Study Of Applicants Applying for Masters and PhD programs at UofR

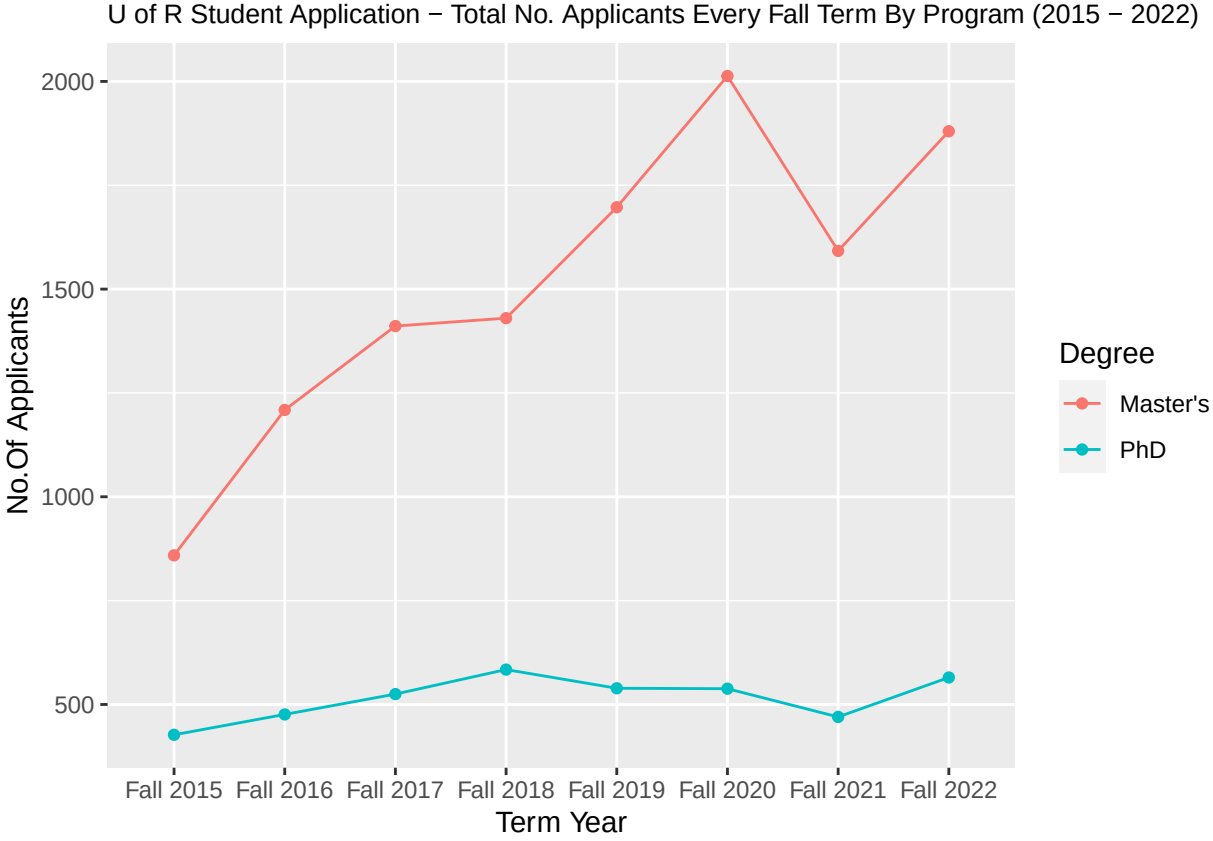
Anik De, Babli Dey, Sharon Gilbert, Veronica Mata

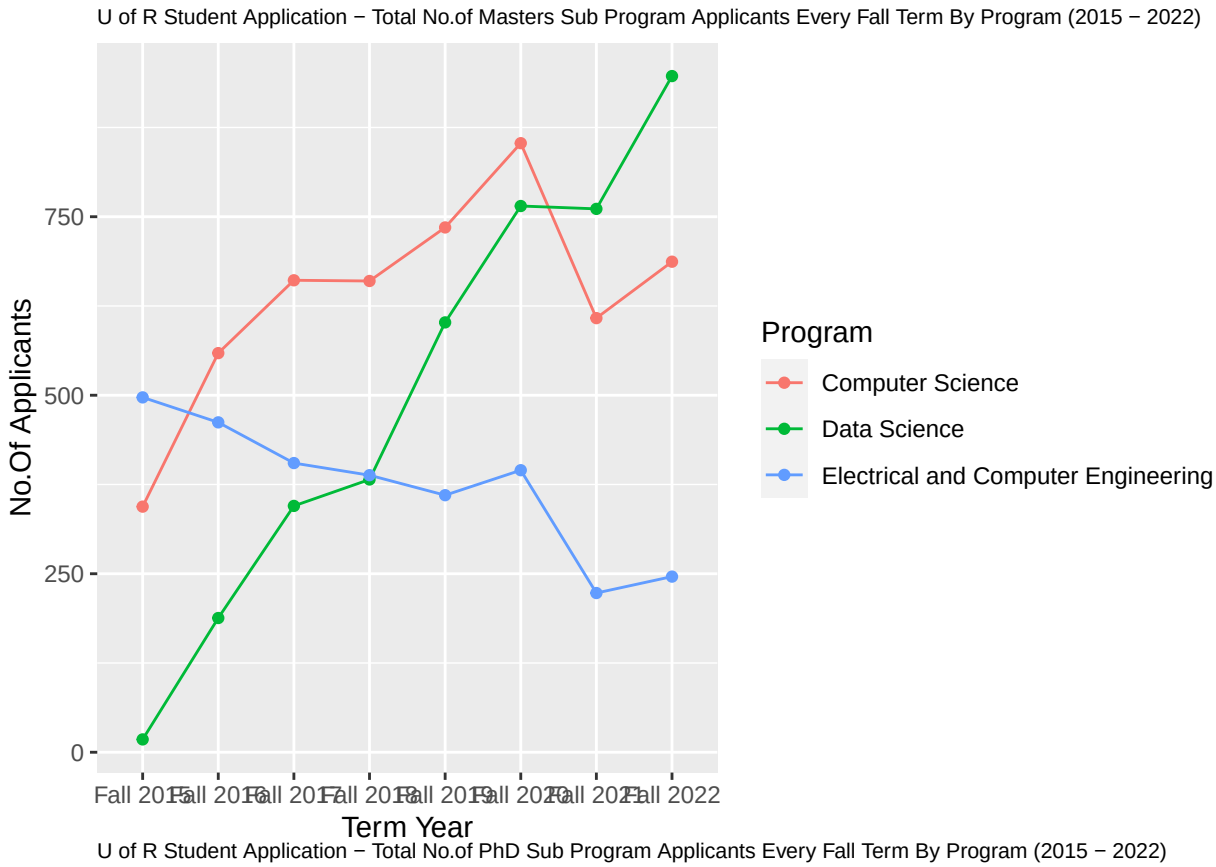
