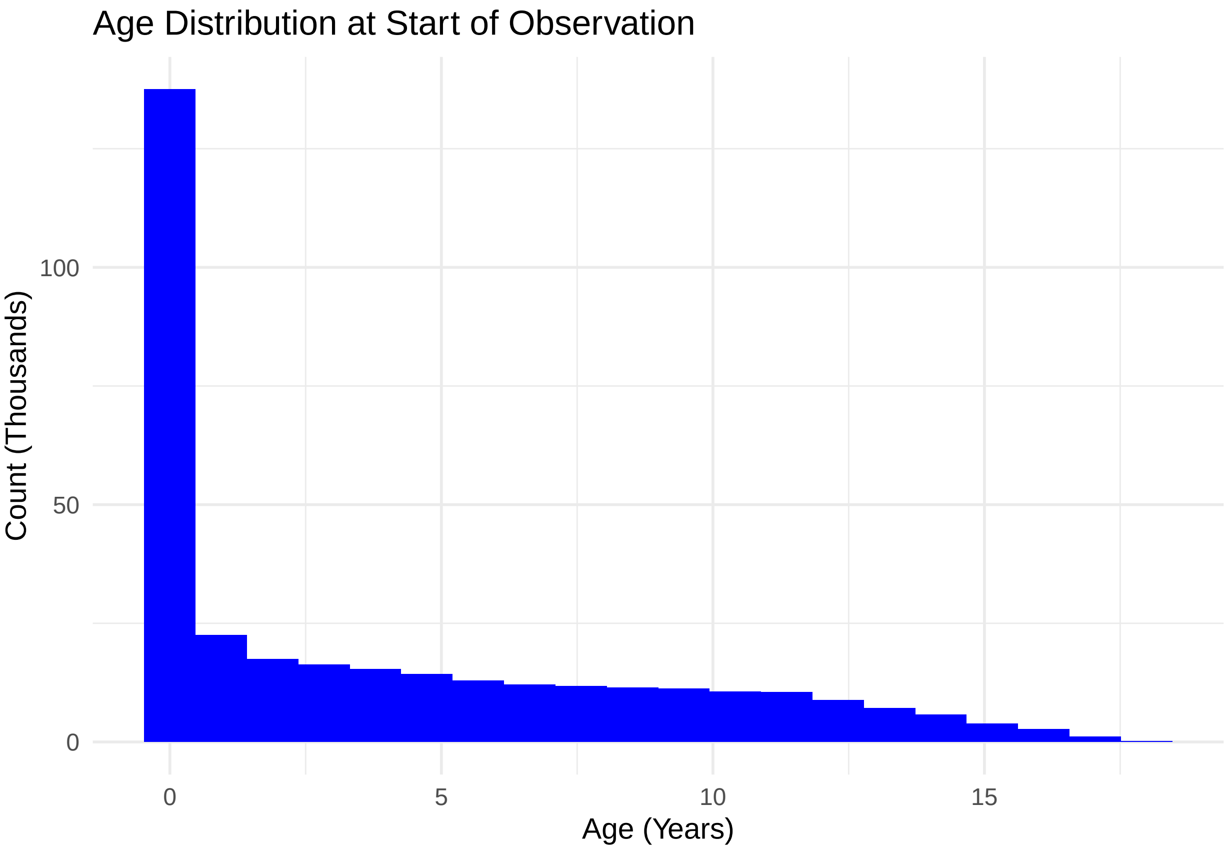
```r
gender_levels <- c("S0 - Male", "S1 - Female")
race_levels <- c("R0 - White", "R1 - Black", "R2 - Asian or Pacific Islander", "R3 - Other", "R4 - Unkn
ethnicity_levels <- c("E0 - Non-Hispanic", "E1 - Hispanic")
payer_levels <- c("P0 - Non-Medicaid", "P1 - Medicaid")

# Convert to factors with explicit levels
allergy_data <- allergy_data %>%
  mutate(
    GENDER_FACTOR = factor(GENDER_FACTOR, levels = gender_levels),
    RACE_FACTOR = factor(RACE_FACTOR, levels = race_levels),
    ETHNICITY_FACTOR = factor(ETHNICITY_FACTOR, levels = ethnicity_levels),
    PAYER_FACTOR = factor(PAYER_FACTOR, levels = payer_levels),
    ATOPIC_MARCH_COHORT = as.logical(ATOPIC_MARCH_COHORT)
  ) %>%
  # Remove rows where AGE_START_YEARS is negative
  filter(AGE_START_YEARS >= 0) %>%
  # Remove treenut-related columns
  select(-matches("TREENUT"))

# Verify the factor levels
str(allergy_data)
```

```
## 'data.frame':    333175 obs. of  48 variables:
##  $ SUBJECT_ID             : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ BIRTH_YEAR             : int  2006 1994 2006 2004 2006 2006 2006 2006 2006 2006 ...
##  $ GENDER_FACTOR          : Factor w/ 2 levels "S0 - Male","S1 - Female": 2 2 1 1 2 1 2 2 1 2 ...
##  $ RACE_FACTOR            : Factor w/ 5 levels "R0 - White","R1 - Black",..: 2 1 1 5 2 1 1 2 1 2 ...
##  $ ETHNICITY_FACTOR       : Factor w/ 2 levels "E0 - Non-Hispanic",..: 1 1 2 2 1 1 1 1 1 1 ...
##  $ PAYER_FACTOR           : Factor w/ 2 levels "P0 - Non-Medicaid",..: 2 1 1 1 1 2 1 1 1 2 ...
##  $ ATOPIC_MARCH_COHORT    : logi  FALSE FALSE TRUE FALSE FALSE FALSE ...
##  $ AGE_START_YEARS        : num  0.0931 12.2327 0.011 2.3984 0.0137 ...
##  $ AGE_END_YEARS          : num  3.16 18.88 6.73 9.11 6.19 ...
##  $ SHELLFISH_ALG_START    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ SHELLFISH_ALG_END      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ FISH_ALG_START         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ FISH_ALG_END           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ MILK_ALG_START         : num  NA NA 1 NA NA ...
##  $ MILK_ALG_END           : num  NA NA 1 NA NA ...
##  $ SOY_ALG_START          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ SOY_ALG_END            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ EGG_ALG_START          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ EGG_ALG_END            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ WHEAT_ALG_START        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ WHEAT_ALG_END          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ PEANUT_ALG_START       : num  NA NA NA NA NA ...
##  $ PEANUT_ALG_END         : num  NA NA NA NA NA ...
##  $ SESAME_ALG_START       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ SESAME_ALG_END         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ WALNUT_ALG_START       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ WALNUT_ALG_END         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ PECAN_ALG_START        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ PECAN_ALG_END          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ PISTACH_ALG_START      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ PISTACH_ALG_END        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ ALMOND_ALG_START       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ ALMOND_ALG_END         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ BRAZIL_ALG_START       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ BRAZIL_ALG_END         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ HAZELNUT_ALG_START     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ HAZELNUT_ALG_END       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ CASHEW_ALG_START       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ CASHEW_ALG_END         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ ATOPIC_DERM_START      : num  NA NA 4.88 NA NA ...
##  $ ATOPIC_DERM_END        : num  NA NA NA NA NA ...
##  $ ALLERGIC_RHINITIS_START: num  NA NA 3.92 NA NA ...
##  $ ALLERGIC_RHINITIS_END  : num  NA NA 6.16 NA NA ...
##  $ ASTHMA_START           : num  NA NA 5.13 NA NA ...
##  $ ASTHMA_END             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ FIRST_ASTHMARX         : num  NA 12.3 1.4 NA NA ...
##  $ LAST_ASTHMARX          : num  NA 18.88 6.16 NA NA ...
##  $ NUM_ASTHMARX           : int  NA 2 4 NA NA NA NA NA NA NA ...
```

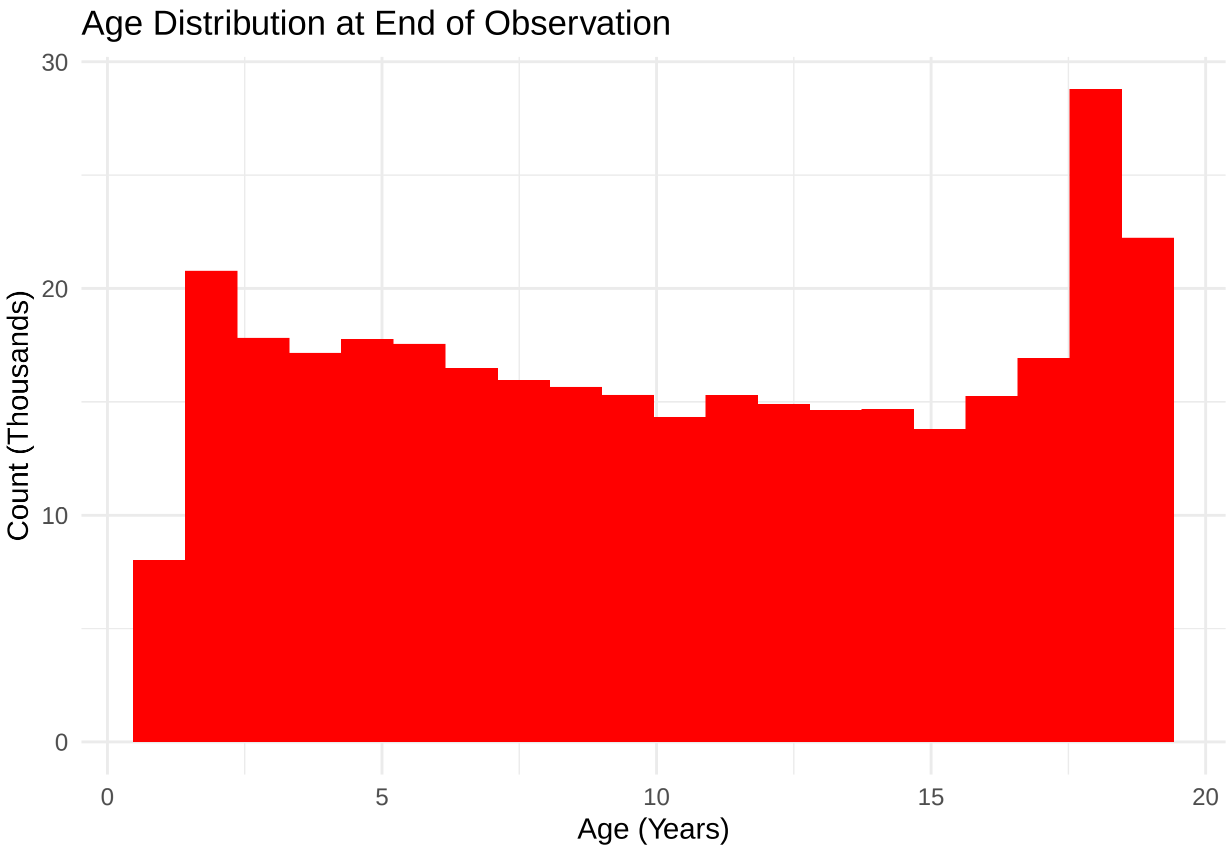
Results

This code creates two histograms showing the age distribution of patients at the start (in blue) and end (in red) of the observation period, with counts displayed in thousands. It also calculates separately the minimum and maximum ages for both time points.

```r
# Age distribution by start visualization (counts in thousands)
ggplot(allergy_data, aes(x = AGE_START_YEARS)) +
  geom_histogram(aes(y = after_stat(count) / 1000), bins = 20, fill = "blue") +
  labs(title = "Age Distribution at Start of Observation",
       x = "Age (Years)",
       y = "Count (Thousands)") +
  theme_minimal()
```



```
# Age distribution by end visualization (counts in thousands)
ggplot(allergy_data, aes(x = AGE_END_YEARS)) +
  geom_histogram(aes(y = after_stat(count) / 1000), bins = 20, fill = "red") +
  labs(title = "Age Distribution at End of Observation",
       x = "Age (Years)",
       y = "Count (Thousands)") +
  theme_minimal()
```



```r
# Print range of age
ageStart_data <- allergy_data$AGE_START_YEARS
ageEnd_data <- allergy_data$AGE_END_YEARS
range_age_start <- range(ageStart_data, na.rm = TRUE)
range_age_end <- range(ageEnd_data, na.rm = TRUE)
print(range_age_start)
```

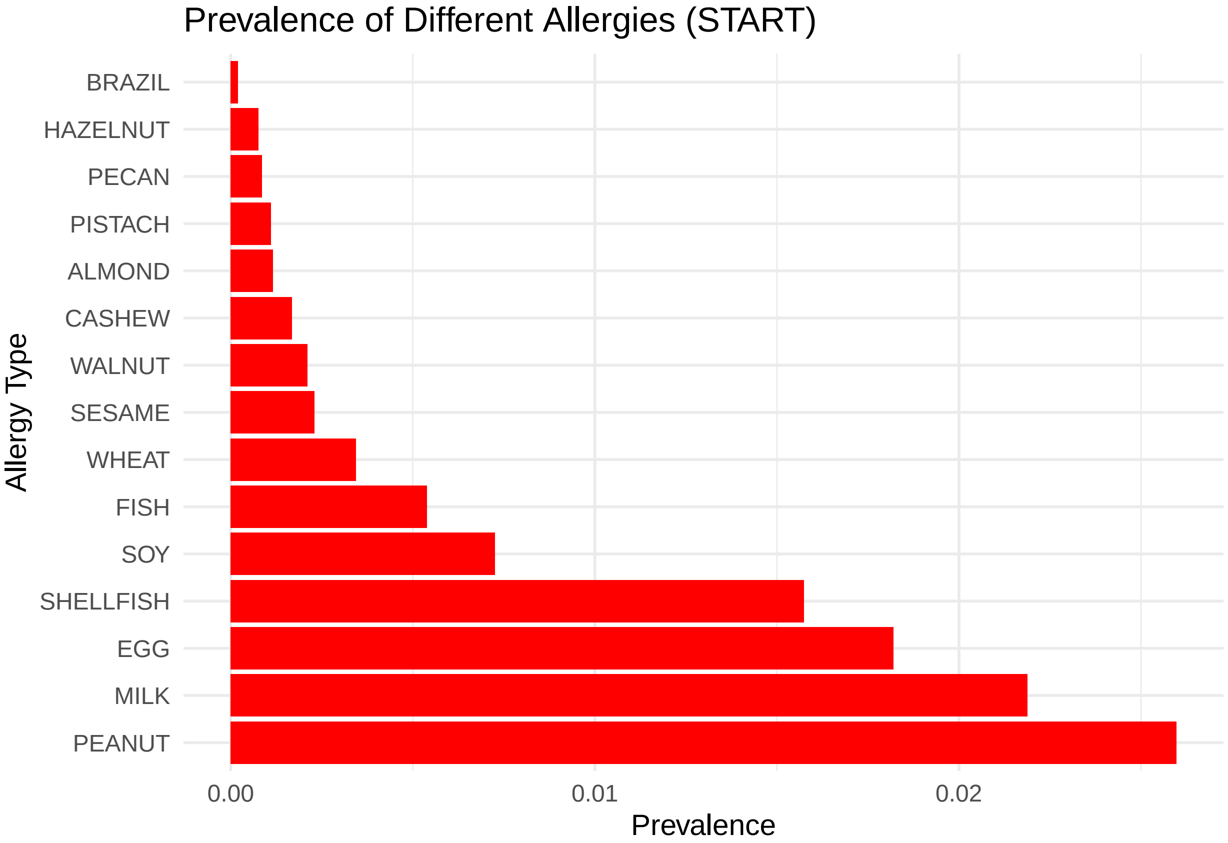
```
## [1]  0.00000 17.98494
```

```r
print(range_age_end)
```

