Study of Applicants Applying for Masters and Phd programs at U of R

Veronica Mata Ramirez

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## 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
```

```
library(readr)
library(ggplot2)
library(dplyr)

data <- read.csv("ProjectforFall2023Dataset20220816204620.csv")

# Summary of the dataset
#summary(data)
str(data)
```

```
## 'data.frame':    17266 obs. of  46 variables:
##  $ Ref                                : int  557508409 330427691 16206600 240979251 678486354 733071123
##  $ Program..ASE.                      : chr  "Computer Science" "Computer Science" "Computer Science" "
##  $ Degree                             : chr  "Master's" "Master's" "Master's" "Master's" ...
##  $ Sub.Category                       : chr  "Theory" "Systems" "" "Health and Biomedical Sciences" ...
##  $ Entry.Term                         : chr  "Fall 2015" "Fall 2015" "Fall 2015" "Fall 2015" ...
##  $ Time.Status                        : chr  "Part Time" "Part Time" "Part Time" "Part Time" ...
##  $ Decision.1                         : chr  "Admit/Accept Offer" "Admit/Defer" "Admit/Accept Offer" "A
##  $ Sex                                : chr  "M" "M" "M" "M" ...
##  $ Birth.Country                      : chr  "United States" "United States" "United States" "United St
##  $ Age.at.App.Submission              : int  22 43 42 57 37 26 23 26 27 30 ...
##  $ Native.Language                    : chr  "EN" "EN" "EN" "EN" ...
##  $ Citizenship                        : chr  "US" "US" "US" "US" ...
##  $ Citizenship1                       : chr  "United States" "United States" "United States" "United St
##  $ Citizenship2                       : chr  "" "" "" "" ...
##  $ Have.you.ever.failed.a.course.     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Ever.Placed.on.Academic.Probation: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Institution.1.Name                 : chr  "Suny College At Geneseo" "University at Buffalo" "The Uni
##  $ Institution.1.Location             : chr  "NY" "NY" "CT" "NY" ...
##  $ Institution.1.Level.of.Study       : chr  "Undergraduate" "Graduate" "Graduate" "Graduate" ...
##  $ Institution.1.Degree               : chr  "Bachelor's" "PhD" "MD" "Master's" ...
```

```
##  $ Institution.1.Major              : chr  "Mathematics" "Geology" "" "Computer and Systems Engineeri
##  $ Institution.2.Name               : chr  "" "University at Buffalo" "University of Saint Joseph" "U
##  $ Institution.2.Location           : chr  "" "NY" "CT" "VA" ...
##  $ Institution.2.Level.of.Study     : chr  "" "Undergraduate" "Undergraduate" "Undergraduate" ...
##  $ Institution.2.Degree             : chr  "" "Bachelor's" "" "Bachelor's" ...
##  $ Institution.2.Major              : chr  "" "Mechanical Engineering" "Post-bacc premedical courses
##  $ Institution.3.Name               : chr  "" "" "New York University" "" ...
##  $ Institution.3.Location           : chr  "" "" "NY" "" ...
##  $ Institution.3.Level.of.Study     : chr  "" "" "Undergraduate" "" ...
##  $ Institution.3.Degree             : chr  "" "" "Bachelor's" "" ...
##  $ Institution.3.Major              : chr  "" "" "Latin American Studies" "" ...
##  $ Job.1.Title                      : chr  "Jr. Analyst/Programmer" "Academic Counselor" "Assistant I
##  $ Job.2.Title                      : chr  "Advanced Developer / Database Liaison (Front and Back End
##  $ Job.3.Title                      : chr  "" "" "Instructor" "Adjunct Instructor" ...
##  $ Recommender.1.Relationship       : chr  "Professor" "Supervisor" "Colleague" "Current Manager/Supe
##  $ Recommender.2.Relationship       : chr  "Boss" "Colleague" "Colleague" "Former co-worker" ...
##  $ Recommender.3.Relationship       : chr  "Undergraduate Advisor and Professor" "" "" "Former manage
##  $ Previously.Applied.              : chr  "No" "No" "No" "No" ...
##  $ Current.Student.                 : chr  "No" "No" "No" "No" ...
##  $ Type                             : chr  "" "" "" "" ...
##  $ Spouse.Studying.Applying         : chr  "No" "No" "No" "No" ...
##  $ Currently.Employed.              : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ How.Applicant.Heard.About.UR     : chr  "Local resident (current or past)" "" "" "Family/Friend"
##  $ How.Applicant.Heard...Other      : chr  "" "" "" "" ...
##  $ Any.Relatives.Listed.            : chr  "No" "No" "No" "Yes" ...
##  $ Other.Schools.Applying.To        : chr  "" "" "" "" ...
```

```
# Number of samples
cat("Number of total samples:", nrow(data), "\n")
```

```
## Number of total samples: 17266
```

```
masters <- data %>% filter(Degree == "Master's")
phd <- data %>% filter(Degree == "PhD")
cat("Number of Master's records:", nrow(masters), "\n")
```