2023-12-09

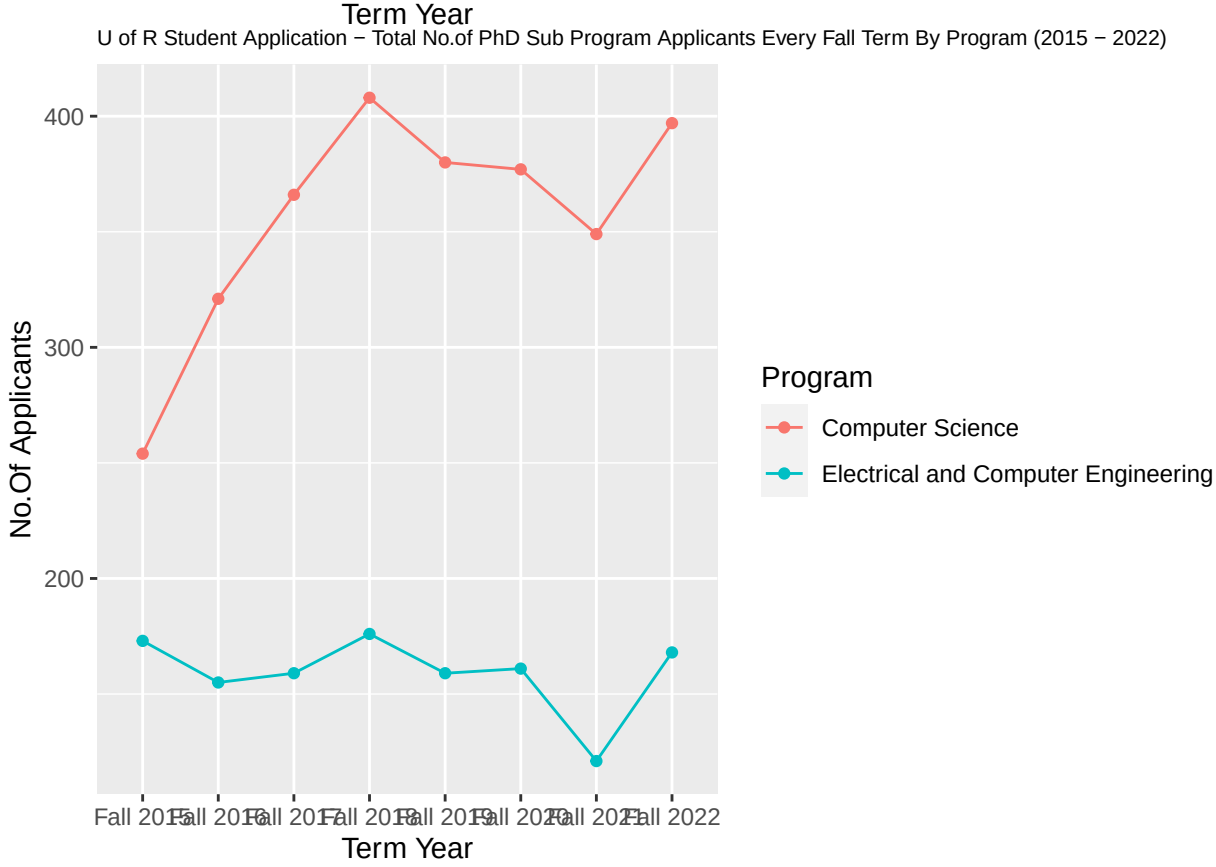
```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Visualization of the data before beginning any analysis (will