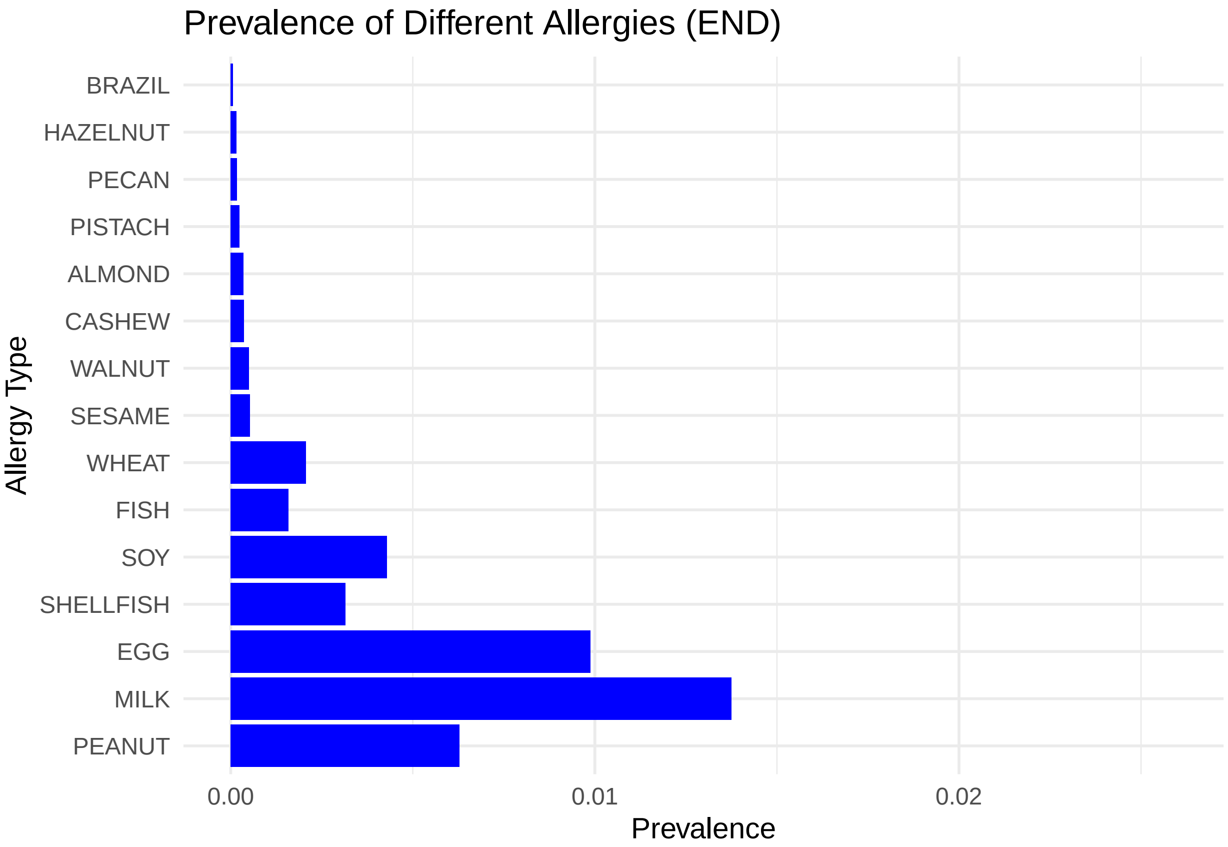
```
## [1]  1.002053 18.997947
```

This code calculates and visualizes the prevalence of different allergies at both the start and end of the observation period, creating two bar plots (red for start, blue for end) with consistent y-axis scaling and ordering of allergens based on their initial prevalence, allowing for direct comparison of how allergy patterns change over time.

```r
# Calculate prevalence of each allergy type (START)
allergy_prevalence_start <- allergy_data %>%
  summarise(across(ends_with("_ALG_START"),
                   ~sum(!is.na(.))/n())) %>%
  gather(key = "allergy_type", value = "prevalence") %>%
  mutate(allergy_type = gsub("_ALG_START", "", allergy_type))

# Calculate prevalence of each allergy type (END)
allergy_prevalence_end <- allergy_data %>%
  summarise(across(ends_with("_ALG_END"),
                   ~sum(!is.na(.))/n()) %>%
  gather(key = "allergy_type", value = "prevalence") %>%
  mutate(allergy_type = gsub("_ALG_END", "", allergy_type))

# Get the maximum prevalence value for consistent scaling
max_prevalence <- max(c(allergy_prevalence_start$prevalence,
                        allergy_prevalence_end$prevalence))

# Get consistent ordering based on START prevalence
allergy_order <- allergy_prevalence_start %>%
  arrange(desc(prevalence)) %>%
  pull(allergy_type)

# Create prevalence plot (START)
ggplot(allergy_prevalence_start,
       aes(x = factor(allergy_type, levels = allergy_order),
           y = prevalence)) +
  geom_bar(stat = "identity", fill = "red") +
  coord_flip() +
  theme_minimal() +
  ylim(0, max_prevalence) +
  labs(title = "Prevalence of Different Allergies (START)",
       x = "Allergy Type",
       y = "Prevalence")
```



```
# Create prevalence plot (END)
ggplot(allergy_prevalence_end,
       aes(x = factor(allergy_type, levels = allergy_order),
           y = prevalence)) +
  geom_bar(stat = "identity", fill = "blue") +
  coord_flip() +
  theme_minimal() +
  ylim(0, max_prevalence) +
  labs(title = "Prevalence of Different Allergies (END)",
       x = "Allergy Type",
       y = "Prevalence")
```

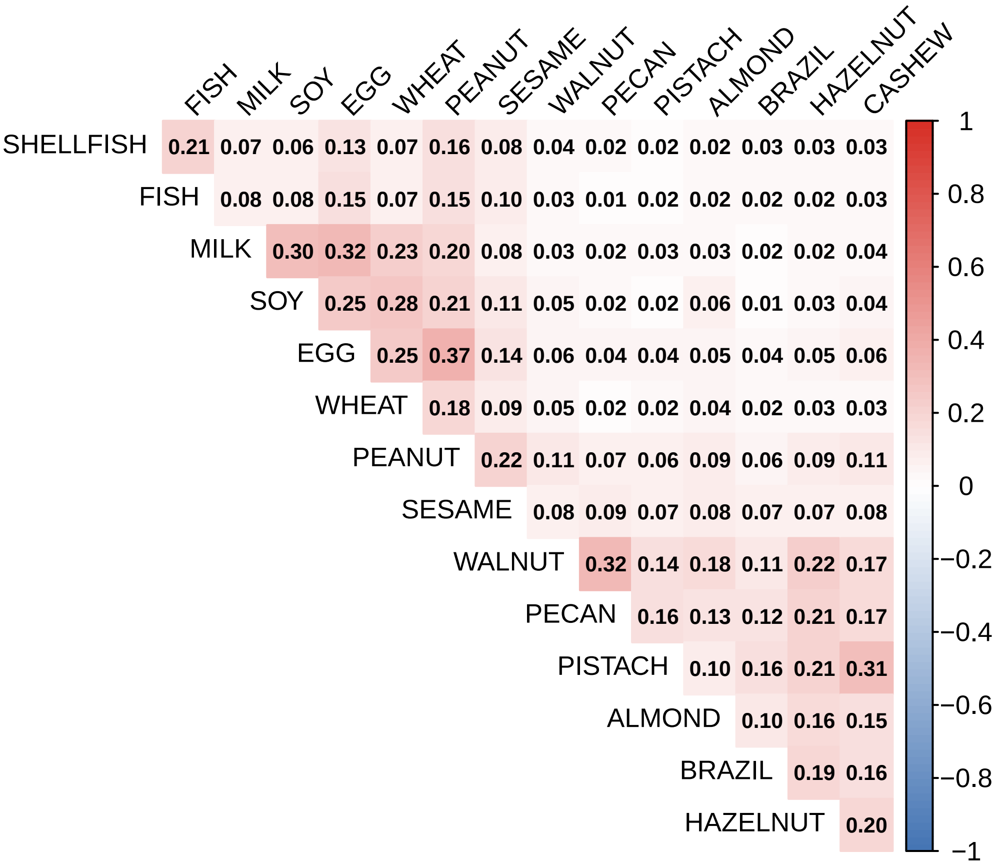


This code creates a correlation matrix heat map showing pairwise relationships between different food allergens, where blue indicates negative correlation, white indicates no correlation, and red indicates positive correlation, with correlation coefficients displayed numerically on the visualization. For the second part, the code calculates and visualizes how often pairs of allergies occur together in children (like milk with egg, or fish with shellfish), displaying the results as a horizontal bar chart with the most frequent combinations at the top. While the heat map shows correlation coefficients between all possible allergy pairs (with values from -1 to 1), the bar chart specifically quantifies how often selected allergy pairs appear together in patients

```
# Create binary columns for each allergy with cleaner names
allergy_binary <- allergy_data %>%
  mutate(across(ends_with("_ALG_START"),
         ~ifelse(is.na(.), 0, 1))) %>%
  select(ends_with("_ALG_START")) %>%
  rename_with(~str_remove(., "_ALG_START"))

# Calculate and visualize correlation matrix
library(corrplot)
```

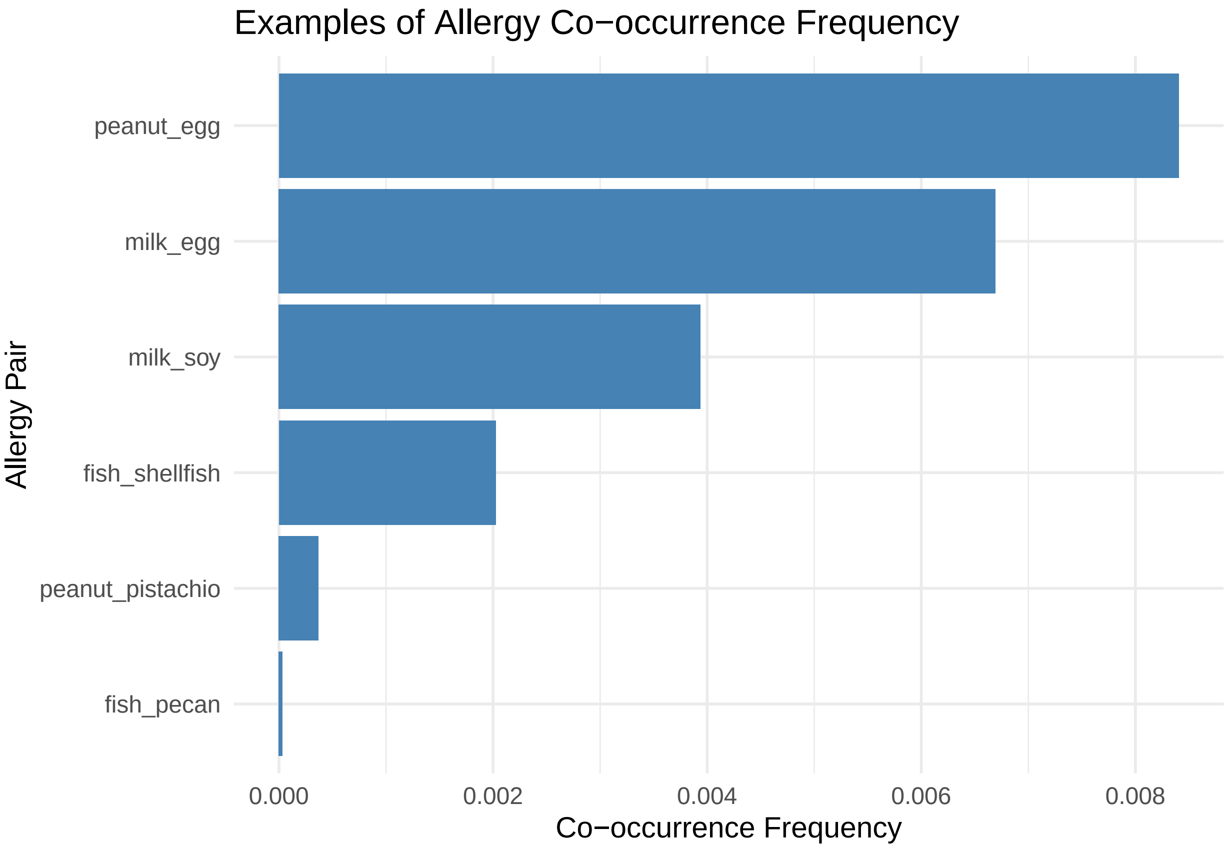
```
## corrplot 0.95 loaded
```

```
allergy_correlation <- cor(allergy_binary)
corrplot(allergy_correlation,
         method = "color",
         type = "upper",
         tl.col = "black",
         addCoef.col = "black",
         number.cex = 0.7,
         tl.srt = 45,
         tl.cex = 0.8,
         col = colorRampPalette(c("#4575B4", "white", "#D73027"))(100),
         diag = FALSE)
```



