

Lab Module 3: Bar Charts

PHW251B: Data Visualization for Public Health

Contents

| | |
|--|-----------|
| Introduction | 1 |
| The Data | 3 |
| Data Pre-processing | 4 |
| Part 1: Bar Chart Basic Anatomy | 5 |
| 1.1: Setting up the initial <code>ggplots</code> layer | 5 |
| 1.2: Titles, captions, axes labels, fonts | 6 |
| Part 2: Color | 8 |
| 2.1: Adding color | 8 |
| 2.2: Color Use Cases | 14 |
| 2.3: Legend | 17 |
| 2.4: Gridlines/background | 19 |
| Part 3: Types of bar charts | 27 |
| 3.1: Stacked bar chart | 28 |
| 3.2: Dodged bar charts | 40 |
| 3.3: Faceted bar charts | 45 |
| 3.4: Horizontal bar chart | 49 |
| Summary/Recap: | 51 |

Introduction

Welcome to Week 3!

In this module, we will explore how to utilize R and the `ggplot2` package to create publication-ready bar charts. Bar charts are useful in visualizing summary of data by categories.

After this module, students should be able to:

- Create basic graphs using `ggplot`
- Use advanced `ggplot` aesthetics to customize graphs
- Apply concepts of effective graphical design to produce professional graphs

Helpful Resources:

- [ggplot Cheatsheet link](#)
- [R Brewer Palettes link](#)
- [HEX Color Palette Generator link](#)
- [ggthemes List link](#)

The Data

For this module, we will use data from a case-control study of esophageal cancer in Ille-et-Vilaine, France using a built-in dataset in R.

There are 5 variables:

- **agegp**: Age group
- **alcgp**: Alcohol consumption in grams/day
- **tobgp**: Tobacco consumption in grams/day
- **ncases**: Number of cases
- **ncontrols**: Number of controls

```
# Load data NOTE: How is loading this built-in data different from loading other  
# datasets?  
data(esoph)  
  
# Let's view the first couple of lines of the data NOTE: What do you notice about  
# the types of variables?  
head(esoph)
```

```
##   agegp      alcgp      tobgp ncases ncontrols  
## 1 25-34 0-39g/day 0-9g/day      0         40  
## 2 25-34 0-39g/day 10-19      0         10  
## 3 25-34 0-39g/day 20-29      0          6  
## 4 25-34 0-39g/day 30+        0          5  
## 5 25-34 40-79 0-9g/day      0         27  
## 6 25-34 40-79 10-19      0          7
```

Data Pre-processing

Based on the data, I am curious to see which age group has the most cases of esophageal cancer. So, to find that out, we need to first clean and subset our data.

```
# The 'agg_agegp' data frame will contain summarized information about esophageal  
# cancer cases by age group  
agg_agegp <- esoph %>%  
  # Group the data by the 'agegp' variable  
  group_by(agegp) %>%  
    # Summarize the grouped data by calculating the total number of cases (ncases)  
    # for each age group  
    summarize(totalcases = sum(ncases))  
# Display or return the 'agg_agegp' data frame, which now contains the summarized  
# information  
agg_agegp
```

```
## # A tibble: 6 x 2  
##   agegp totalcases  
##   <ord>      <dbl>  
## 1 25-34         1  
## 2 35-44         9  
## 3 45-54        46  
## 4 55-64        76  
## 5 65-74        55  
## 6 75+         13
```

We have the total number of cases by age group, but it may be helpful to also get a percentage of the total cases between age groups.

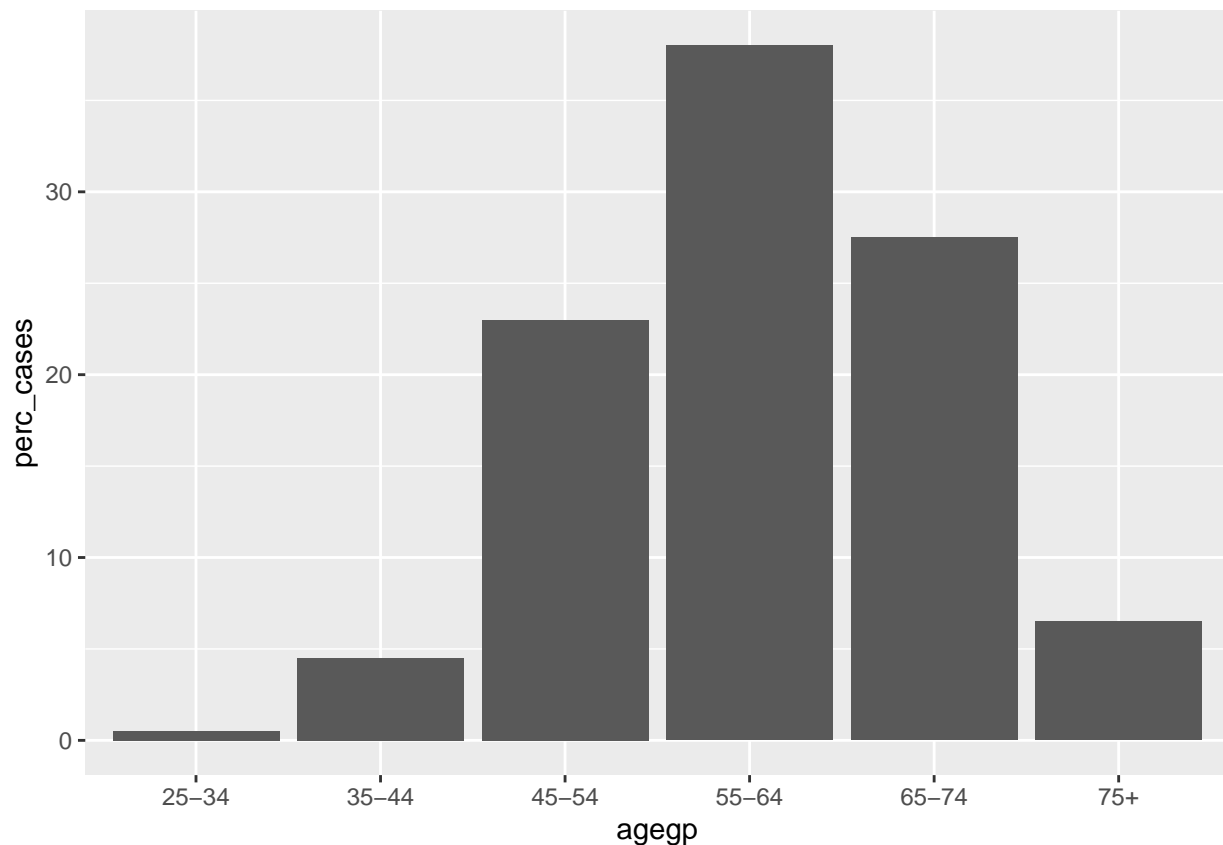
```
# Add a new column 'perc_cases' to the data frame Calculate the percentage of total  
# cases for each age group The calculation is done by dividing 'totalcases' by the  
# sum of all 'totalcases' values, and then multiplying by 100  
agg_agegp <- agg_agegp %>%  
  mutate(perc_cases = 100 * totalcases/sum(totalcases))
```

To get a better understanding of this trend, let's visualize this data using a barchart.

Part 1: Bar Chart Basic Anatomy

1.1: Setting up the initial ggplots layer

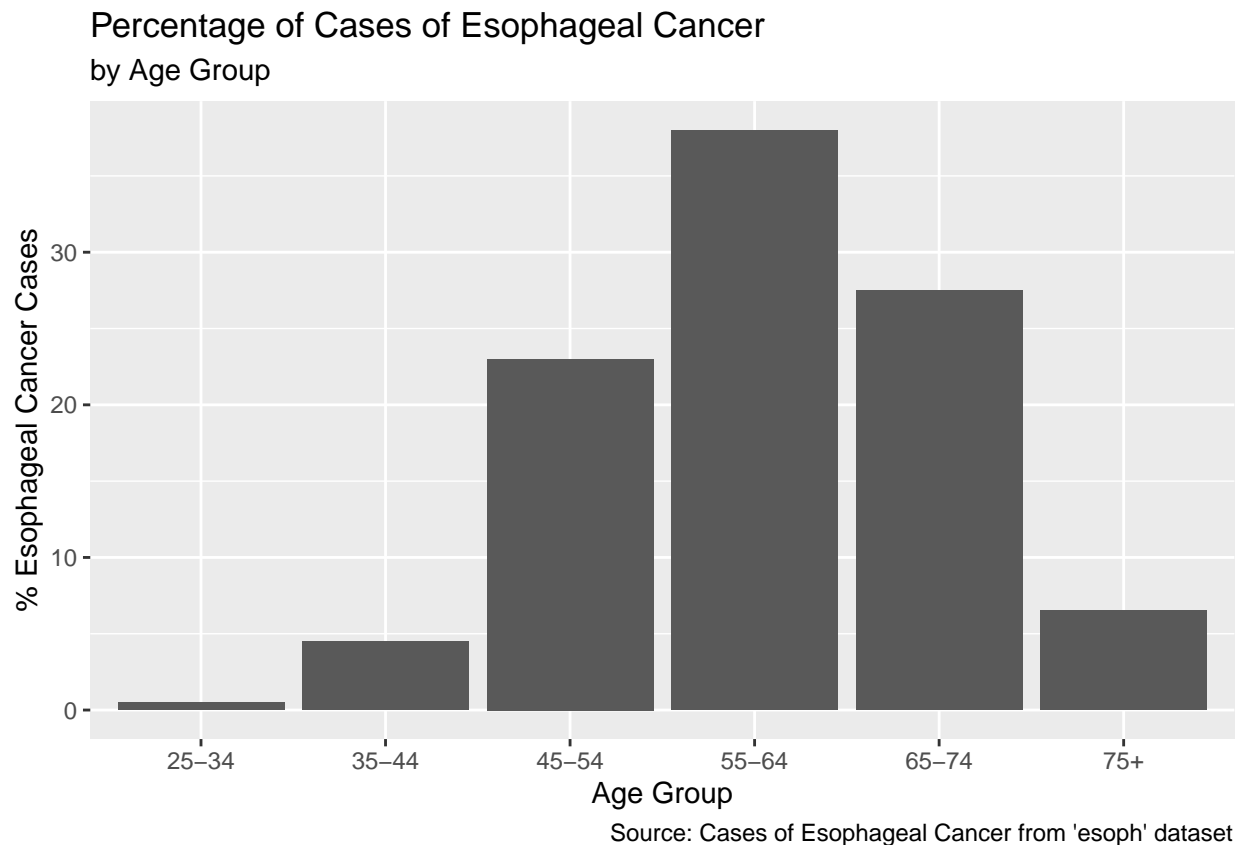
```
# Initialize the 'barchart' object and specify the data frame 'agg_agegp' as the data source  
# Use 'aes()' to define the aesthetics (mapping of variables to visual properties)  
barchart <- ggplot(data = agg_agegp, # Selecting the data that will be fed into the plot  
                  aes(x = agegp, # 'x=agegp' maps the 'agegp' variable to the x-axis  
                    y = perc_cases)) + # 'y=perc_cases' maps the 'perc_cases' variable to the y-axis  
  
  # Add a bar layer to the plot using 'geom_bar()'  
  # 'stat="identity"' means the heights of the bars correspond to the actual data values  
  geom_bar(stat = "identity") # 'geom_bar' specifies that this will be a bar chart  
  
# Display the bar chart  
barchart
```



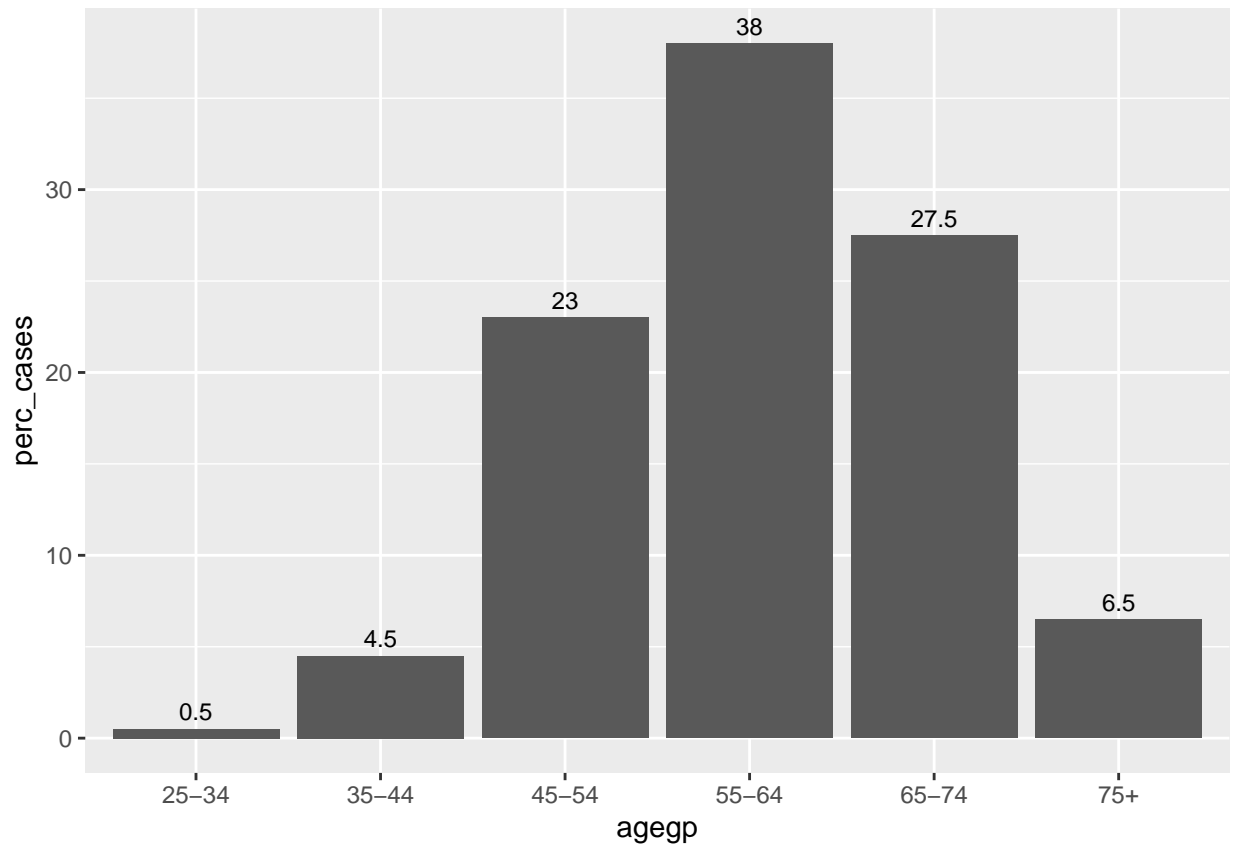
Alright, so we have a basic bar chart! What correlations do you see between the cases of esophageal cancer and age group? We will now move into various aspects of the graph you may want to modify to make it more presentable to an audience.

1.2: Titles, captions, axes labels, fonts

```
# Add custom labels to the 'barchart'
barchart +
  labs(title = "Percentage of Cases of Esophageal Cancer", # Title
        subtitle = "by Age Group", # Subtitle
        x = "Age Group", # X-axis label
        y = "% Esophageal Cancer Cases", # Y-axis label
        caption = "Source: Cases of Esophageal Cancer from 'esoph' dataset") # Caption
```



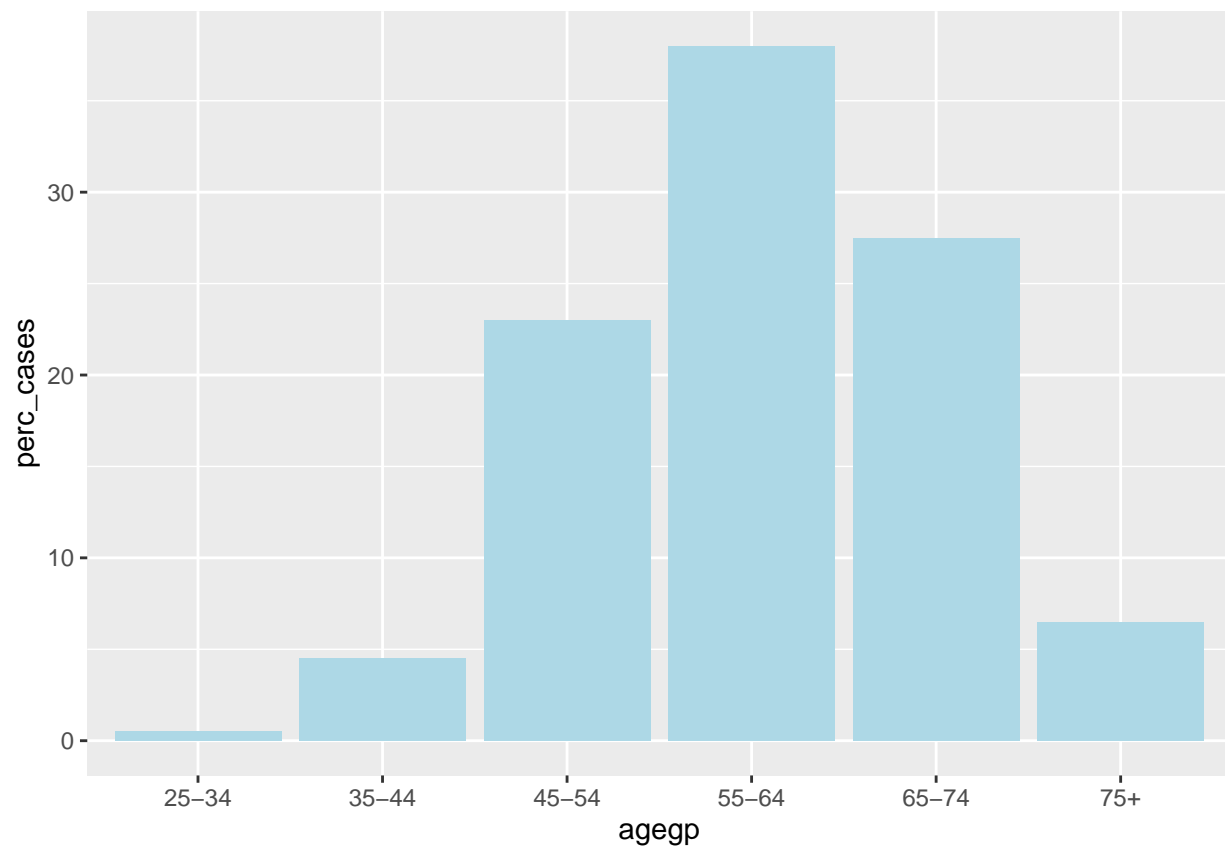
```
# Add text labels above each bar using the 'perc_cases' values as labels
barchart +
  geom_text(aes(label = perc_cases), # Add text labels to each bar
            vjust = -0.5, # Adjust vertical placement (- is up, + is down)
            size = 3, # Adjust text size
            color = "black", # Set color
            family = "sans" # Set font
  )
```



