ERC Walk-In Lab Report

Carolyn Silverman

Contents

1 Introduction

The following report is based on Empirical Reasoning Center walk-in data from Fall 2016, Spring 2017.

1.1 Background

What is Empirical Reasoning?

Empirical reasoning is the process of thinking critically about organizing, analyzing, and visualizing qualitative, quantitative, and/or geospatial data.

What is the Empirical Reasoning Center?

The ERC, located in Barnard College's library, provides assistance to students, teachers, staff, and alumni of Barnard College and the larger Columbia University community in three main areas:

1. **Training and Technical Assistance**: The ERC offers training for statistical analysis, textual analysis, and geographical information systems software. We support both Macs and PCs.
2. **Individual Guidance**: The ERC can help individuals through each step of the research process from basic research design and formulating a hypothesis to data analysis and visualization to interpreting and presenting results.
3. **Classroom Support**: The ERC supports courses with supplementary training sessions focusing on using analysis softwares, finding appropriate datasets, and interpreting and understanding the narrative of quantative and qualitative data.

The ERC walk-in center is staffed by undergraduate fellows and graduate assistants who work one-on-one with students to help them with coursework and research projects grounded in empirical reaosning. All visitors to the ERC are required to fill out a paper sign-in sheet if they receive help from a fellow or use our resources. A copy of the sign-in sheet can be found **here**. ERC fellows later copy the information from the sign-in sheets verbatim into a Qualtrics survey. At any point, we can obtain a CSV that contains the results of the Qualtrics survey to date.

1.2 Summary of Methods

This report analyzes the data produced by the Qualtrics survey. The long-term goal of the project is described below:

With only a small amount of additional data cleaning, the CSV for a new semester of ERC walk-in data can be read into R and a comprehensive report will automatically be produced via code written in R markdown. The report begins with simple descriptive statistics about single variables of interest.

> Who are the students visiting the ERC: class year? major category? school?
>
> Who are they coming to see: undergraduate fellows? graduate students?
>
> In which software packages and programming languages are they seeking support: R? GIS? Excel? MATLAB? Stata?
>
> What types of problems are they facing: data analysis? visualization? finding data?
>
> Which courses are they taking: intro level? project-based?
>
> When are they most likely to visit the ERC: week? day? time?

This information is helpful for training new fellows, providing workshops for students, improving outreach methods, and scheduling fellows during peak hours. It will also help track our progress over time and provide insight into how we can accommodate the changing needs and interests of our students.

We further break down some of the summary statistics by school, namely Barnard College versus the other Columbia University undergraduate schools (CC/SEAS/GS). We are particulary interested in the courses that our Columbia visitors take and the programs they use.

The report also provides information about the length of each visit and determines the categorical variables that have a statistically significant relationship with length of stay. Finally, we assign a student ID (sid) to each unique visitor in order to track repeat visitors over time. We determine the variables associated with the students who visit the ERC multiple times via chi-squared tests. Detailed methods are outlined at the beginning of each relevant section.

The code currently runs on 2 or more semesters of data (and we will soon generalize it to run on only one input data set).

Cleaning Scripts

Each CSV is first pre-processed and cleaned via a script unique to the relevant semester of data. Unique cleaning scripts are used because the fellows, courses we support, and sign-in sheet questions and answers change from year to year.

For data sets from semesters after Spring 2017, the clean_s17.R script can be used as a template. The user will have to adjust the sections that clean the open-ended questions of the Qualtrics survey (hear about and course), as well as program (if the ERC begins to support new programs), semester, and fellow. And of course, if changes are made to the Qualtrics survey, these changes must be reflected in the cleaning scripts.

1.3 Summary of Findings for Fall 2016, Spring 2017.

A crude summary of this report's findings is outlined below:

-Total number of visits to the ERC in Fall 2016 was 473 and in Spring 2017 was 359

-Significantly more upperclassmen than underclassmen visitors

-By far the largest number of social science students compared to other major categories

-Most popular softwares/languages are Excel and R, followed by GIS, Stata, and MATLAB

-Saw a drop in the number of EXCEL and R users and an increase in the number of GIS and STATA users from Fall 2016 to Spring 2017

-3 most common problems visitors face are method, data visualization/charts, and data analysis

-Most visitors hear about the ERC through class or a professor, but we did see an increase in the number of students who heard about us though word of mouth and other students

-More students came to see our undergraduate fellows than our graduate fellows, even normalizing for the number of each type of fellow

-Students from a handful of courses dominate our walk-in hours, the most notable being General Chemistry Lab, Programming for the Behavioral Sciences, and Social Research Methods

-The ERC generally receives the most visitors in the morning and early afternoon, and the number of visitors decreases as the week progresses

-On a broad scale, the ERC receives the most visitors a few weeks into the semester and a few weeks before the semester ends

-Average length of stay increased from 43 minutes in Fall 2016 to 61 minutes in Spring 2017

-Based on ANOVA results, the mean length of stay had statistically significant differences when broken down by major category, program, school, and class year

    -Notable differences of mean length of stay were between STEM and Social Science students, Barnard and CC/SEAS/GS students, and first years and students of all other class years

-Columbia students most commonly seek assistance with R and GIS and the majority of them come from the Social Research Methods course

-Total number of unique visitors in Fall 2016 was 262 and in Spring 2017 was 152

-Software program/language, school, and whether the student attended an ERC workshop all had a statistically significant relationship with the manyVisits dummy variable (which was 1 if the students came to the ERC multiple times and 0 otherwise)

-The odds of a student visiting the ERC many times were

    -2.04 times higher for STEM majors than humanities majors

    -3.26 times higher for Banard students than Columbia students

    -1.52 times higher for workshop attendees

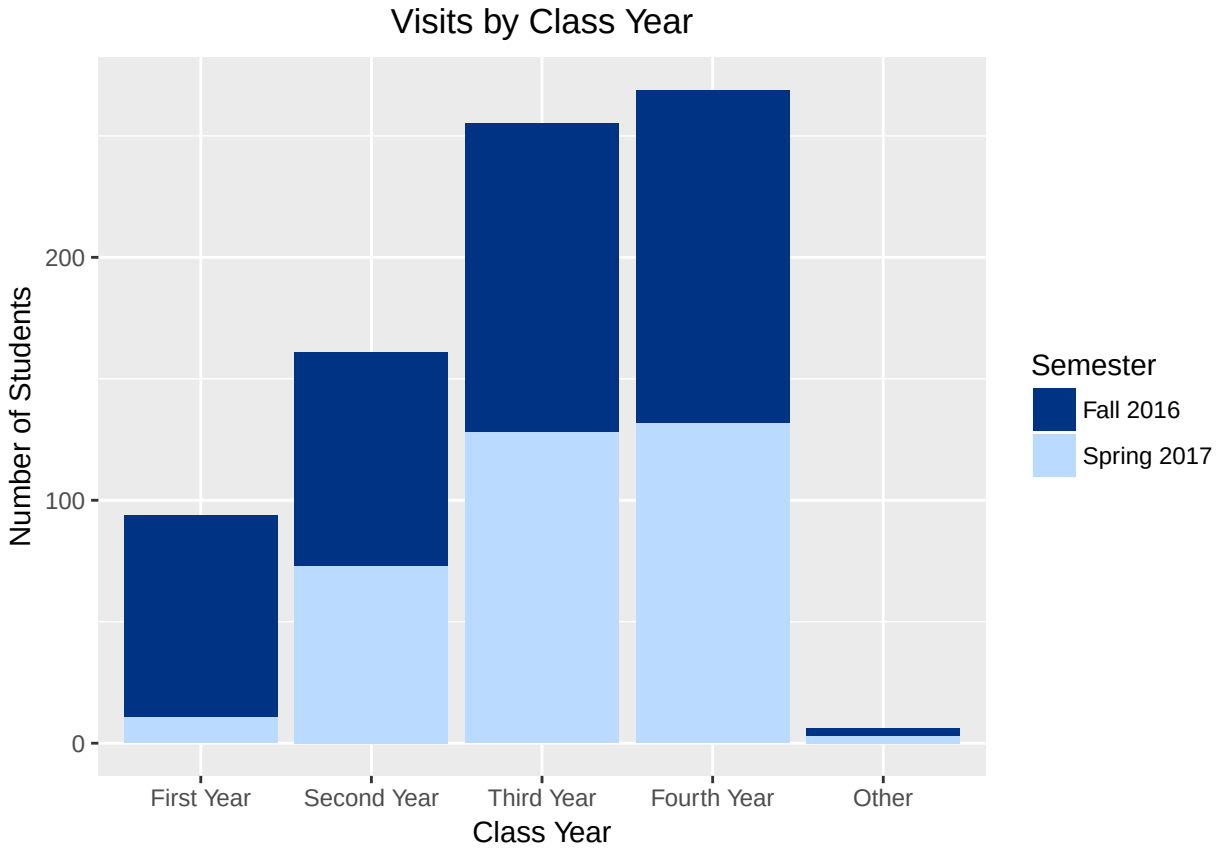
2 Descriptive Statistics For Single Variables

We start by providing descriptive statistics and plots for single variables. We find the number of visitors that fall into each level of a given categorical variable, as well as the proportion of all visitors that this number represents. Each plot is broken down by semester, while each table contains information for the aggregate data set. Tables with counts and proportions for individual semesters can be easily produced from the R code. Although these results are not explicitly presented in the report, many of the descriptions of findings will cite these statistics. NA's have been omitted in the analysis of all single variables.