	FISH	MILK	SOY	EGG	WHEAT	PEANUT	SESAME	WALNUT	PECAN	PISTACH	ALMOND	BRAZIL	HAZELNUT	CASHEW
SHELLFISH	0.21	0.07	0.06	0.13	0.07	0.16	0.08	0.04	0.02	0.02	0.02	0.03	0.03	0.03
FISH		0.08	0.08	0.15	0.07	0.15	0.10	0.03	0.01	0.02	0.02	0.02	0.02	0.03
MILK			0.30	0.32	0.23	0.20	0.08	0.03	0.02	0.03	0.03	0.02	0.02	0.04
SOY				0.25	0.28	0.21	0.11	0.05	0.02	0.02	0.06	0.01	0.03	0.04
EGG					0.25	0.37	0.14	0.06	0.04	0.04	0.05	0.04	0.05	0.06
WHEAT						0.18	0.09	0.05	0.02	0.02	0.04	0.02	0.03	0.03
PEANUT							0.22	0.11	0.07	0.06	0.09	0.06	0.09	0.11
SESAME								0.08	0.09	0.07	0.08	0.07	0.07	0.08
WALNUT									0.32	0.14	0.18	0.11	0.22	0.17
PECAN										0.16	0.13	0.12	0.21	0.17
PISTACH											0.10	0.16	0.21	0.31
ALMOND												0.10	0.16	0.15
BRAZIL													0.19	0.16
HAZELNUT														0.20

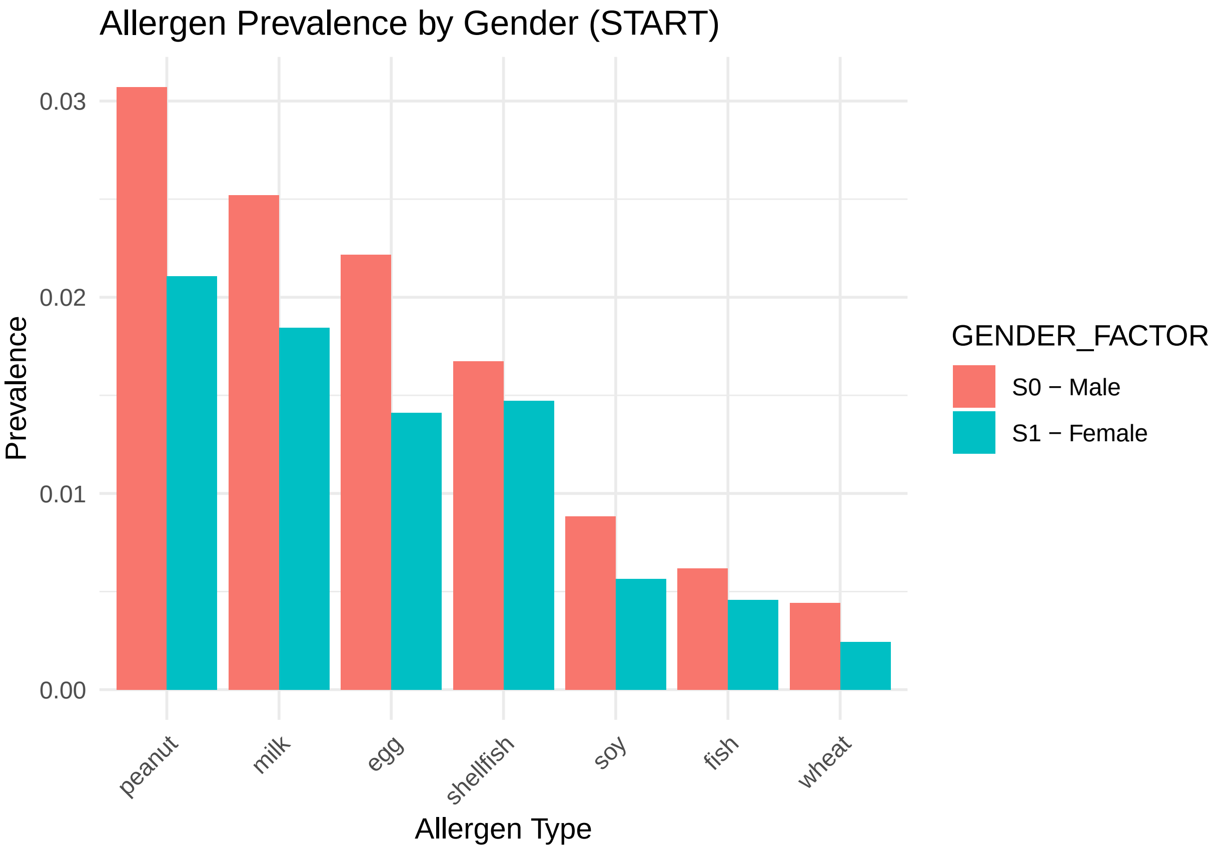
```
# Calculate co-occurrence frequencies
allergy_pairs <- allergy_binary %>%
  summarise(
    milk_egg = sum(MILK == 1 & EGG == 1)/n(),
    milk_soy = sum(MILK == 1 & SOY == 1)/n(),
    peanut_egg = sum(PEANUT == 1 & EGG == 1)/n(),
    peanut_pistachio = sum(PEANUT == 1 & PISTACH == 1)/n(),
    fish_shellfish = sum(FISH == 1 & SHELLFISH == 1)/n(),
    fish_pecan = sum(FISH == 1 & PECAN == 1)/n()
  )

# Visualize co-occurrence
allergy_pairs %>%
  gather(key = "pair", value = "frequency") %>%
  ggplot(aes(x = reorder(pair, frequency), y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Examples of Allergy Co-occurrence Frequency",
       x = "Allergy Pair",
       y = "Co-occurrence Frequency")
```



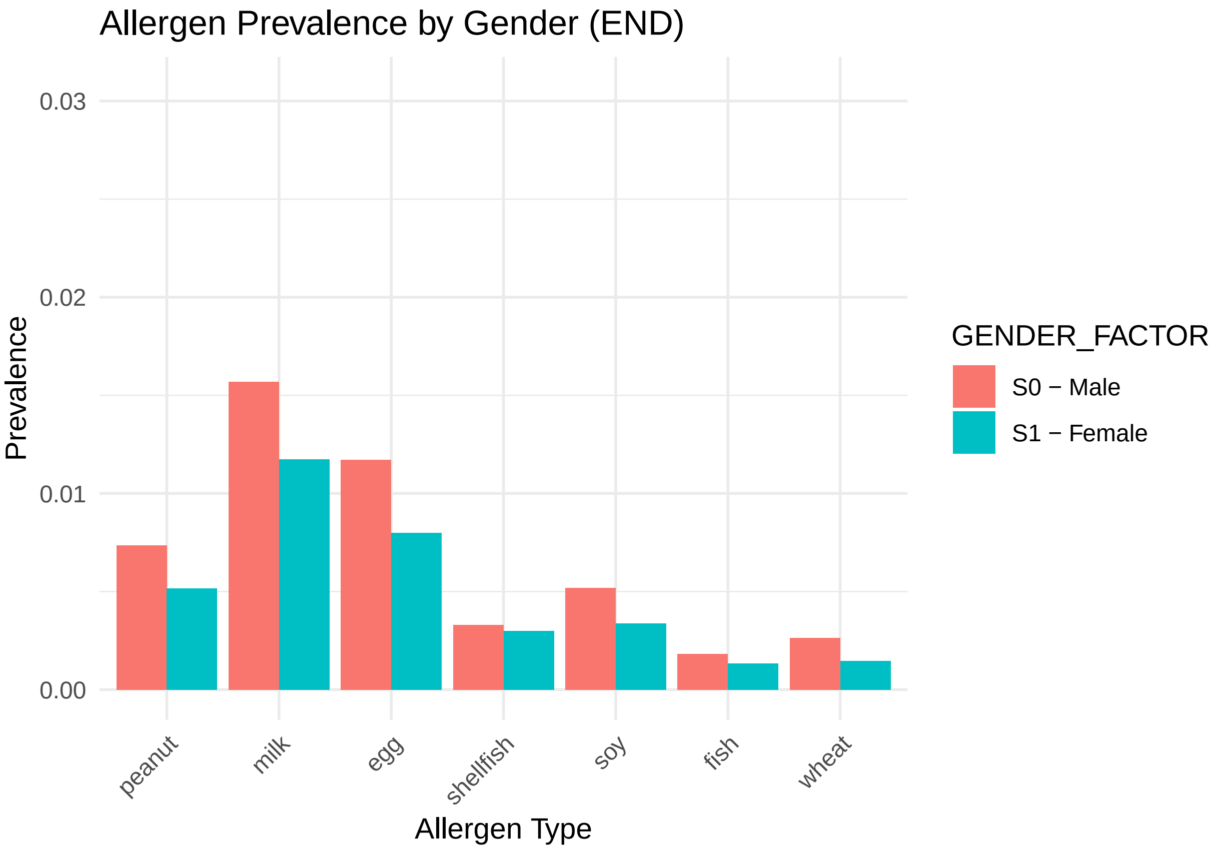
This code calculates and visualizes the prevalence of different allergies by gender at both the start and end of the observation period, creating side-by-side bar charts that show how allergy patterns differ between males and females, with consistent scaling and ordering of allergens based on their initial prevalence.

```
# Calculate allergen presence by gender with types (START)
gender_allergen_analysis_start <- allergy_data %>%
  group_by(GENDER_FACTOR) %>%
  summarise(
    peanut = sum(!is.na(PEANUT_ALG_START))/n(),
    egg = sum(!is.na(EGG_ALG_START))/n(),
    milk = sum(!is.na(MILK_ALG_START))/n(),
    wheat = sum(!is.na(WHEAT_ALG_START))/n(),
    shellfish = sum(!is.na(SHELLFISH_ALG_START))/n(),
    fish = sum(!is.na(FISH_ALG_START))/n(),
    soy = sum(!is.na(SOY_ALG_START))/n()
  )

# Calculate allergen presence by gender with types (END)
gender_allergen_analysis_end <- allergy_data %>%
```

```r
  group_by(GENDER_FACTOR) %>%
  summarise(
    peanut = sum(!is.na(PEANUT_ALG_END))/n(),
    egg = sum(!is.na(EGG_ALG_END))/n(),
    milk = sum(!is.na(MILK_ALG_END))/n(),
    wheat = sum(!is.na(WHEAT_ALG_END))/n(),
    shellfish = sum(!is.na(SHELLFISH_ALG_END))/n(),
    fish = sum(!is.na(FISH_ALG_END))/n(),
    soy = sum(!is.na(SOY_ALG_END))/n()
  )

# Get maximum prevalence value for consistent scaling
max_prevalence <- max(c(
  gather(gender_allergen_analysis_start, key = "allergen", value = "prevalence", -GENDER_FACTOR)$prevale
  gather(gender_allergen_analysis_end, key = "allergen", value = "prevalence", -GENDER_FACTOR)$prevalen
))

# Get consistent allergen ordering based on START prevalence
allergen_order <- gender_allergen_analysis_start %>%
  gather(key = "allergen", value = "prevalence", -GENDER_FACTOR) %>%
  group_by(allergen) %>%
  summarise(total_prev = sum(prevalence)) %>%
  arrange(desc(total_prev)) %>%
  pull(allergen)

# Create visualization for types (START)
ggplot(gender_allergen_analysis_start %>%
       gather(key = "allergen", value = "prevalence", -GENDER_FACTOR)) +
  geom_bar(aes(x = factor(allergen, levels = allergen_order),
               y = prevalence, fill = GENDER_FACTOR),
           stat = "identity", position = "dodge") +
  theme_minimal() +
  ylim(0, max_prevalence) +
  labs(title = "Allergen Prevalence by Gender (START)",
       x = "Allergen Type",
       y = "Prevalence") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

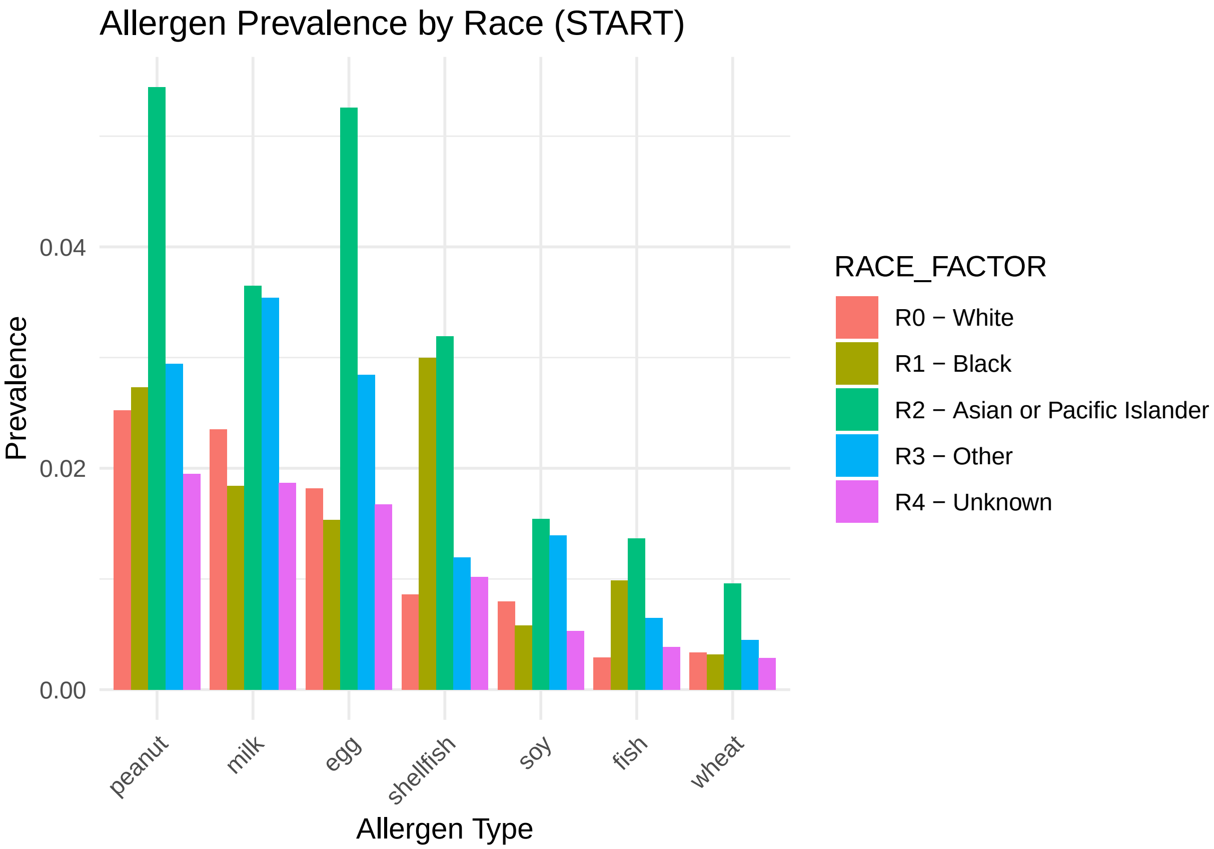