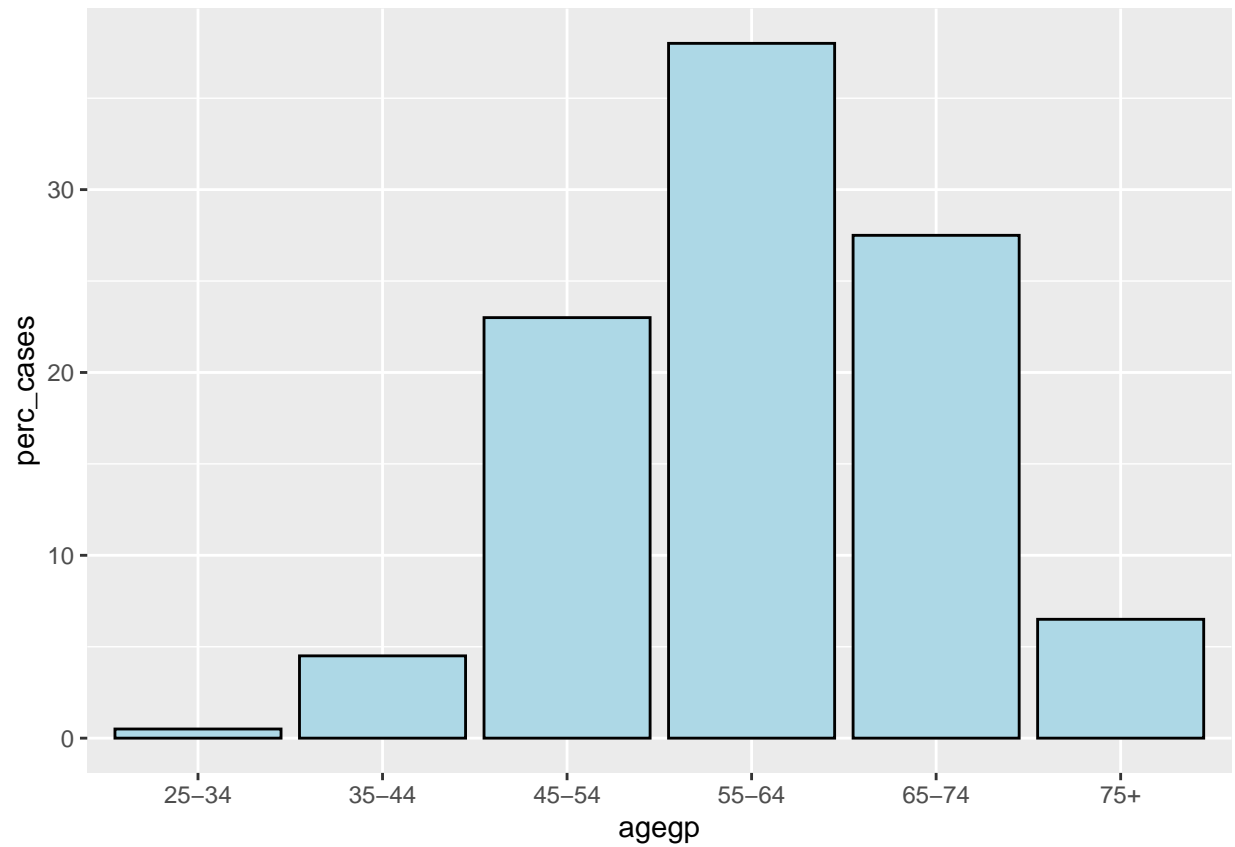
Part 2: Color

2.1: Adding color

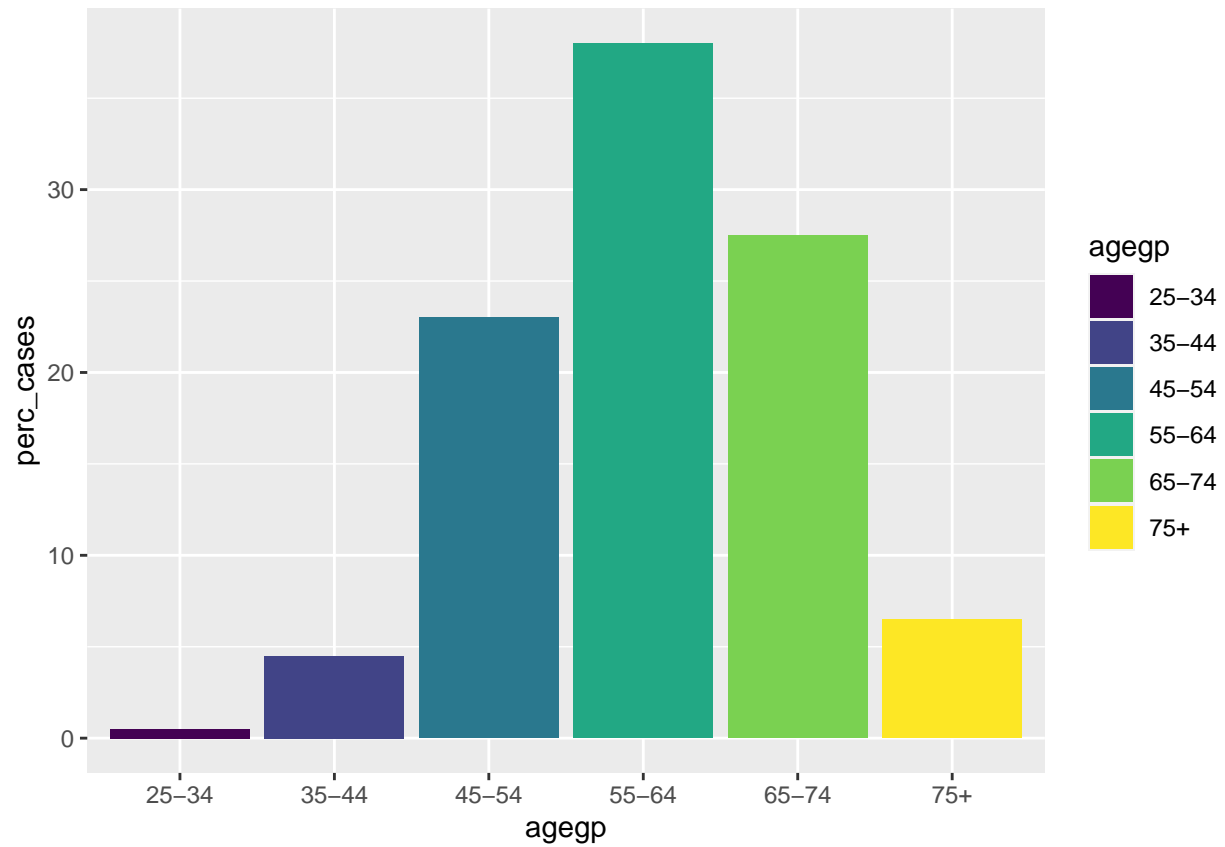
```
# One color for all bars
barchart +
  geom_bar(stat="identity", # Directly plots provided data values as bar heights, without transformation
          fill="lightblue") # Filling in the bars
```



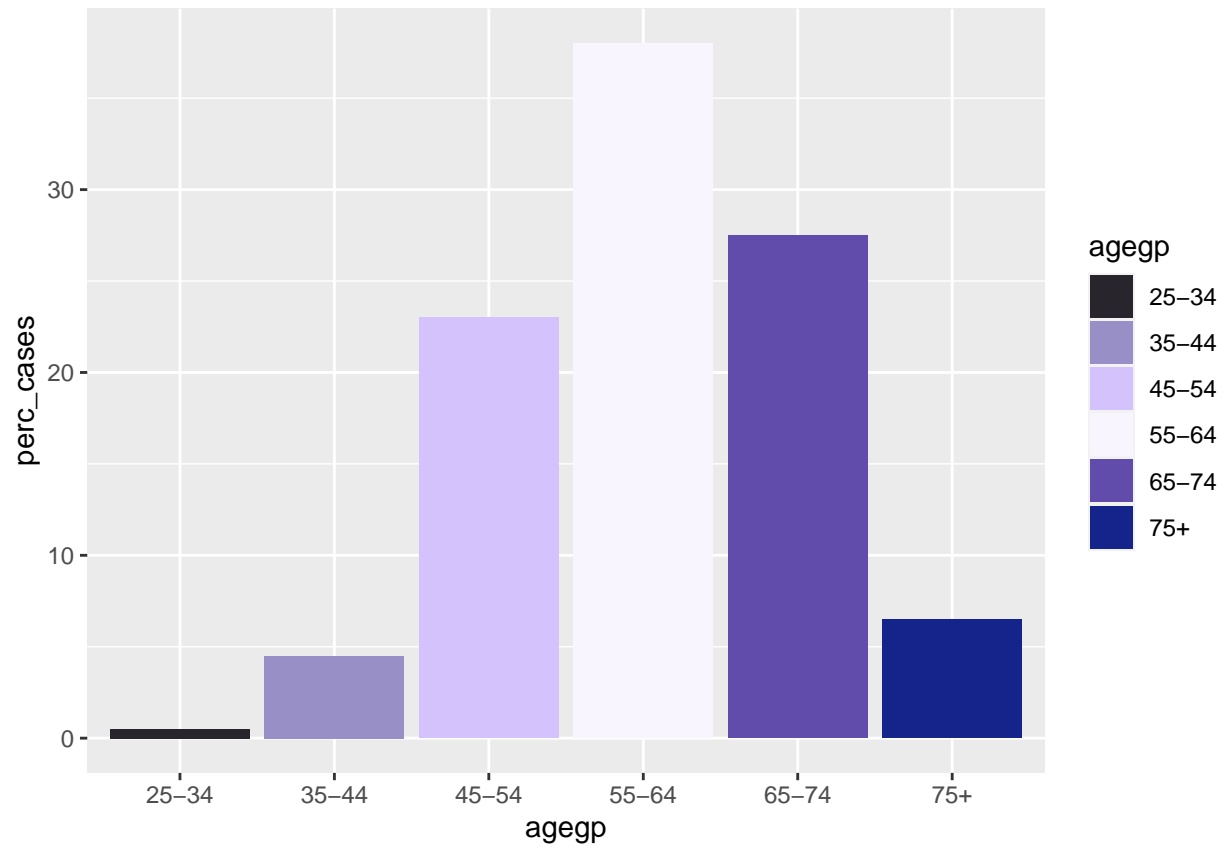
```
# Outline and fill in the bars
barchart +
  geom_bar(stat = "identity",
          fill = "lightblue",
          color = "black") # Outlining the bars
```

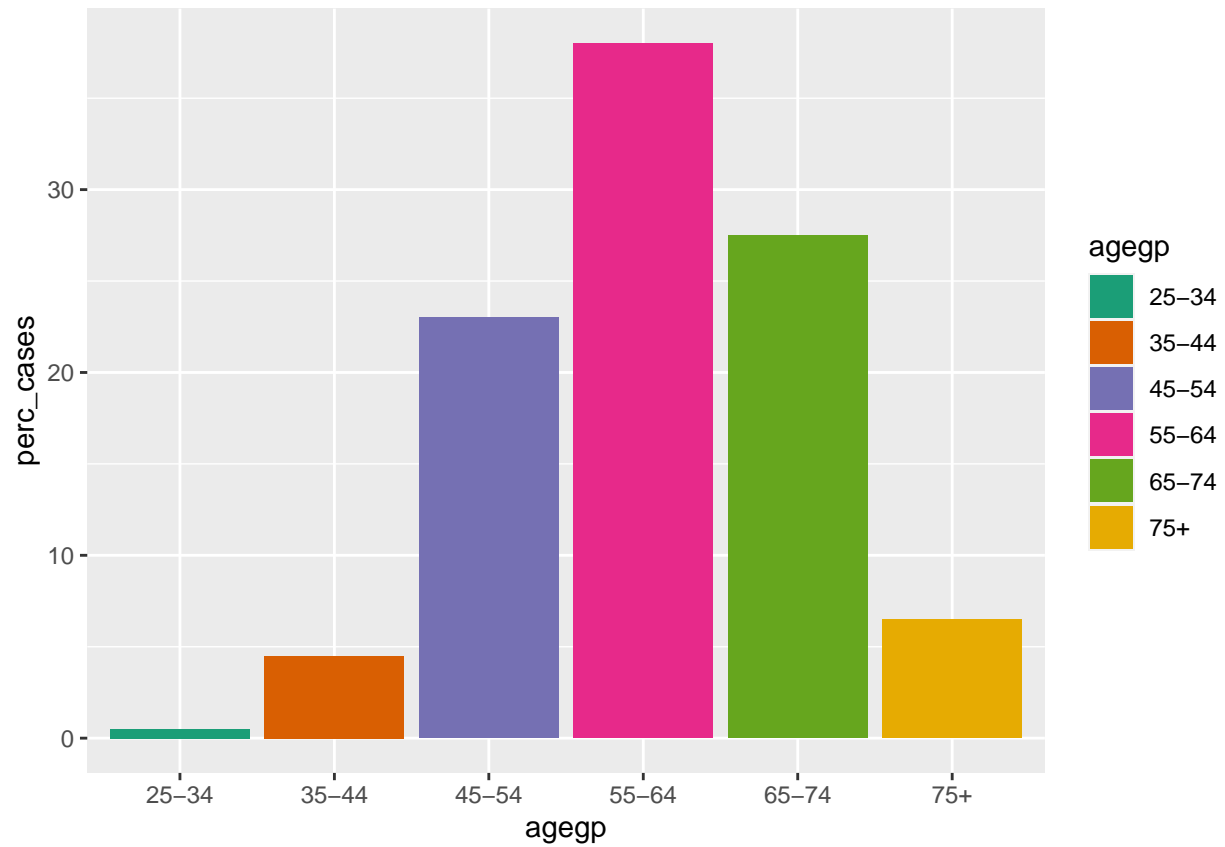
```
# Color by Group (More useful for stacked barcharts which are covered later in this module)  
  
# Use default colors:  
barchart2 <- ggplot(agg_agegp, aes(x=agegp, y=perc_cases, fill = agegp)) +  
  geom_bar(stat="identity")  
barchart2
```



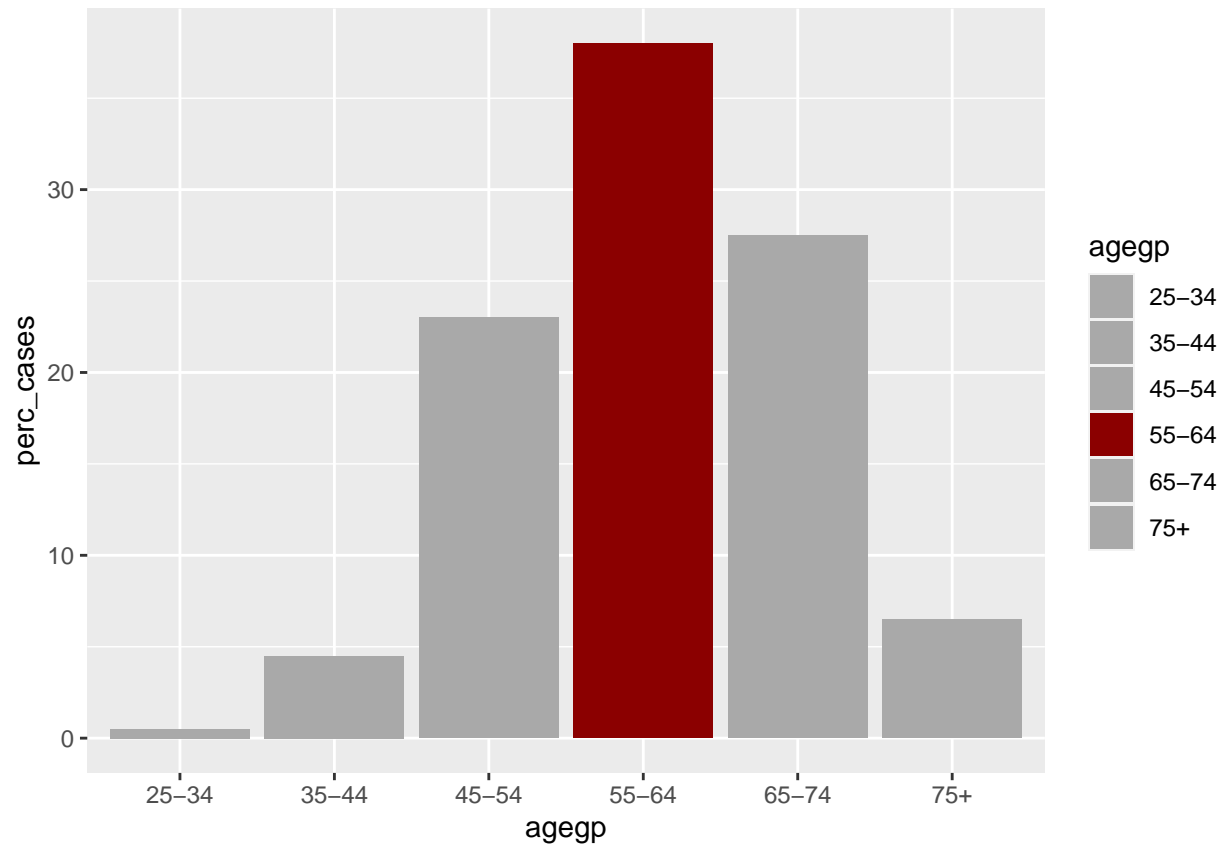
```
# We can also manually select the fill colors:  
# (1) HEX Colors  
barchart2 + scale_fill_manual(values=c("#28262C", "#998FC7", "#D4C2FC",  
                                         "#F9F5FF", "#624CAB", "#14248A"))
```



```
# (2) Color Palettes (i.e. Brewer's)  
barchart2 + scale_fill_brewer(palette="Dark2")
```



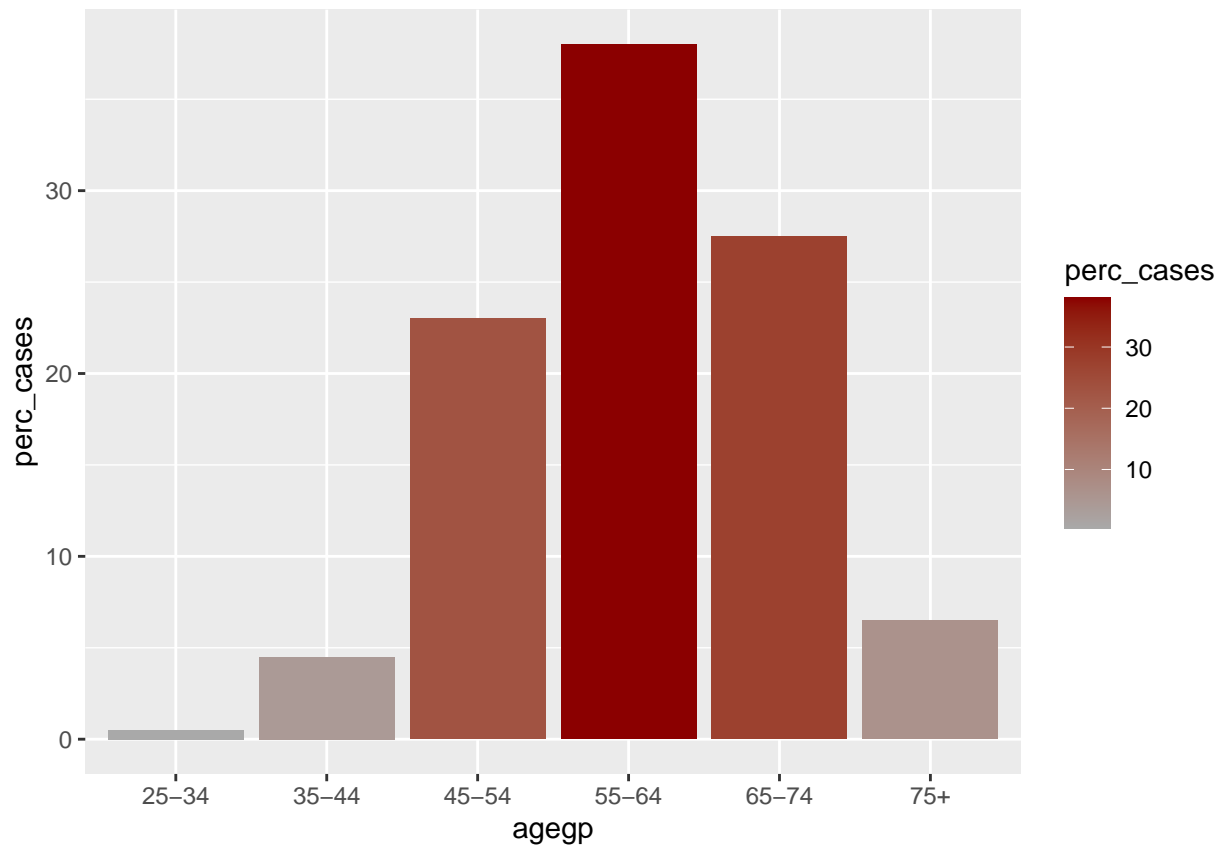
```
# A useful use case is to emphasize one of the groups  
barchart2 + scale_fill_manual(values=c("darkgrey", "darkgrey", "darkgrey",  
                                         "darkred", "darkgrey", "darkgrey"))
```