```
## Number of Master's records: 12851
```

```
cat("Number of PhD records:", nrow(phd), "\n")
```

```
## Number of PhD records: 4236
```

```
print("Count of missing values by column")
```

```
## [1] "Count of missing values by column"
```

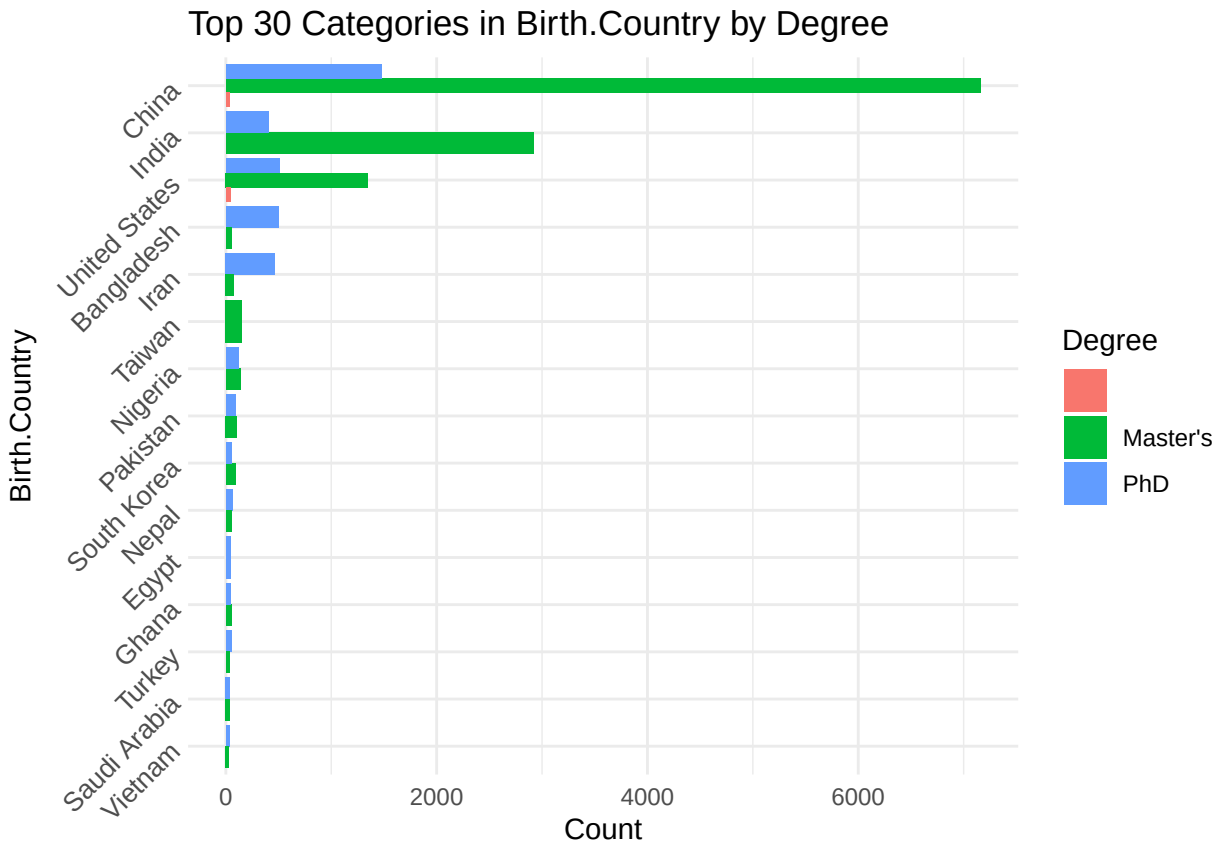
```
empty_or_na_counts <- sapply(data, function(x) sum(x == "" | is.na(x)))
sorted_counts <- sort(empty_or_na_counts, decreasing = TRUE)
print(sorted_counts)
```