```
# Create visualization for types (END)
ggplot(gender_allergen_analysis_end %>%
       gather(key = "allergen", value = "prevalence", -GENDER_FACTOR)) +
  geom_bar(aes(x = factor(allergen, levels = allergen_order),
               y = prevalence, fill = GENDER_FACTOR),
           stat = "identity", position = "dodge") +
  theme_minimal() +
  ylim(0, max_prevalence) +
  labs(title = "Allergen Prevalence by Gender (END)",
       x = "Allergen Type",
       y = "Prevalence") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

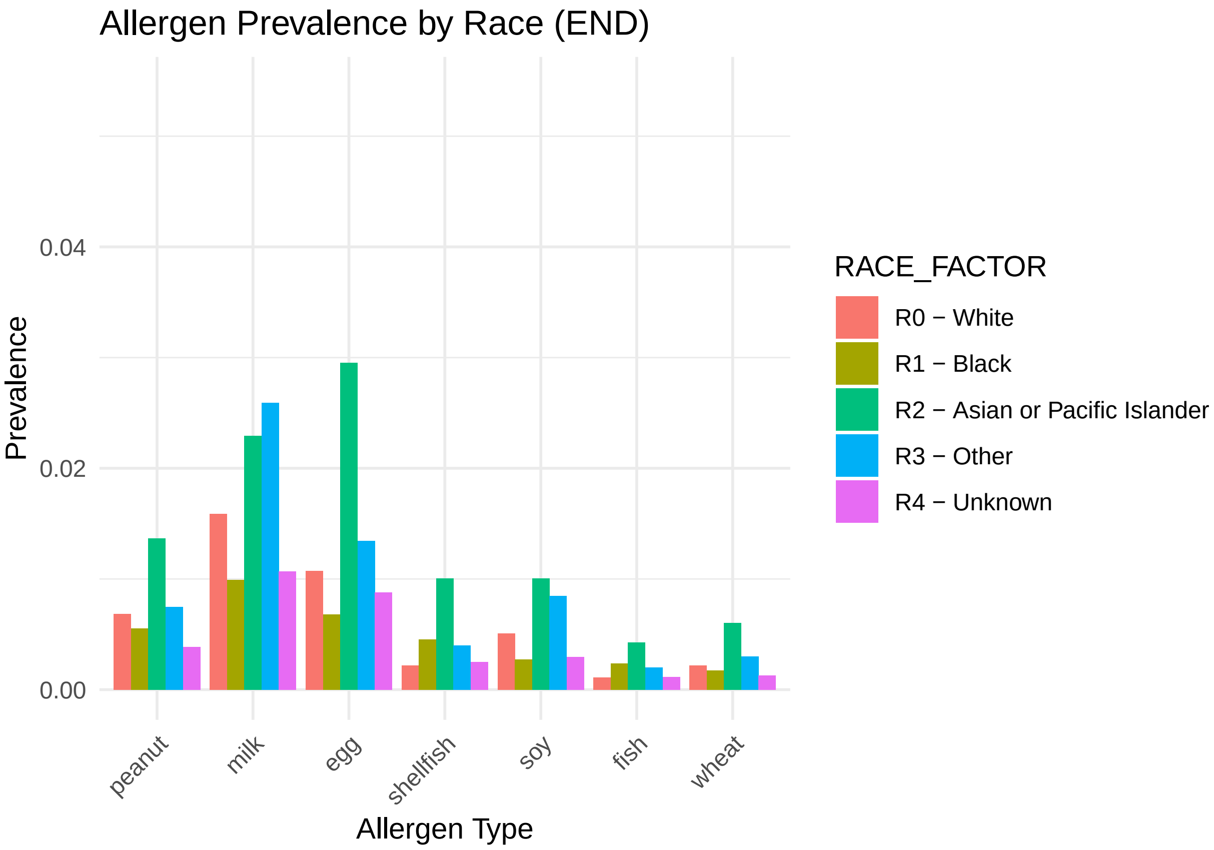


This code calculates and visualizes the prevalence of different allergies at both the start and end of the observation period, creating bar charts that show how allergy patterns vary by race (White, Black, Asian/Pacific Islander, Other, Unknown) and ethnicity (Hispanic, Non-Hispanic). It keeps consistency with scaling and ordering of allergens based on their initial prevalence.

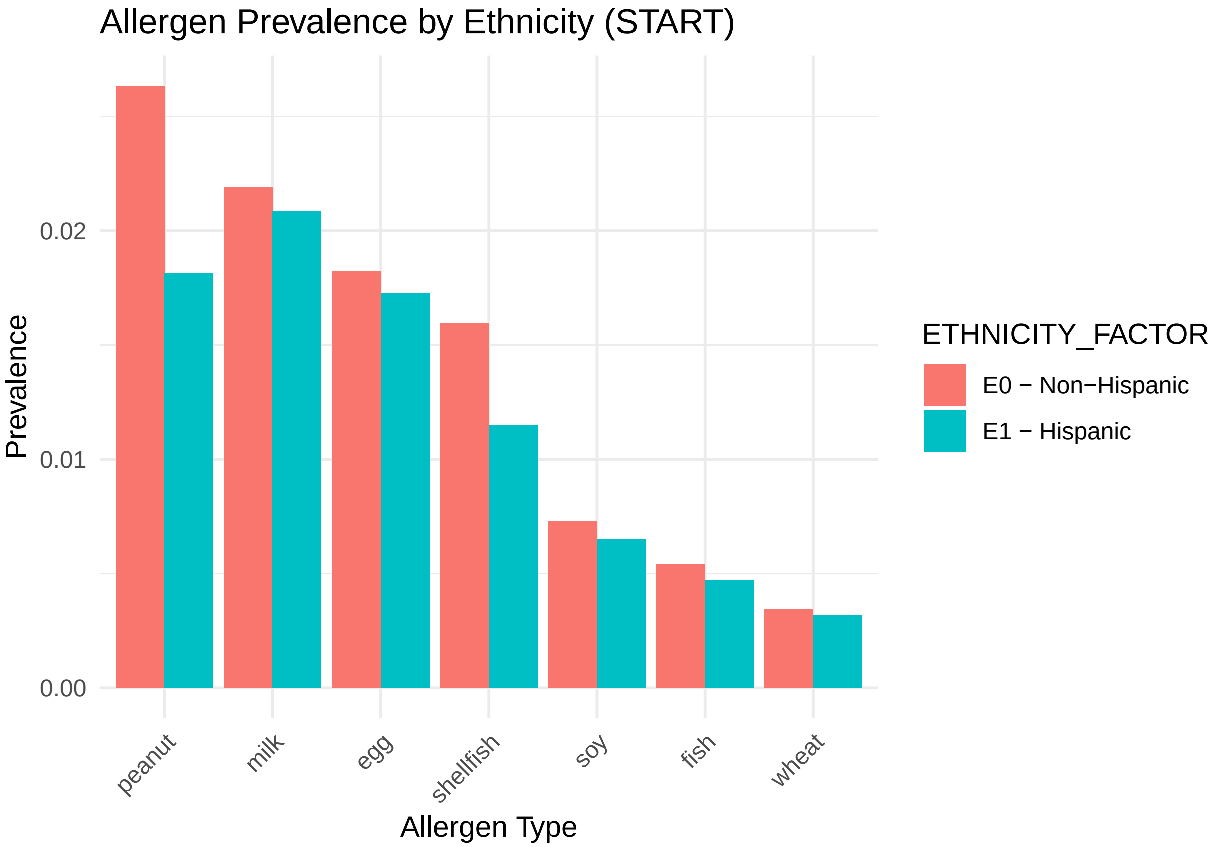
```r
# Calculate allergen presence by race (START)
race_allergen_analysis_start <- allergy_data %>%
  group_by(RACE_FACTOR) %>%
  summarise(
    peanut = sum(!is.na(PEANUT_ALG_START))/n(),
    egg = sum(!is.na(EGG_ALG_START))/n(),
    milk = sum(!is.na(MILK_ALG_START))/n(),
    wheat = sum(!is.na(WHEAT_ALG_START))/n(),
    shellfish = sum(!is.na(SHELLFISH_ALG_START))/n(),
    fish = sum(!is.na(FISH_ALG_START))/n(),
    soy = sum(!is.na(SOY_ALG_START))/n()
  )

# Calculate allergen presence by race (END)
race_allergen_analysis_end <- allergy_data %>%
  group_by(RACE_FACTOR) %>%
  summarise(
    peanut = sum(!is.na(PEANUT_ALG_END))/n(),
    egg = sum(!is.na(EGG_ALG_END))/n(),
    milk = sum(!is.na(MILK_ALG_END))/n(),
    wheat = sum(!is.na(WHEAT_ALG_END))/n(),
    shellfish = sum(!is.na(SHELLFISH_ALG_END))/n(),
    fish = sum(!is.na(FISH_ALG_END))/n(),
    soy = sum(!is.na(SOY_ALG_END))/n()
  )

# Get maximum prevalence and consistent ordering for race
max_prevalence_race <- max(c(
  gather(race_allergen_analysis_start, key = "allergen", value = "prevalence", -RACE_FACTOR)$prevalence
  gather(race_allergen_analysis_end, key = "allergen", value = "prevalence", -RACE_FACTOR)$prevalence
))

allergen_order_race <- race_allergen_analysis_start %>%
  gather(key = "allergen", value = "prevalence", -RACE_FACTOR) %>%
  group_by(allergen) %>%
  summarise(total_prev = sum(prevalence)) %>%
  arrange(desc(total_prev)) %>%
  pull(allergen)

# Visualizations for race
ggplot(race_allergen_analysis_start %>%
       gather(key = "allergen", value = "prevalence", -RACE_FACTOR)) +
  geom_bar(aes(x = factor(allergen, levels = allergen_order_race),
               y = prevalence, fill = RACE_FACTOR),
           stat = "identity", position = "dodge") +
  theme_minimal() +
  ylim(0, max_prevalence_race) +
  labs(title = "Allergen Prevalence by Race (START)",
       x = "Allergen Type",
       y = "Prevalence") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



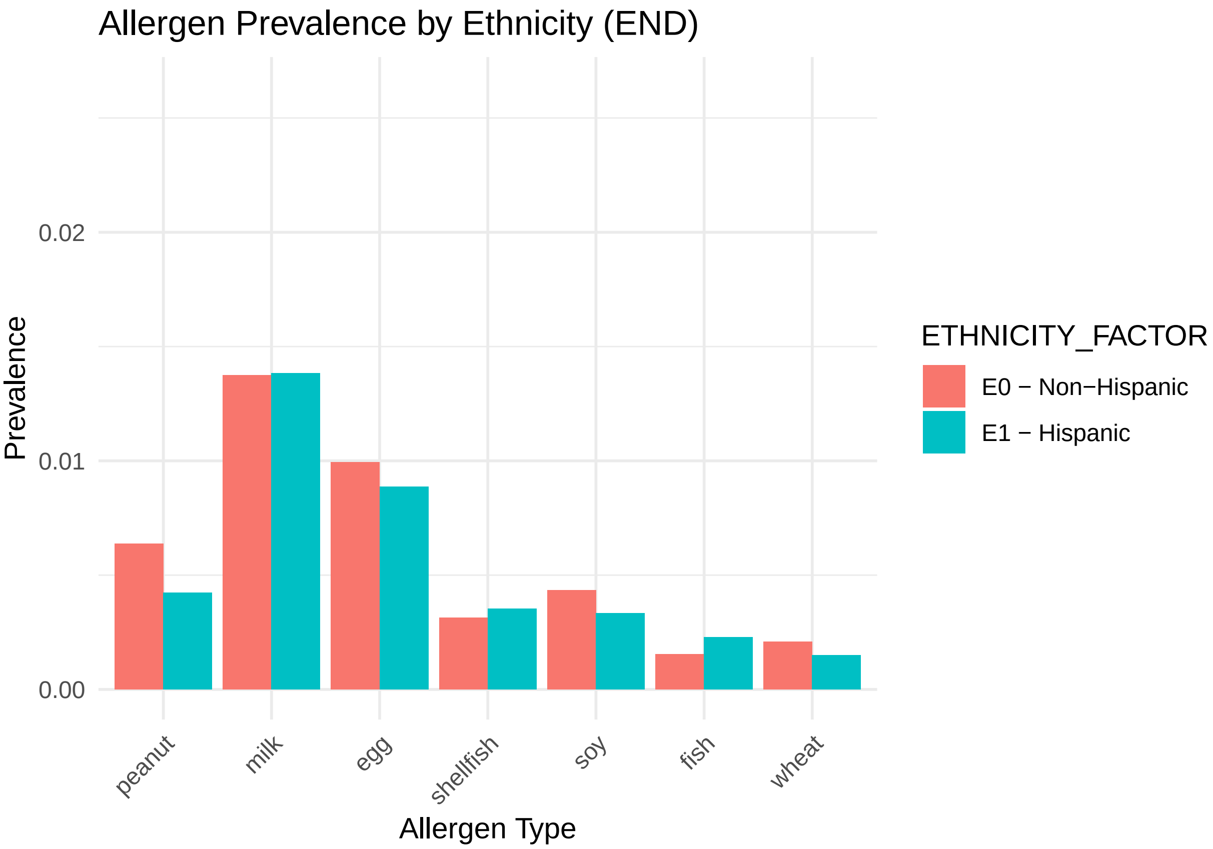
```
ggplot(race_allergen_analysis_end %>%
       gather(key = "allergen", value = "prevalence", -RACE_FACTOR)) +
  geom_bar(aes(x = factor(allergen, levels = allergen_order_race),
               y = prevalence, fill = RACE_FACTOR),
           stat = "identity", position = "dodge") +
  theme_minimal() +
  ylim(0, max_prevalence_race) +
  labs(title = "Allergen Prevalence by Race (END)",
       x = "Allergen Type",
       y = "Prevalence") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
# Calculate allergen presence by ethnicity (START and END)
ethnicity_allergen_analysis_start <- allergy_data %>%
  group_by(ETHNICITY_FACTOR) %>%
  summarise(
    peanut = sum(!is.na(PEANUT_ALG_START))/n(),
    egg = sum(!is.na(EGG_ALG_START))/n(),
    milk = sum(!is.na(MILK_ALG_START))/n(),
    wheat = sum(!is.na(WHEAT_ALG_START))/n(),
    shellfish = sum(!is.na(SHELLFISH_ALG_START))/n(),
    fish = sum(!is.na(FISH_ALG_START))/n(),
    soy = sum(!is.na(SOY_ALG_START))/n()
  )

ethnicity_allergen_analysis_end <- allergy_data %>%
  group_by(ETHNICITY_FACTOR) %>%
  summarise(
    peanut = sum(!is.na(PEANUT_ALG_END))/n(),
    egg = sum(!is.na(EGG_ALG_END))/n(),
    milk = sum(!is.na(MILK_ALG_END))/n(),
    wheat = sum(!is.na(WHEAT_ALG_END))/n(),
    shellfish = sum(!is.na(SHELLFISH_ALG_END))/n(),
    fish = sum(!is.na(FISH_ALG_END))/n(),
    soy = sum(!is.na(SOY_ALG_END))/n()
  )