The total number of visits to the ERC in Fall 2016 was 473 and in Spring 2017 was 359. Details about unique visitors are analyzed in a later section. For context, the Barnard College student body consists of about 2,500 students.

2.1 Class Year

Class year generally applies to students enrolled in one of the undergraduate colleges of Columbia University (BC, CC, SEAS, GS). If the visitor is a graduate student, professor, alumnus, etc, she will fall into the "Other" category.



Findings:

In both Fall 2016 and Spring 2017, we received substantially more upperclassmen visitors than underclassmen. Precisely 2.055 more third and fourth years came into the ERC than first and second years during the full academic year. We also saw a substanital drop in the number of first years from Fall 2016 to Spring 2017. This decline is largely due to the number of General Chemistry students we assist in the fall semester every year. Otherwise, the trends appear similar across the two semesters. The number of visitors in each category increases with class year, with first years accounting for only 12.0% of visitors and fourth years accounting for

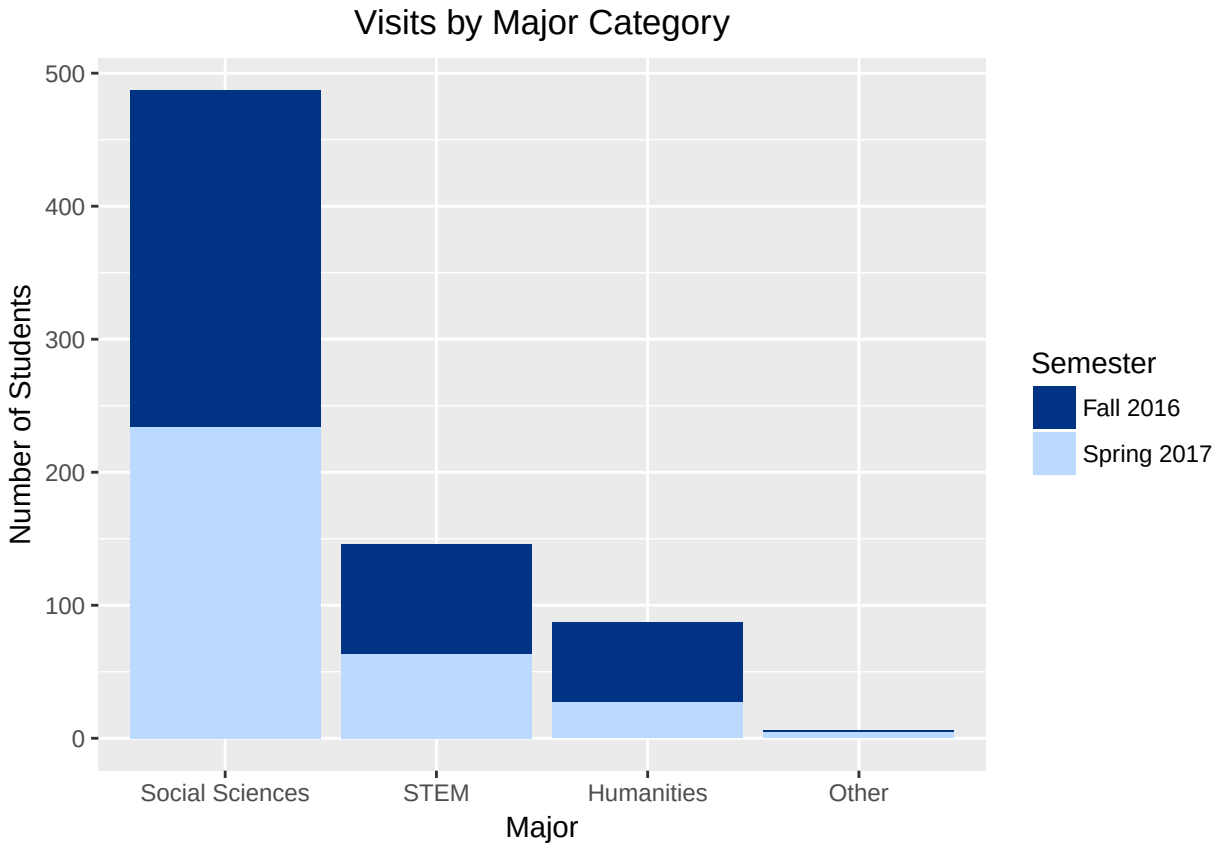
Class Year	Visitors	Proportion
First Year	94	0.120
Second Year	161	0.205
Third Year	255	0.325
Fourth Year	269	0.343
Other	6	0.008

Table 1: Visits by Class Year (Aggregate)

34.3% of visitors in the aggregate data set. The small number of visitors in the "Other" category indicates that nearly all of our walk-in visitors are undergraduate students, though it should be noted that professors and alumni may omit this category when they fill out the survey.

2.2 Major Category

Rather than obtaining counts for specific majors, it is more useful to group our visitors into three categories based on major: Social Sciences, STEM, and Humanities. Visitors who are double majors or who specify a minor in the survey are counted twice in the analysis. For a comprehensive list of majors that fall into each category, see the appendix (will add appendix).



Findings:

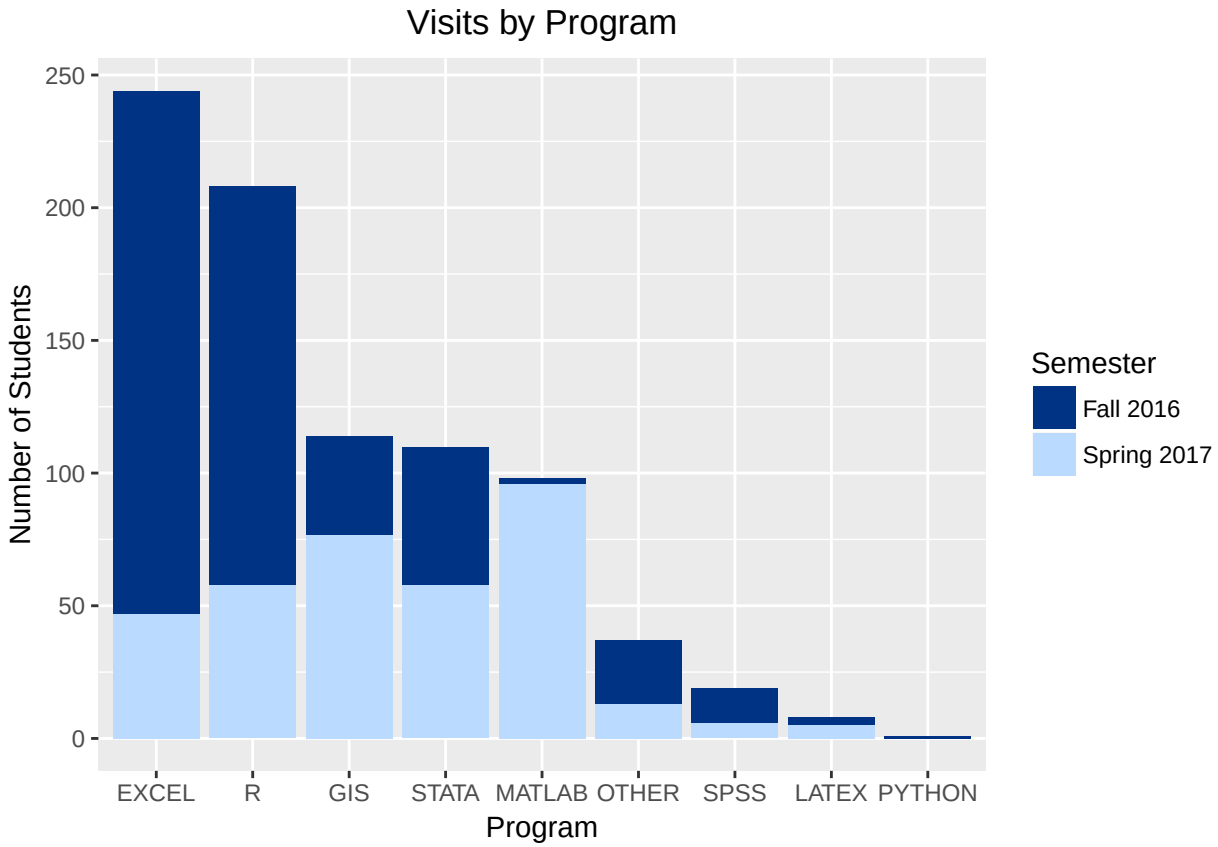
Of the three major categories, we receive by far the most students studying the social sciences. Over both semesters, these students made up 67.1% of our visitors. We saw a slight drop in the number of humanities students from 60 (~15%) in Fall 2016 to 27 (~8%) in Spring 2017. The proportions of social science and STEM students remained similar.