```
##                              Type                       Citizenship2
##                             17162                              17060
##       How.Applicant.Heard...Other               Institution.3.Degree
##                             16860                              16603
##               Institution.3.Major             Institution.3.Location
##                             16307                              16256
##      Institution.3.Level.of.Study                 Institution.3.Name
##                             16255                              16030
```

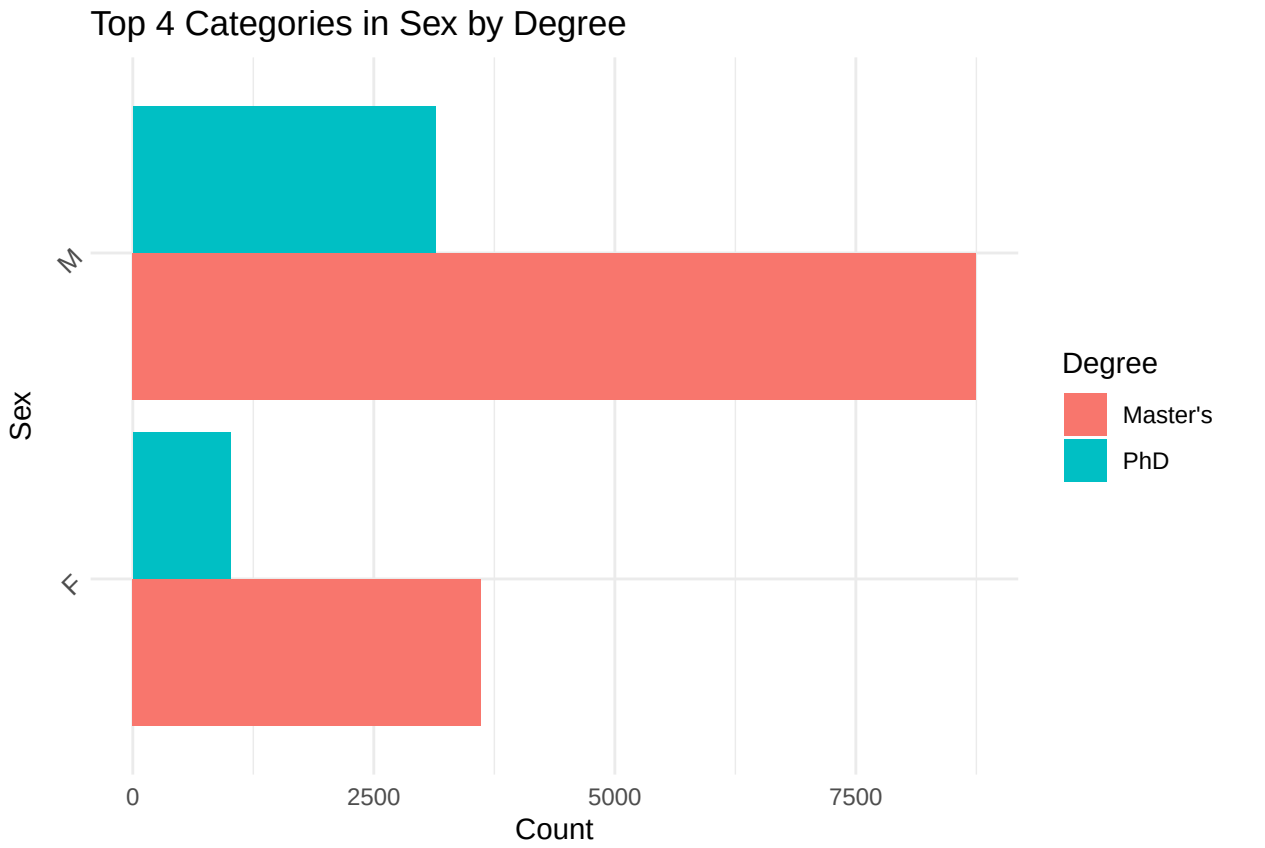
```
##                       Job.3.Title               Institution.2.Degree
##                             13854                              13600
##               Institution.2.Major            Institution.2.Location
##                             12468                              12307
##      Institution.2.Level.of.Study                Institution.2.Name
##                             12300                              11734
##         Other.Schools.Applying.To                       Job.2.Title
##                             11536                              10578
##                       Job.1.Title        Recommender.3.Relationship
##                              5564                               3720
##                        Decision.1             Age.at.App.Submission
##                              3529                               3481
##        Recommender.2.Relationship        Recommender.1.Relationship
##                              3211                               3106
##              Institution.1.Degree      How.Applicant.Heard.About.UR
##                              3032                               2141
##               Institution.1.Major            Institution.1.Location
##                              1586                               1451
##      Institution.1.Level.of.Study                Institution.1.Name
##                              1413                               1335
##    Have.you.ever.failed.a.course. Ever.Placed.on.Academic.Probation
##                               901                                883
##                    Native.Language                       Citizenship1
##                               749                                690
##                        Citizenship                       Sub.Category
##                               663                                653
##                               Sex                      Birth.Country
##                               624                                250
##                       Time.Status                             Degree
##                               146                                128
##                               Ref                      Program..ASE.
##                                 0                                  0
##                        Entry.Term                Previously.Applied.
##                                 0                                  0
##                  Current.Student.           Spouse.Studying.Applying
##                                 0                                  0
##               Currently.Employed.              Any.Relatives.Listed.
##                                 0                                  0
```

```r
# Function to create braplot
create_bar_plot <- function(data, column_name, top_n = 15, fill_column = "Entry.Term") {
  # Check if the column names exist in the data frame
  if (!column_name %in% names(data)) {
    stop("The specified column does not exist in the data frame.")
  }
  if (!fill_column %in% names(data)) {
    stop("The specified fill column does not exist in the data frame.")
  }

  prepared_data <- data %>%
    filter(!is.na(!!sym(column_name)) & !!sym(column_name) != "") %>%
    group_by(!!sym(column_name), !!sym(fill_column)) %>%
    summarise(Count = n(), .groups = "drop") %>%
    arrange(desc(Count)) %>%
```

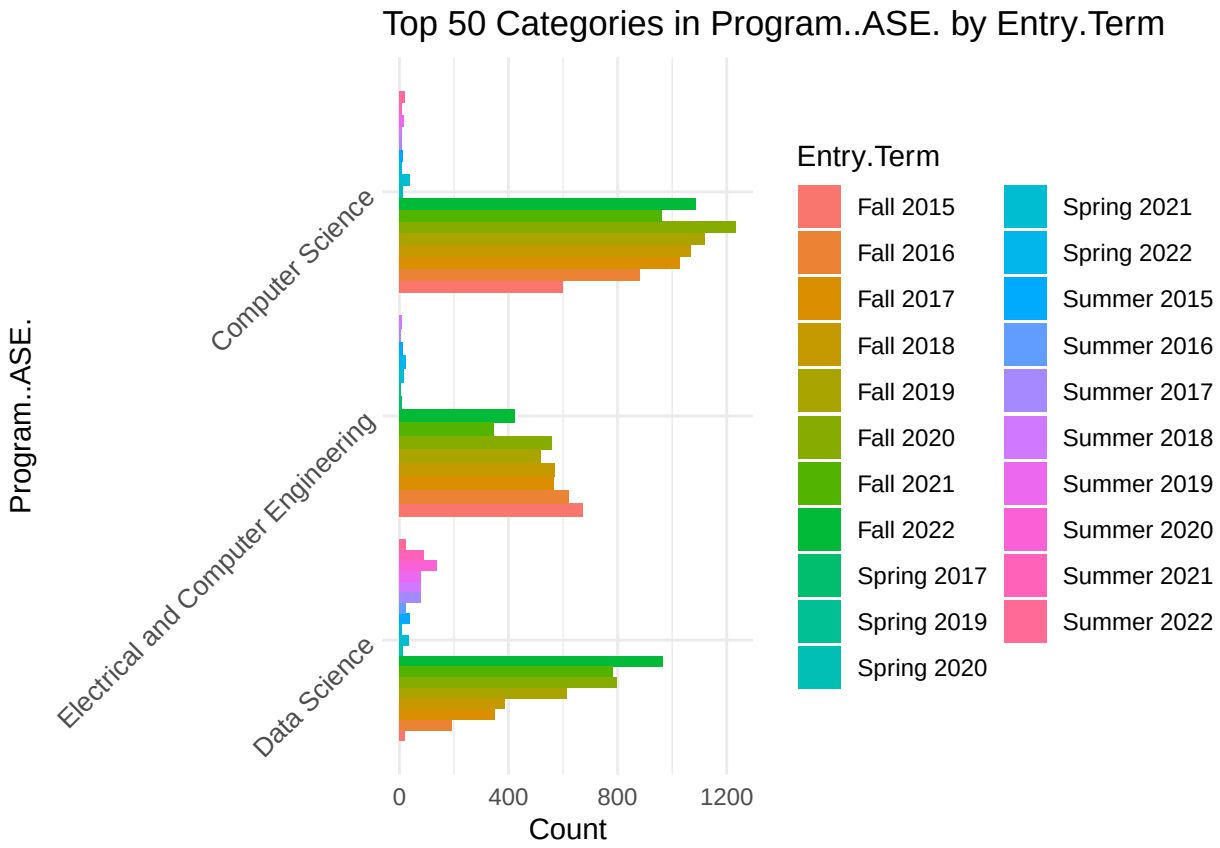
```
    top_n(top_n, wt = Count) # Select top N based on Count