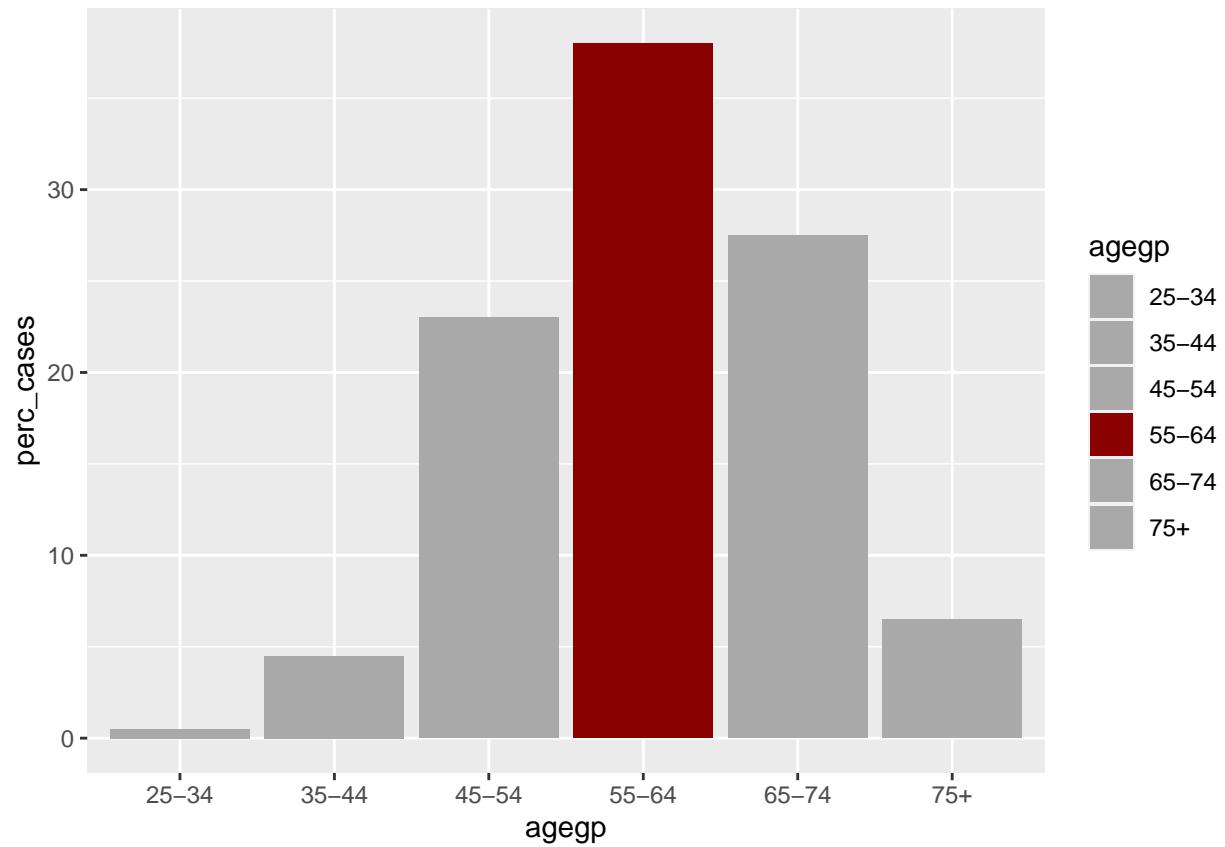
2.2: Color Use Cases

2.2.1: Color to depict quantity (When you want the emphasis to be on a continuous value)

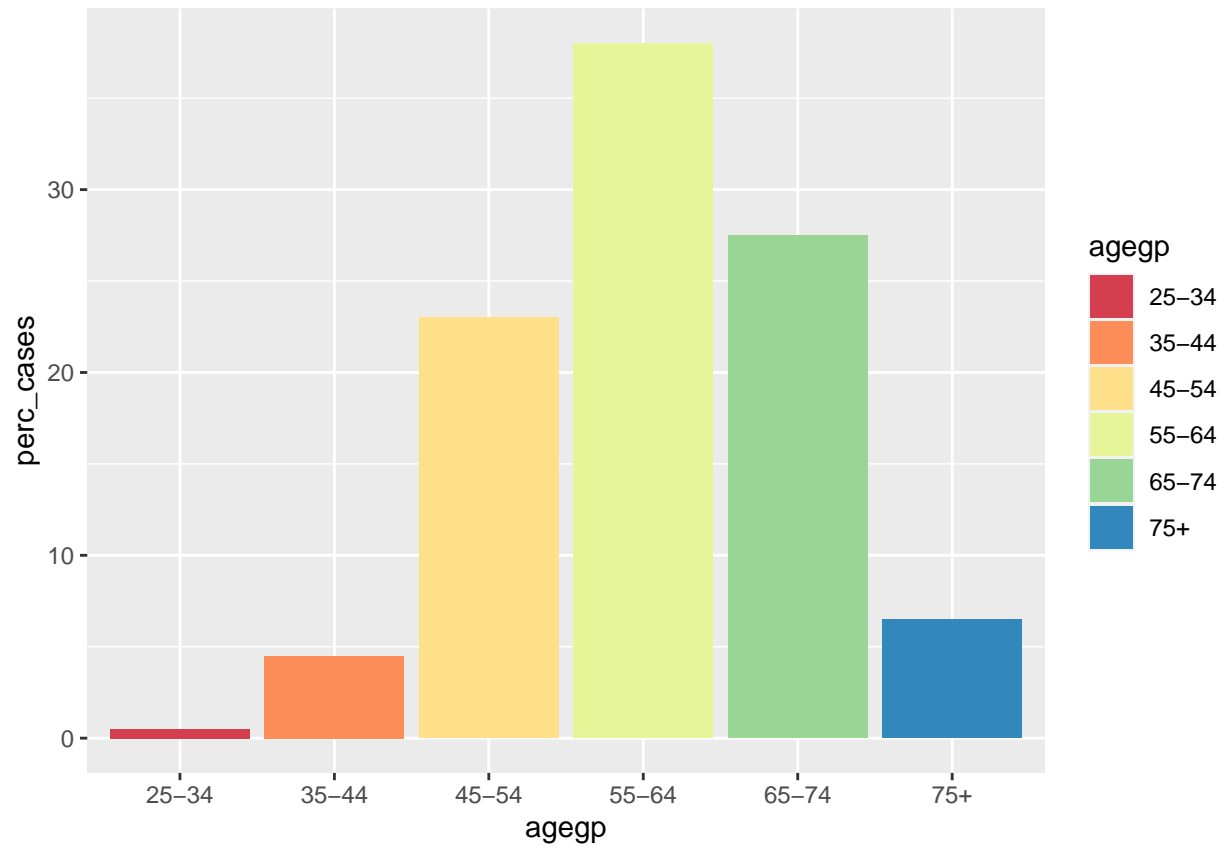
```
barchart_continuous <- ggplot(agg_agegp, aes(x = agegp,  
                                              y = perc_cases)) + # Key is to fill with continuous value  
  geom_bar(stat = "identity", aes(fill = perc_cases)) +  
  scale_fill_gradient(low = "darkgrey", high = "darkred") # Apply color gradient  
barchart_continuous
```



```
# Color to highlight a specific group (When you want to emphasize one of the groups)  
barchart2 + scale_fill_manual(values = c("darkgrey", "darkgrey", "darkgrey", "darkred",  
                                          "darkgrey", "darkgrey"))
```



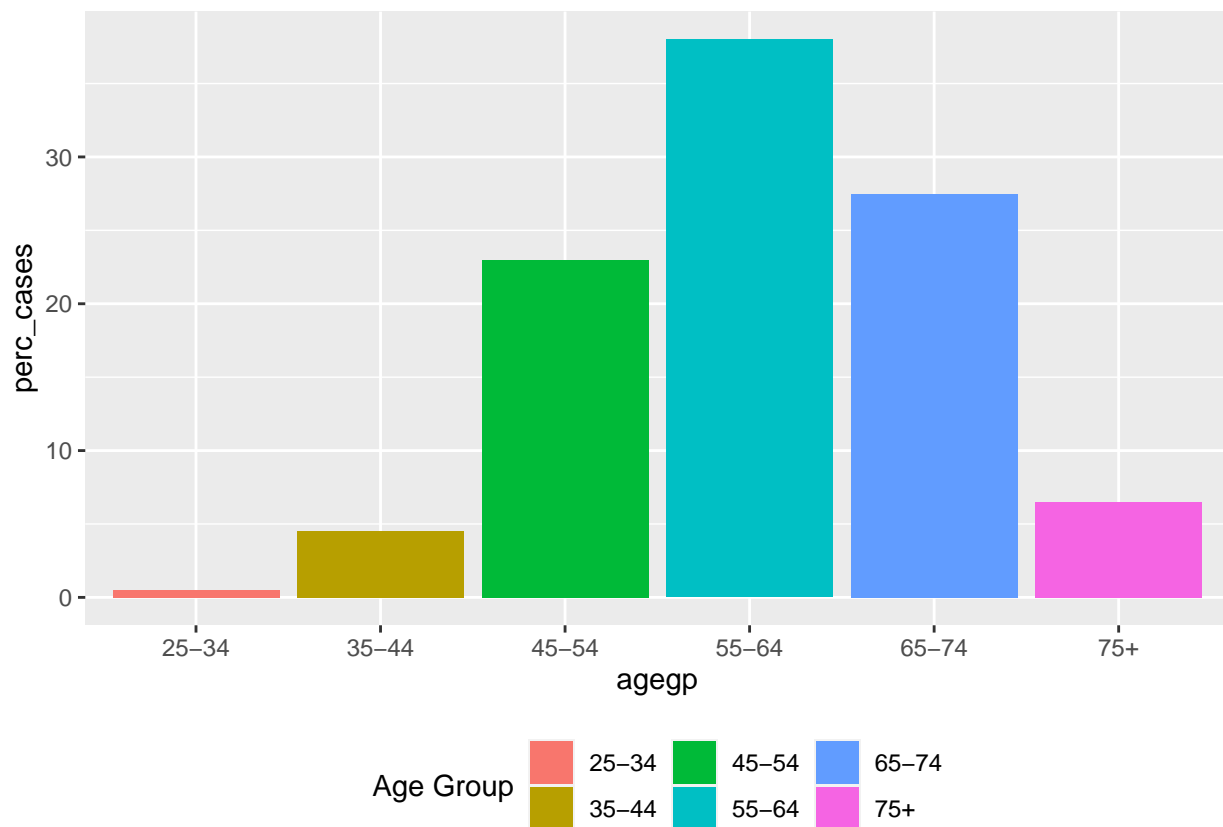
```
# Color to distinguish between groups (Will become more useful in Stacked & Dodged
# bar plots!!)
barchart2 + scale_fill_brewer(palette = "Spectral") # Use 'qualitative' or 'diverging' palettes for ca
```



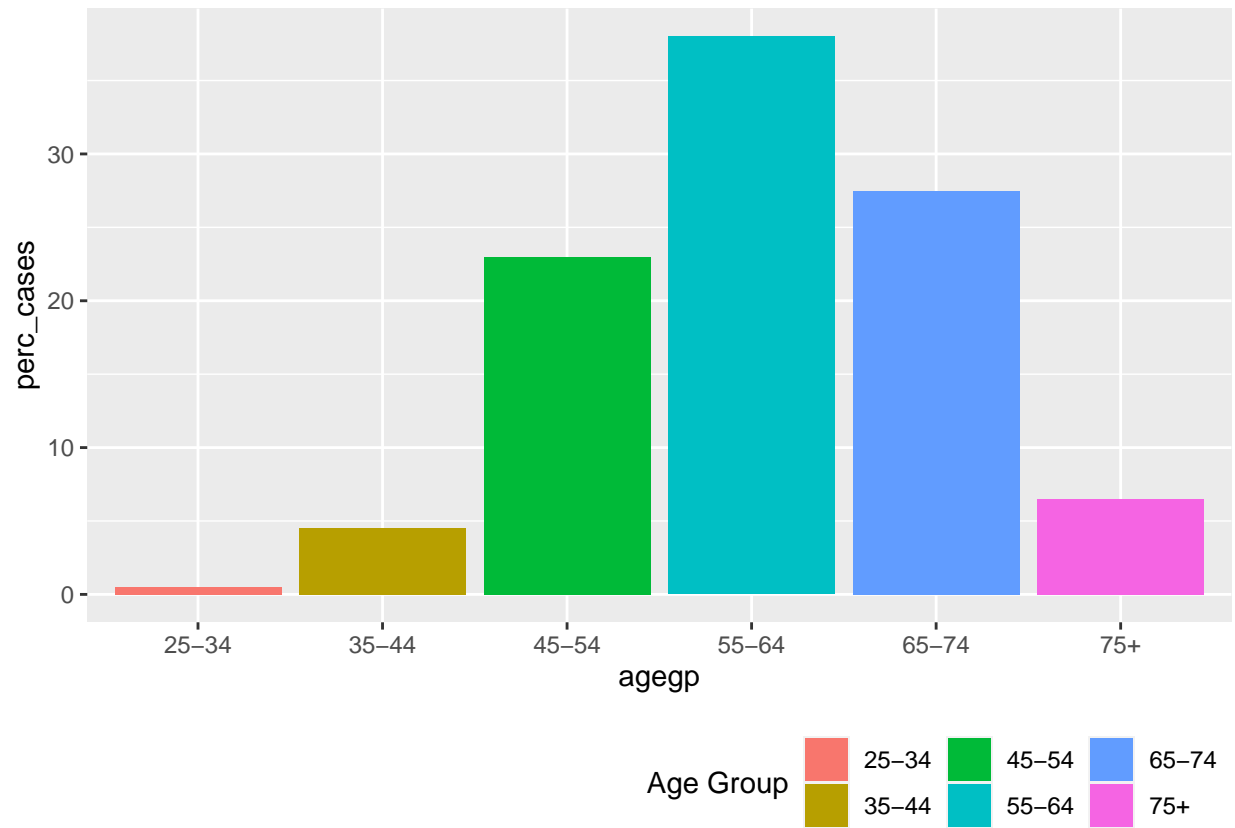
2.3: Legend

To modify the Legend:

```
barchart2 +  
  scale_fill_brewer(palette="Accent") +  
  scale_fill_discrete(name = "Age Group") + # Change name of Legend  
  theme(legend.position="bottom") # Change position of legend (left, right, top, bottom)
```



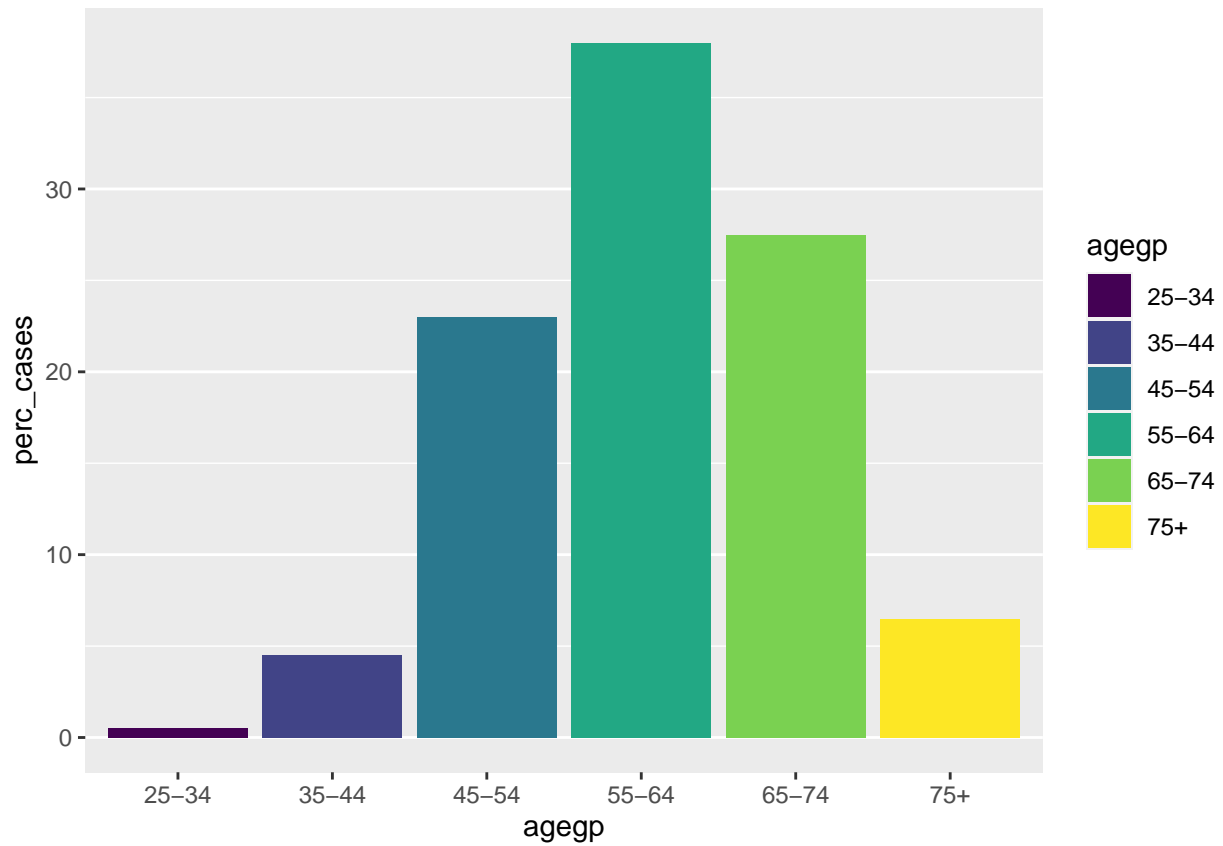
```
barchart2 +  
  scale_fill_brewer(palette = "Accent") +  
  scale_fill_discrete(name = "Age Group") + # Change name of Legend  
  theme(legend.position = "bottom",  
        legend.justification = "right") # Another directioning position for combinations (top & right)
```



2.4: Gridlines/background

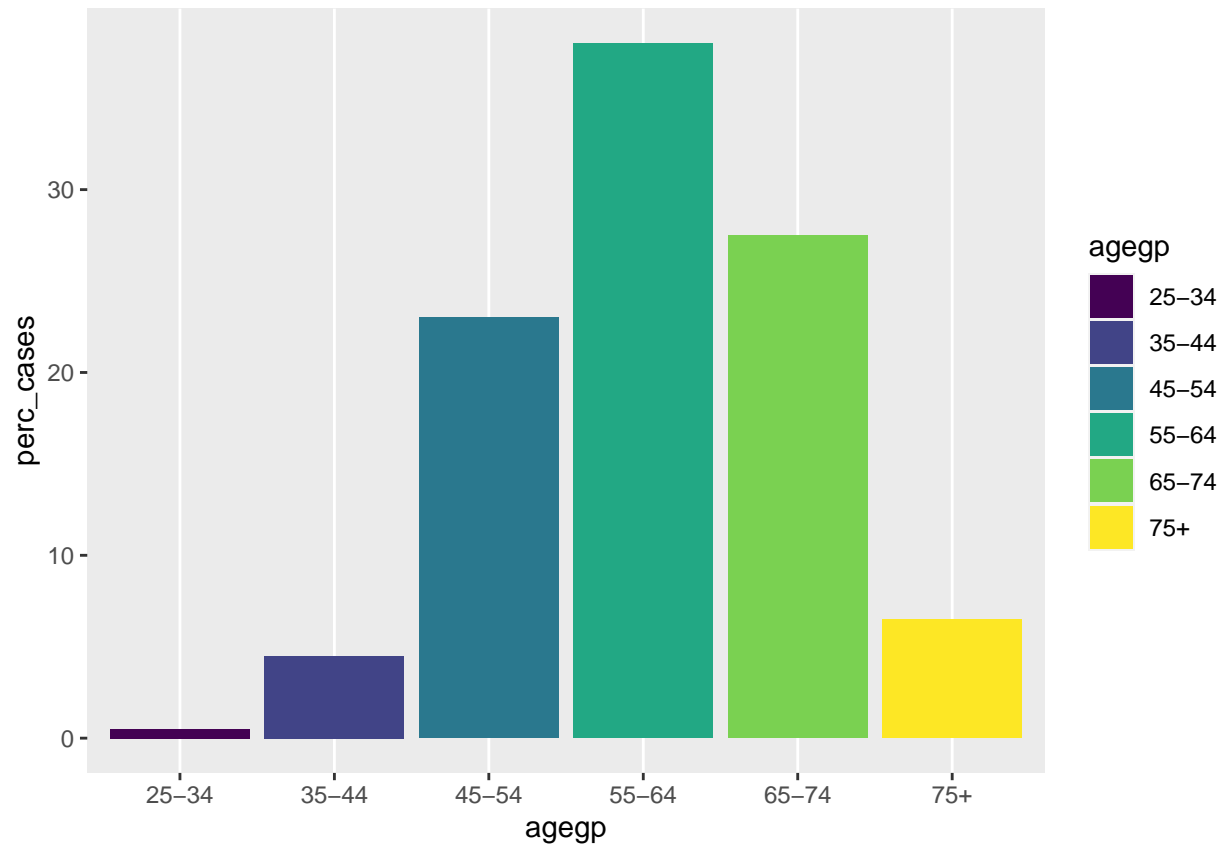
```
# Remove Vertical Lines
```

```
barchart2 + theme(panel.grid.major.x = element_blank(), panel.grid.minor.x = element_blank())
```