include unknown data) - Masters v/s PhD:

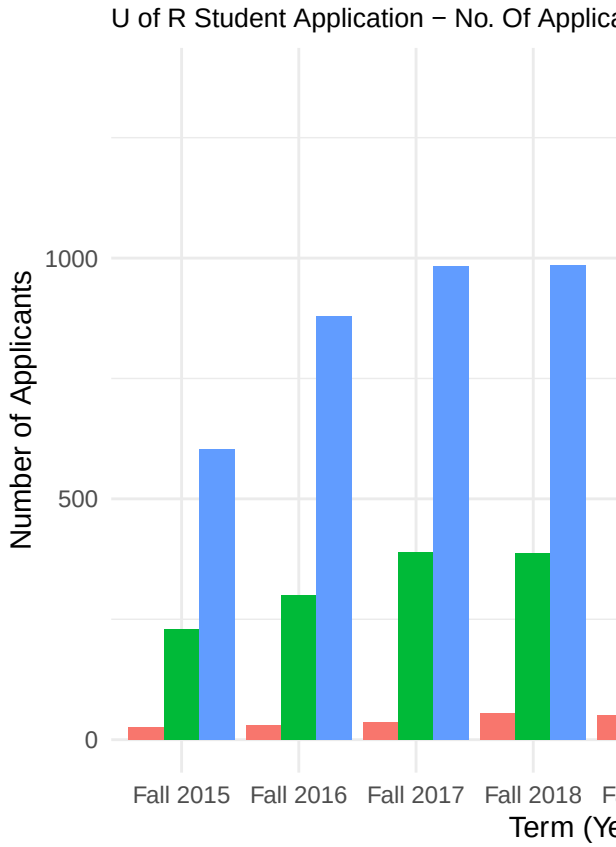
1.What is the annual total no. of applicants, which term recorded the highest and lowest number of applicants?