Major	Visitors	Proportion
Social Sciences	487	0.671
STEM	146	0.201
Humanities	87	0.120
Other	6	0.008

Table 2: Visits by Major Category (Aggregate)

2.3 Program

Since one of the main purposes of the Empirical Reasoning Center is to provide support for statistical software programs, Program is a main variable of interest. Students seeking help with multiple programs are counted multiple times. The GIS category consists of all GIS-related softwares including arcGIS and QGIS.



Findings:

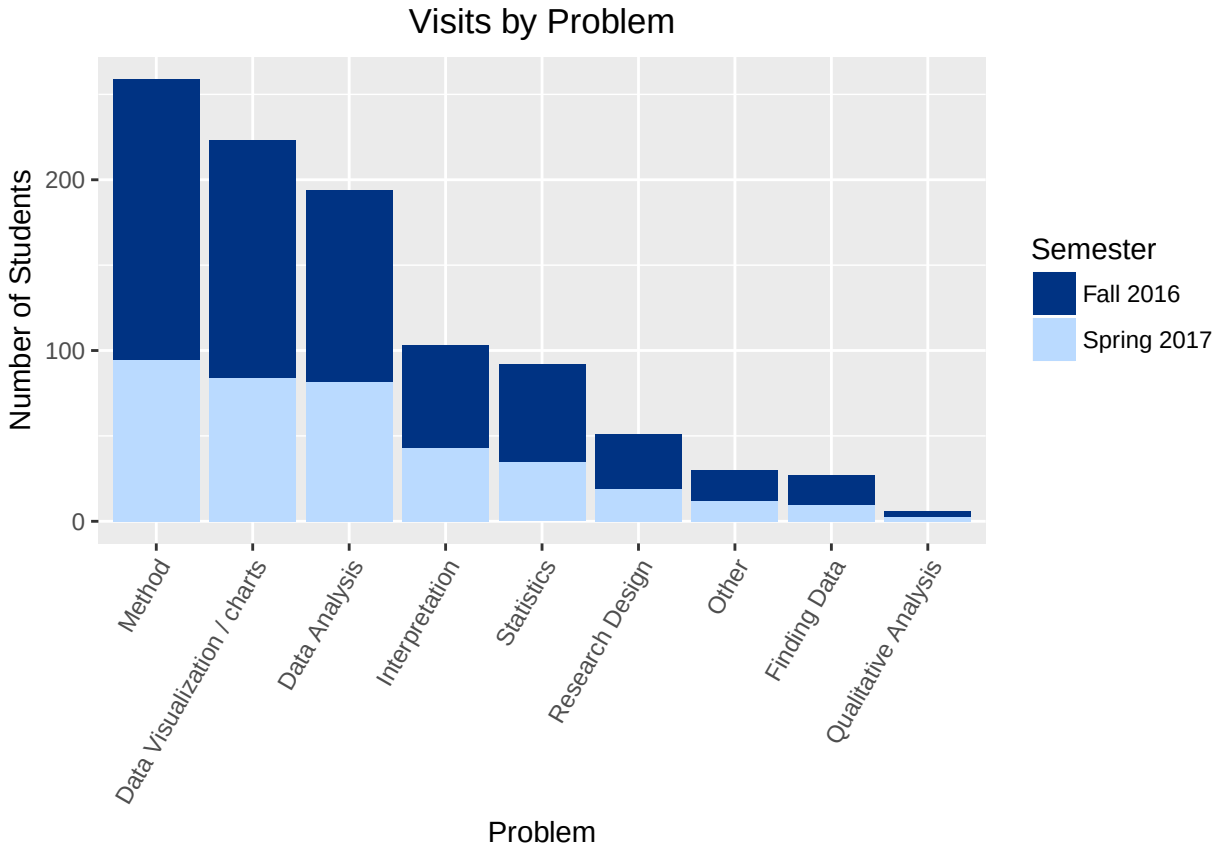
From Fall 2016 to Spring 2017, there was a big drop in the number of EXCEL and R users and an increase in the number of GIS and STATA users. These changes are likely due to the classes we supported each semester, specifically General Chemistry (EXCEL) in the fall, and Programming for the Beahvorial Sciences (MATLAB) in the spring. In Fall 2016, we made a big push to switch our supported classes from STATA to R, so it is somewhat concerning that the number of R users in Spring 2017 fell so drastically and the number of STATA users increased.

Program	Visitors	Proportion
EXCEL	244	0.291
R	208	0.248
GIS	114	0.136
STATA	110	0.131
MATLAB	98	0.117
OTHER	37	0.044
SPSS	19	0.023
LATEX	8	0.010
PYTHON	1	0.001

Table 3: Visits by Program (Aggregate)

2.4 Problem

"Nature of the question" is a multiple choice question in our survery (with multiple selections allowed). Fellows, not visitors, fill out this question. Again, a visitor with multiple problems is counted multiple times.



Findings:

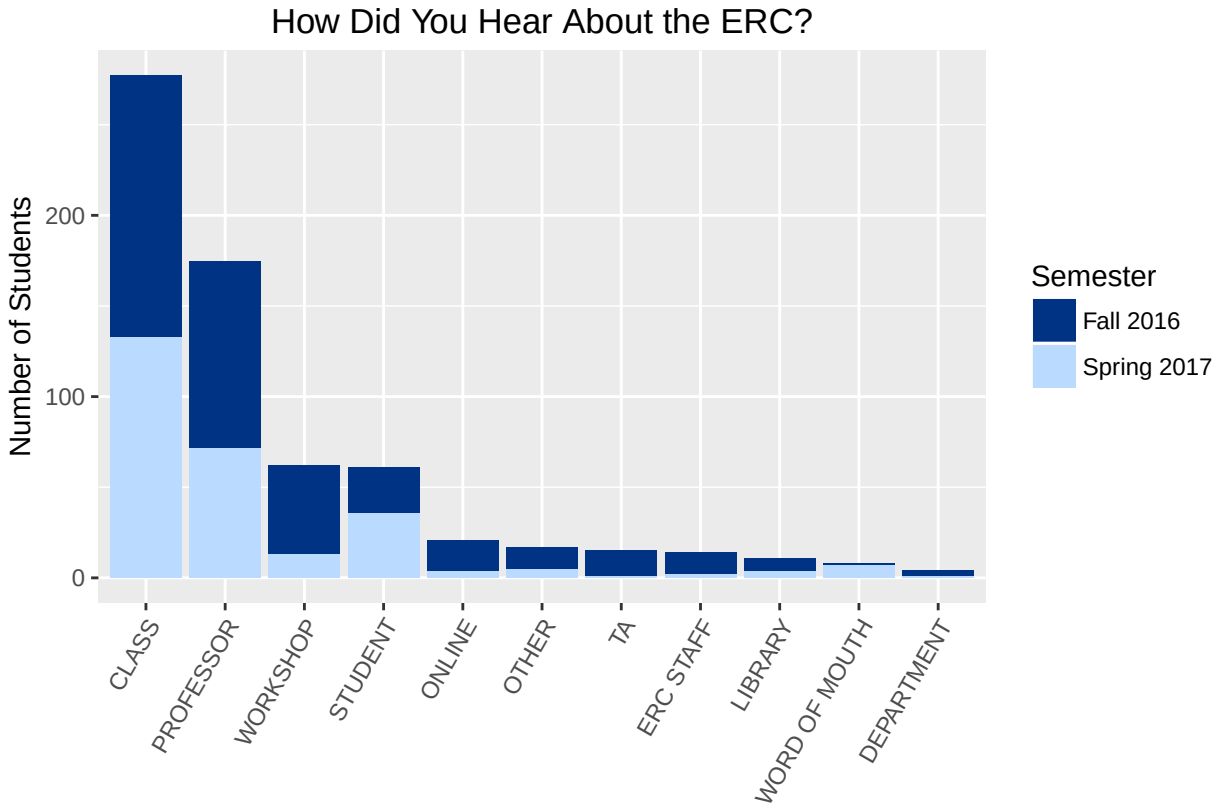
We see very similar trends across semesters. The three most prominent categories are Method, Data Visualization/Charts, and Data Analysis, making up 68.6% of the observations. The results generally agree with our mission: we aim to offer more assistance with software packages and the data analysis process and less with theoretical statistics and interpretation. Professors often want students to work in the latter areas without help. In the future, we may want to find ways to get our visitors more interested in their own research projects so that we see improvements in the Research Design and Finding Data categories.