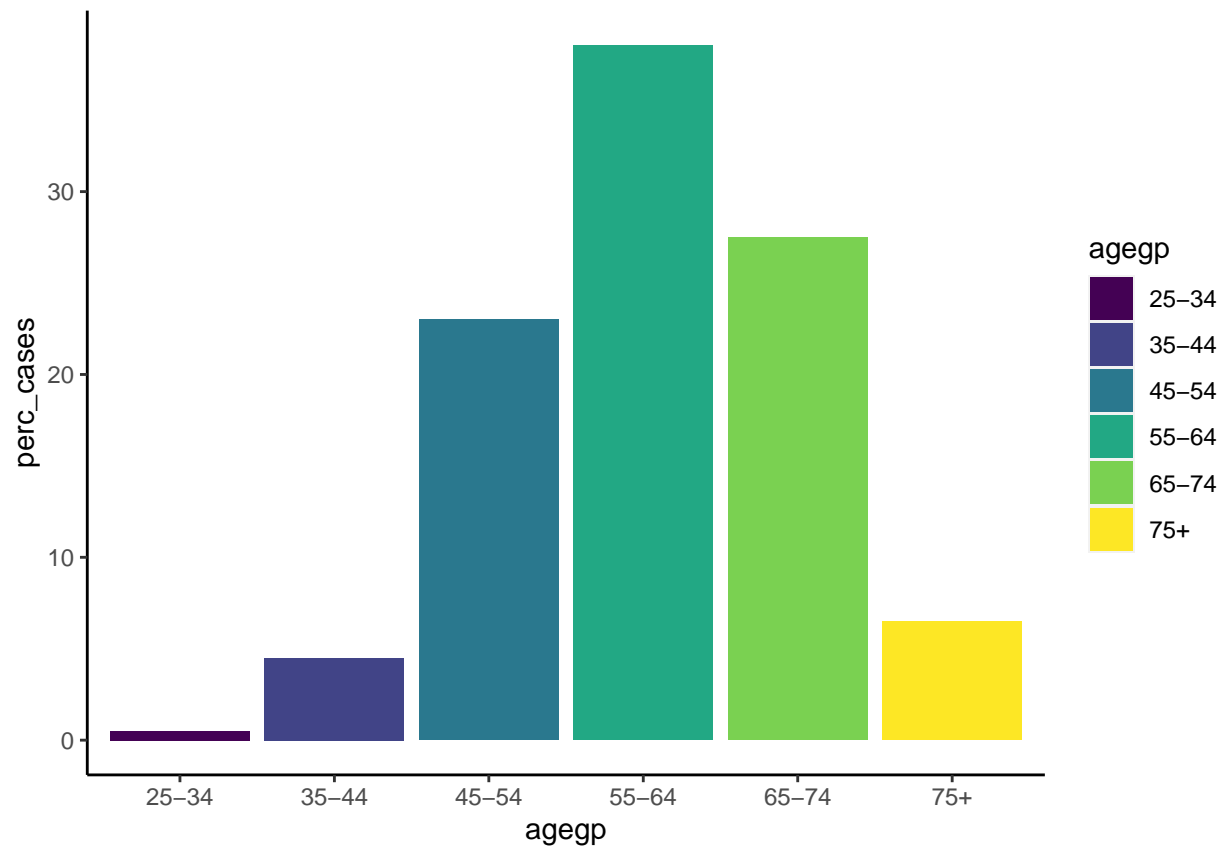


```
# Remove Horizontal Lines
```

```
barchart2 + theme(panel.grid.major.y = element_blank(), panel.grid.minor.y = element_blank())
```

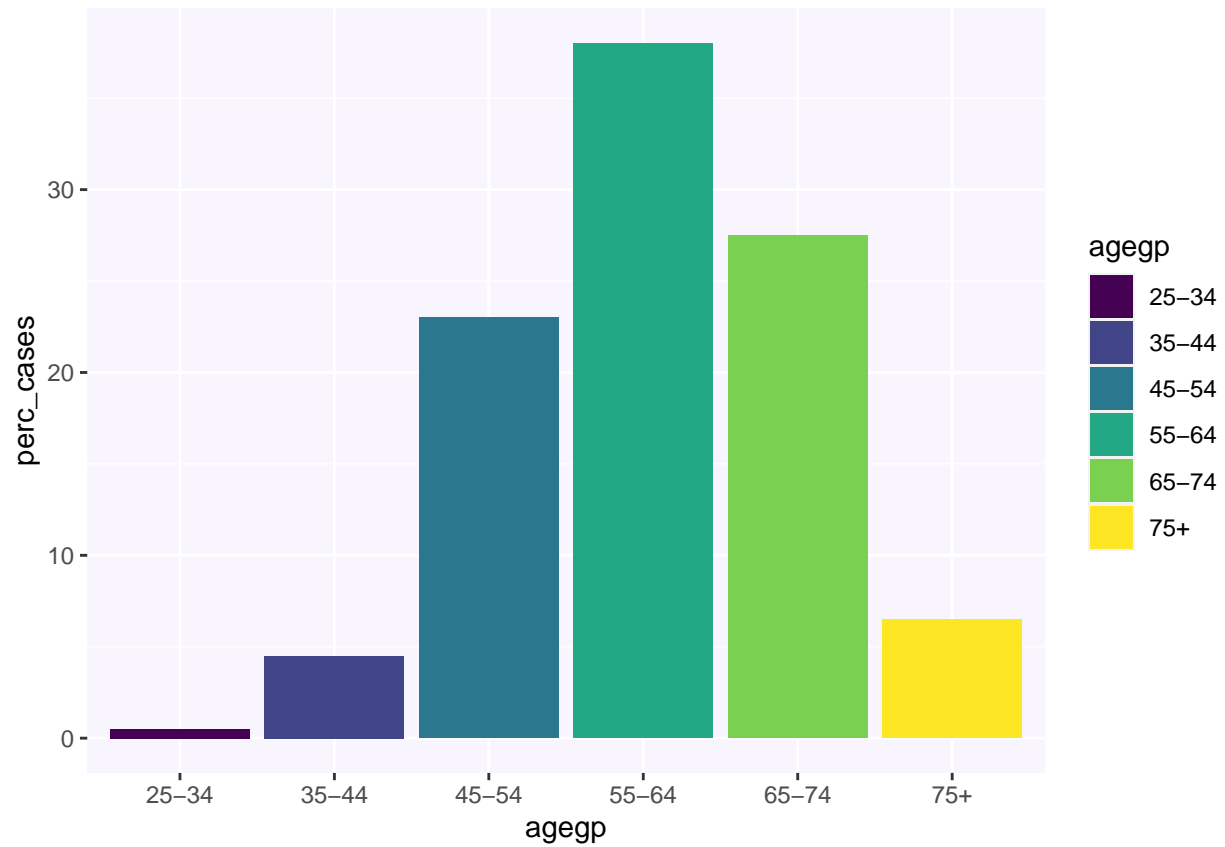


```
# Remove both the lines and background
barchart2 + theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
  panel.background = element_blank(), axis.line = element_line(colour = "black"))
```

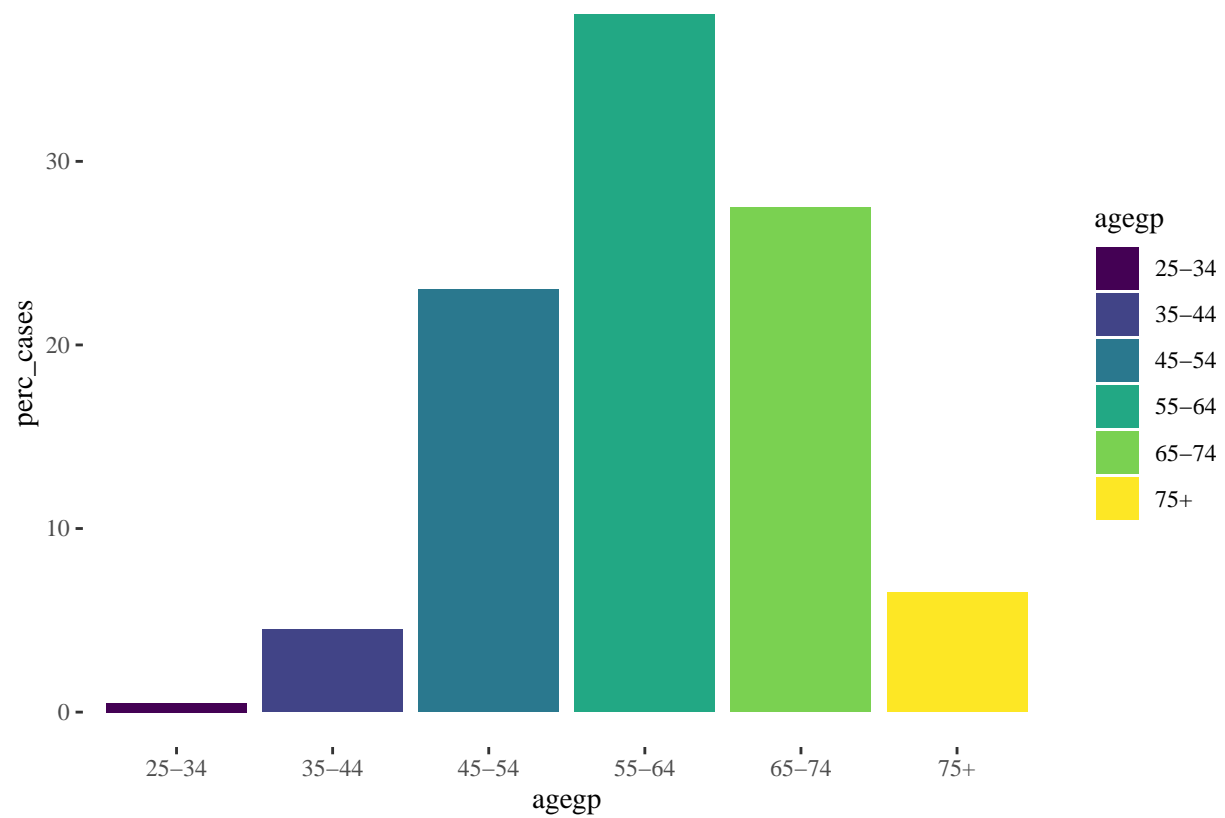


```
# Add Custom background color
barchart2 + theme(panel.background = element_rect(fill = "#F9F5FF", size = 2, linetype = "solid"))
```

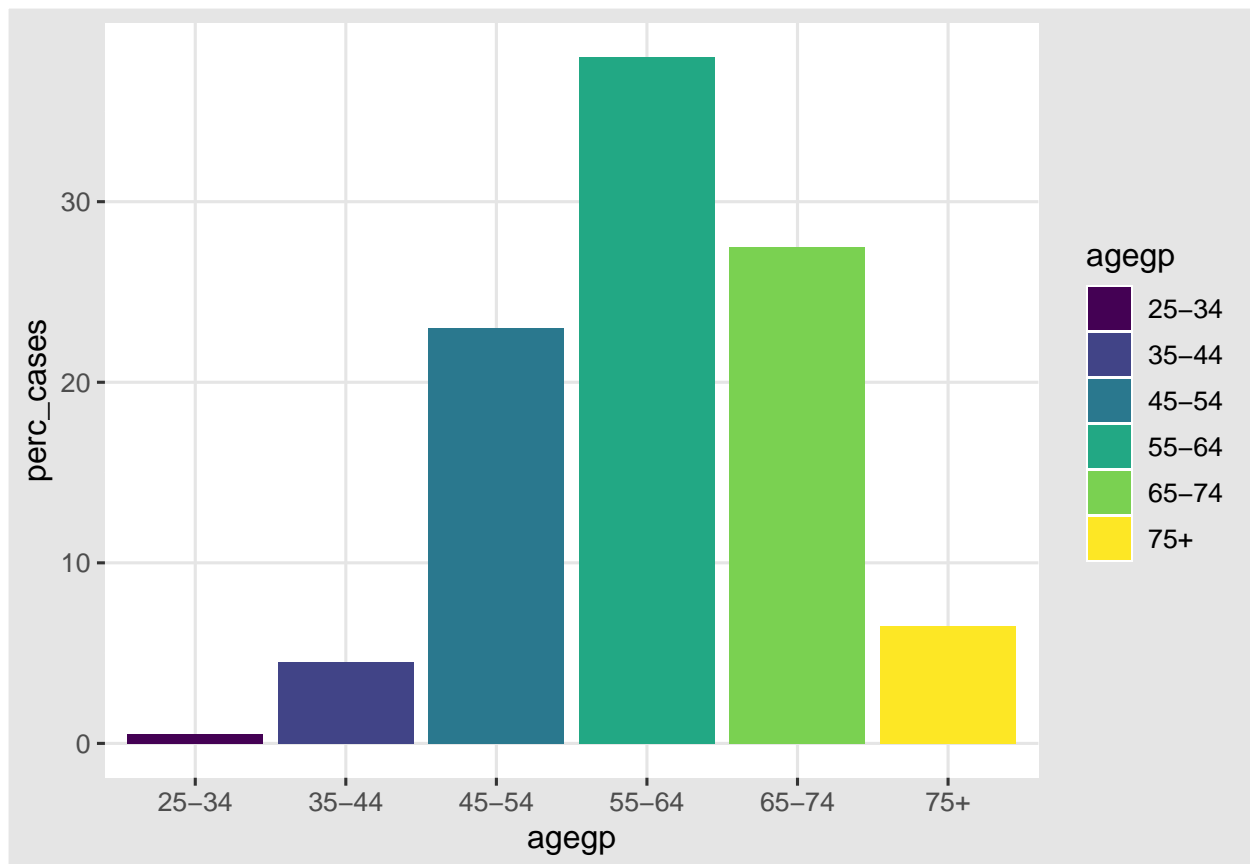
```
## Warning: The 'size' argument of 'element_rect()' is deprecated as of ggplot2 3.4.0.
## i Please use the 'linewidth' argument instead.
```



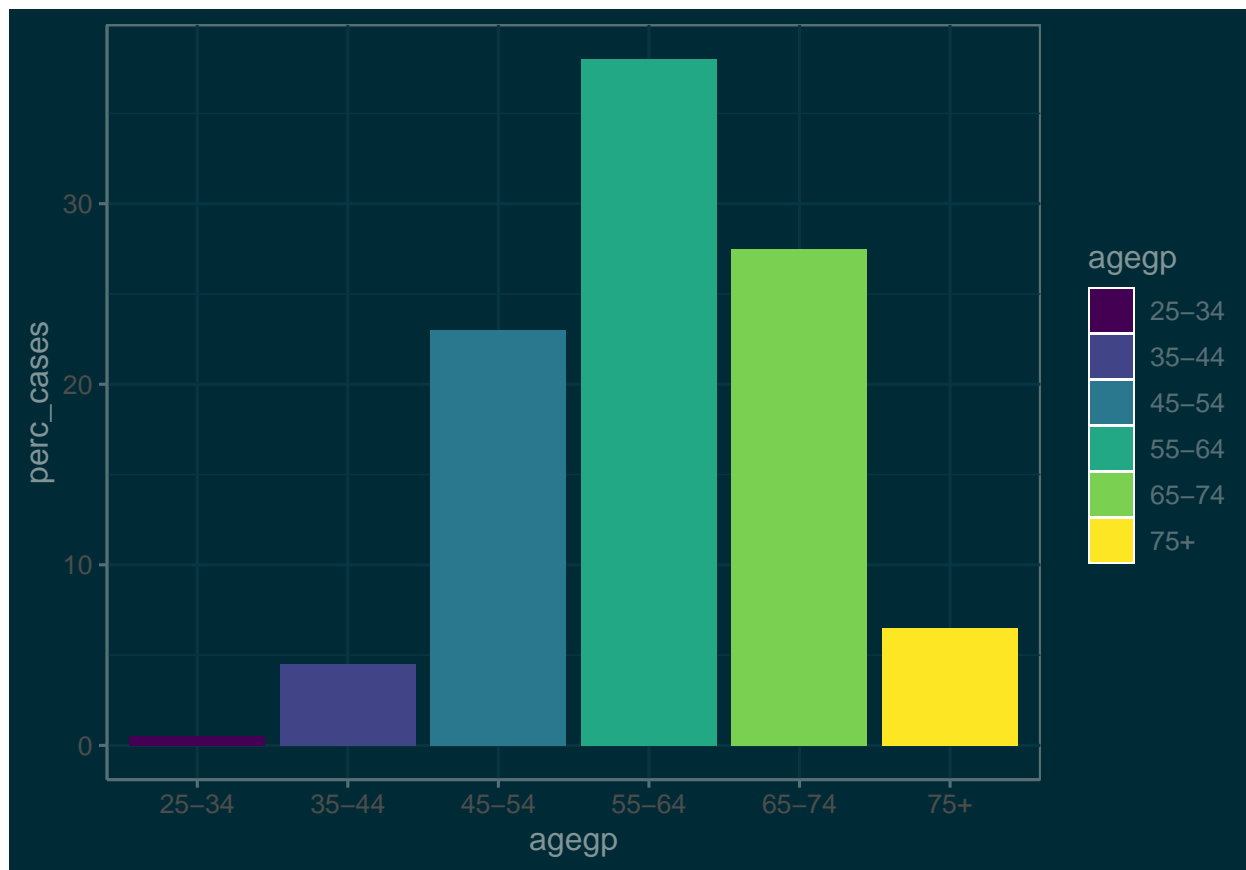
```
# Use ggthemes to add/remove grid lines, change color of background, and change plot  
# themes  
  
# Example 1: Minimalist Theme  
barchart2 + theme_tufte()
```



```
# Example 2: Inverse Gray Theme  
barchart2 + theme_igray()
```

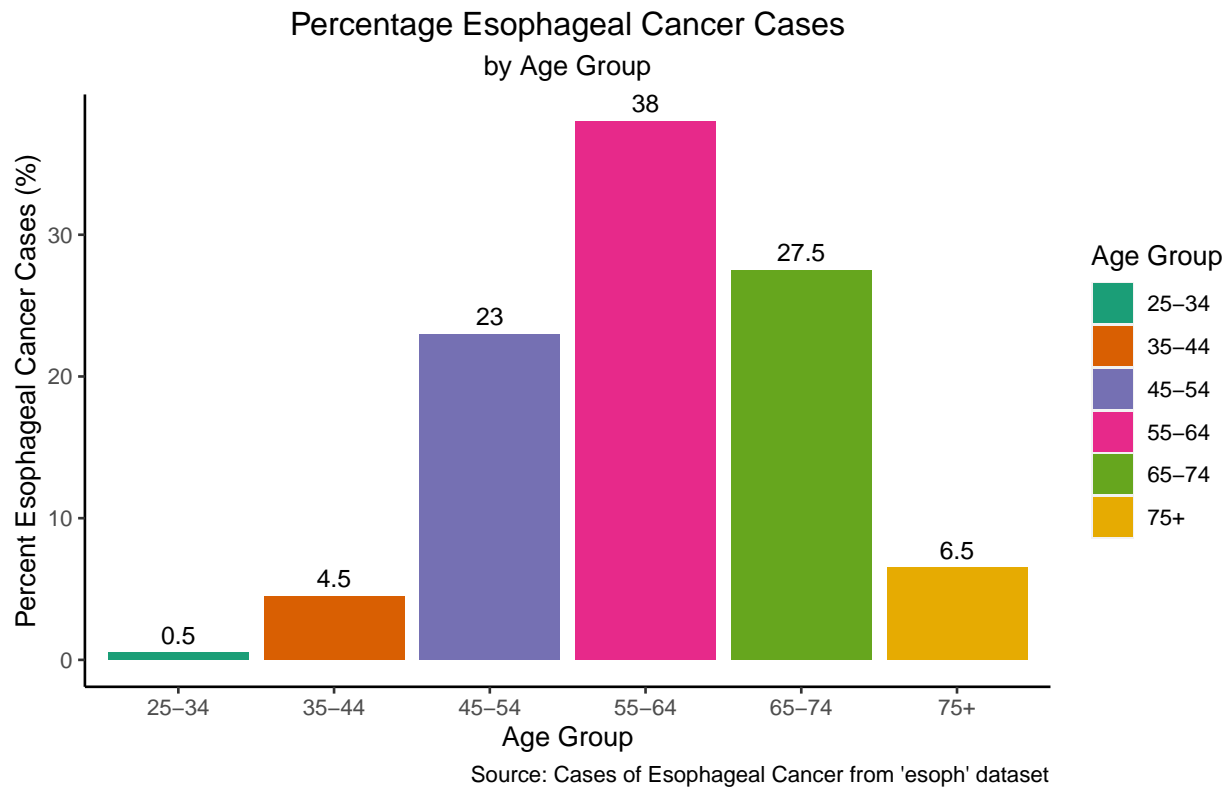


```
# Example 3: Dark Theme  
barchart2 + theme_solarized(light = FALSE)
```