  ggplot(prepared_data, aes(x = reorder(!!sym(column_name), Count), y = Count, fill = !!sym(fill_column
    geom_bar(stat = "identity", position = position_dodge(), width = 0.9) +
    theme_minimal() +
    labs(title = paste("Top", top_n, "Categories in", column_name, "by", fill_column),
         x = column_name,
         y = "Count") +
    coord_flip() +
    theme(axis.text.y = element_text(size = 10, angle = 45))
}


create_bar_plot(data, "Birth.Country", 30,  "Degree")
```



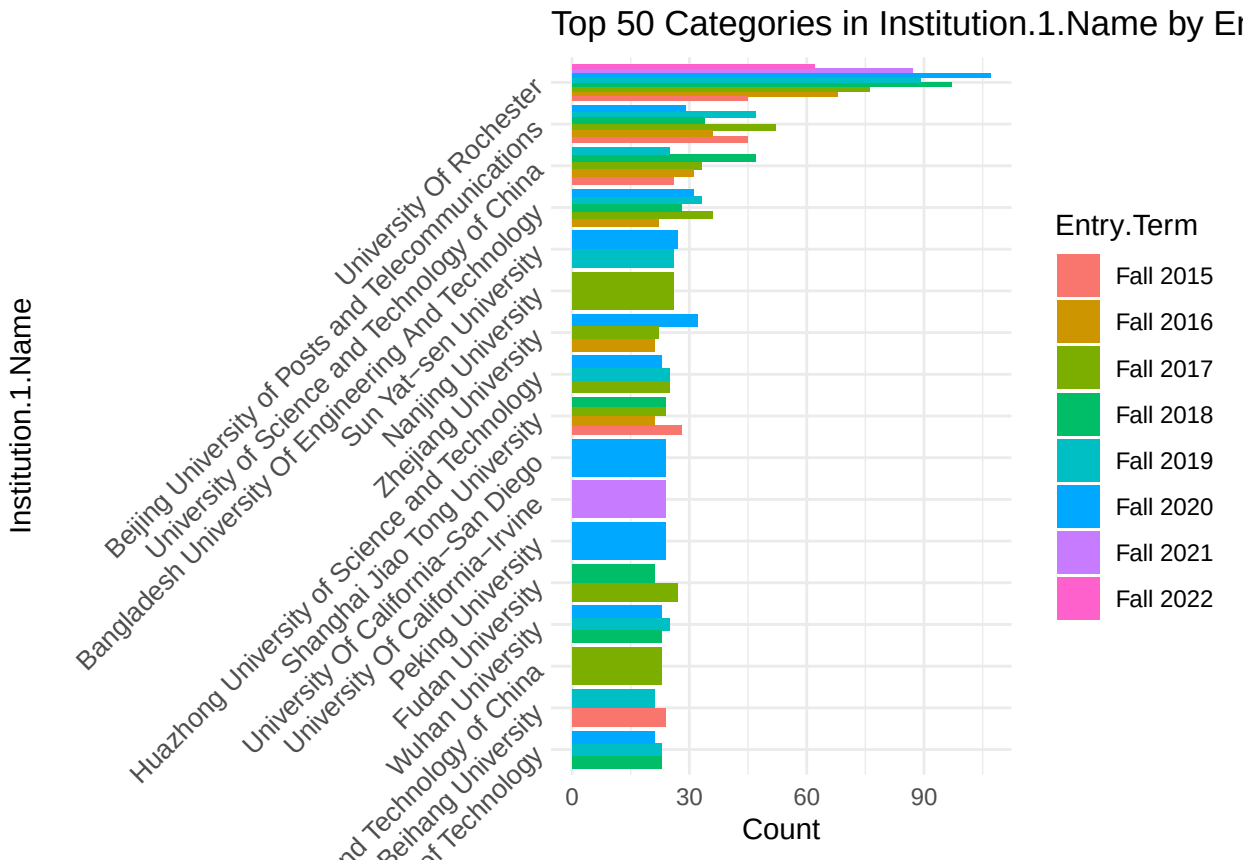