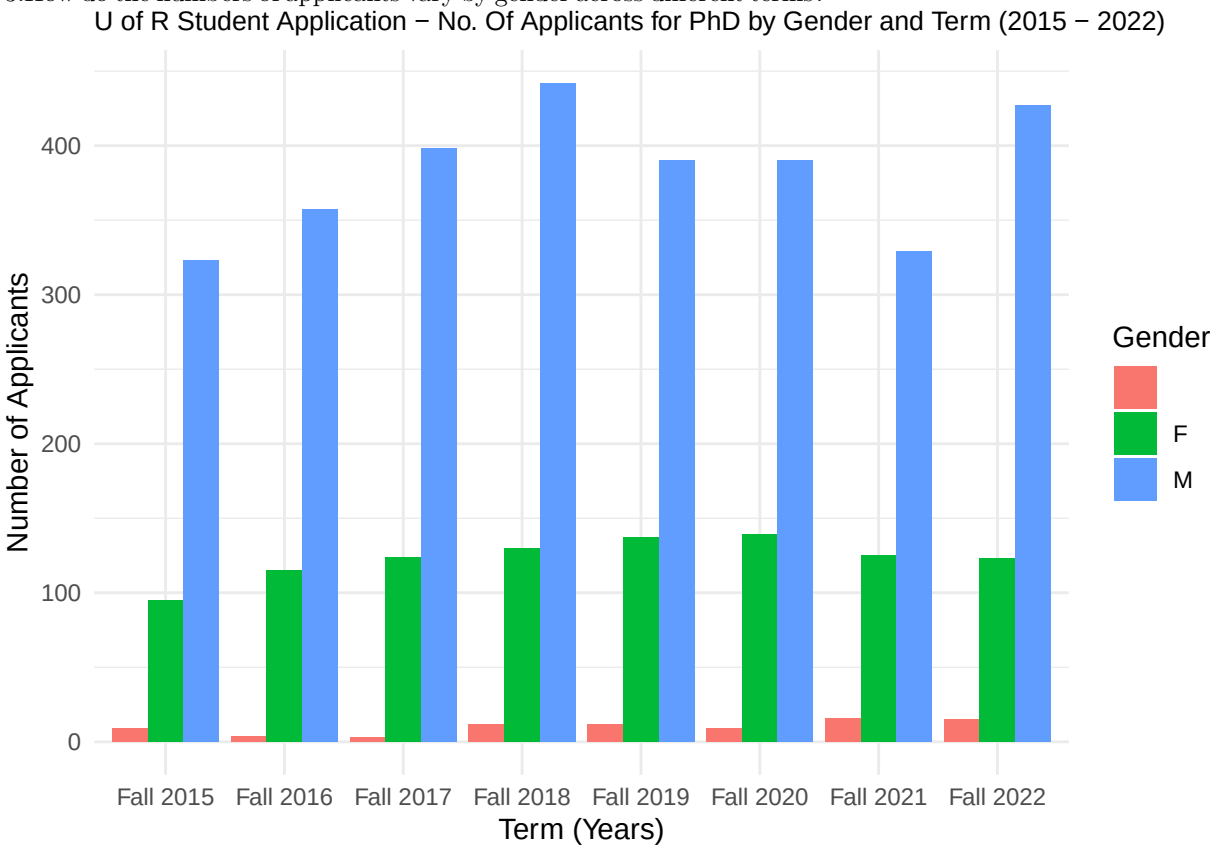


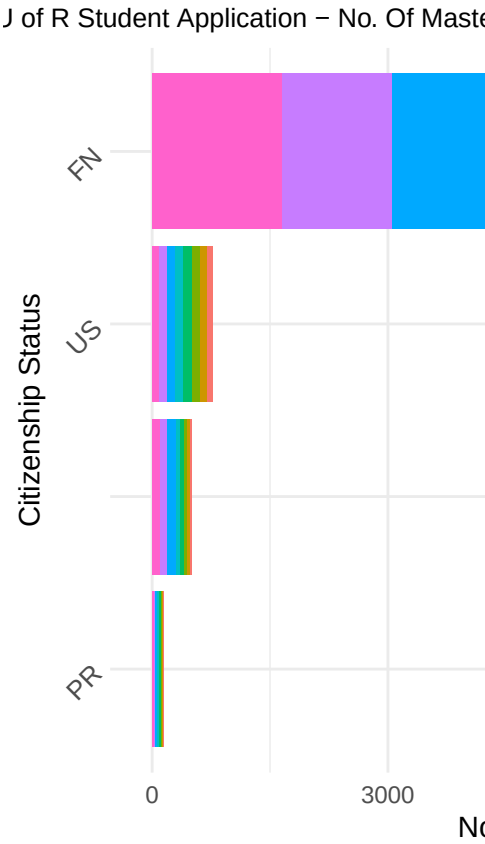


2.Identify the top 10 undergraduate institutes from which applicants are applying.