Problem	Visitors	Proportion
Method	259	0.263
Data Visualization / charts	223	0.226
Data Analysis	194	0.197
Interpretation	103	0.105
Statistics	92	0.093
Research Design	51	0.052
Other	30	0.030
Finding Data	27	0.027
Qualitative Analysis	6	0.006

Table 4: Visits by Problem (Aggregate)

2.5 How did you Hear About the ERC?

This is currently a free response question on our survey, so the responses had to be heavily cleaned in order to categorize them.



Source	Number of Students	Proportion
CLASS	277	0.333
PROFESSOR	175	0.210
WORKSHOP	62	0.075
STUDENT	61	0.073
ONLINE	21	0.025
OTHER	17	0.020
TA	15	0.018
ERC STAFF	14	0.017
LIBRARY	11	0.013
WORD OF MOUTH	8	0.010
DEPARTMENT	4	0.005

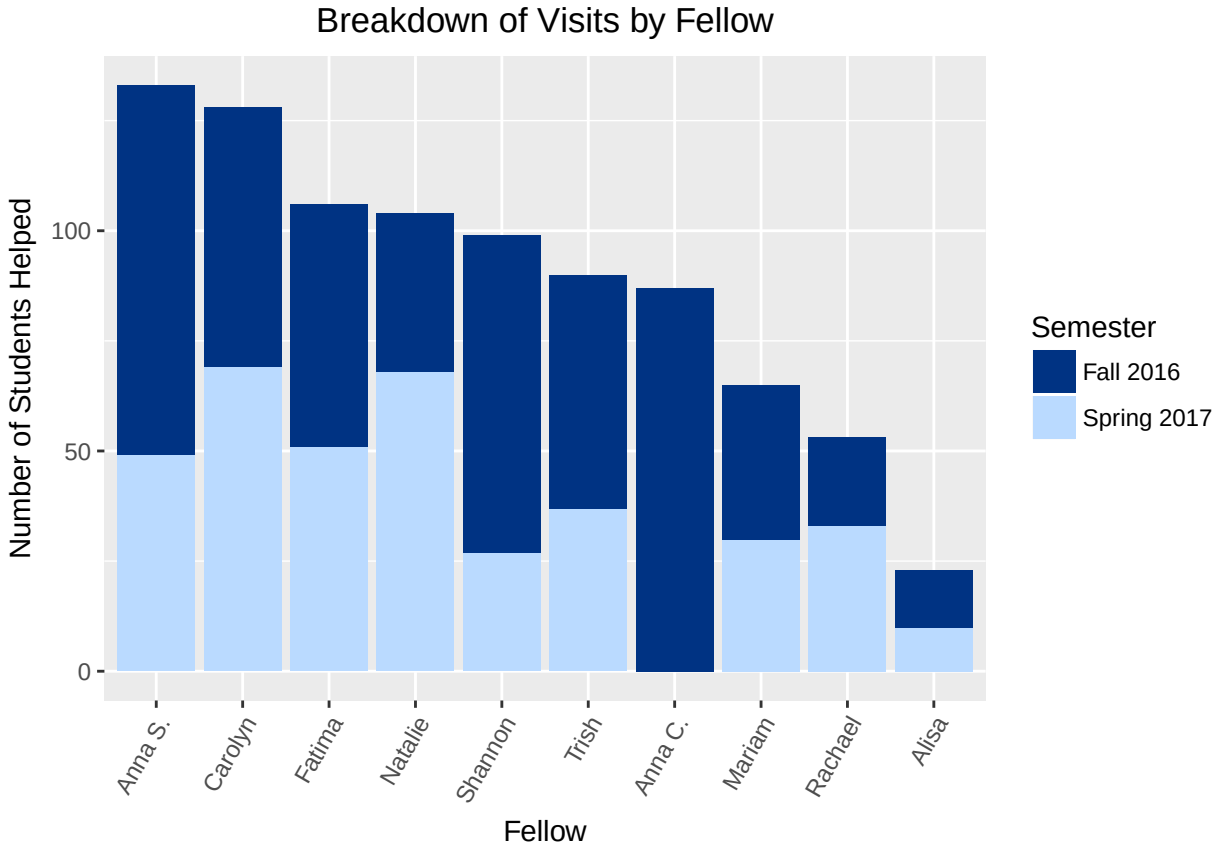
Table 5: How did you hear about the ERC? (Aggregate)

Findings:

Overwhelmingly, visitors heard about the ERC through class or a professor. These two categories are clearly not mutually exclusive, but we have no way of better classifying them based on the responses. In Fall 2016, more people learned about the center through workshops than in the spring (49 Fall vs 13 Spring), and in Spring 2017 more people came to us through other students and word of mouth than the previous semester (26 Fall vs 43 Spring). This finding is encouraging, as it indicates that our name is beginning to circulate more amongst the student body.