```
create_bar_plot(data, "Sex", 4,  "Degree")
```




```
create_bar_plot(data, "Program..ASE.", 50,  "Entry.Term")
```

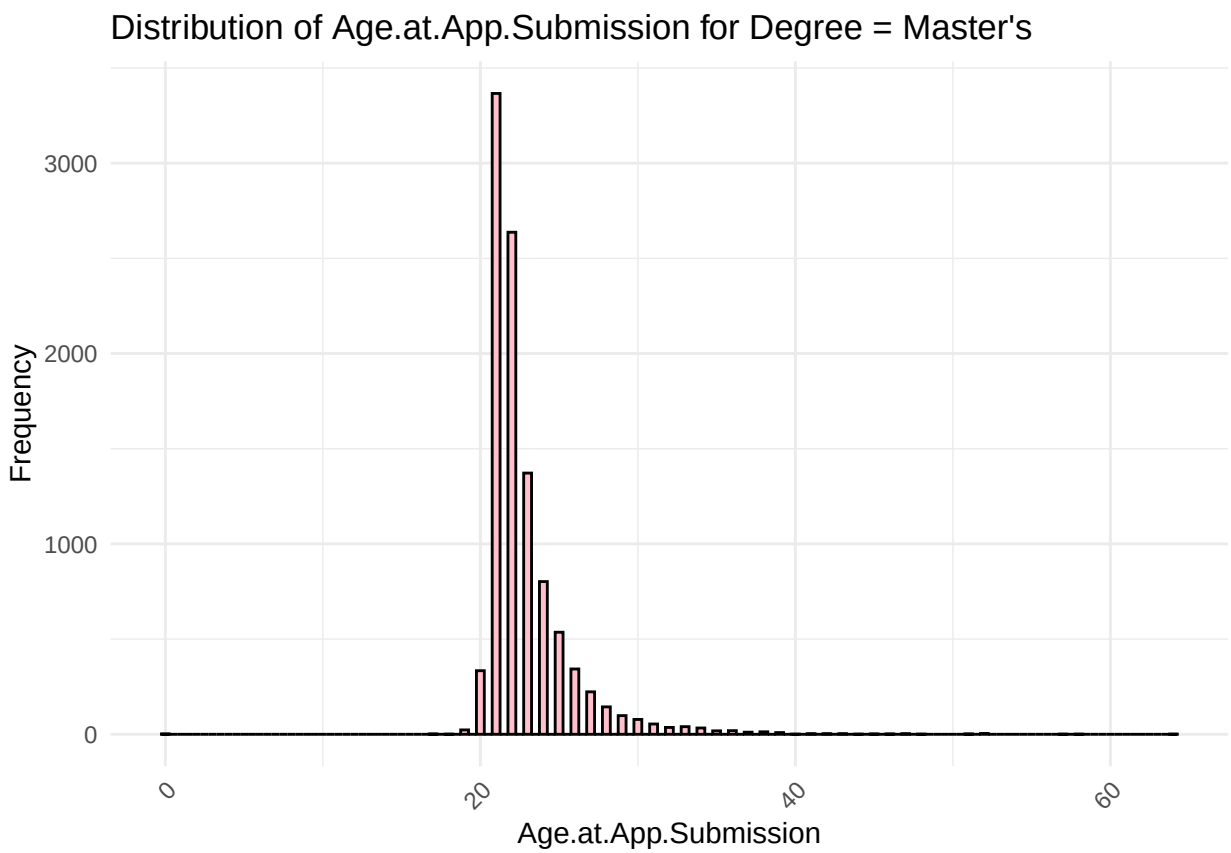



```
create_bar_plot(data, "Institution.1.Name", 50,  "Entry.Term")
```



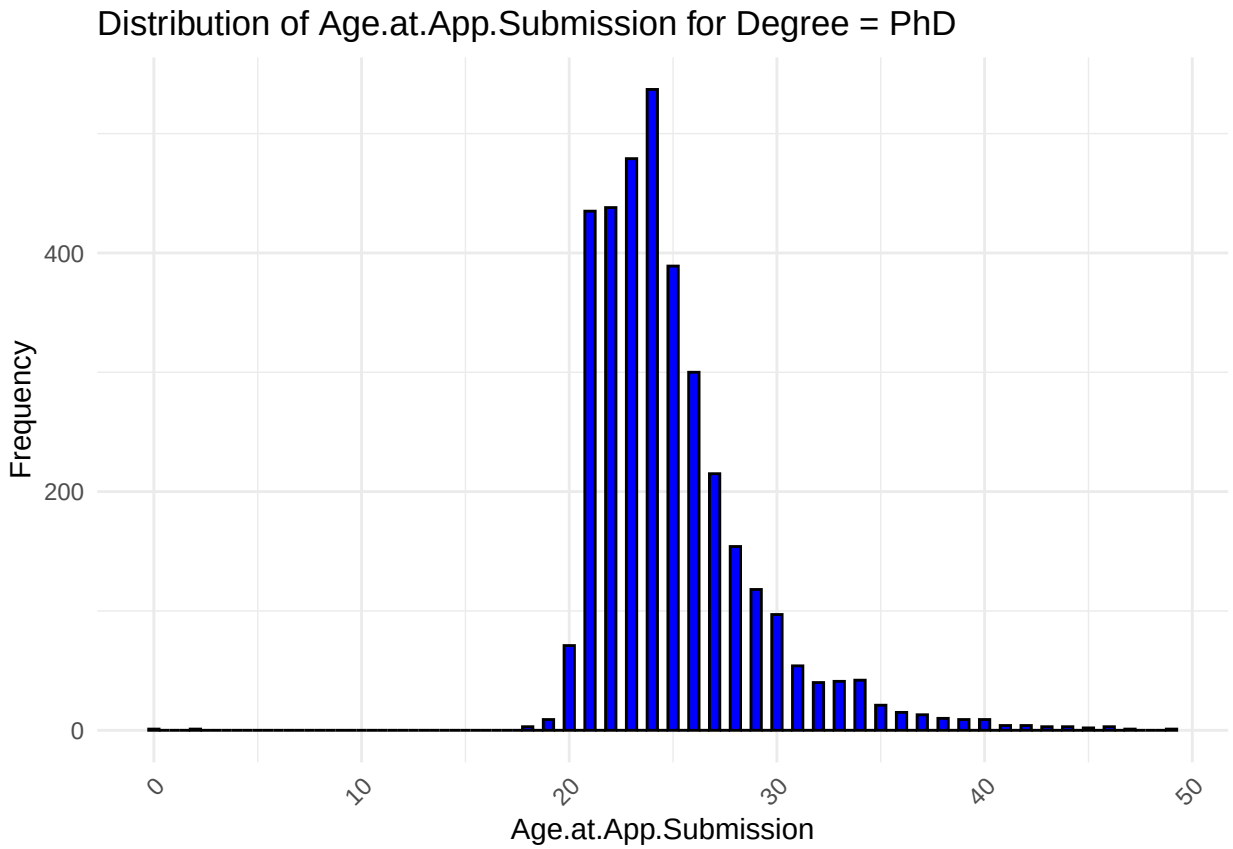