# Get maximum prevalence and consistent ordering for ethnicity
max_prevalence_ethnicity <- max(c(
  gather(ethnicity_allergen_analysis_start, key = "allergen", value = "prevalence", -ETHNICITY_FACTOR)$p
  gather(ethnicity_allergen_analysis_end, key = "allergen", value = "prevalence", -ETHNICITY_FACTOR)$pre
))

allergen_order_ethnicity <- ethnicity_allergen_analysis_start %>%
  gather(key = "allergen", value = "prevalence", -ETHNICITY_FACTOR) %>%
  group_by(allergen) %>%
  summarise(total_prev = sum(prevalence)) %>%
  arrange(desc(total_prev)) %>%
  pull(allergen)

# Visualizations for ethnicity
ggplot(ethnicity_allergen_analysis_start %>%
       gather(key = "allergen", value = "prevalence", -ETHNICITY_FACTOR)) +
  geom_bar(aes(x = factor(allergen, levels = allergen_order_ethnicity),
               y = prevalence, fill = ETHNICITY_FACTOR),
           stat = "identity", position = "dodge") +
  theme_minimal() +
  ylim(0, max_prevalence_ethnicity) +
  labs(title = "Allergen Prevalence by Ethnicity (START)",
       x = "Allergen Type",
       y = "Prevalence") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



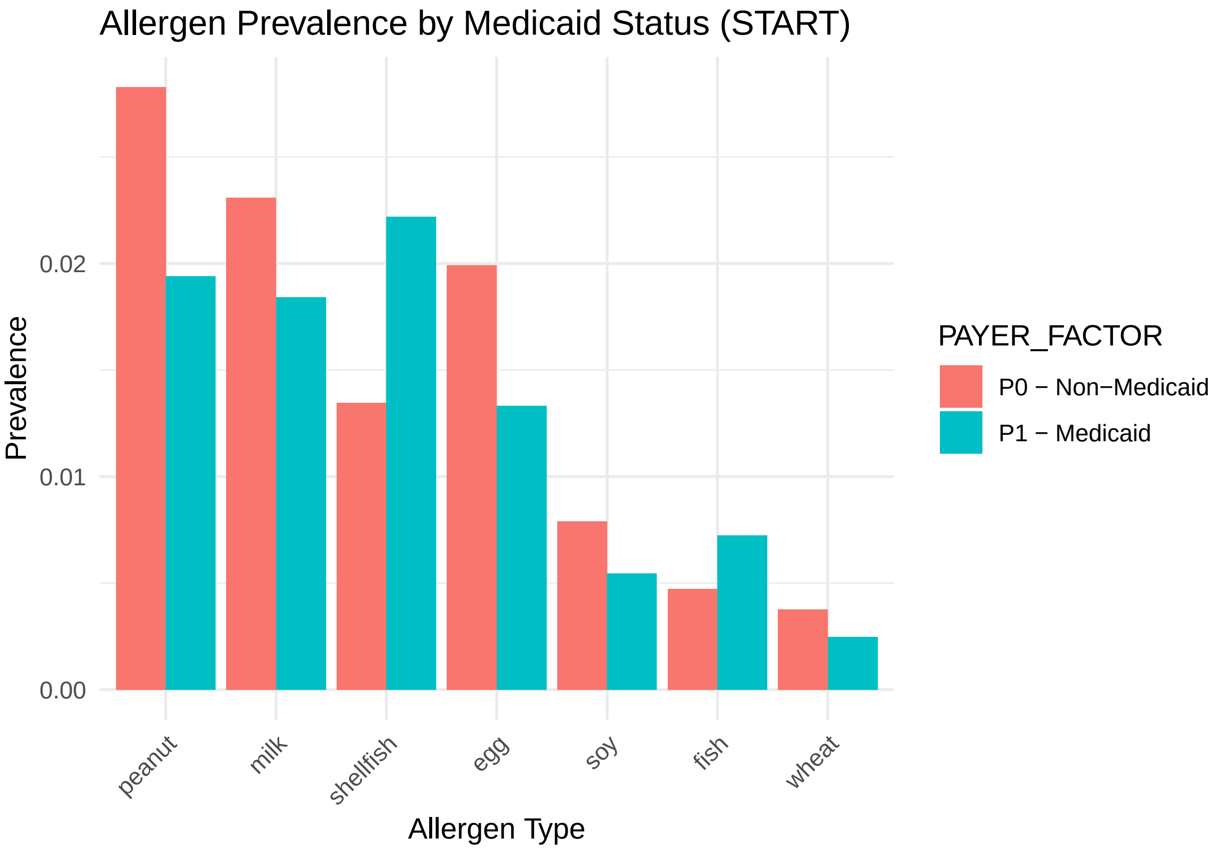
```
ggplot(ethnicity_allergen_analysis_end %>%
       gather(key = "allergen", value = "prevalence", -ETHNICITY_FACTOR)) +
  geom_bar(aes(x = factor(allergen, levels = allergen_order_ethnicity),
              y = prevalence, fill = ETHNICITY_FACTOR),
          stat = "identity", position = "dodge") +
  theme_minimal() +
  ylim(0, max_prevalence_ethnicity) +
  labs(title = "Allergen Prevalence by Ethnicity (END)",
      x = "Allergen Type",
      y = "Prevalence") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



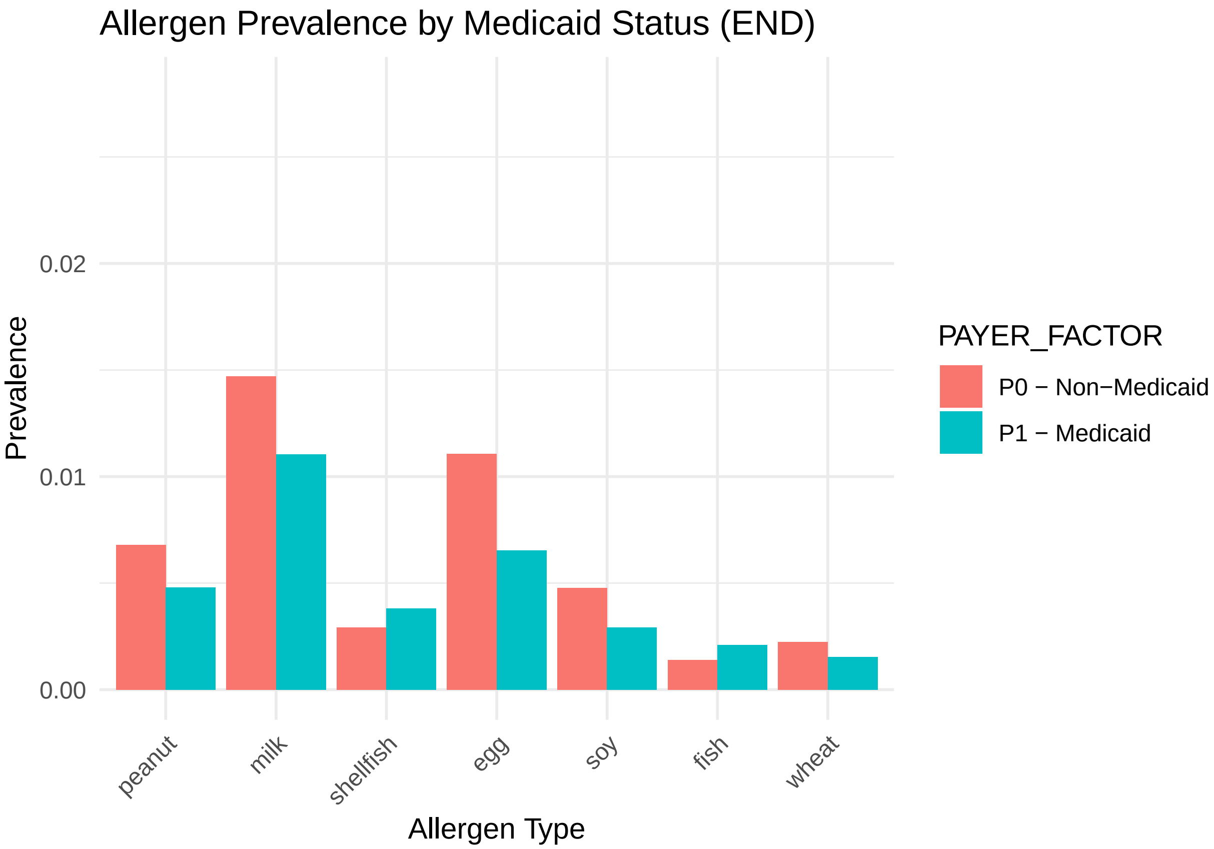
This code calculates and visualizes the prevalence of different allergies between Medicaid and non-Medicaid patients at both the start and end of the observation period, creating side-by-side bar charts that show how allergy patterns differ based on insurance status. It keeps consistency with scaling and ordering of allergens based on their initial prevalence.

```
# Calculate allergen presence by payer status (START)
payer_allergen_analysis_start <- allergy_data %>%
  group_by(PAYER_FACTOR) %>%
  summarise(
    peanut = sum(!is.na(PEANUT_ALG_START))/n(),
    egg = sum(!is.na(EGG_ALG_START))/n(),
    milk = sum(!is.na(MILK_ALG_START))/n(),
    wheat = sum(!is.na(WHEAT_ALG_START))/n(),
    shellfish = sum(!is.na(SHELLFISH_ALG_START))/n(),
    fish = sum(!is.na(FISH_ALG_START))/n(),
    soy = sum(!is.na(SOY_ALG_START))/n()
  )
```

```
# Calculate allergen presence by payer status (END)
payer_allergen_analysis_end <- allergy_data %>%
  group_by(PAYER_FACTOR) %>%
  summarise(
    peanut = sum(!is.na(PEANUT_ALG_END))/n(),
    egg = sum(!is.na(EGG_ALG_END))/n(),
    milk = sum(!is.na(MILK_ALG_END))/n(),
    wheat = sum(!is.na(WHEAT_ALG_END))/n(),
    shellfish = sum(!is.na(SHELLFISH_ALG_END))/n(),
    fish = sum(!is.na(FISH_ALG_END))/n(),
    soy = sum(!is.na(SOY_ALG_END))/n()
  )

# Get maximum prevalence and consistent ordering
max_prevalence_payer <- max(c(
  gather(payer_allergen_analysis_start, key = "allergen", value = "prevalence", -PAYER_FACTOR)$prevalen
  gather(payer_allergen_analysis_end, key = "allergen", value = "prevalence", -PAYER_FACTOR)$prevalence
))

allergen_order_payer <- payer_allergen_analysis_start %>%
  gather(key = "allergen", value = "prevalence", -PAYER_FACTOR) %>%
  group_by(allergen) %>%
  summarise(total_prev = sum(prevalence)) %>%
  arrange(desc(total_prev)) %>%
  pull(allergen)

# Create visualization (START)
ggplot(payer_allergen_analysis_start %>%
       gather(key = "allergen", value = "prevalence", -PAYER_FACTOR)) +
  geom_bar(aes(x = factor(allergen, levels = allergen_order_payer),
               y = prevalence, fill = PAYER_FACTOR),
           stat = "identity", position = "dodge") +
  theme_minimal() +
  ylim(0, max_prevalence_payer) +
  labs(title = "Allergen Prevalence by Medicaid Status (START)",
       x = "Allergen Type",
       y = "Prevalence") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