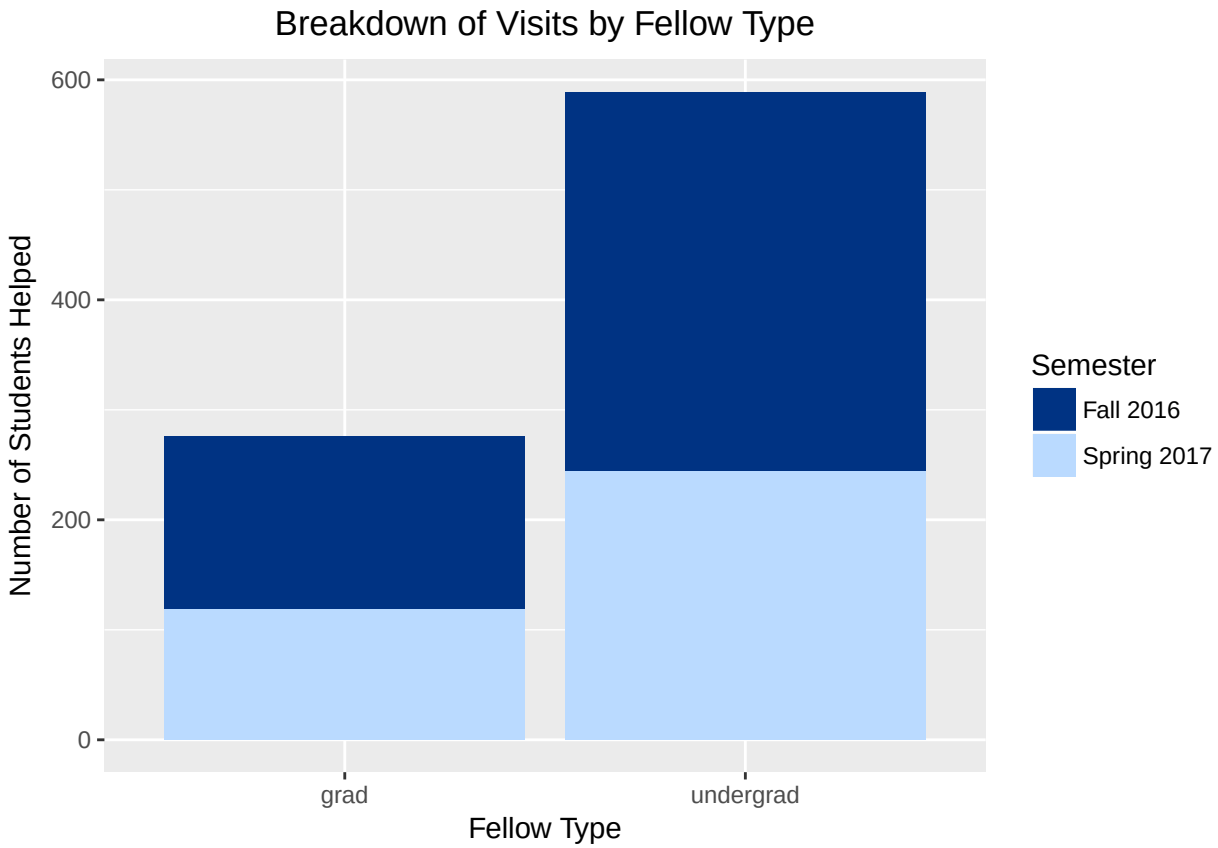
2.6 Fellow

We display tables and plots for both individual fellows and fellow "type", that is undergraduate or graduate student. In the 2016-17 academic year, the ERC had 3 graduate fellows (Trish, Anna S., and Rachael) and 5 (Fall) / 6 (Spring) undergraduate fellows.



Fellow	Visitors	Proportion
Anna S.	133	0.150
Carolyn	128	0.144
Fatima	106	0.119
Natalie	104	0.117
Shannon	99	0.111
Trish	90	0.101
Anna C.	87	0.098
Mariam	65	0.073
Rachael	53	0.060
Alisa	23	0.026

Table 6: Visits by Fellow (Aggregate)



Fellow Type	Visitors	Proportion
Alisa R.	23	0.026
grad	276	0.311
undergrad	589	0.663

Table 7: Visits by Fellow Type (Aggregate)

Findings:

In both Fall 2016 and Spring 2017, more vistors came to see our undergraduate fellows as a whole than our graduate fellows. Even normalizing for the number of each type of fellow, each graduate fellow received an average of 92 visitors and each undergraduate fellow received an average of 107 visitors over the course of the full academic year.

2.7 Course List

The following tables show the number of visits to the ERC by course for each semester. The tables exclude courses with fewer than five visitors. It should also be noted that this question is listed on the back of our survey, so there is a large number of NA's for this variable.

GENERAL CHEMISTRY LAB	71
SOCIAL RESEARCH METHODS	55
INTRO ECON	32
PSYCH STATS	25
DEVELOPMENT ECONOMICS	19
INTRO GIS METHODS	16
GIS METHODS	15
AMERICAN POLITICAL PARTIES	14
AMERICAN ELECTIONS	11
ECONOMETRICS	10
PHYSICS LAB	10
INDEPENDENT STUDY	9
QUANTITATIVE POLITICAL RESEARCH	7
APPLIED LINEAR REGRESSION ANALYSIS	6
PERSONALITY PSYCH LAB	5

Table 8: Number of Visitors by Course, Fall 2016

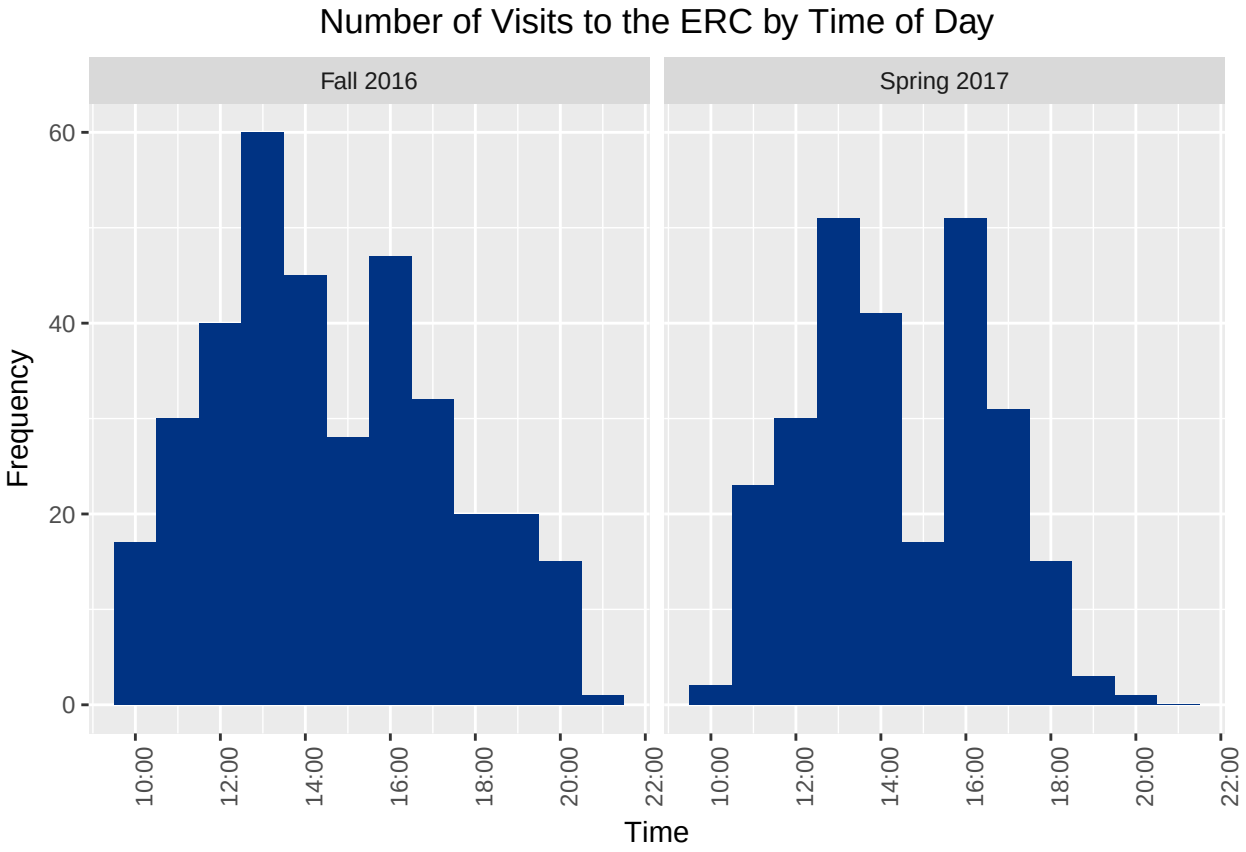
PROGRAMMING FOR BEHAVIORAL SCIENCES	63
ENVIRONMENTAL ECONOMICS	34
THESIS	22
APPLIED STATISTICAL COMPUTING	20
PUBLIC OPINION	19
EMPIRICAL DEVELOPMENT ECONOMICS	17
HAPPINESS ECONOMICS SEMINAR	7
INDEPENDENT STUDY	7
POLITICAL SCIENCE RESEARCH METHODS	7
ECONOMETRICS	5