```
# Function to see distributions
create_filtered_histogram <- function(data, plot_column, filter_column, filter_value, binwidth = 0.5, fi
  # Filter the data
  filtered_data <- data %>%
    filter(!!sym(filter_column) == filter_value)

  # Create the histogram
  ggplot(filtered_data, aes(x = !!sym(plot_column))) +
    geom_histogram(binwidth = binwidth, fill = fill_color, color = "black") +
    theme_minimal() +
    labs(title = paste("Distribution of", plot_column, "for", filter_column, "=", filter_value),
         x = plot_column,
         y = "Frequency") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}


# To help respond 2nd sub question of th primary question
create_filtered_histogram(data, "Age.at.App.Submission", "Degree", "Master's", 0.5, "pink")
```

```
## Warning: Removed 2629 rows containing non-finite values (`stat_bin()`).
```



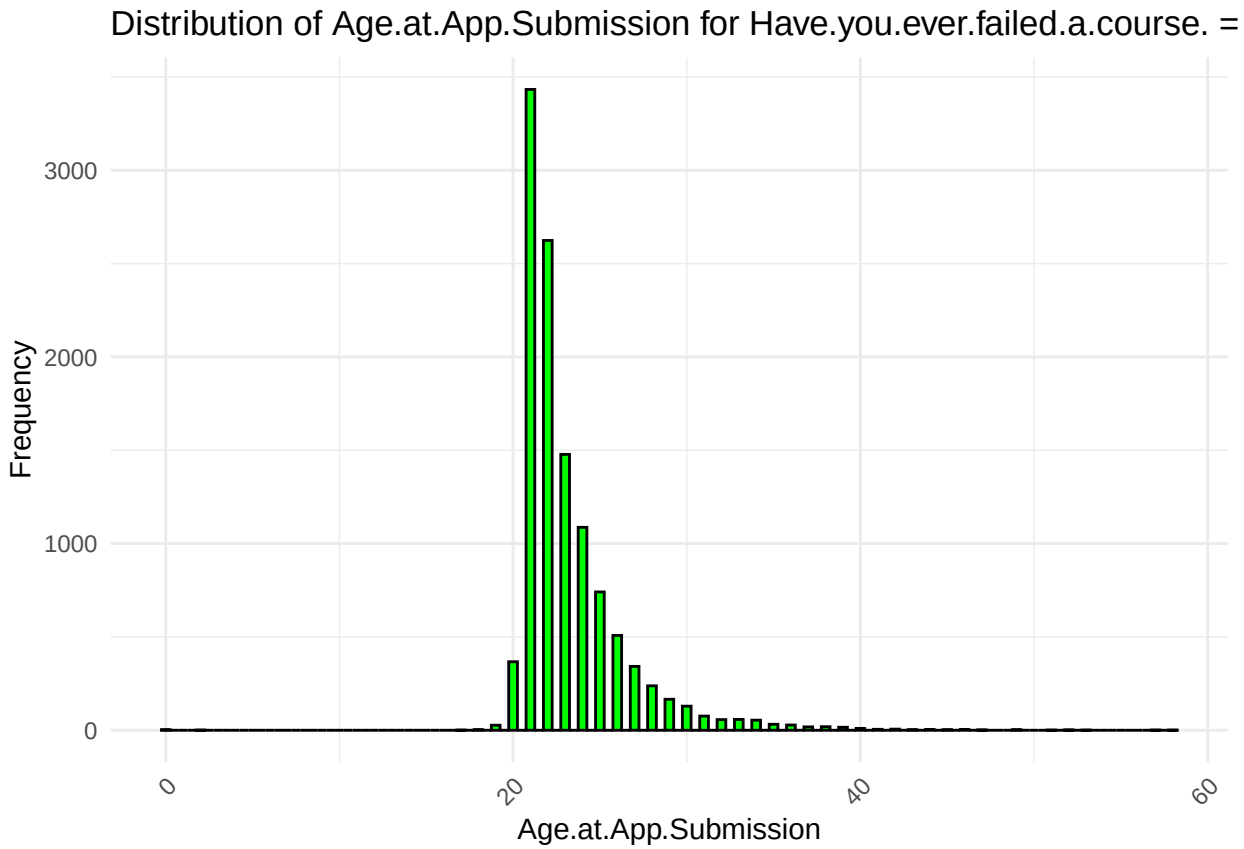

```
create_filtered_histogram(data, "Age.at.App.Submission", "Degree", "PhD", 0.5, "blue")
```