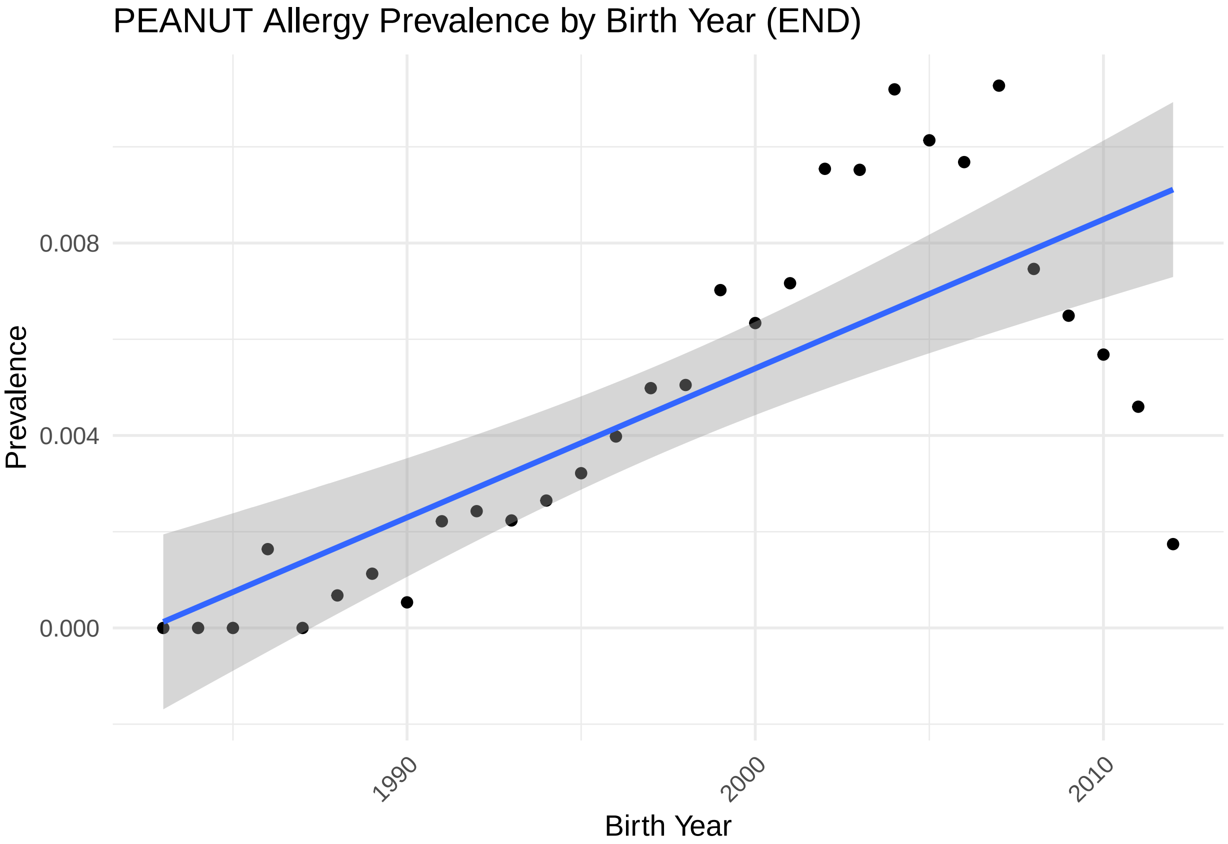
```r
# Create visualization (END)
ggplot(payer_allergen_analysis_end %>%
       gather(key = "allergen", value = "prevalence", -PAYER_FACTOR)) +
  geom_bar(aes(x = factor(allergen, levels = allergen_order_payer),
               y = prevalence, fill = PAYER_FACTOR),
           stat = "identity", position = "dodge") +
  theme_minimal() +
  ylim(0, max_prevalence_payer) +
  labs(title = "Allergen Prevalence by Medicaid Status (END)",
       x = "Allergen Type",
       y = "Prevalence") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



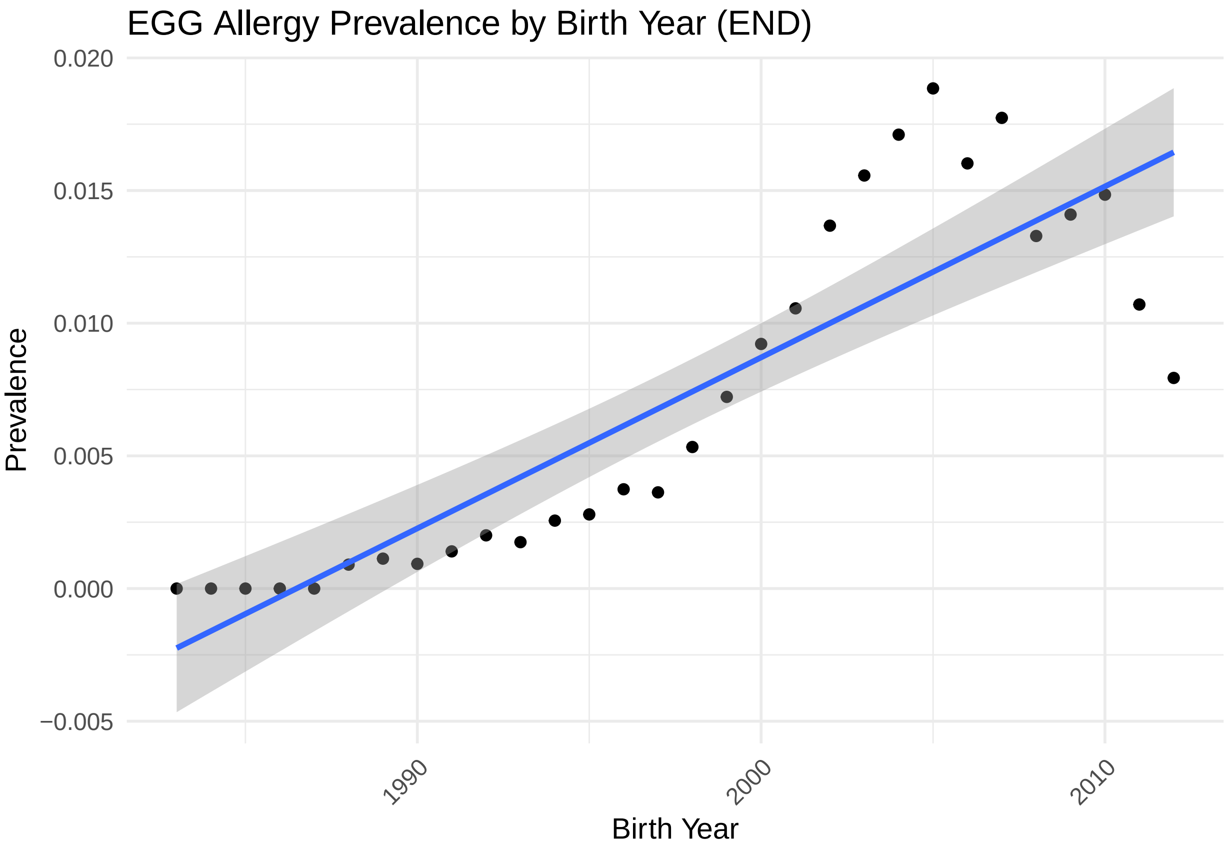
This code creates individual scatter plots for each type of allergy (peanut, egg, milk, etc.) showing how their prevalence changes across different birth years, with each plot including a linear trend line and confidence interval to visualize the overall pattern of allergy rates over time.

```
# List of allergens
allergens <- c("PEANUT", "EGG", "MILK", "WHEAT", "SHELLFISH", "FISH", "SOY")

# Function to create a plot for a single allergen with trend line
create_allergen_plot <- function(allergen) {
  allergy_data %>%
    group_by(BIRTH_YEAR) %>%
    summarise(prevalence = sum(!is.na(get(paste0(allergen, "_ALG_END"))))/n()) %>%
    ggplot(aes(x = BIRTH_YEAR, y = prevalence)) +
    geom_point() +
    geom_smooth(method = "lm", se = TRUE) +  # Add trend line with confidence interval
    theme_minimal() +
    labs(title = paste(allergen, "Allergy Prevalence by Birth Year (END)"),
         x = "Birth Year",
         y = "Prevalence") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}