TRY-IT-YOURSELF: We have individually gone through the methods to customize and elevate our plots. So, to produce a final, professional bar chart, can you combine all of the modifications (i.e. Captions/Labels, Color, and Legend) above into a single call?

```
#Type your answer here
custom_barchart <- ggplot(data=agg_agegp, aes(x=agegp, y=perc_cases, fill= agegp)) +
  geom_bar(stat="identity") +
  ggtitle("Percentage Esophageal Cancer Cases",
    subtitle = "by Age Group") +
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5)) +
  labs(x = "Age Group",
    y = "Percent Esophageal Cancer Cases (%)",
    caption = "Source: Cases of Esophageal Cancer from 'esoph' dataset") +
  guides(fill=guide_legend(title="Age Group")) +
  geom_text(aes(label = perc_cases), vjust = -0.50, size = 3.4, color = "black") +
  scale_fill_brewer(palette="Dark2") +
  theme(panel.grid.major = element_blank(), # Same as theme_classic()
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "black"))
custom_barchart
```



Part 3: Types of bar charts

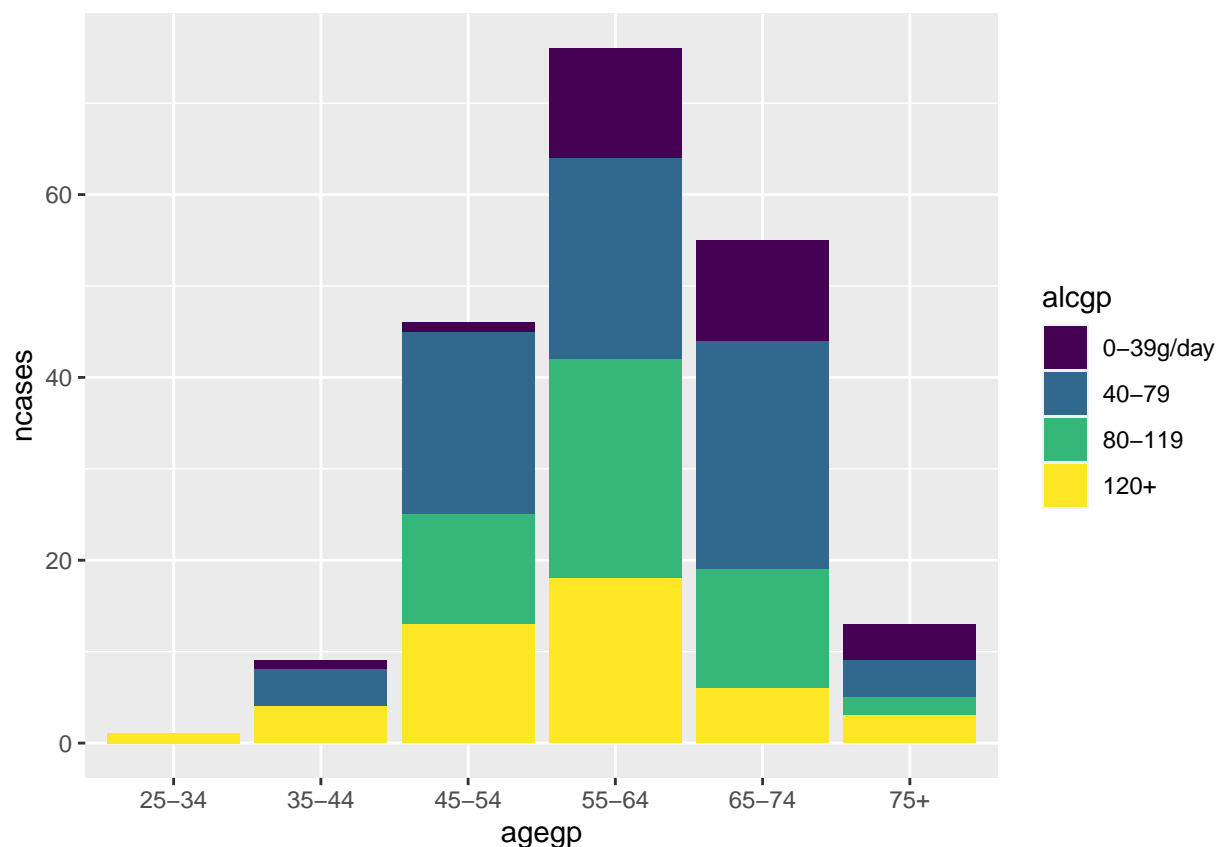
Alright, so we have explored how to modify regular bar charts to make them more professional and presentable.

In this part of the module, we will dive into creating more complex bar charts such as grouped and stacked bar charts.

3.1: Stacked bar chart

Earlier, we compared the number and percent of cases of esophageal cancer by age group and noticed that some age groups tended to have a higher percentage of cases than others. To explore this relationship more, let's look at how the number of cases are broken down by each age group's alcohol consumption.

```
stacked_bc <- ggplot(esoph, aes(x = agegp, y = ncases, fill = alcgp)) + geom_bar(stat = "identity")
stacked_bc
```

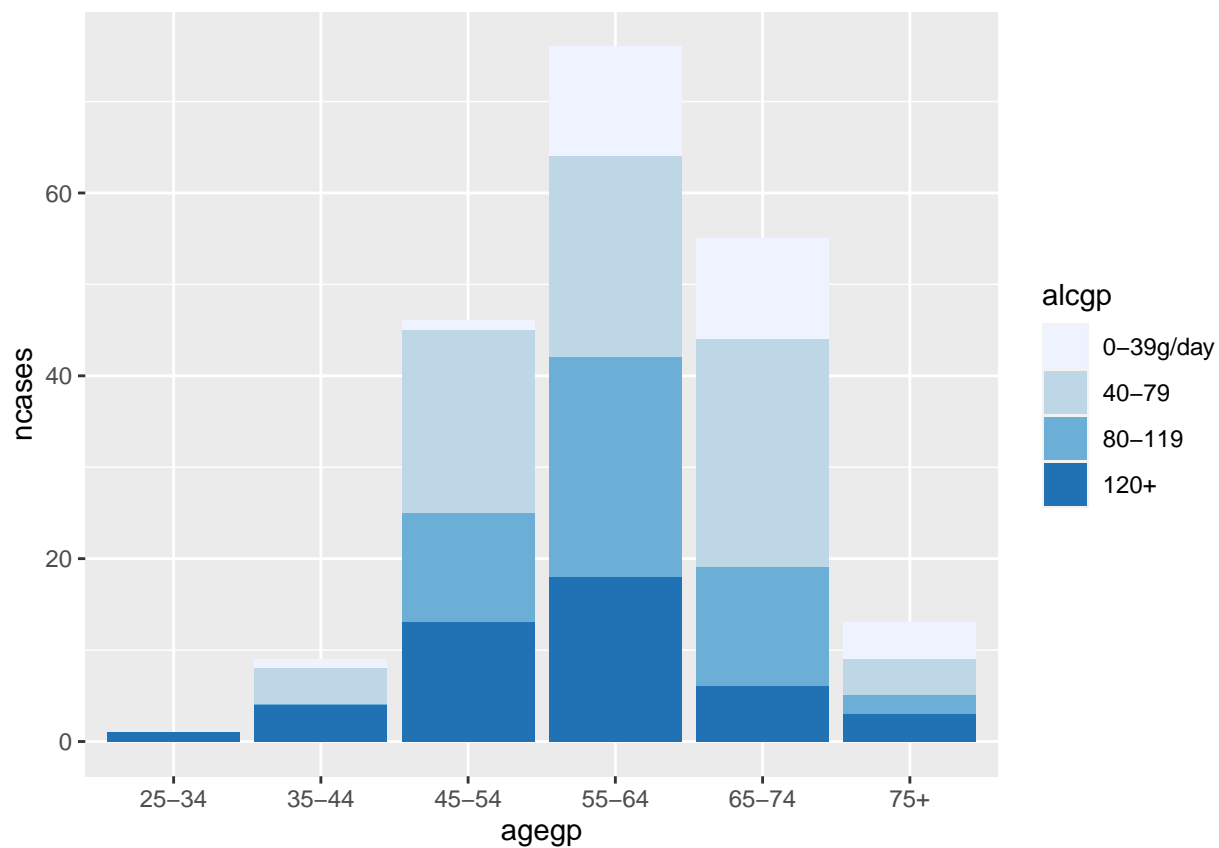


What do you notice about the distribution of alcohol consumption across the age groups?

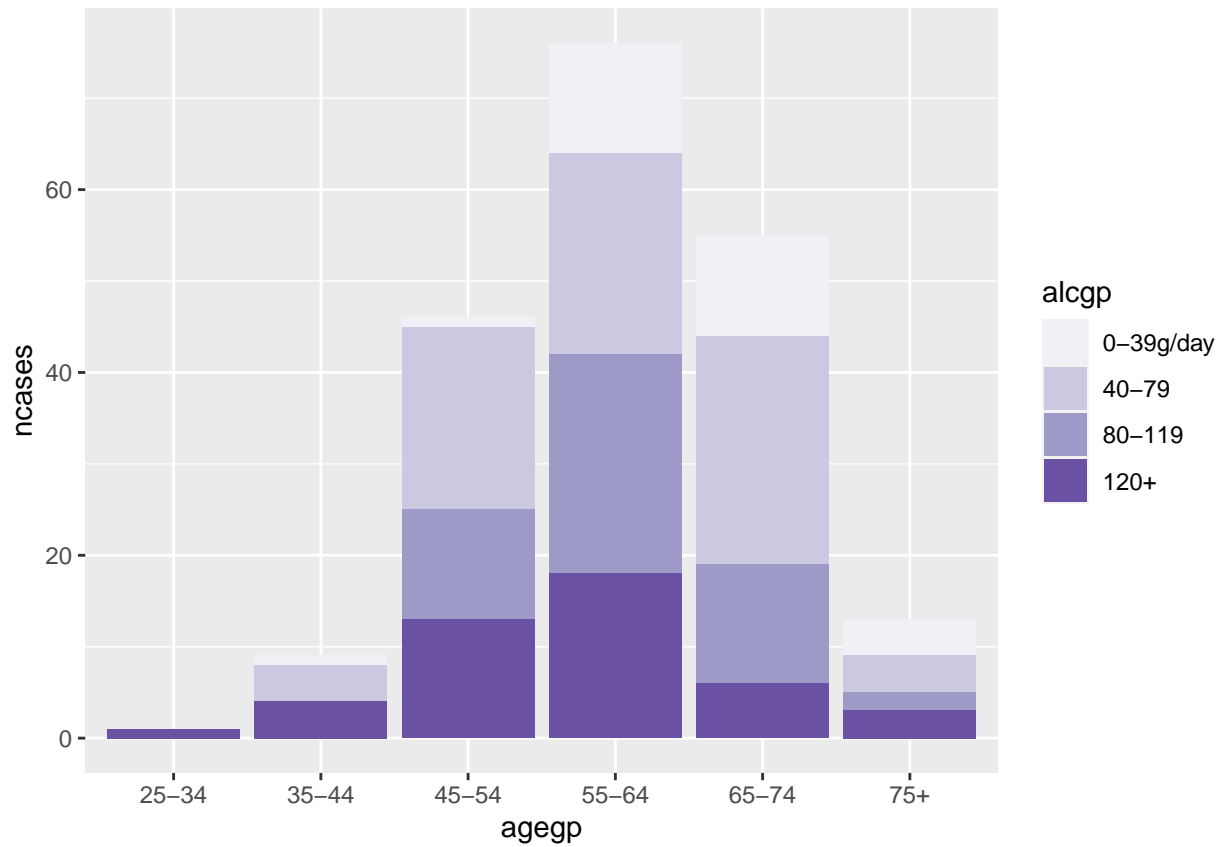
Currently, our stacked bar chart is using the default colors which may or may not look great or fit with our theme.

3.1.1: Color

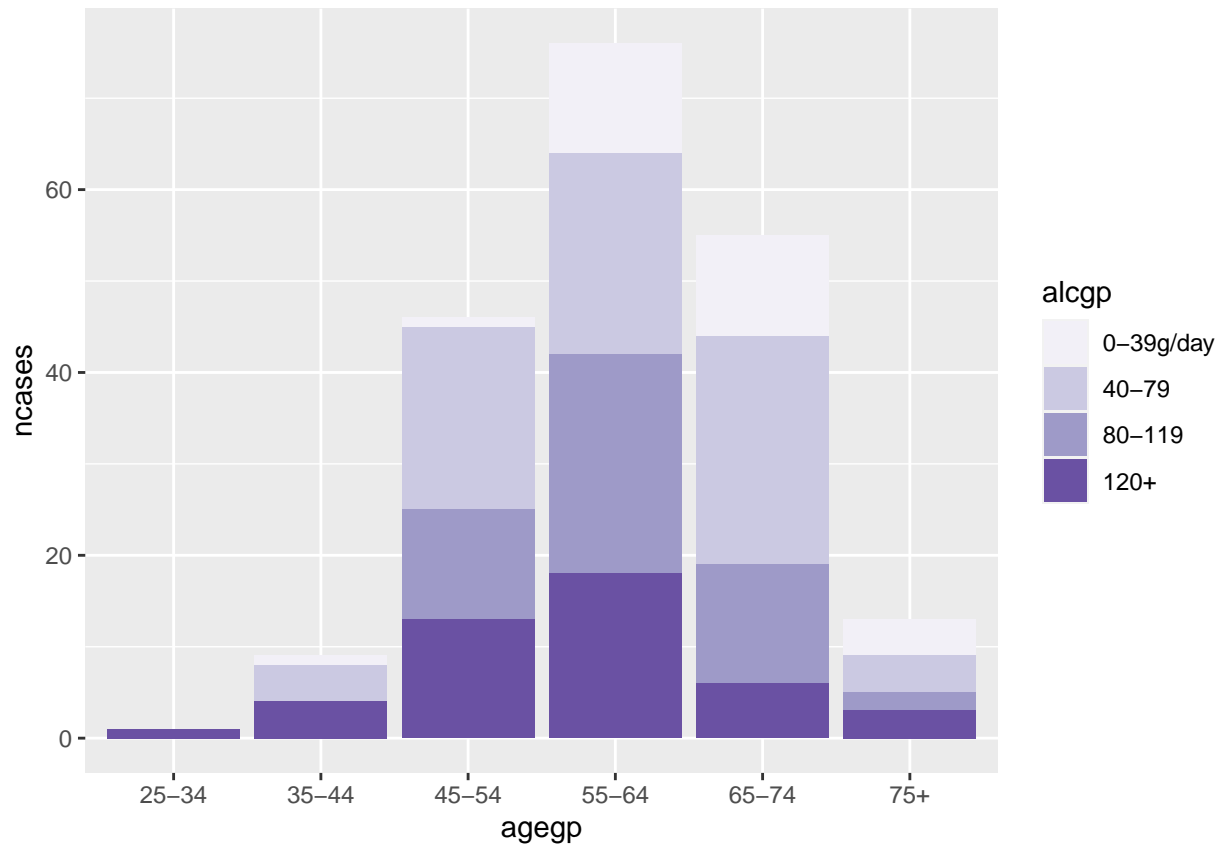
```
# Color Palettes  
stacked_bc + scale_fill_brewer() #Default Brewer color palette
```



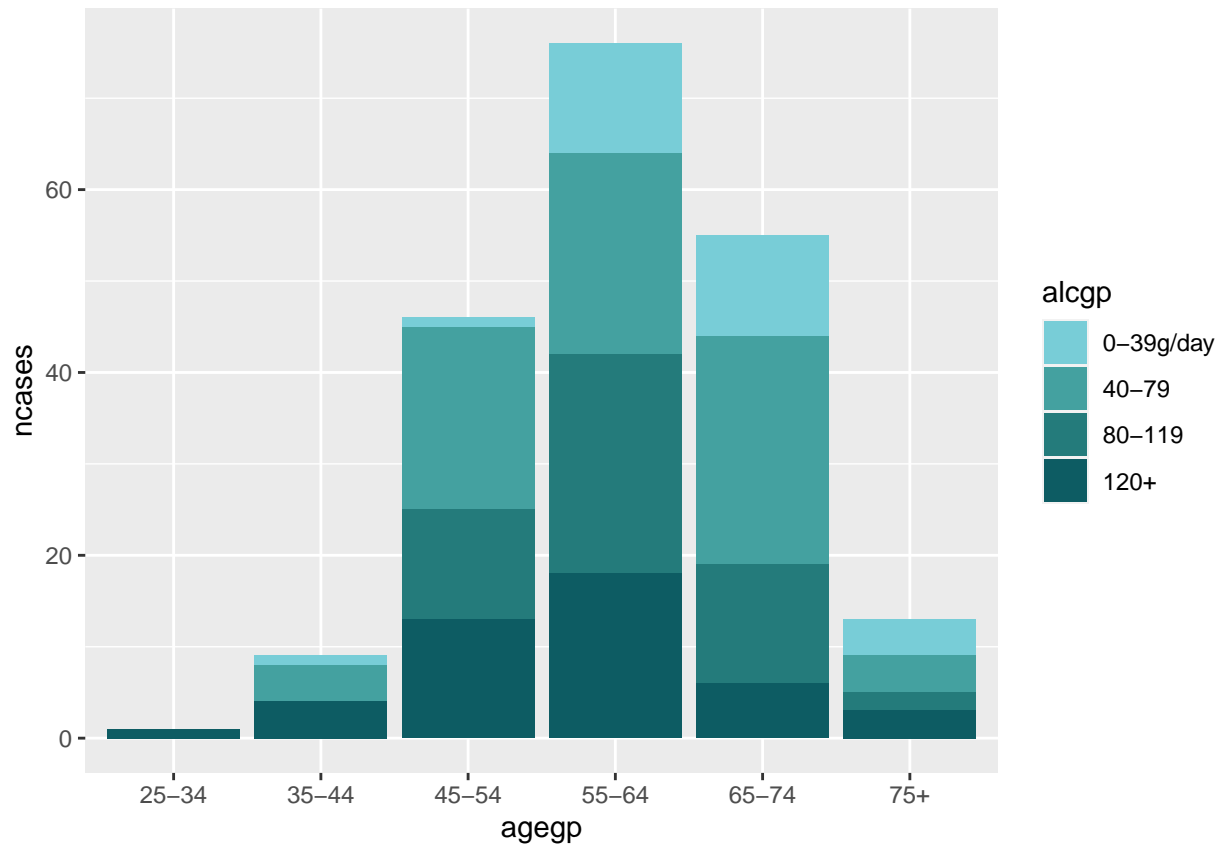
```
# Can specify the palette by palette name or number  
stacked_bc + scale_fill_brewer(palette = 12) #OR
```



```
stacked_bc + scale_fill_brewer(palette = "Purples")
```