Table 9: Number of Visitors by Course, Spring 2017

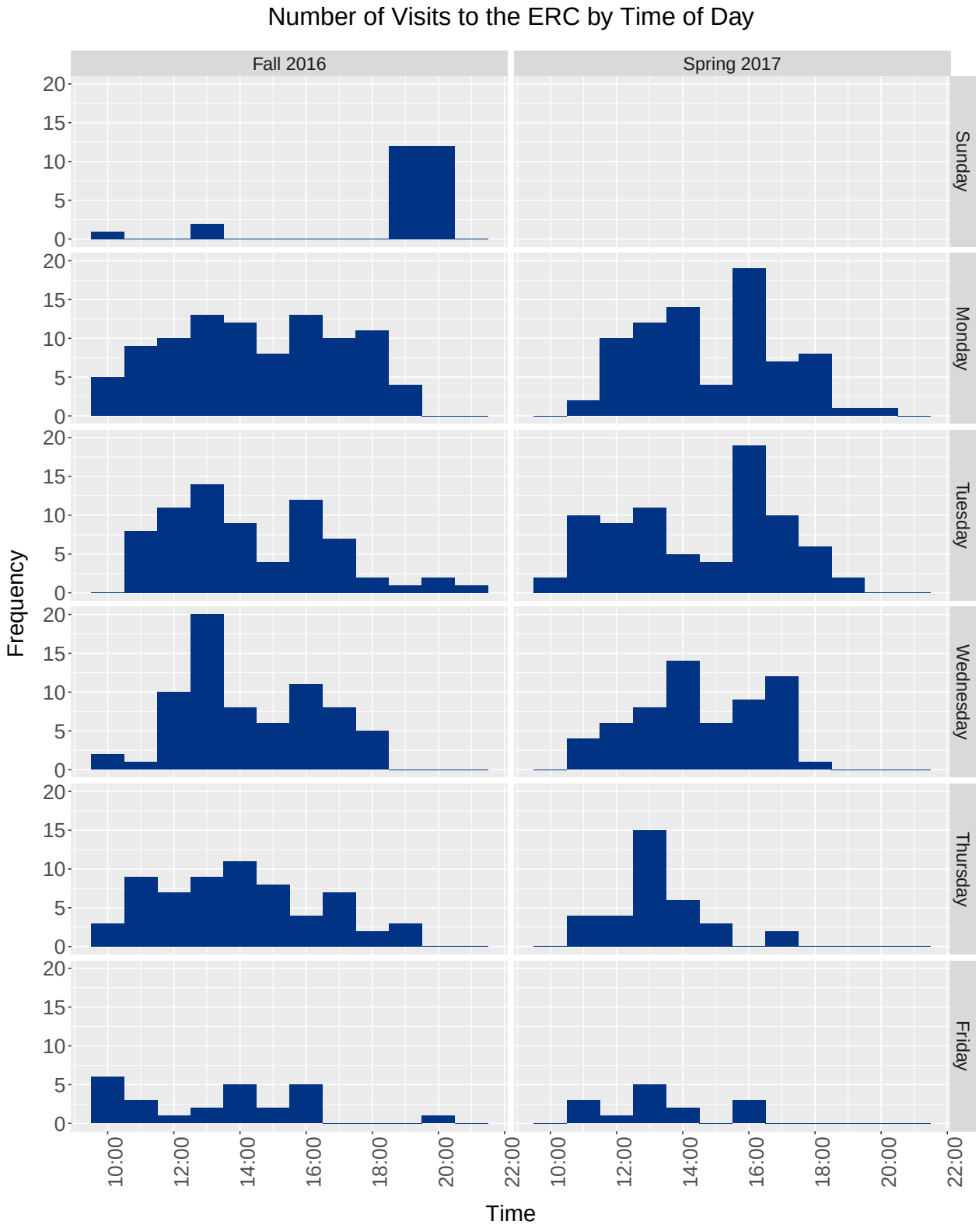
For both semesters, students from only a handful of courses dominated our walk-in hours. In Fall 2016, these students came from General Chemistry Lab, Social Research Methods, Intro to Economic Reasoning, and Statistics for Psychology, while in Spring 2017 they came from Programming for the Behavioral Sciences, Environmental Economics, Applied Statistical Computing, and various Senior Thesis seminars. We provided official support (i.e. the professor reached out to us or vice versa) for nearly all of these courses.

2.8 Time of Day

The following plots break down the number of visits to the ERC by time of day based on the time that the student entered. If the student stayed for multiple hours, it is not reflected in the histogram. The first set of histograms aggregates across all days of the week, while the second set has a unique plot for each day. The bulk of our walk-in hours are offered Monday through Thursday, with limited hours offered by appointment on Friday. In Fall 2016, we also offered walk-in hours on Sunday evenings. For a current calendar of our hours, **click here**.

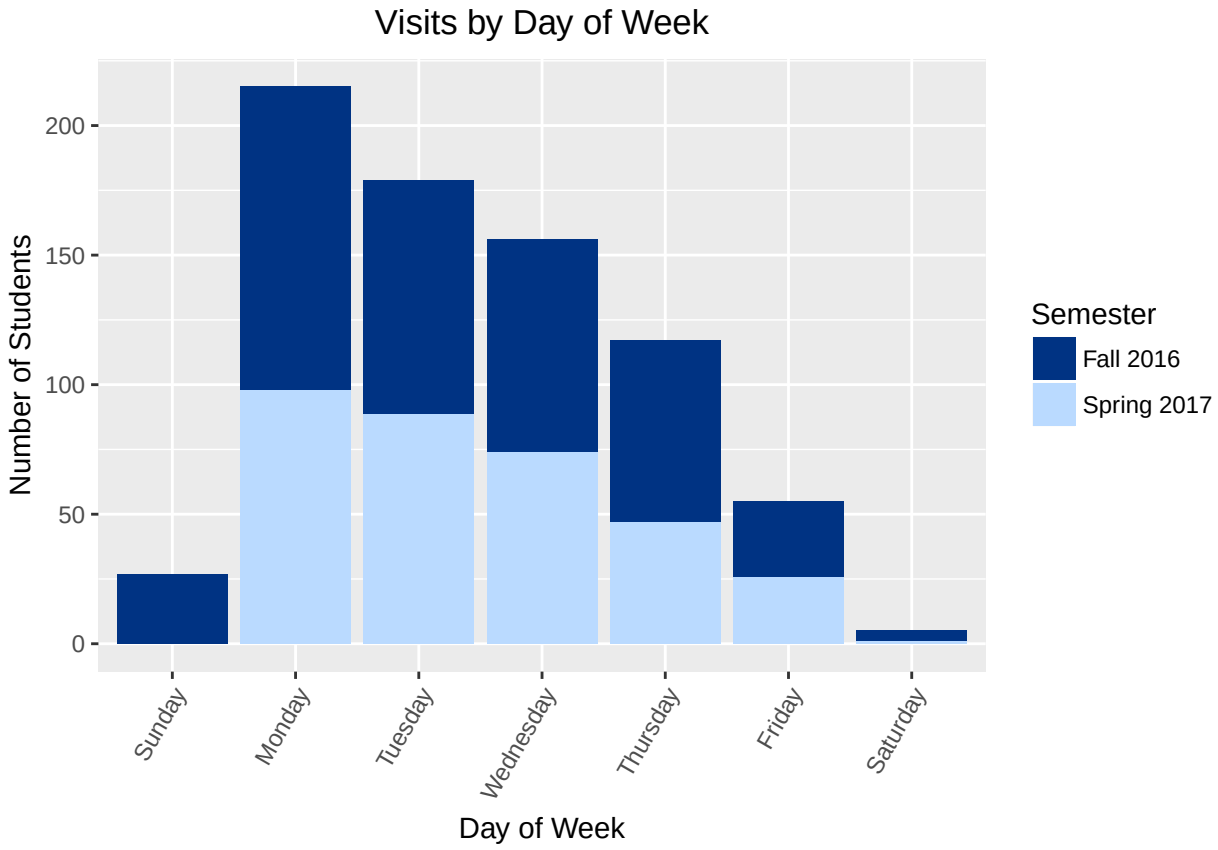


Findings: The distributions have slightly different shapes, but in general, we see a trend toward a bimodal distribution with peaks in the early and late afternoon.


Number of Visits to the ERC by Time of Day
Fall 2016
Spring 2017
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Frequency
0
5
10
15
20
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time

2.9 Day of Week

The Day of Week plot displays the number of visitors to the ERC on each day for Fall 2016 and Spring 2017.