# Create and display plots
allergen_plots <- lapply(allergens, create_allergen_plot)
for (plot in allergen_plots) {
  print(plot)
}
```

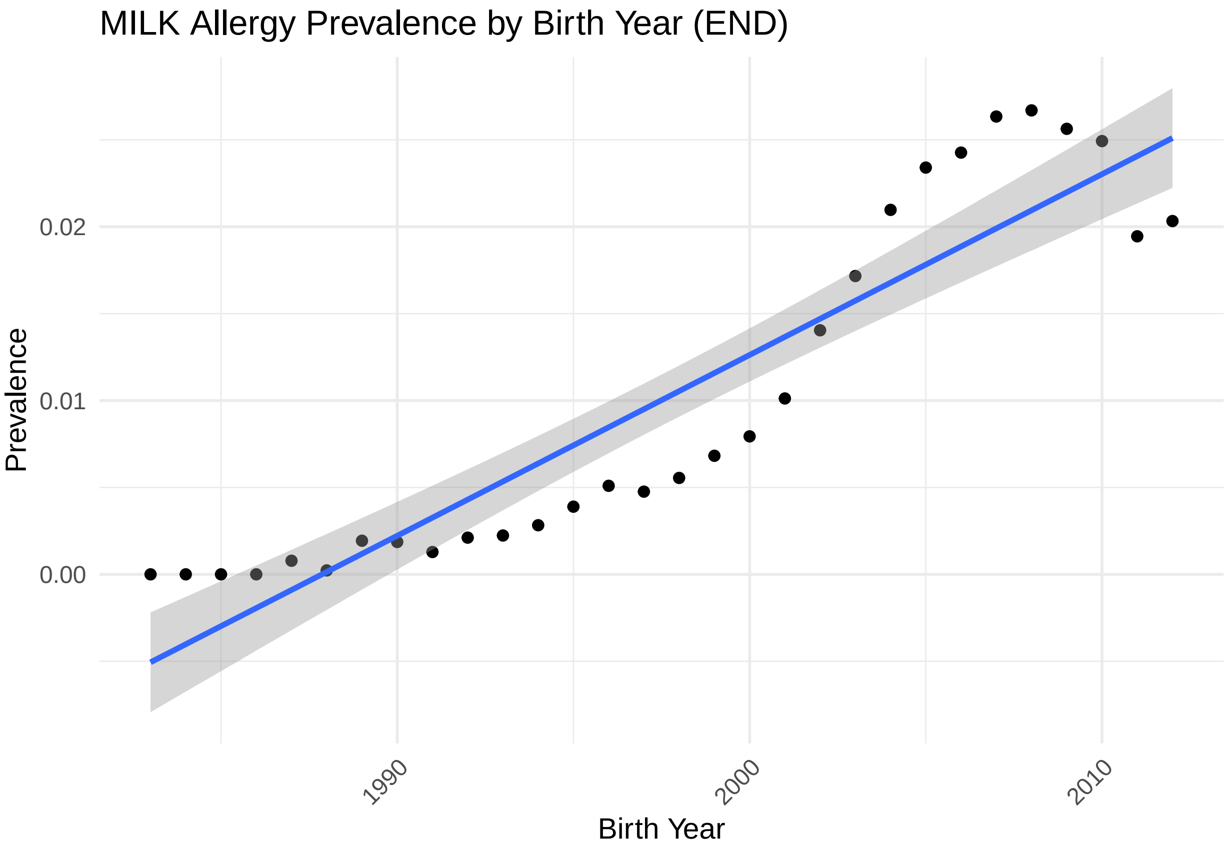
```
## `geom_smooth()` using formula = 'y ~ x'
```



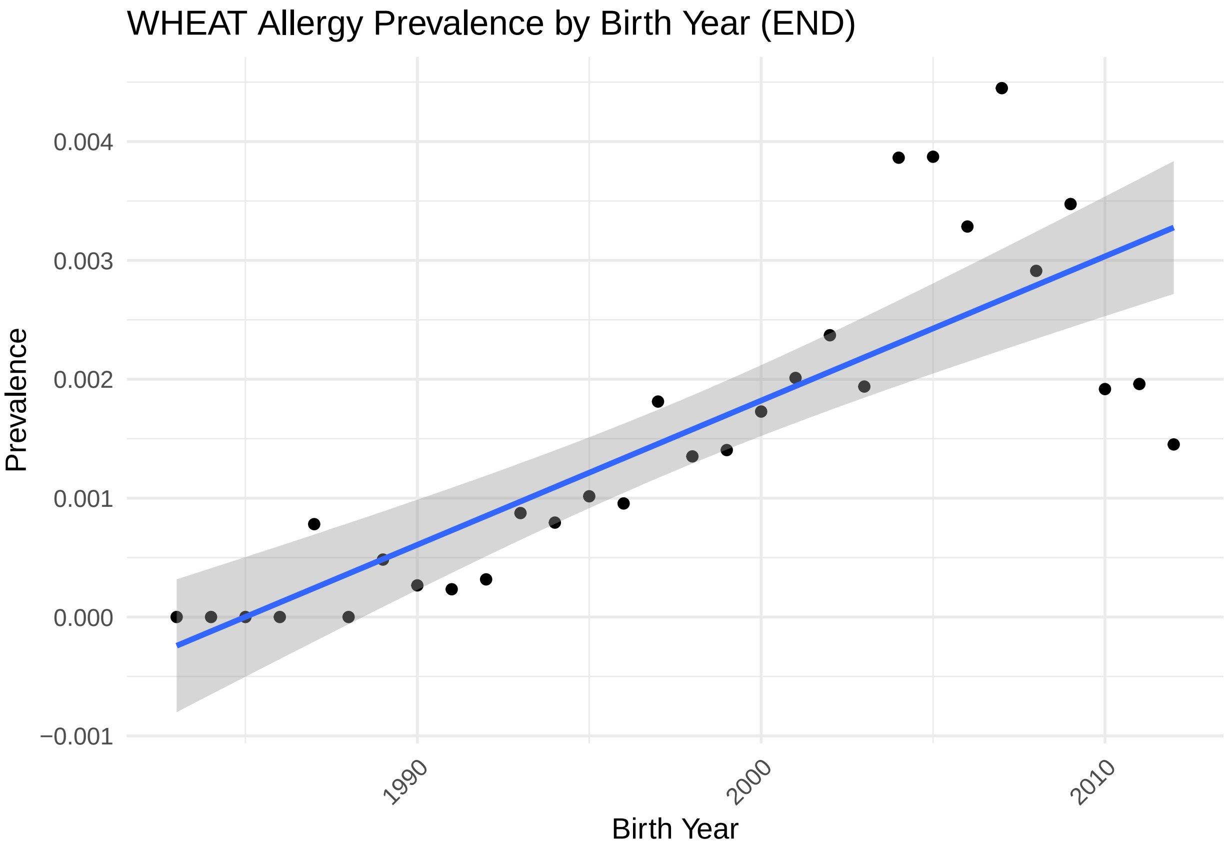
```
## `geom_smooth()` using formula = 'y ~ x'
```



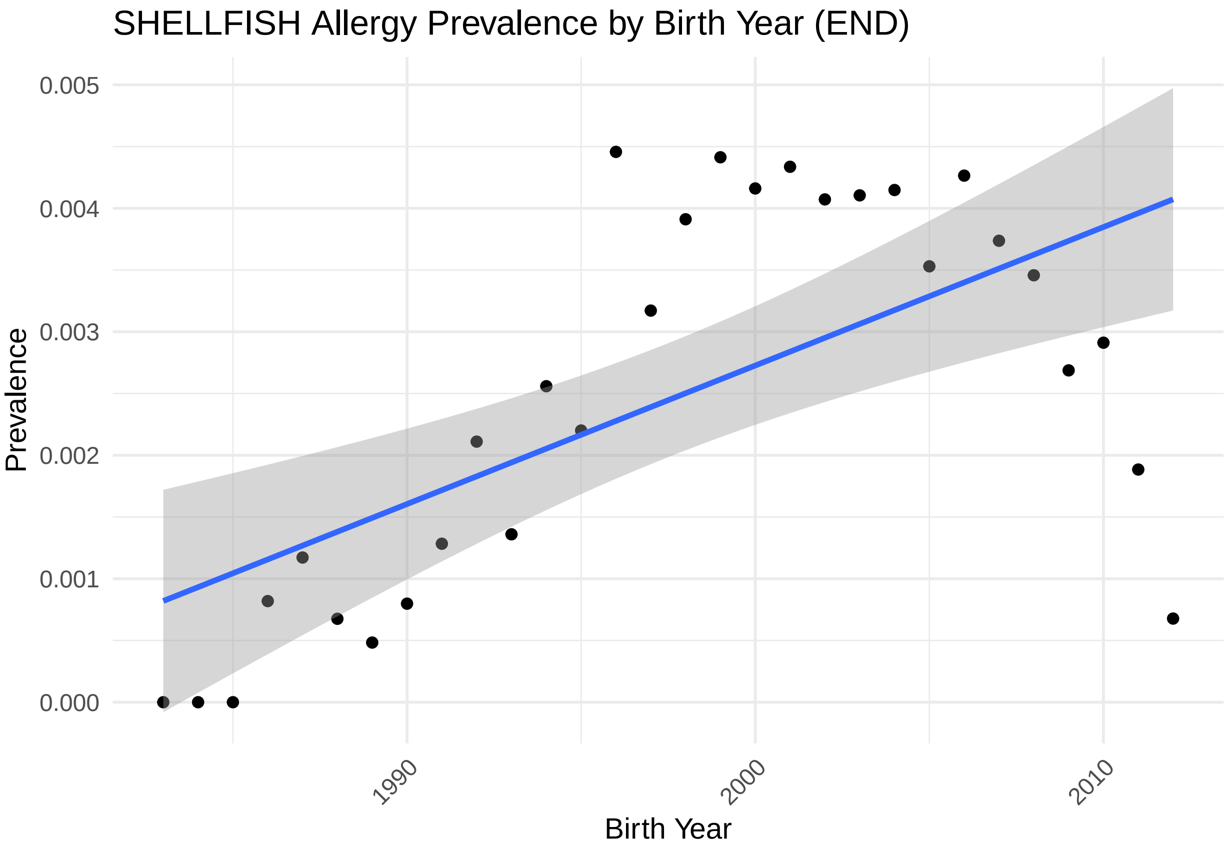
```
## `geom_smooth()` using formula = 'y ~ x'
```



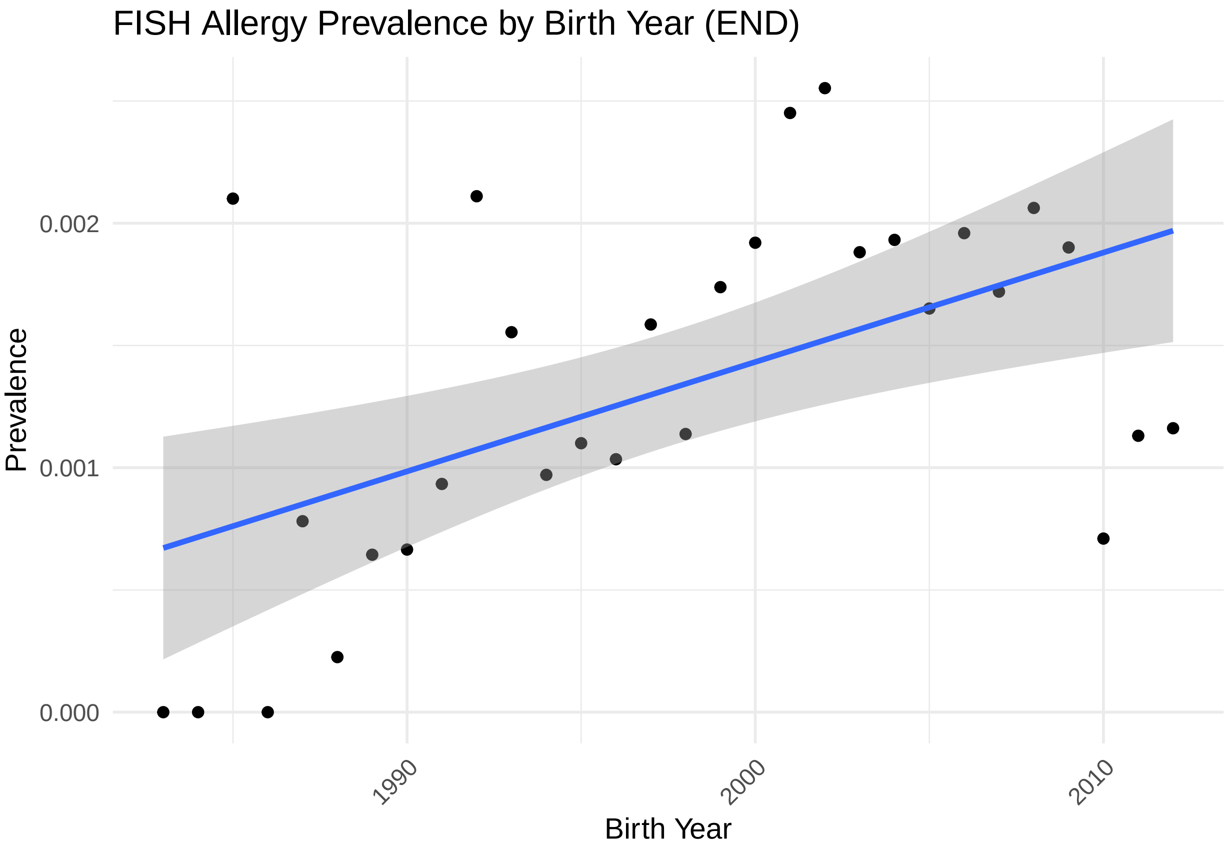
```
## `geom_smooth()` using formula = 'y ~ x'
```



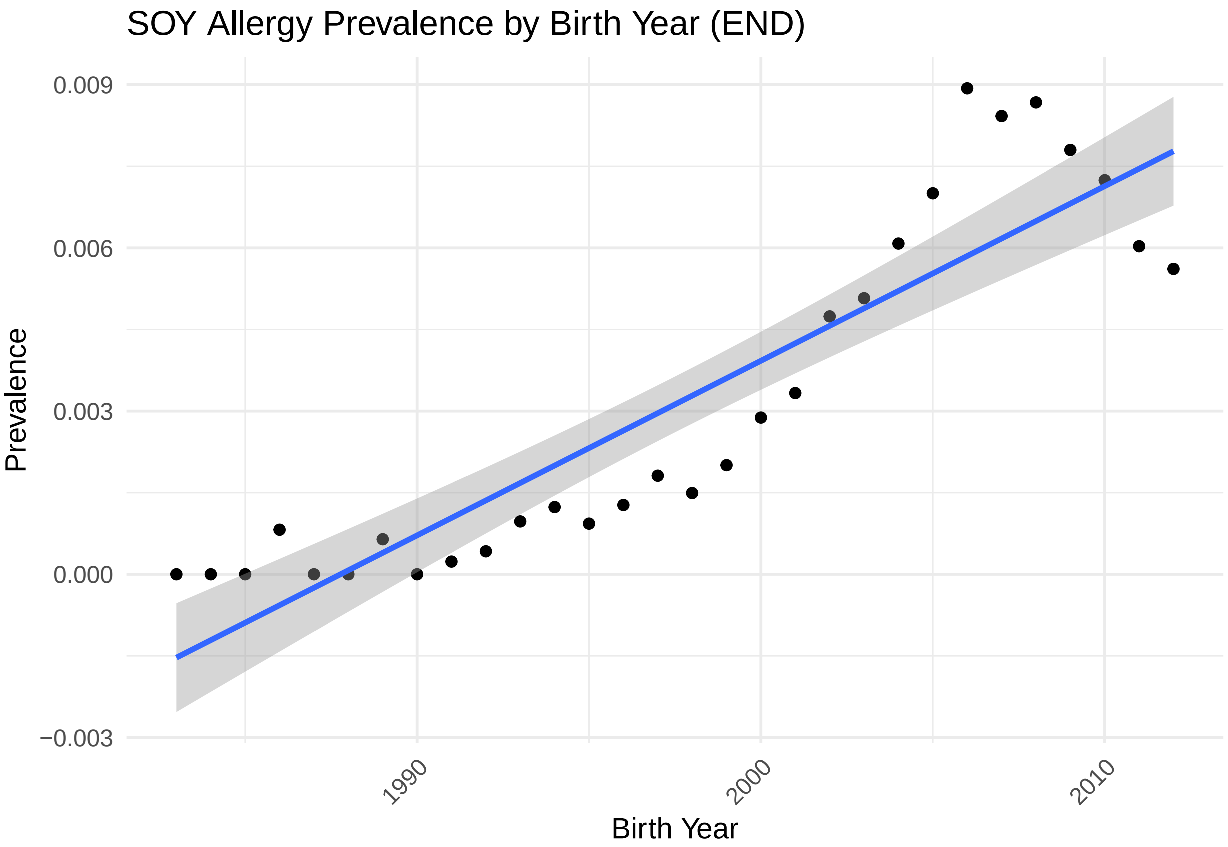
```
## `geom_smooth()` using formula = 'y ~ x'
```



```
## `geom_smooth()` using formula = 'y ~ x'
```

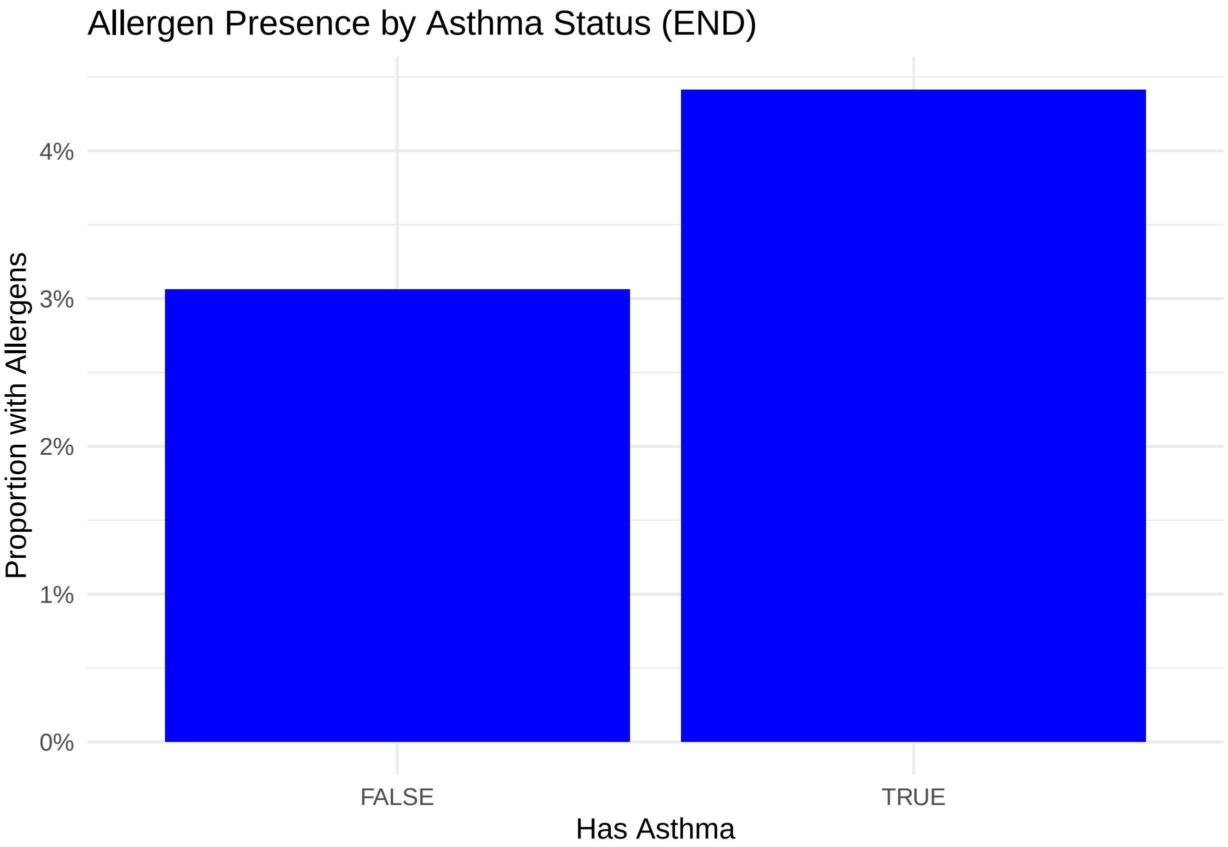


```
## `geom_smooth()` using formula = 'y ~ x'
```

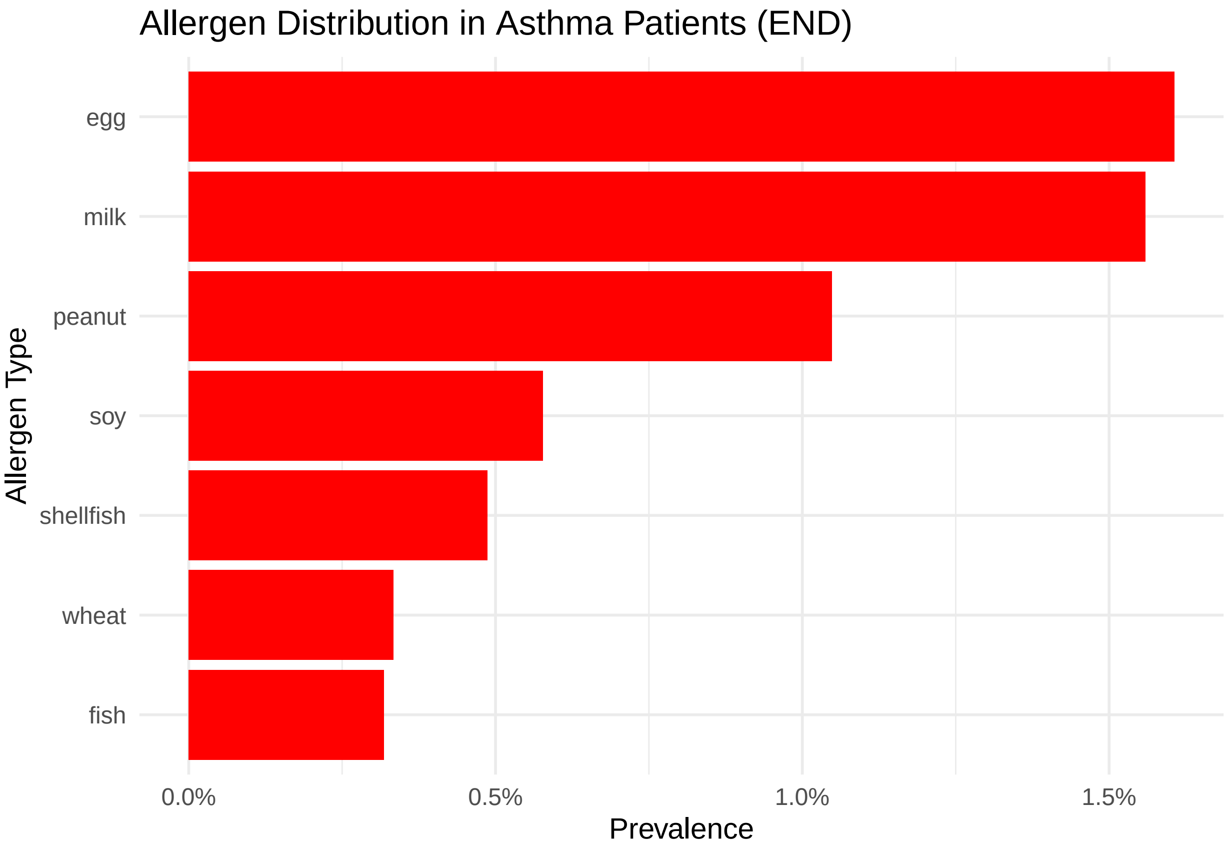


This code analyzes and visualizes the relationship between asthma and allergies at the end of the observation period in two ways: first by showing the overall proportion of allergens in asthma vs non-asthma patients, and then breaking down the specific types of allergies present in asthma patients, with both visualizations using percentage scales for clear interpretation.