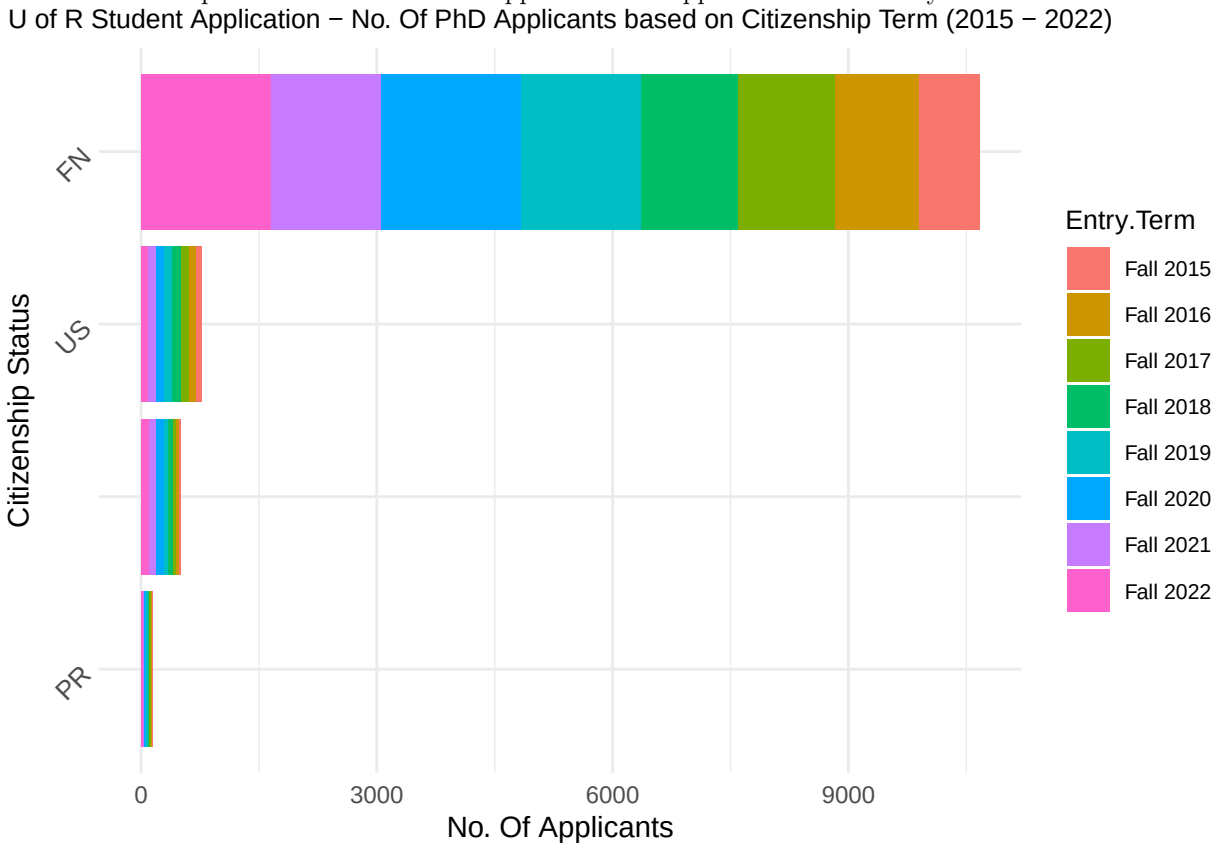


3.How do the numbers of applicants vary by gender across different terms?





4.What are the top 10 countries from which applicants have applied to the university?



5.Determine the total number of applicants who were placed on probation.

6.Determine the total number of applicants who failed a course.

Data Cleaning: Comment on the steps and also mention if you have modified the data set

1.Apply a filter to the dataset based on the 24 specified dimensions.

2.Change column names to lowercase and use underscores or simpler terms for better interpretability.

```
##  [1] "reference"              "program"                "degree"
##  [4] "entry_term"             "time_status"            "decision"
##  [7] "sex"                    "birth_country"          "age"
## [10] "native_language"        "citizenship_status"     "citizenship"
## [13] "failed_course"          "academic_probation"     "inst_1_name"
## [16] "inst_1_location"        "inst_1_study_level"     "inst_1_degree"
## [19] "inst_1_major"           "currently_employed"     "heard_about_UR"
## [22] "other_schools_applying" "Num_Two_Recs"           "Num_Three_Recs"
```

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

3.Segment the dataset into two parts: one for master's programs and another for Ph.D. programs.

4.For empty values, categorize them as "Unknown" for categorical data. For age, fill in with the average age corresponding to the respective program.

Descriptive Analysis: Master's and PhD

1. Calculate the mean, variance, and standard deviation of age, and visualize the distribution using histograms and box plots.
2. Examine the acceptance rate based on gender across different terms.
3. Explore the acceptance rate by country in relation to different terms.
4. Investigate the acceptance rate for applicants who were placed on probation.
5. Investigate the acceptance rate for applicants who failed a course.
6. Investigate the age distribution between applicants who failed a course v/s applicants who did not fail a course.

Inferential Analysis:

1. Is there a notable distinction in age variance between applicants who have experienced course failure and those who haven't, considering each degree type?
2. Is there a discernible difference in the average age between master's and Ph.D. students?
3. Compare the acceptance rate for students who learned about the university through family/friends versus those who found it through an internet search.

Logistic Regression:

1. Can we use factors such as age, citizenship, gender, recommender relationships, current employment status, and program choice to predict the likelihood of a candidate being accepted?