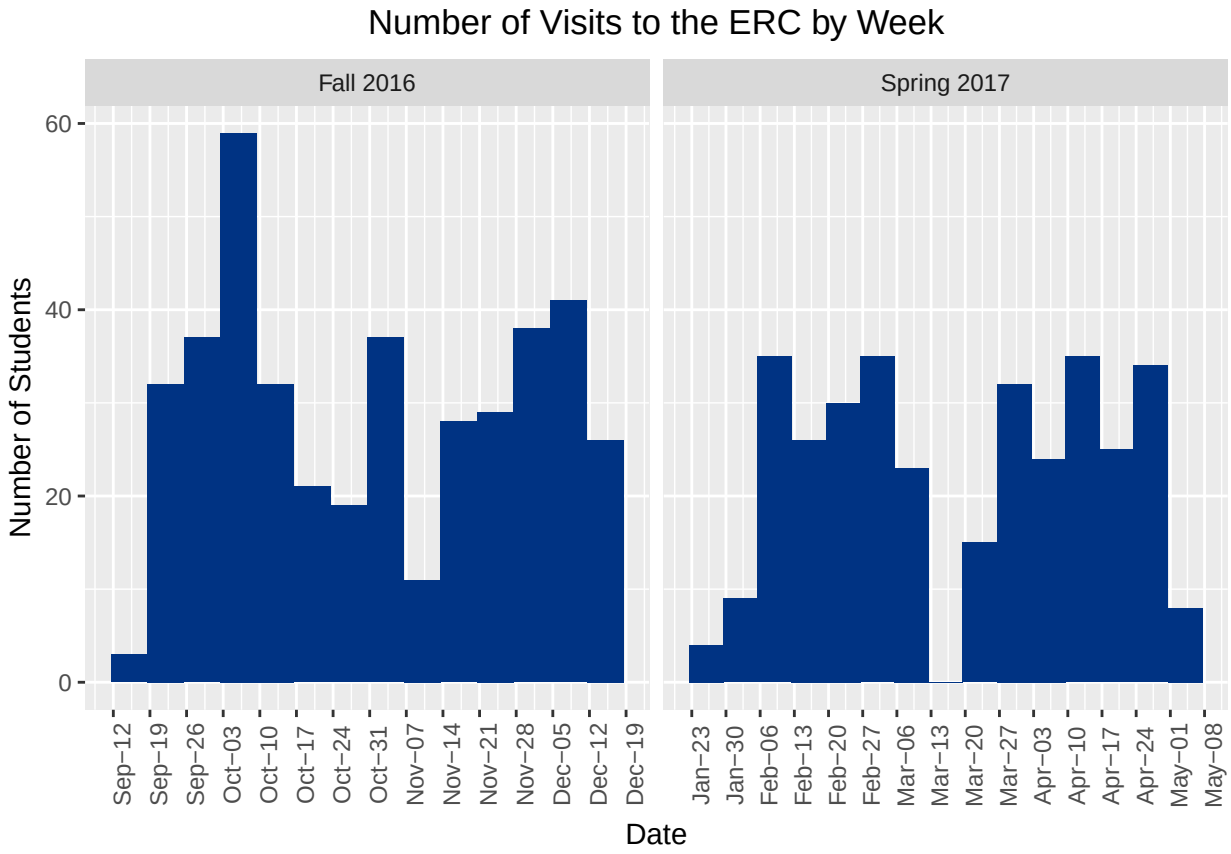


Findings:

Interestingly, for both semesters, we see a gradual decrease in the number of students as the week progresses. Because of this trend, we may want to schedule more hours earlier in the week to accomodate the needs of our visitors.

2.10 Week of Year

The final set of histograms displays the number of visitors to the ERC on a macro level for each semester.



Findings:

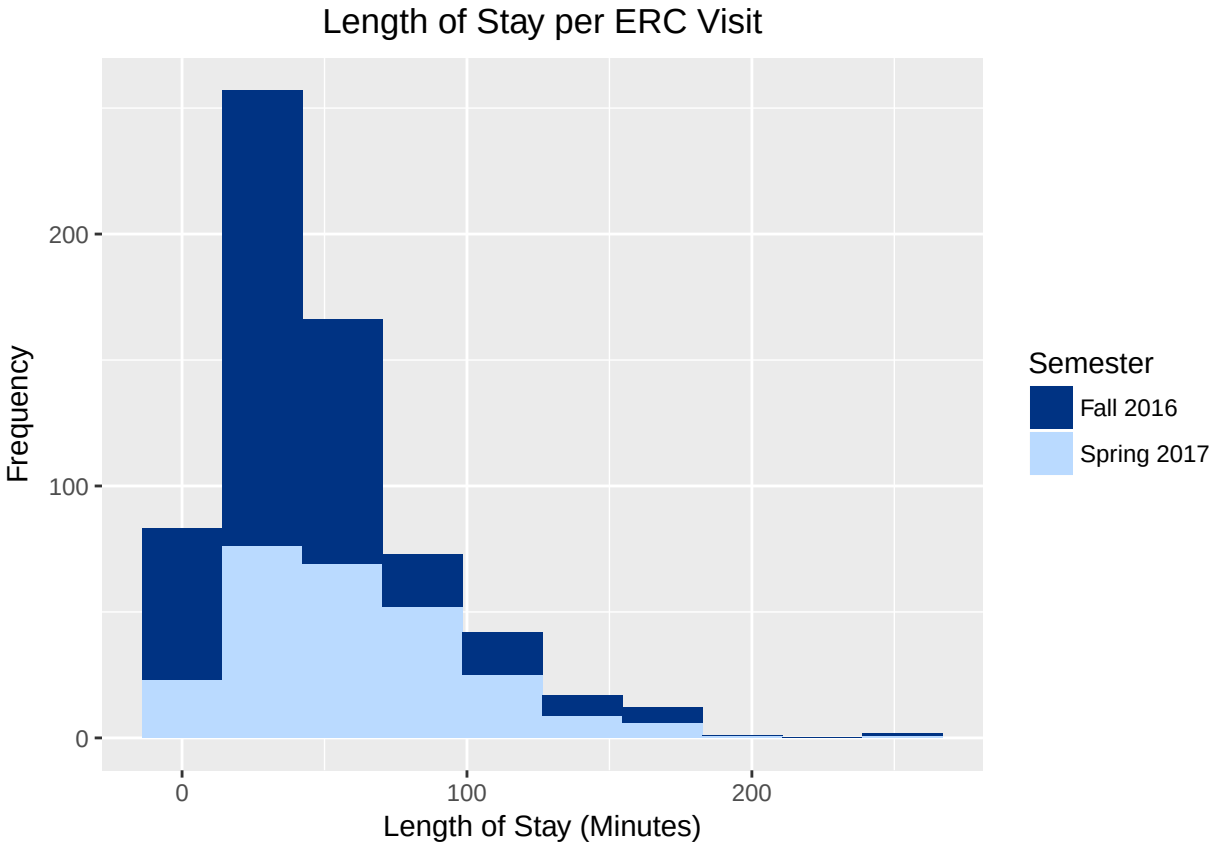
There appears to be a tendency toward more visitors a few weeks into the semester and a few weeks before the semester ends, with a slight drop in the middle (partially due to mid-semester breaks). The spike in Fall 2016 is likely due to the influx of General Chemistry students.

3 Length of Visit

In this section, we analyze the length of each visit to the ERC in minutes and present graphs, tables, ANOVA, and paired t-test results for length of stay broken down by select variables (Major Category, Program, School, and Class Year).