```
# Calculate asthma and allergen presence for END
asthma_allergen_analysis_end <- allergy_data %>%
  mutate(
    has_asthma = !is.na(ASTHMA_END),
    has_allergen = rowSums(!is.na(select(., ends_with("_ALG_END")))) > 0
  ) %>%
  group_by(has_asthma) %>%
  summarise(
    allergen_presence = mean(has_allergen),
    count = n(),
    .groups = 'drop'
  )

# Create visualizations for END
ggplot(asthma_allergen_analysis_end,
       aes(x = factor(has_asthma), y = allergen_presence)) +
  geom_bar(stat = "identity", fill = "blue") +
  theme_minimal() +
  labs(title = "Allergen Presence by Asthma Status (END)",
       x = "Has Asthma",
       y = "Proportion with Allergens") +
  scale_y_continuous(labels = scales::percent)
```

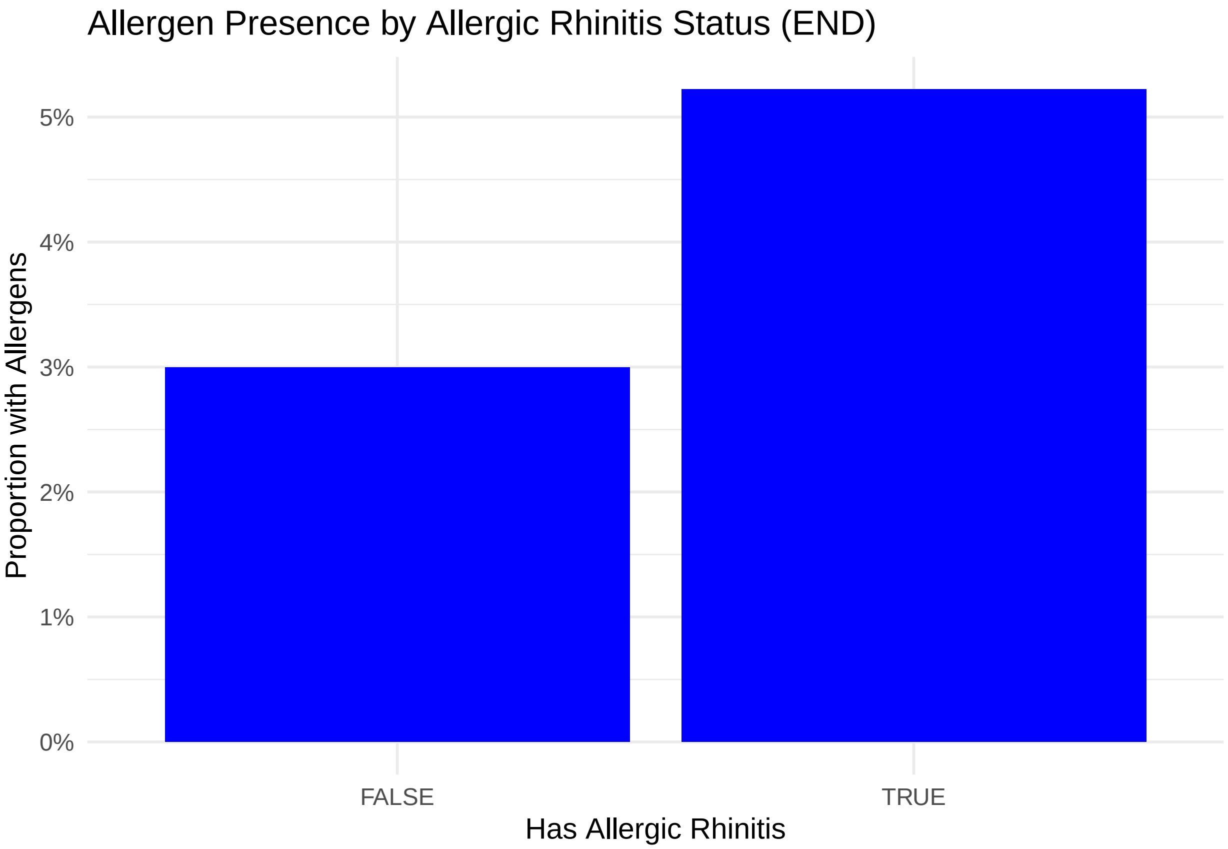


```r
# Detailed allergen breakdown for asthma patients (END)
asthma_specific_allergens_end <- allergy_data %>%
  filter(!is.na(ASTHMA_END)) %>%
  summarise(
    peanut = sum(!is.na(PEANUT_ALG_END))/n(),
    egg = sum(!is.na(EGG_ALG_END))/n(),
    milk = sum(!is.na(MILK_ALG_END))/n(),
    wheat = sum(!is.na(WHEAT_ALG_END))/n(),
    shellfish = sum(!is.na(SHELLFISH_ALG_END))/n(),
    fish = sum(!is.na(FISH_ALG_END))/n(),
    soy = sum(!is.na(SOY_ALG_END))/n()
  ) %>%
  gather(key = "allergen", value = "prevalence")

# Create visualizations for specific allergens (END)
ggplot(asthma_specific_allergens_end,
       aes(x = reorder(allergen, prevalence), y = prevalence)) +
  geom_bar(stat = "identity", fill = "red") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Allergen Distribution in Asthma Patients (END)",
       x = "Allergen Type",
       y = "Prevalence") +
  scale_y_continuous(labels = scales::percent)
```

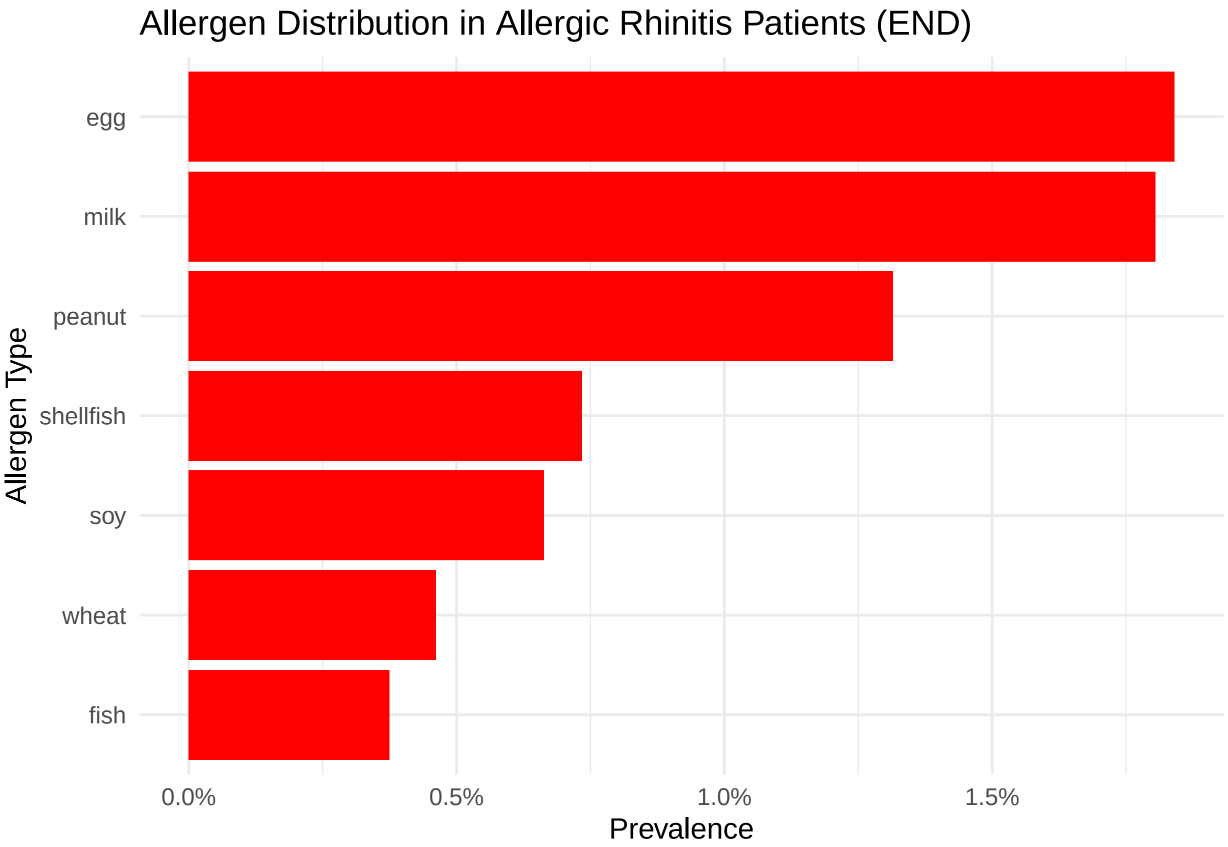


This code analyzes and visualizes the relationship between allergic rhinitis and allergies at the end of the observation period in two ways: first showing the overall proportion of allergens in rhinitis vs non-rhinitis patients, and then breaking down the specific types of allergies present in rhinitis patients, with both visualizations using percentage scales for clear interpretation.

```r
# Calculate rhinitis and allergen presence for END
rhinitis_allergen_analysis_end <- allergy_data %>%
  mutate(
    has_rhinitis = !is.na(ALLERGIC_RHINITIS_END),
    has_allergen = rowSums(!is.na(select(., ends_with("_ALG_END")))) > 0
  ) %>%
  group_by(has_rhinitis) %>%
  summarise(
    allergen_presence = mean(has_allergen),
    count = n(),
    .groups = 'drop'
  )

# Create visualizations for END
ggplot(rhinitis_allergen_analysis_end,
       aes(x = factor(has_rhinitis), y = allergen_presence)) +
  geom_bar(stat = "identity", fill = "blue") +
  theme_minimal() +
  labs(title = "Allergen Presence by Allergic Rhinitis Status (END)",
       x = "Has Allergic Rhinitis",
       y = "Proportion with Allergens") +
  scale_y_continuous(labels = scales::percent)
```



```
# Detailed allergen breakdown for rhinitis patients (END)
rhinitis_allergen_analysis_end <- allergy_data %>%
  filter(!is.na(ALLERGIC_RHINITIS_END)) %>%
  summarise(
    peanut = sum(!is.na(PEANUT_ALG_END))/n(),
    egg = sum(!is.na(EGG_ALG_END))/n(),
    milk = sum(!is.na(MILK_ALG_END))/n(),
    wheat = sum(!is.na(WHEAT_ALG_END))/n(),
    shellfish = sum(!is.na(SHELLFISH_ALG_END))/n(),
    fish = sum(!is.na(FISH_ALG_END))/n(),
    soy = sum(!is.na(SOY_ALG_END))/n()
  ) %>%
  gather(key = "allergen", value = "prevalence")

# Create visualizations for specific allergens (END)
ggplot(rhinitis_allergen_analysis_end,
      aes(x = reorder(allergen, prevalence), y = prevalence)) +
  geom_bar(stat = "identity", fill = "red") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Allergen Distribution in Allergic Rhinitis Patients (END)",
      x = "Allergen Type",
      y = "Prevalence") +
  scale_y_continuous(labels = scales::percent)
```



Conclusion

The analysis of food allergies and their relationships reveals several key findings:

Overall Prevalence Patterns

The prevalence analysis shows peanut allergies are most common (>0.02 prevalence), followed by milk and egg allergies, while Brazil nut and other tree nut allergies show the lowest prevalence (<0.005). This hierarchy of allergen prevalence remains somewhat consistent from start to end of observation.

Demographic Distribution

- Males show consistently higher prevalence of all food allergies compared to females
- Asian/Pacific Islander populations show notably higher rates of egg and milk allergies (around 0.02 prevalence difference) compared to other racial groups
- Racial disparities are particularly evident in egg allergies, where Asian/Pacific Islanders show significantly higher prevalence than the rest of the population

Temporal Trends

The allergy prevalence analysis by birth year shows a clear upward trend from 1980's through the 2010's, with prevalence increasing from nearly 0 to upwards of 0.025, suggesting environmental or diagnostic factors may be influencing allergy development over time.

Clinical Implications

These findings support the original conclusions while adding new insights:

- Food allergy prevalence matches previous estimates.
- Higher rates of peanut, milk, egg, shellfish, and soy allergies than prior studies
- Food allergies has a connection to respiratory allergy risks
- The data supports targeted screening strategies, particularly for: Male patients, Asian/Pacific Islander populations, patients born in more recent years, and those with multiple allergies

This analysis provides valuable insights for healthcare providers in predicting allergy patterns and managing patient care, particularly in identifying high-risk populations.

References

Zenodo

Hill, David A et al. "The epidemiologic characteristics of healthcare provider-diagnosed eczema, asthma, allergic rhinitis, and food allergy in children: a retrospective cohort study." BMC pediatrics vol. 16 133. 20 Aug. 2016, doi:10.1186/s12887-016-0673-z

Link to PMC Article