```
## Warning: Removed 714 rows containing non-finite values (`stat_bin()`).
```



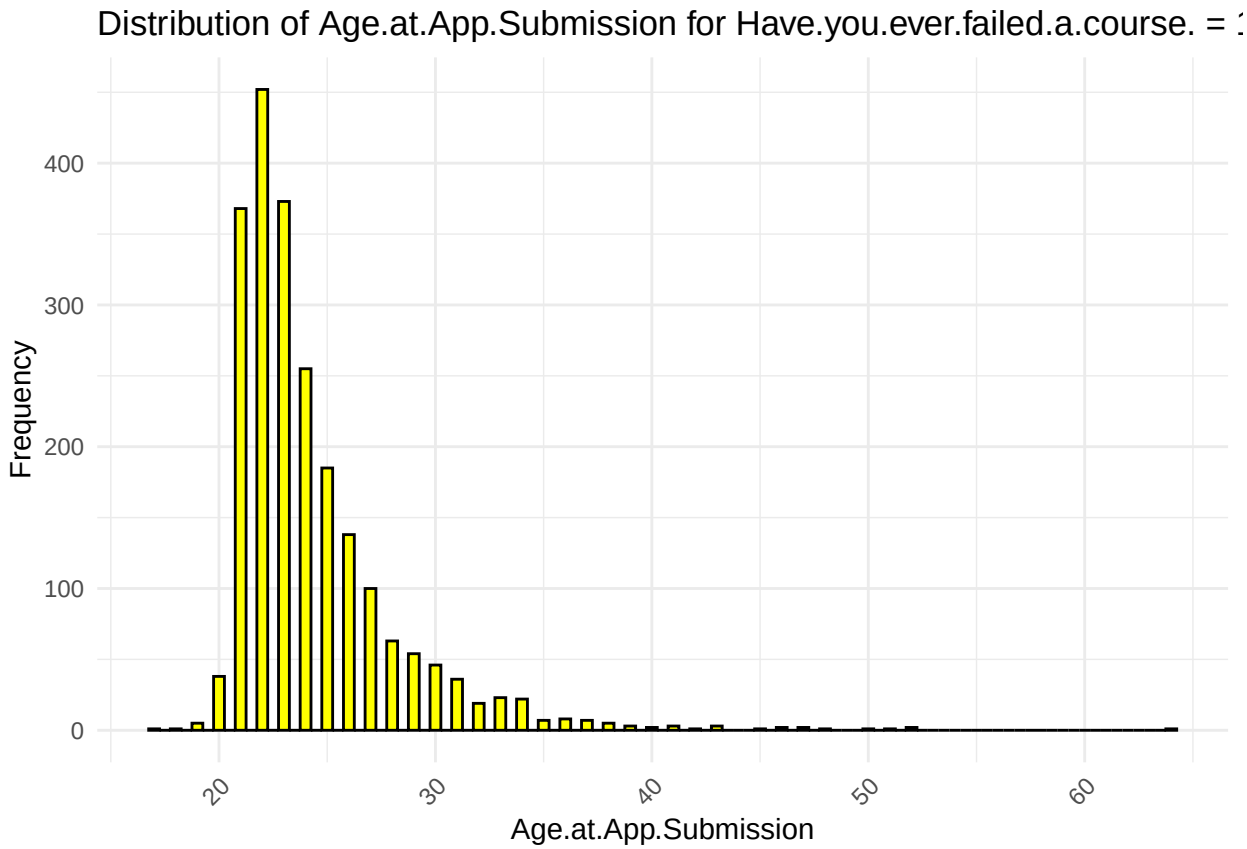

```
# To help respond 3rd sub question of th primary question
create_filtered_histogram(data, "Age.at.App.Submission", "Have.you.ever.failed.a.course.", 0, 0.5, "gre
```

```
## Warning: Removed 2125 rows containing non-finite values (`stat_bin()`).
```



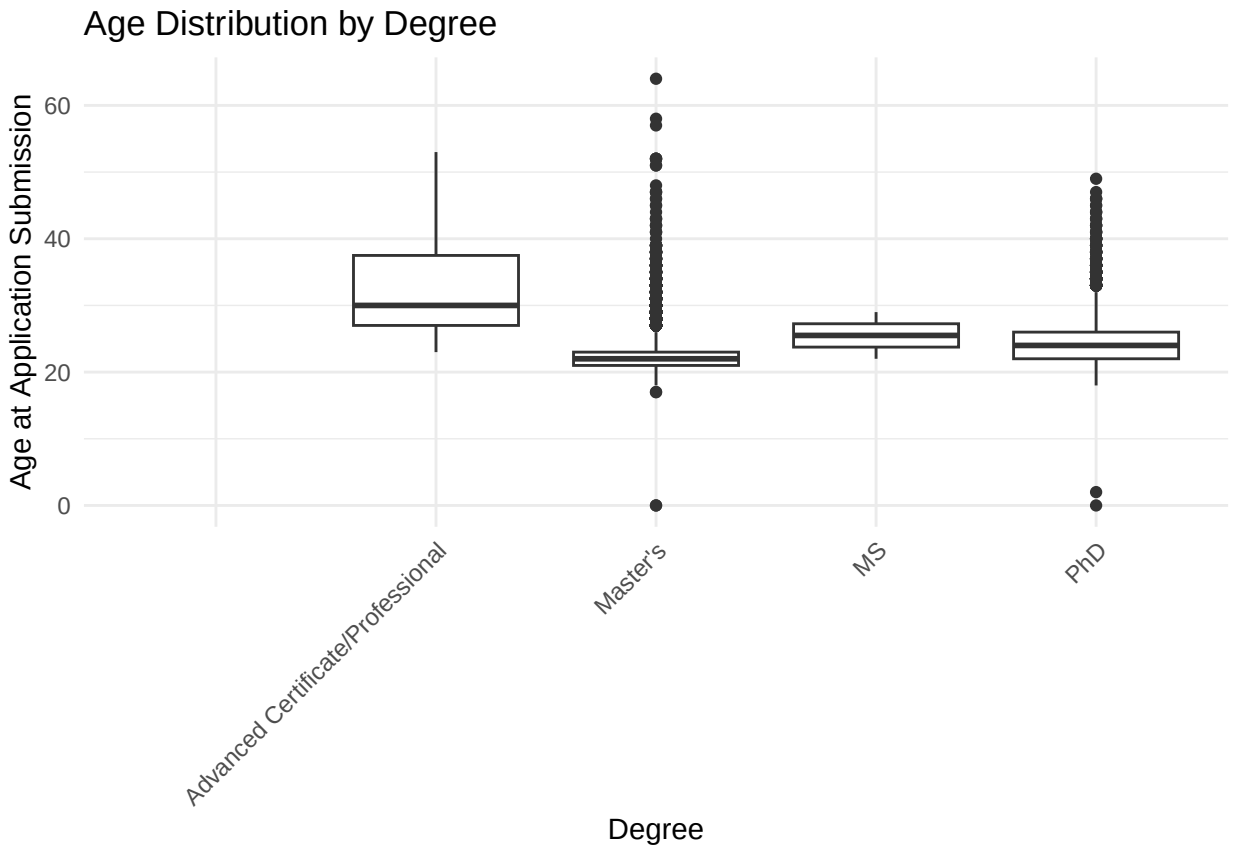