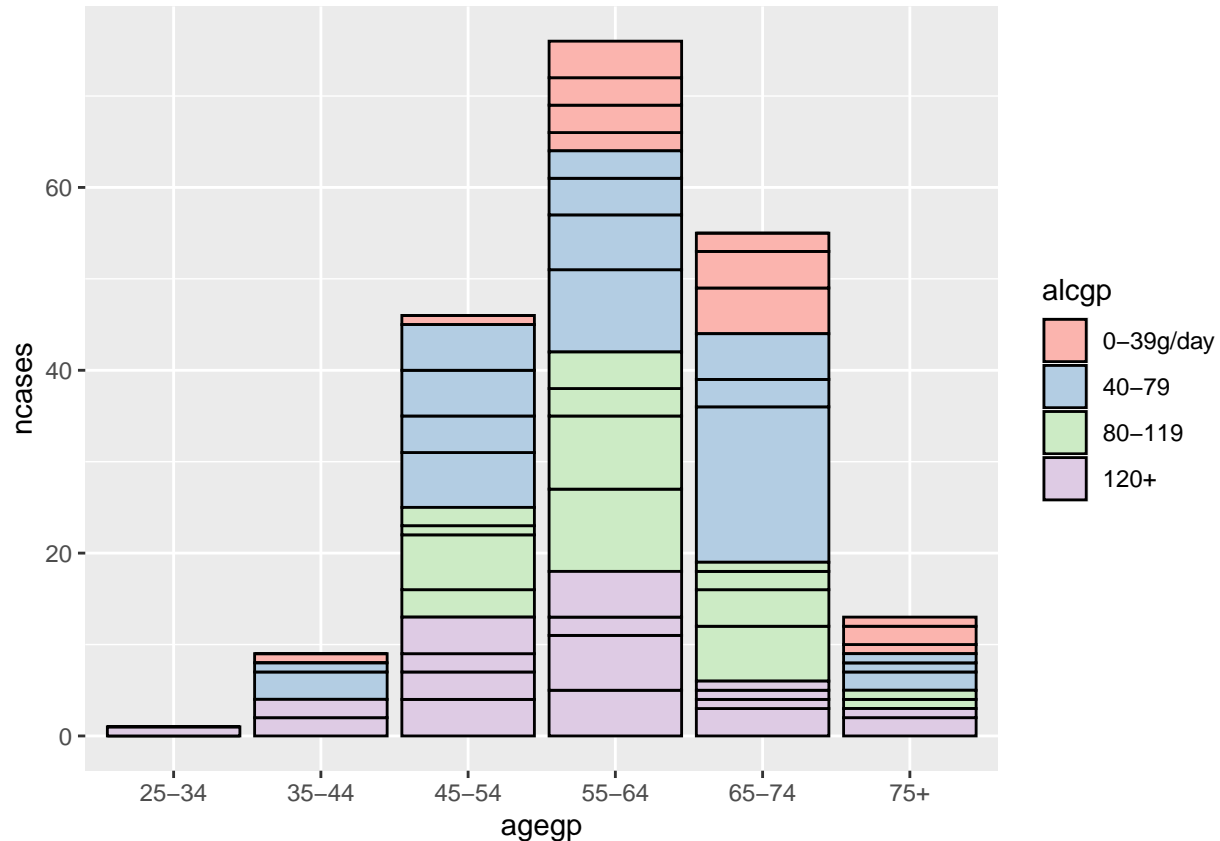


```
# Manually Choose Colors for what you are stratifying by (i.e. alcgp)
stacked_bc + scale_fill_manual(values = c("#78CDD7", "#44A1A0", "#247B7B", "#0D5C63"))
```



3.1.2: Border

```
# Add Border and Color
ggplot(esoph, aes(x = agegp, y = ncases, fill = alcgp)) +
  geom_bar(stat = "identity", color = "black") + #Specify Border by "color ="
  scale_fill_brewer(palette = "Pastel1")
```



```
# The borders look a bit weird due to the way the data is represented in our dataset
# (We have the same alcgp category duplicated multiple times for each agegp)
```

```
# To get the proper border, we will need to clean our data such that
# the alcgp category is not repeated for each agegp
```

```
alc_cases <- esoph %>%
  select(agegp, ncases) %>%
  group_by(agegp, alcgp) %>%
  summarize(total_cases = sum(ncases))
```

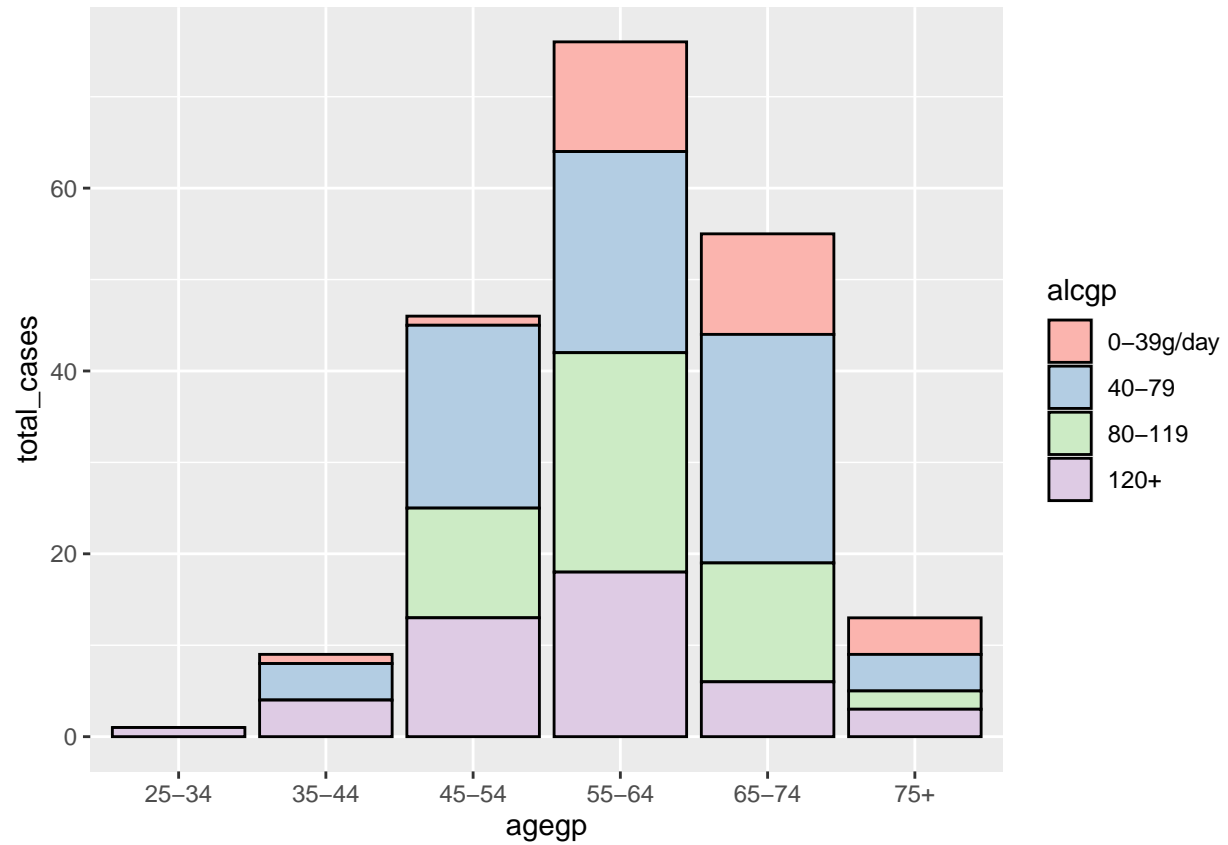
```
## 'summarise()' has grouped output by 'agegp'. You can override using the
## '.groups' argument.
```

```
# Compare this new clean data with the original dataset.
# What do you notice about the ways agegp, alcgp, and ncases are represented?
```

```
barchart3 <- ggplot(alc_cases, aes(x = agegp, y = total_cases, fill = alcgp)) +
```

```
geom_bar(stat = "identity", color = "black") +  
scale_fill_brewer(palette = "Pastell1")
```

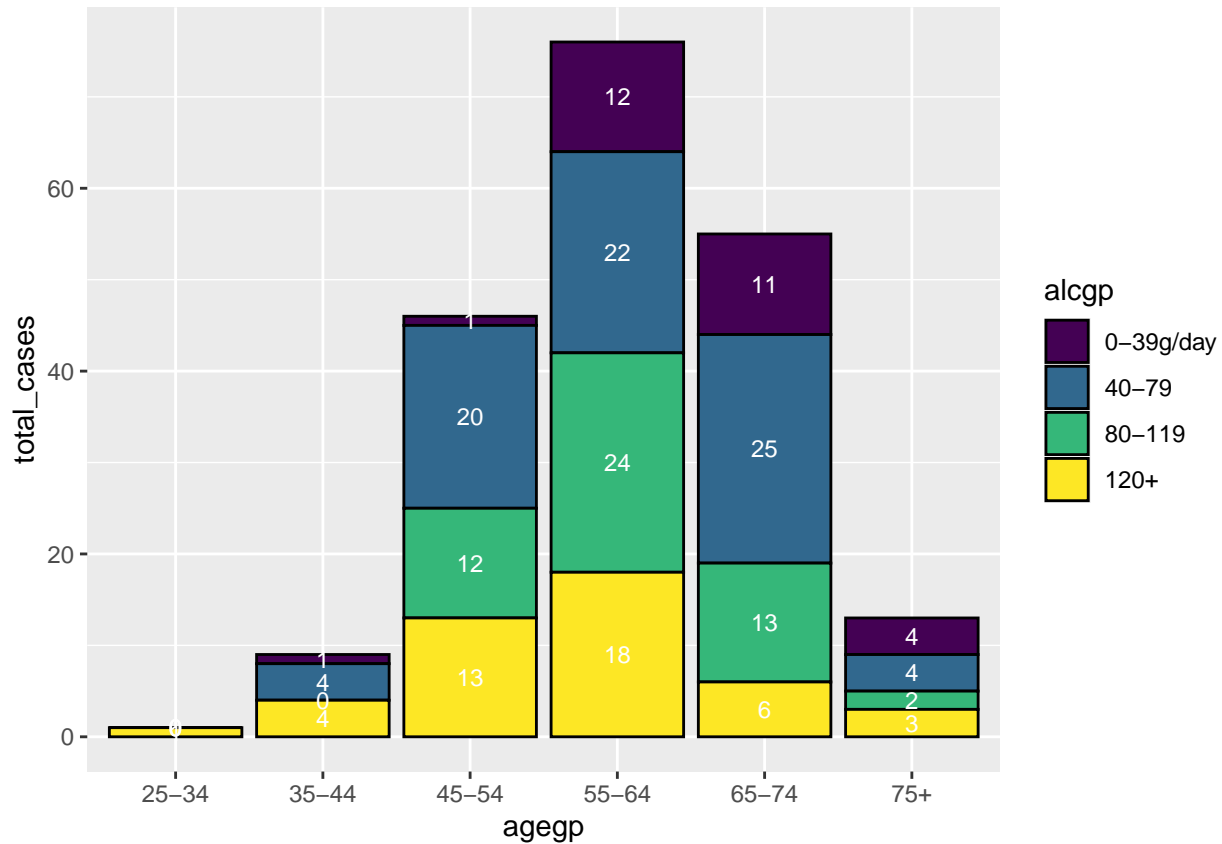
barchart3



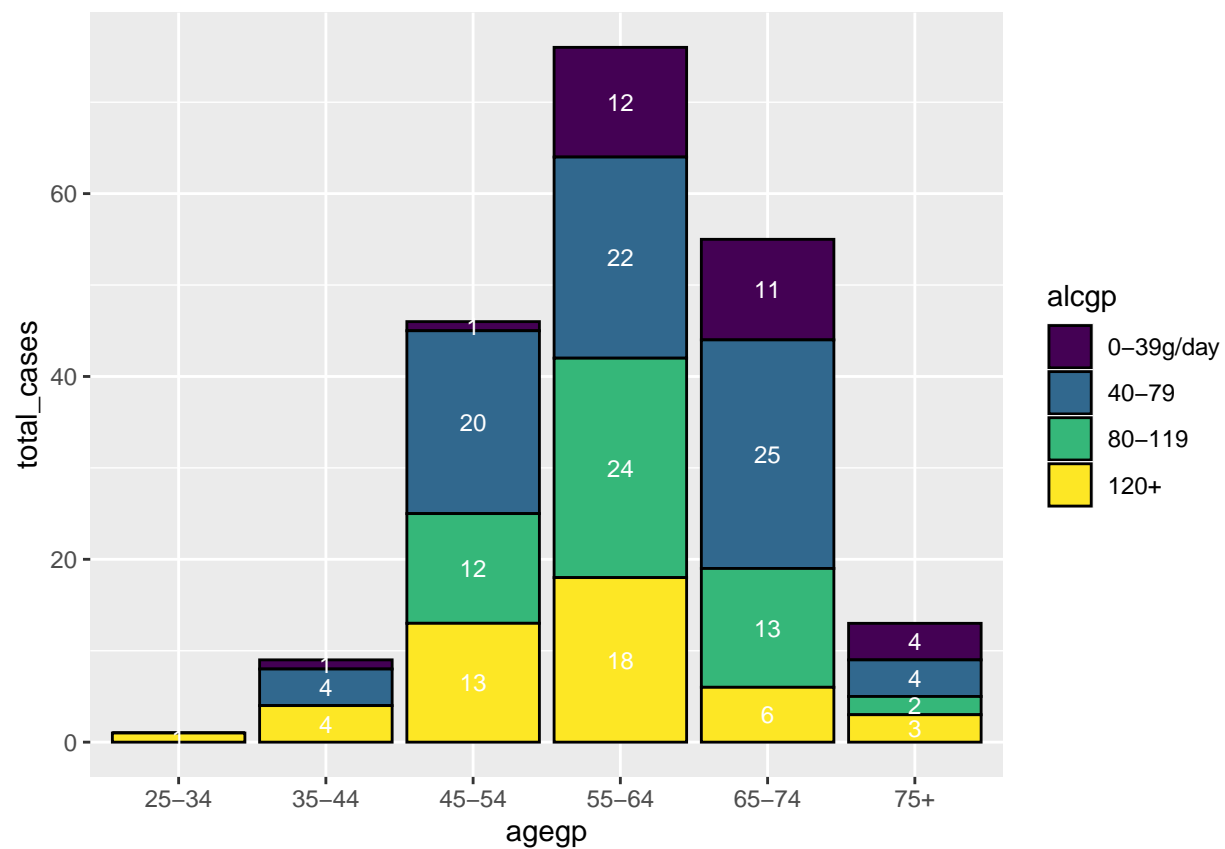
Now the borders look good!

3.1.3: Labels

```
ggplot(alc_cases, aes(x = agegp, y = total_cases, fill = alcgp,  
                      label = total_cases)) + # Need to specify the label  
  geom_bar(stat = "identity", color = "black") +  
  geom_text(position = position_stack(vjust = 0.5), size = 3, color = "#ffffff")
```

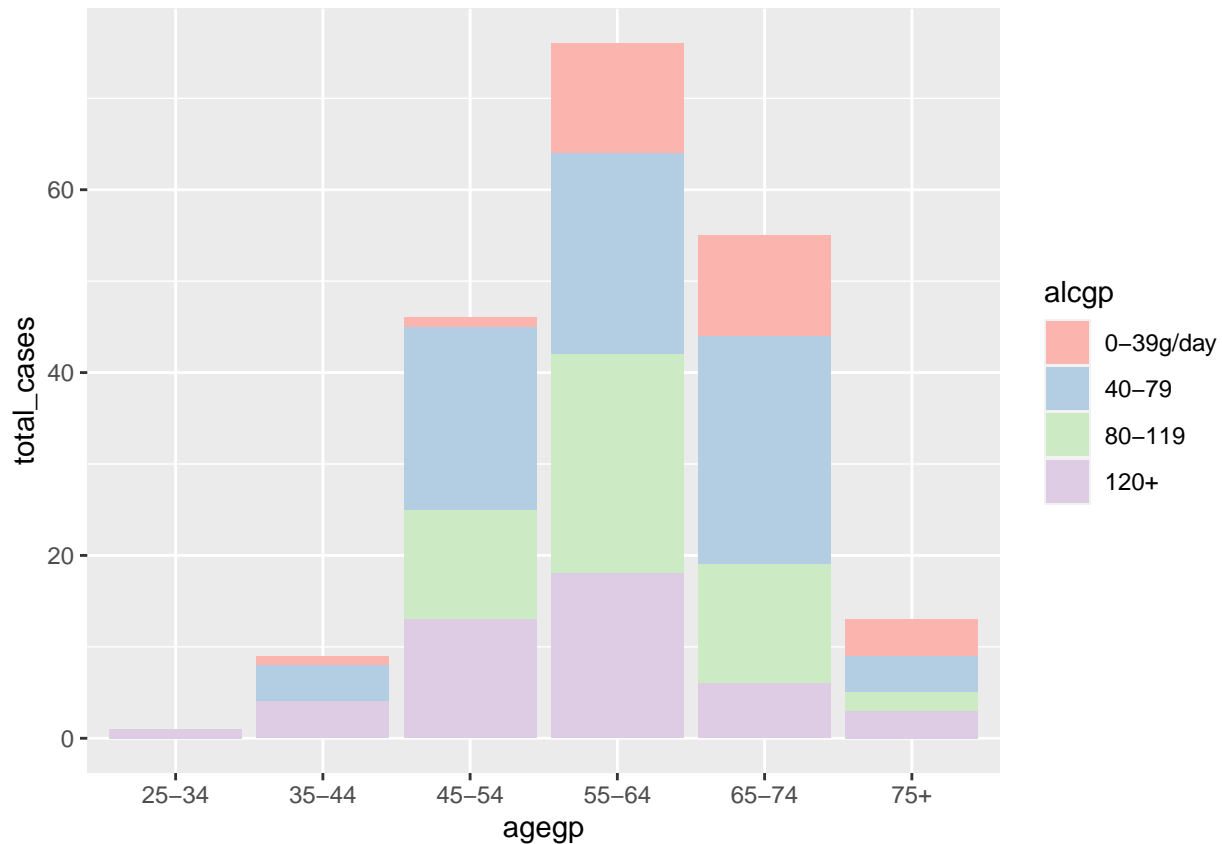


```
# need to mention position = position_stack(); vjust = 0.5 is for centering  
  
# In this case, this barchart is showing the values of 0, which we don't want,  
# so we can modify our code  
ggplot(alc_cases, aes(x=agegp, y=total_cases, fill = alcgp,  
                      label = total_cases)) +  
  geom_bar(stat="identity", color="black")+  
  geom_text(data=subset(alc_cases, total_cases != 0),  
            position = position_stack(vjust = 0.5), size=3, color = "#ffffff")
```



3.1.4: Legend

```
# Change the Category Labels
ggplot(alc_cases, aes(x = agegp, y = total_cases, fill = alcgp)) +
  geom_bar(stat = "identity") +
  scale_fill_brewer(palette = "Pastel1")
```

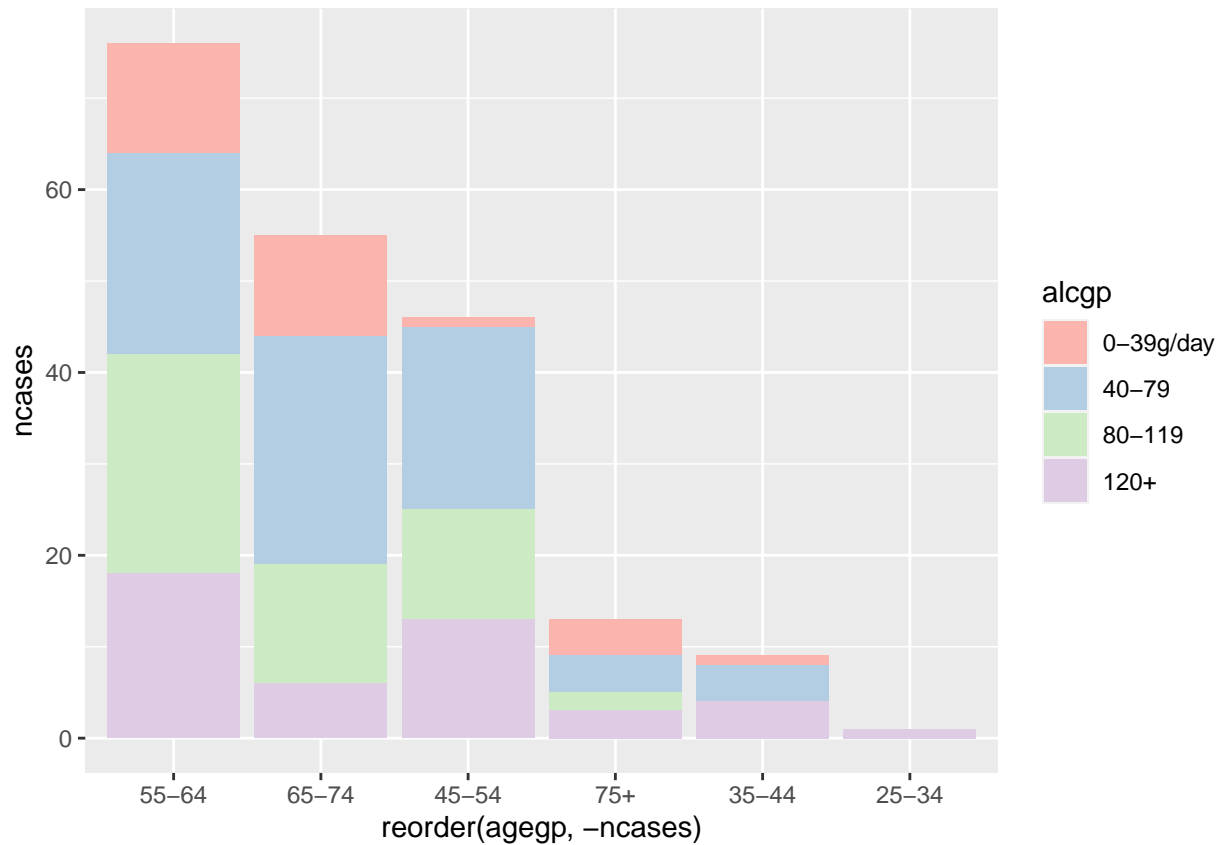


```
guides(fill=guide_legend( # 'Guide legend' allows to manually input a legend
  title="Alcohol Use \n      (g/day)")) + # '\n' adds a new line and then add spaces to center '(g/day)'
scale_fill_discrete(labels = c("<= 39", "40-79", "80-119", ">= 120"))
```

NULL

3.1.5: Ordering Bars

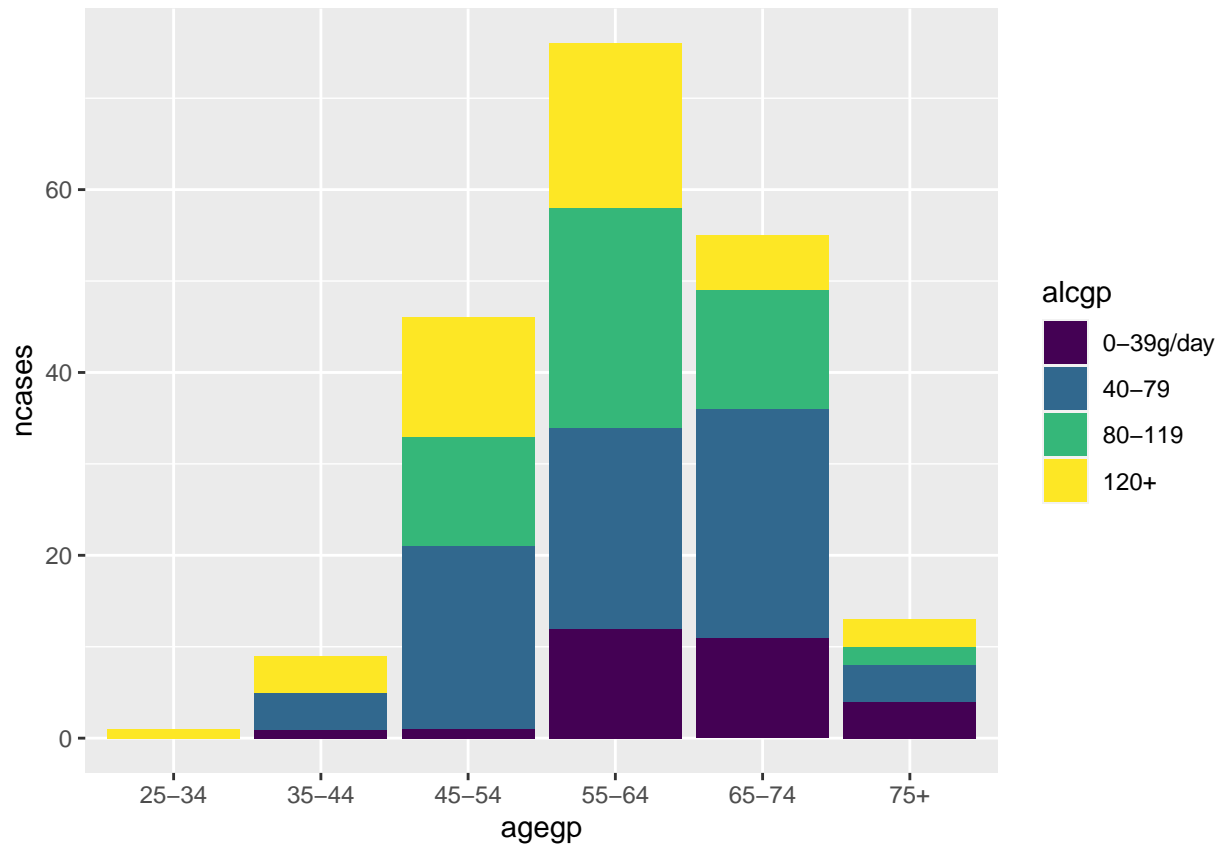
```
# Reordering the bars in ascending/descending order
ggplot(esoph, aes(x = reorder(agegp, -ncases), y = ncases, fill = alcgp)) + geom_bar(stat = "identity")
  scale_fill_brewer(palette = "Pastel1")
```



```
# Syntax: x = reorder(X variable, +/-Y variable); + = ascending, - = descending NOTE:
# It does not make sense to reorder the bars in this context as the age group
# categories are out of order
```

3.1.6: Reverse bar stacking

```
# Reverse the stacking of the bars
ggplot(esoph, aes(x = agegp, y = ncases, fill = alcgp)) + geom_bar(stat = "identity",
  position = position_stack(reverse = TRUE)) # Reversing stacking
```

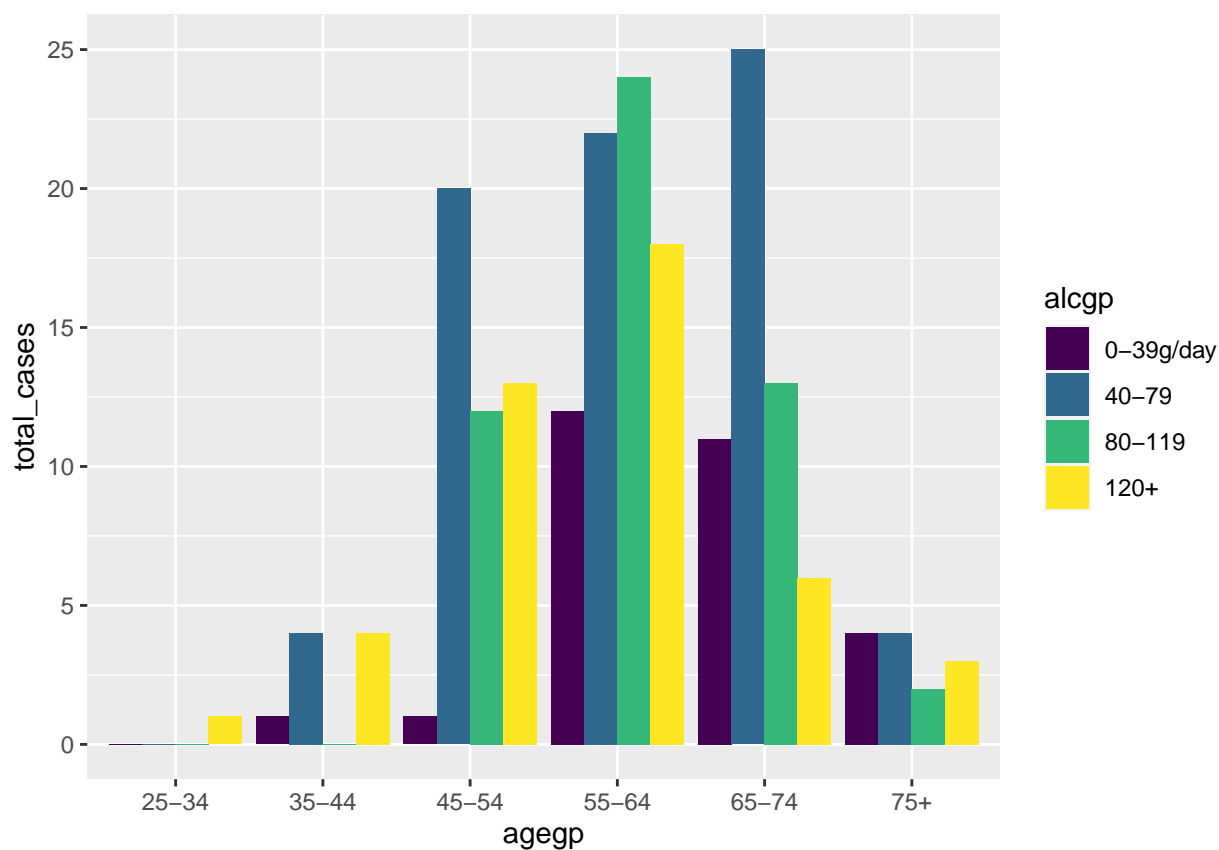


3.2: Dodged bar charts

Sometimes a stacked barchart may not be easy to understand or interpret. Well, there is a solution to that: Dodged bar charts!

Dodged bar charts are very similar to stacked bar charts, with very minor changes in code syntax. We will look at grouped bar charts with our previous example of alcohol consumption.

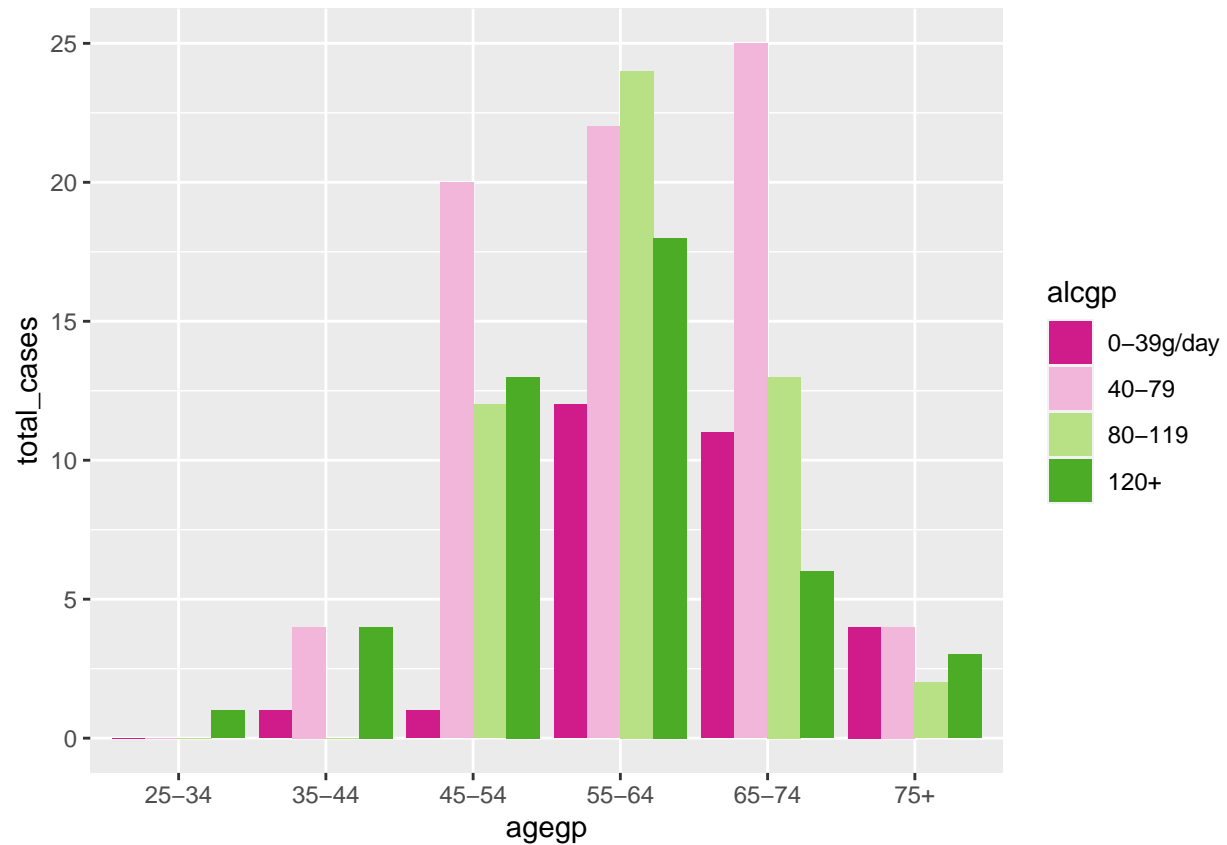
```
dodged_bc <- ggplot(alc_cases, aes(x = agegp, y = total_cases, fill = alcgp)) + geom_bar(stat = "identity", position = "dodge") # Need to specify position = 'dodge' to group bars next to each other
dodged_bc
```



3.2.1: Color (Redundant Maybe Delete? Move somewhere else?)

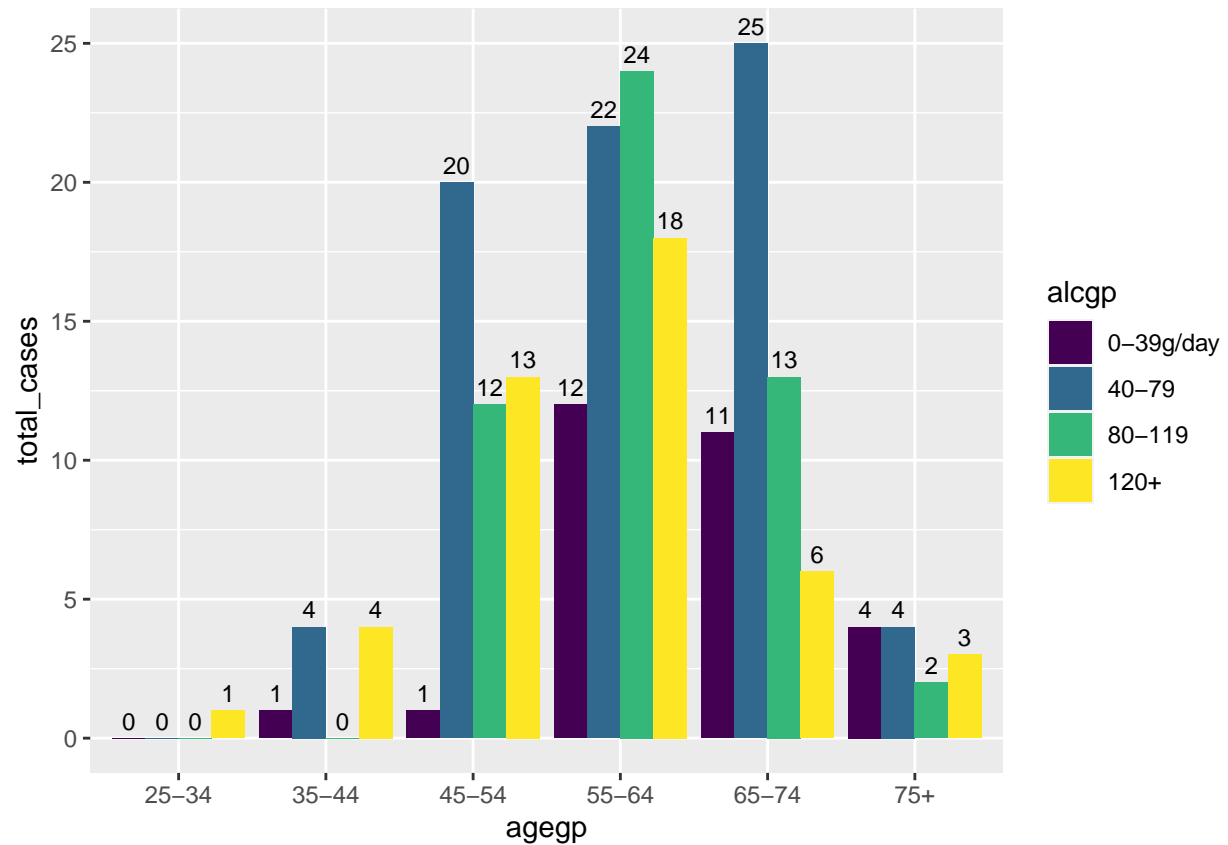
```
# Change color the same way you did for stacked barcharts with either
# scale_fill_manual or scale_fill_brewer

dodged_bc + scale_fill_brewer(palette = "PiYG")
```

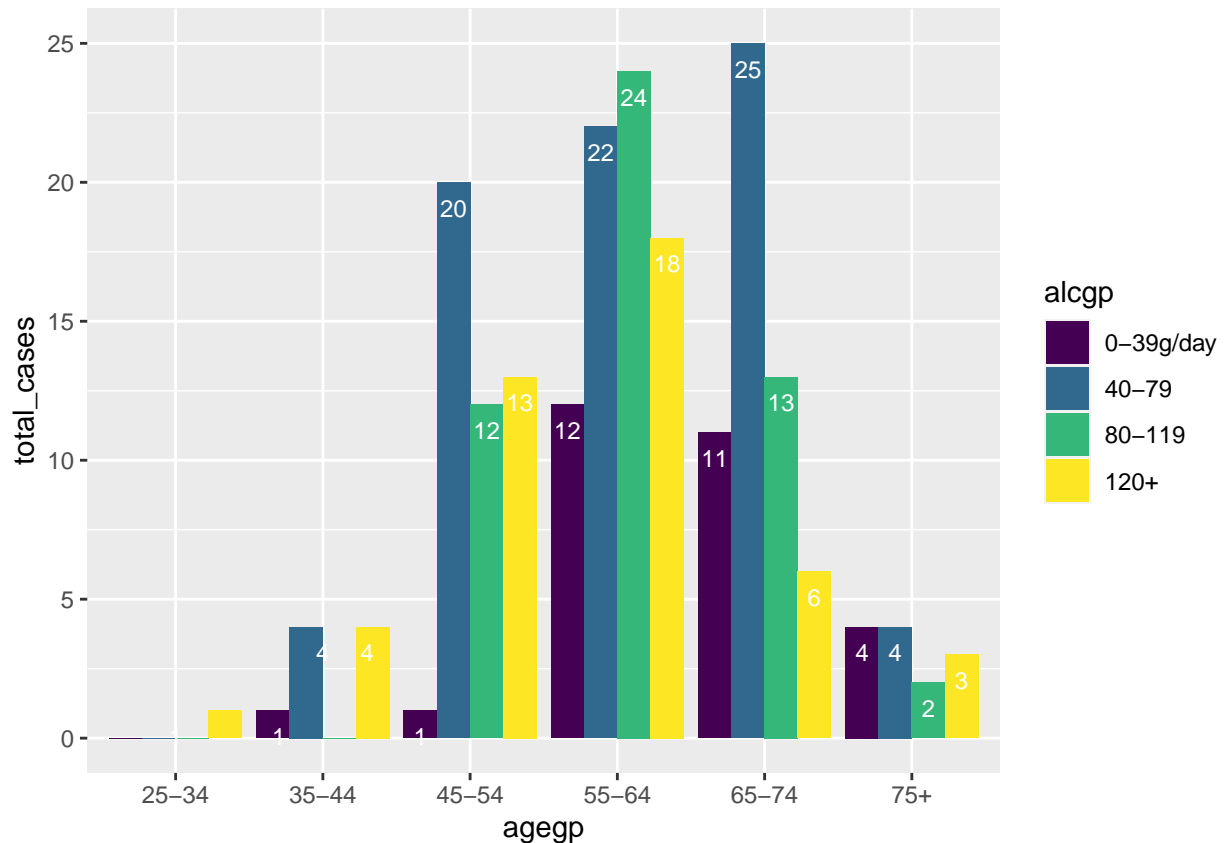



3.2.2: Labels (Redundant Maybe Delete? Move somewhere else?)

```
# If you would like to include the empty bar space along with the values
dodged_bc + geom_text(aes(label = total_cases), position = position_dodge(0.9), vjust = -0.5,
  size = 3, color = "black")
```