```
create_filtered_histogram(data, "Age.at.App.Submission", "Have.you.ever.failed.a.course.", 1, 0.5, "yel
```

```
## Warning: Removed 459 rows containing non-finite values (`stat_bin()`).
```



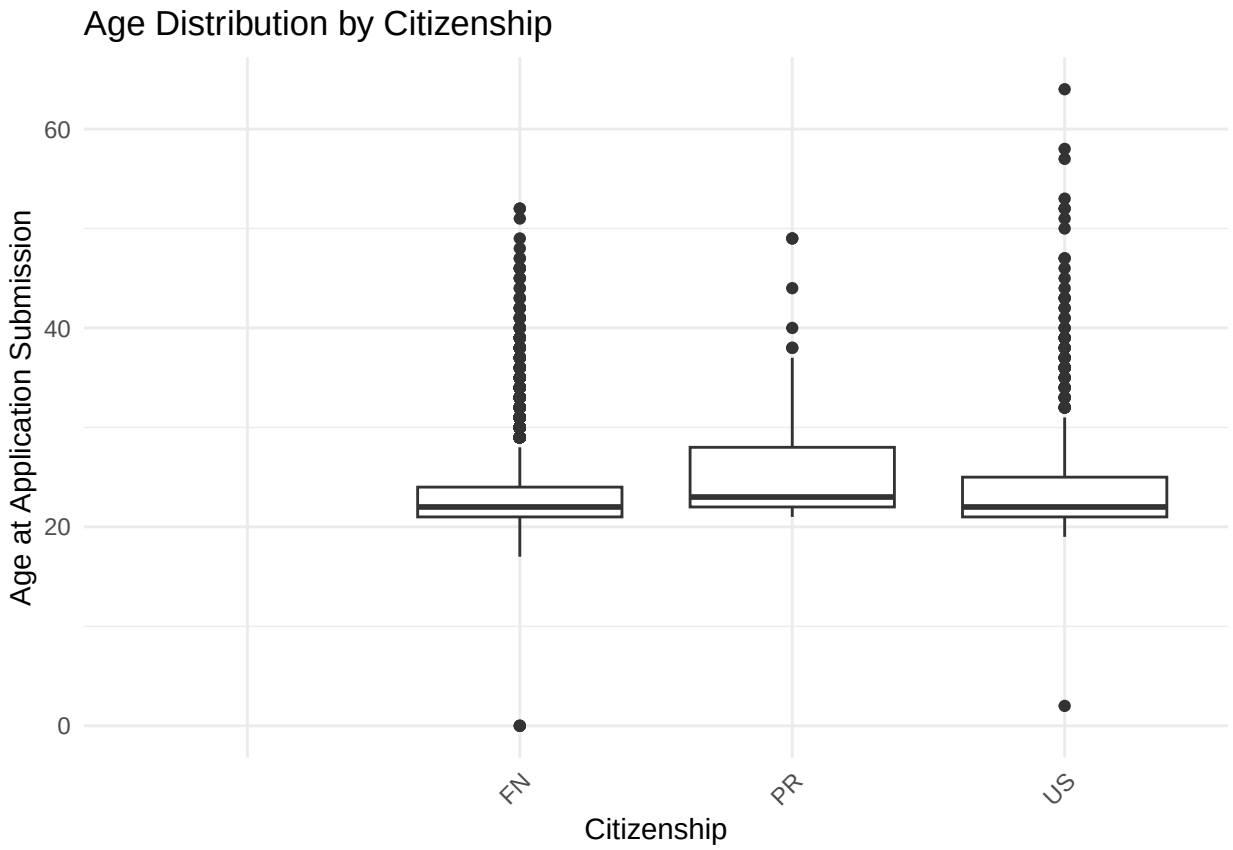

```
ggplot(data, aes(x = Degree, y = Age.at.App.Submission)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Age Distribution by Degree",
       x = "Degree",
       y = "Age at Application Submission") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 3481 rows containing non-finite values (`stat_boxplot()`).
```




```
ggplot(data, aes(x = Citizenship, y = Age.at.App.Submission)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Age Distribution by Citizenship",
       x = "Citizenship",
       y = "Age at Application Submission") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 3481 rows containing non-finite values (`stat_boxplot()`).
```


Age Distribution by Citizenship
Age at Application Submission
60
40
20
0
FN
PR
US
Citizenship