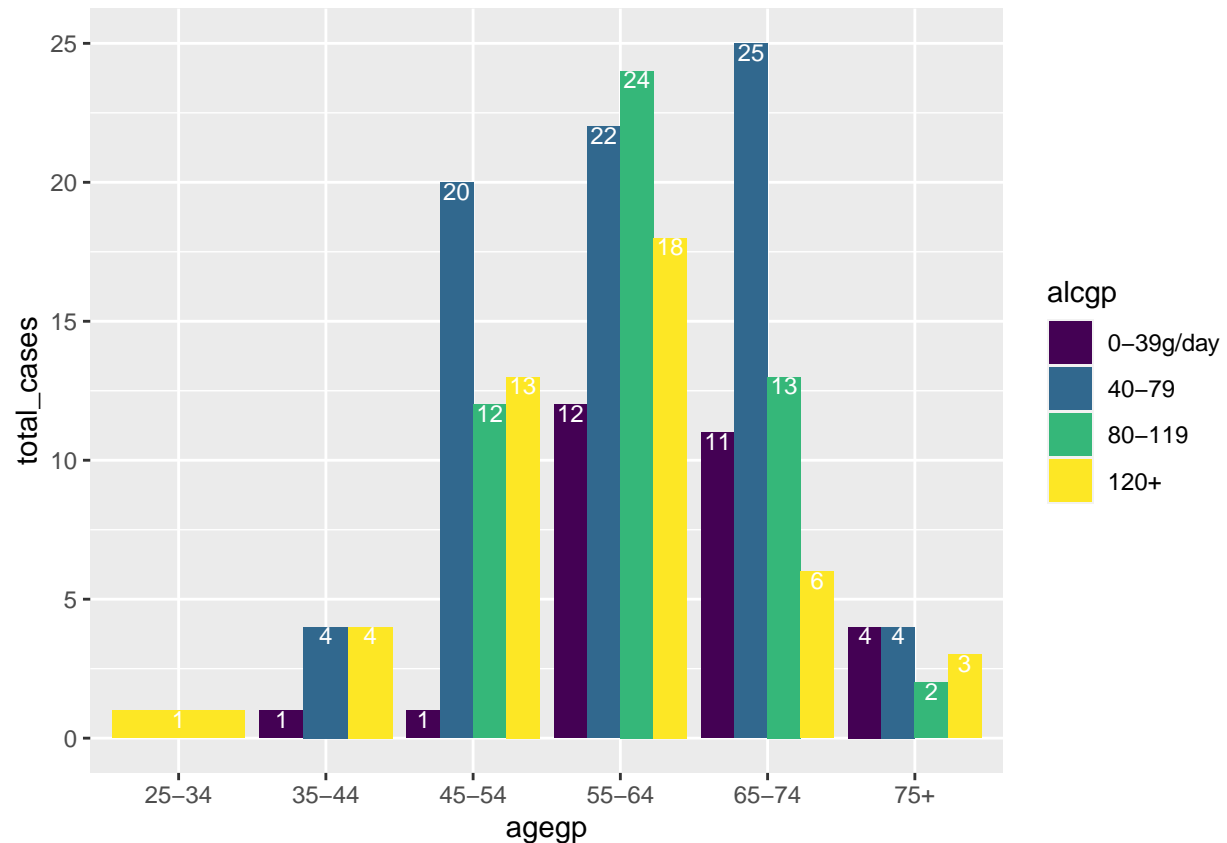
```
# If you do not want values of 0 to show up on the graph
dodged_bc + geom_text(data = subset(alc_cases, total_cases != 0), aes(label = total_cases),
  position = position_dodge(0.9), vjust = 2, size = 3, color = "#ffffff")
```



*# Notice that this graph above shows the empty bar space that we don't want and
therefore some of the numbers are not formatted on the bar. To fix this, we need
to remove any rows where the total_cases is 0 from our dataset and plot again*

```
alc_cases2 <- alc_cases %>%
  filter(total_cases > 0)

ggplot(alc_cases2, aes(x = agegp, y = total_cases, fill = alcgp)) + geom_bar(stat = "identity",
  position = "dodge") + geom_text(aes(label = total_cases), position = position_dodge(0.9),
  vjust = 1, size = 3, color = "#ffffff") # Much better!
```



The code for everything else for dodged bar charts is similar to what has been previously covered in this module.

TRY-IT-YOURSELF: You have already created a stacked bar chart for tobacco consumption. Similarly, now that we have gone through grouped bar charts, produce a professional grouped barchart that shows the number of cases of esophageal cancer by each age group's tobacco consumption.

Type your answer here

3.3: Faceted bar charts

3.3.0: Data Import

The **esoph** data set is too small and has too few features to support quality faceting therefore we will be leveraging the **NHANES** database to demonstrate **faceted bar charts**.

```
# Install the NHANES package if you need to install it install.packages('NHANES')

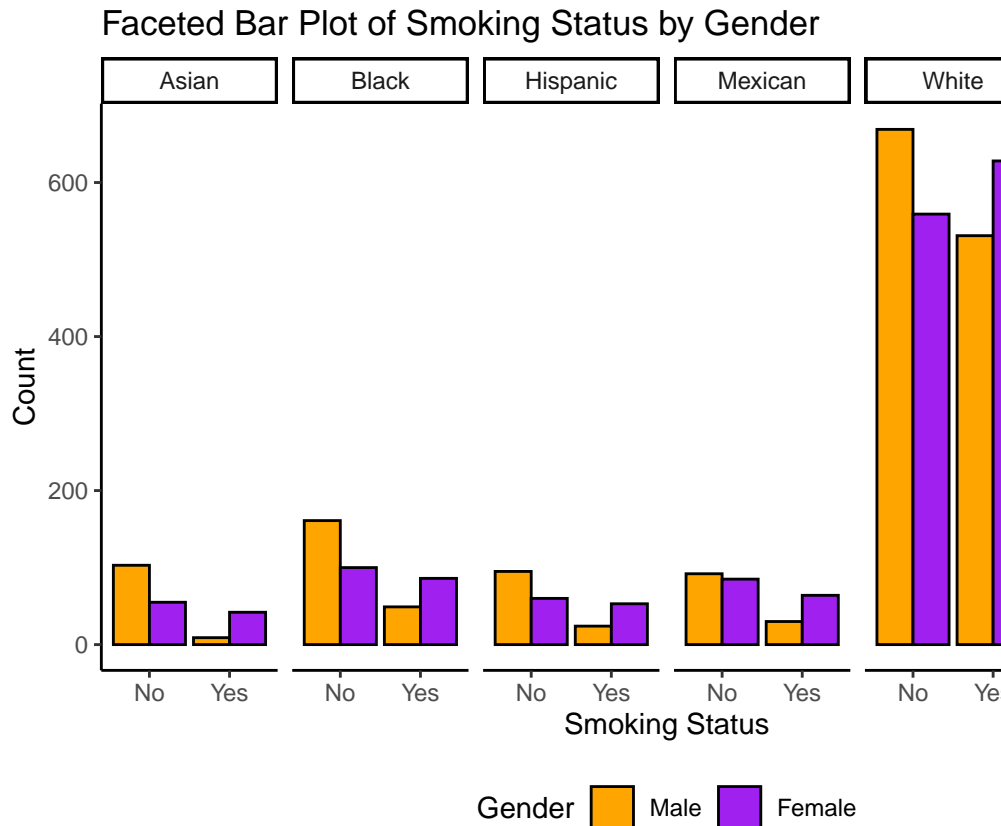
# Load NHANES library
library(NHANES)

# Load the 'NHANES' dataset
data(NHANES)

# Dropping NA values
NHANES_df <- NHANES %>%
  drop_na(Race3, Education)
```

3.3.1: Method 1: facet_grid()

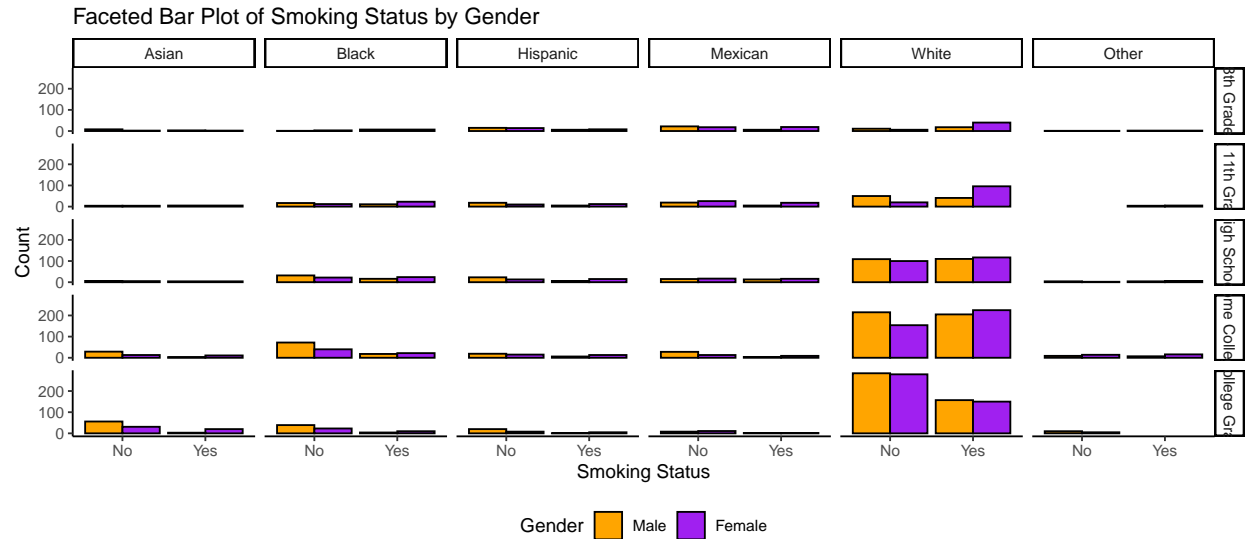
```
# Making it look cleaner
ggplot(NHANES_df, aes(x = Smoke100, fill = Gender)) +
  geom_bar(position = "dodge", color = "black") +
  labs(x = "Smoking Status", y = "Count", title = "Faceted Bar Plot of Smoking Status by Gender") +
  scale_fill_manual(values = c("male" = "purple", "female" = "orange"),
                    labels = c("Male", "Female")) + # Be mindful of assigning colors to gender
  facet_grid(~ Race3) + # 1 Feature: Education by Race
  theme_classic() +
  theme(legend.position = "bottom")
```



3.3.1.1: Faceting by one feature

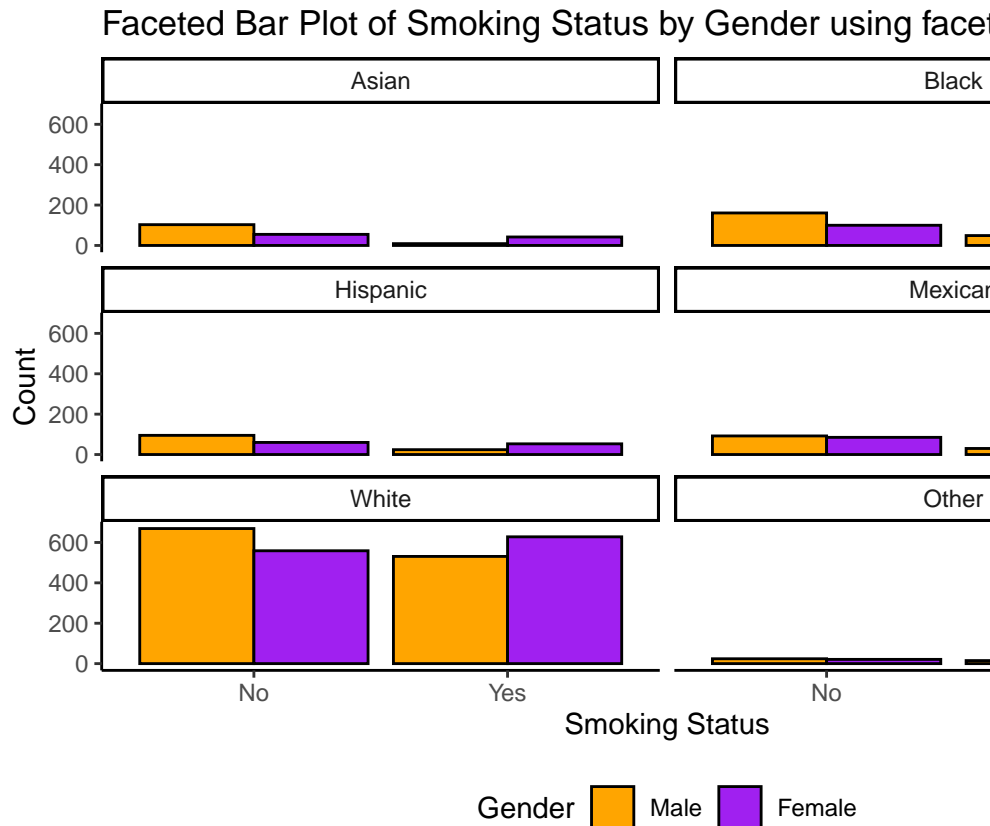
3.3.1.2: Faceting by two features

```
ggplot(NHANES_df, aes(x = Smoke100, fill = Gender)) +
  geom_bar(position = "dodge", color="black") +
  labs(x = "Smoking Status", y = "Count", title = "Faceted Bar Plot of Smoking Status by Gender") +
  scale_fill_manual(values = c("male" = "purple", "female" = "orange"),
                    labels = c("Male", "Female")) +
  facet_grid(Education ~ Race3) + # 2 Features: Education by Race
  theme_classic() +
  theme(legend.position = "bottom")
```



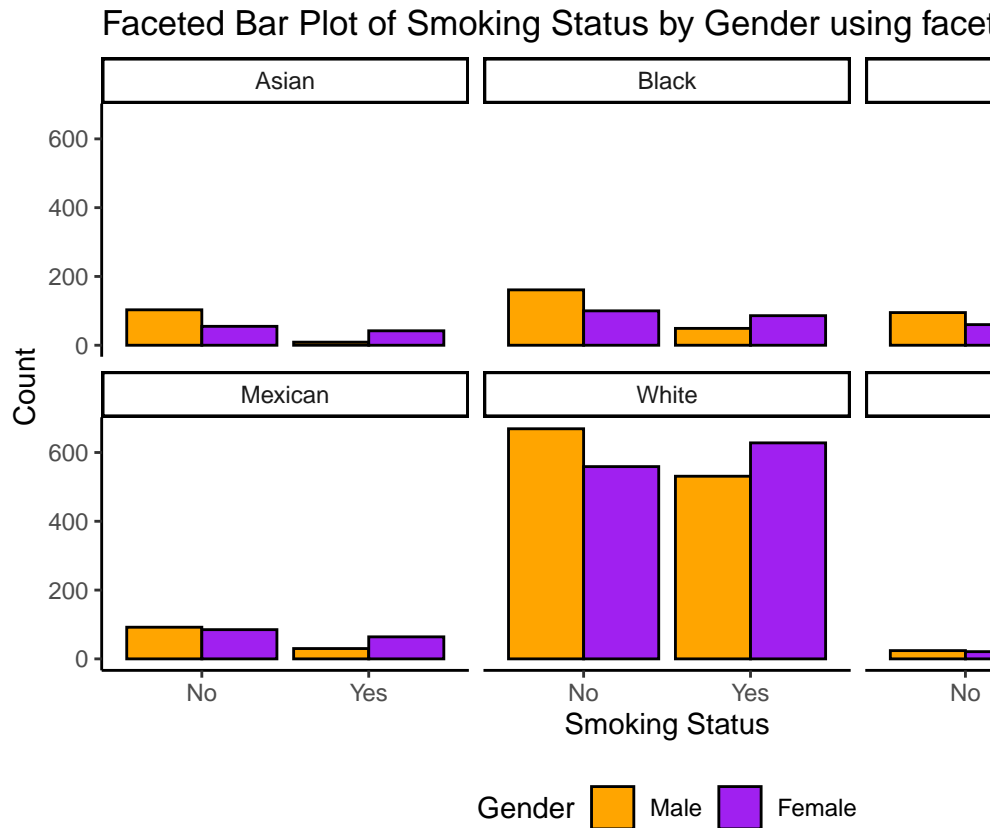
3.3.2: Method 2: facet_wrap()

```
# Create a faceted bar plot of smoking status by gender using facet_wrap with 2 columns
ggplot(NHANES_df, aes(x = Smoke100, fill = Gender)) +
  geom_bar(position = "dodge", color = "black") +
  labs(x = "Smoking Status", y = "Count", title = "Faceted Bar Plot of Smoking Status by Gender using facet_wrap") +
  scale_fill_manual(values = c("male" = "purple", "female" = "orange"),
                    labels = c("Male", "Female")) +
  facet_wrap(~ Race3, ncol = 2) + # 2 columns
  theme_classic() +
  theme(legend.position = "bottom")
```



3.3.2.1: Faceting with 2 columns

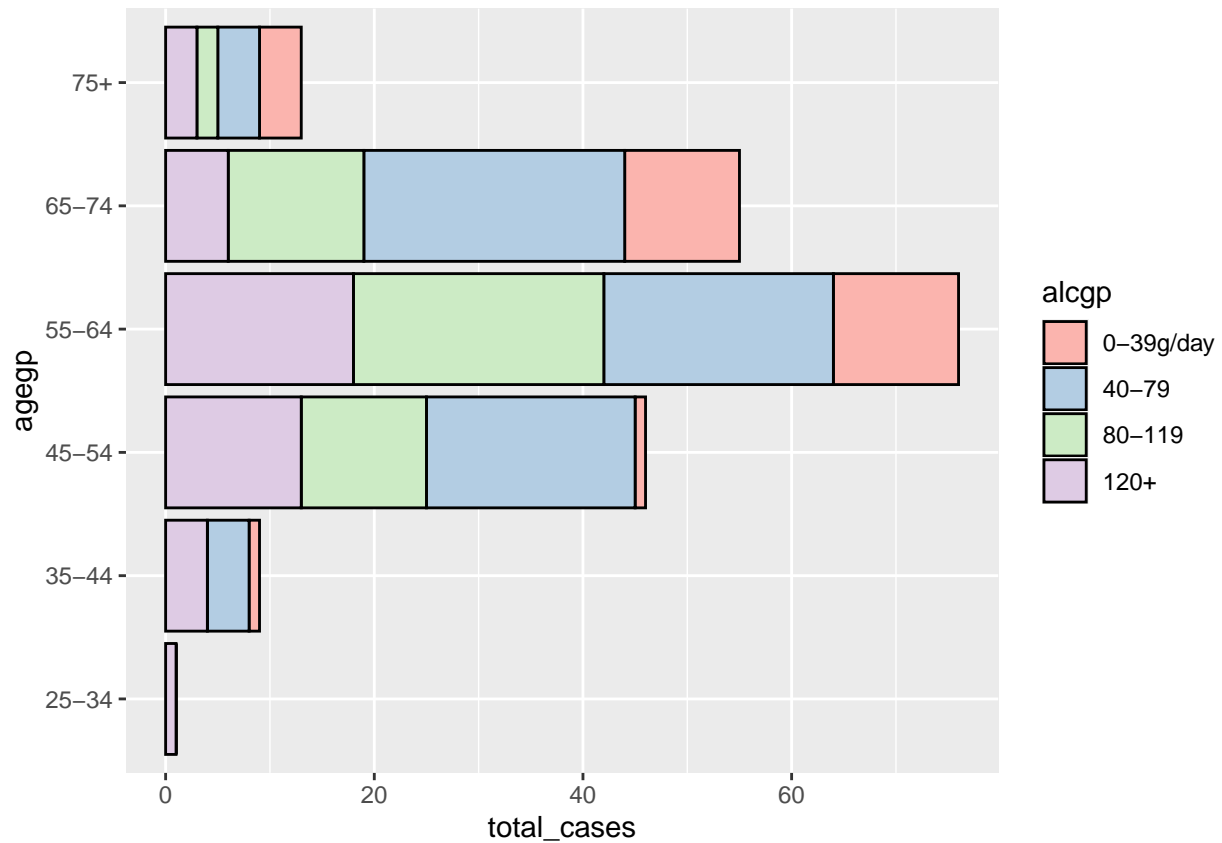
```
# Create a faceted bar plot of smoking status by gender using facet_wrap
ggplot(NHANES_df, aes(x = Smoke100, fill = Gender)) +
  geom_bar(position = "dodge", color = "black") +
  labs(x = "Smoking Status", y = "Count", title = "Faceted Bar Plot of Smoking Status by Gender using facet_wrap") +
  scale_fill_manual(values = c("male" = "purple", "female" = "orange"),
    labels = c("Male", "Female")) +
  facet_wrap(~ Race3, ncol = 3) + # 3 columns
  theme_classic() +
  theme(legend.position = "bottom")
```

3.3.2.2: Faceting with 3 columns

3.4: Horizontal bar chart

```
# Flipping the axis of the barchart
barchart3 + coord_flip() # Mention this to swap the x and y axes
```



TRY-IT-YOURSELF: Now that we have reviewed stacked barcharts, it's time for you to create and modify one! Produce a professional stacked barchart that shows the number of cases of esophageal cancer by each age group's tobacco consumption.

Type your answer here

Summary/Recap:

- We have reviewed how to customize `ggplot` figures for publications or presentation to a professional audience
- We explored modifications to color, labels, appearance, and legends
- You should be able to produce a polished bar chart that is publication-worthy (run the code below for a professional bar chart)