3.1 Summary Statistics and Distribution

Plot of Distribution



Summary of Distribution by Semester

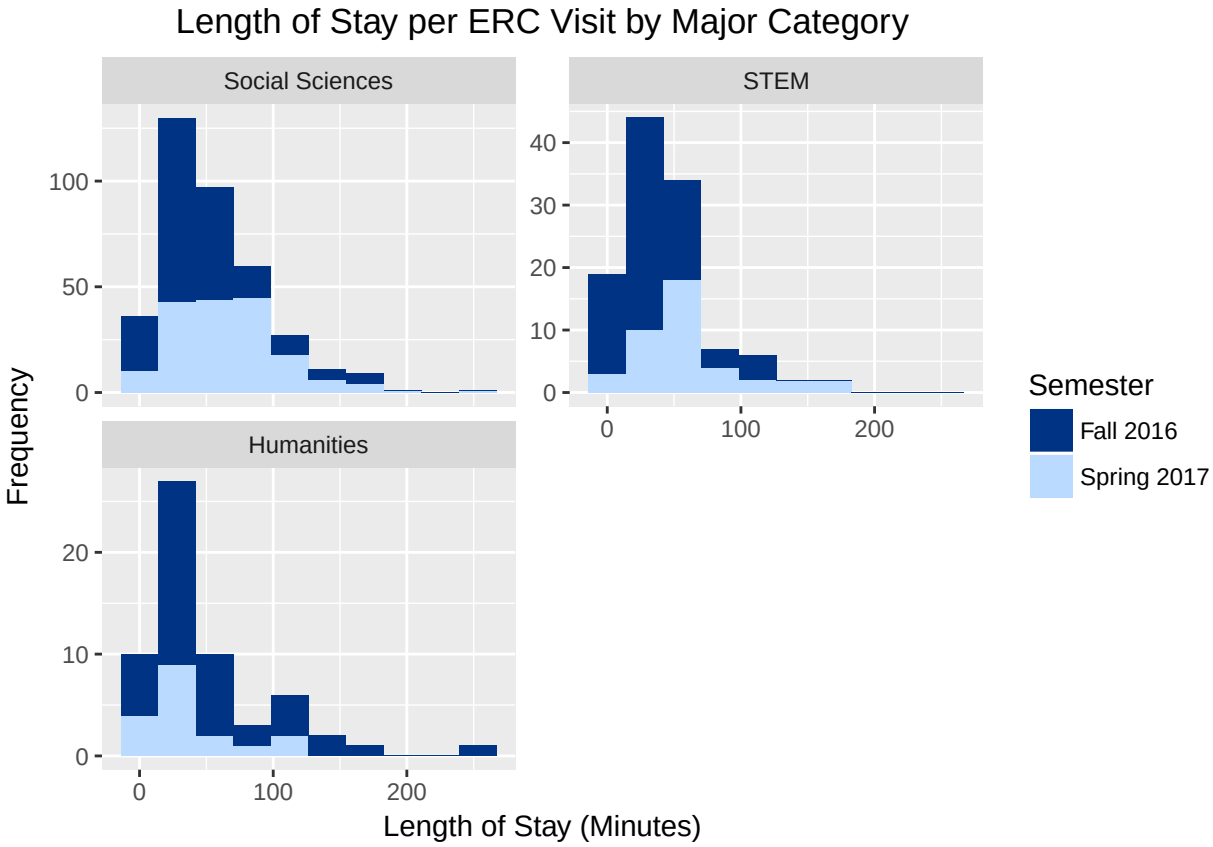
```
Fall 2016 :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   2.00   17.50   30.00   42.59   60.00  240.00      82
Spring 2017 :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   4.00   30.00   60.00   61.39   85.00  255.00      97
```

Findings:

The average length of stay in Fall 2016 was ~43 minutes and in Spring 2017 was ~61 minutes. The plot shows that the shapes of the distributions for the two semesters are similar, but there were many more students coming into the ERC for about 30 minutes in the fall compared to the spring. This finding is probably related to the decrease in visitors using Excel from Fall 2016 to Spring 2017, as previously mentioned.

3.2 Length of Stay by Major Category

We start by plotting the distributions for length of stay by major category, excluding "Other" and NA's. We then display the average length of stay in minutes by major category. Finally, we run an Analysis of Variance (ANOVA) test to determine if there is a statistically significant difference in the mean length of stay for each major category (again excluding "Other" and NA's). If the p-value is significant at the $p < .05$ level, we run pairwise t-tests with bonferroni corrections for multiple testing to determine which pairs of group means have a statistically significant difference. The p-values are displayed in Table 12.



Social Sciences	56.01
STEM	44.39
Humanities	49.17
Other	78.00

Table 10: Average Length of Stay by Major

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
majorCat	2	12667.60	6333.80	3.88	0.0213
Residuals	543	886701.56	1632.97		

Table 11: ANOVA Results

	Social Sciences	STEM
STEM	0.02	
Humanities	0.67	1.00

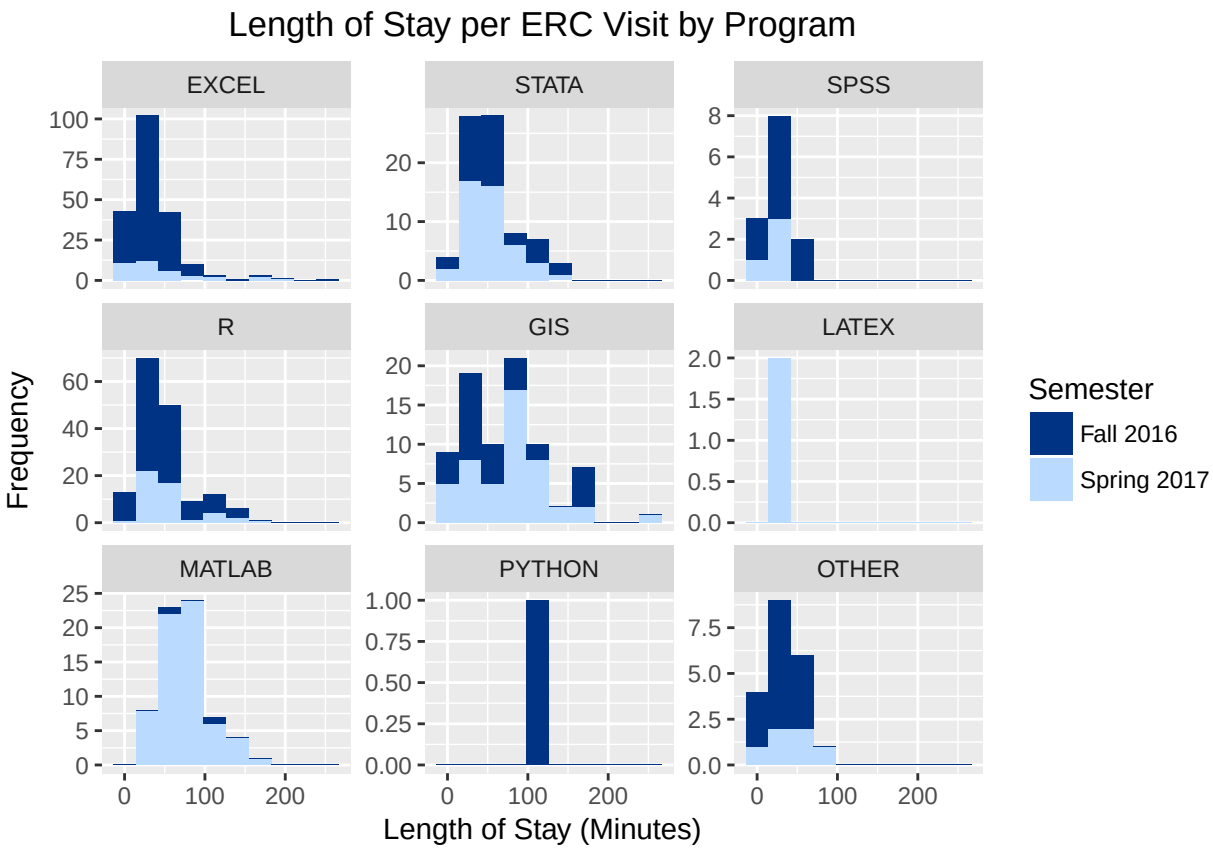
Table 12: p-values for pairwise t-tests with pooled SD and bonferroni adjustment

Findings:

Excluding students with majors in the "Other" category, Social Science students are most likely to stay the longest per visit averaging 56 minutes, while STEM students are likely to stay the shortest averaging 44 minutes. For the ANOVA test, we reject the null hypothesis that all major categories have the same mean length of stay. The pairwise t-tests show that the only statistically significant difference of group means is between STEM and Social Sciences ($p < .05$).

3.3 Length of Stay by Program

The same procedure described above for major category is applied to program. For simplicity, we only consider the first program listed by the visitor if he/she listed multiple programs. This simplification will have little effect on the results, since the majority of our visitors come in for assistance with one program at a time. The ANOVA and paired t-tests exclude programs with fewer than ten observations, as well as "Other."



EXCEL	36.47
STATA	54.96
SPSS	26.85
R	49.04
GIS	73.47
LATEX	25.00
MATLAB	76.76
PYTHON	120.00
OTHER	30.45

Table 13: Average Length of Stay by Program

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
program1	5	137201.58	27440.32	20.31	0.0000
Residuals	598	807773.53	1350.79		

Table 14: ANOVA Results

Findings:

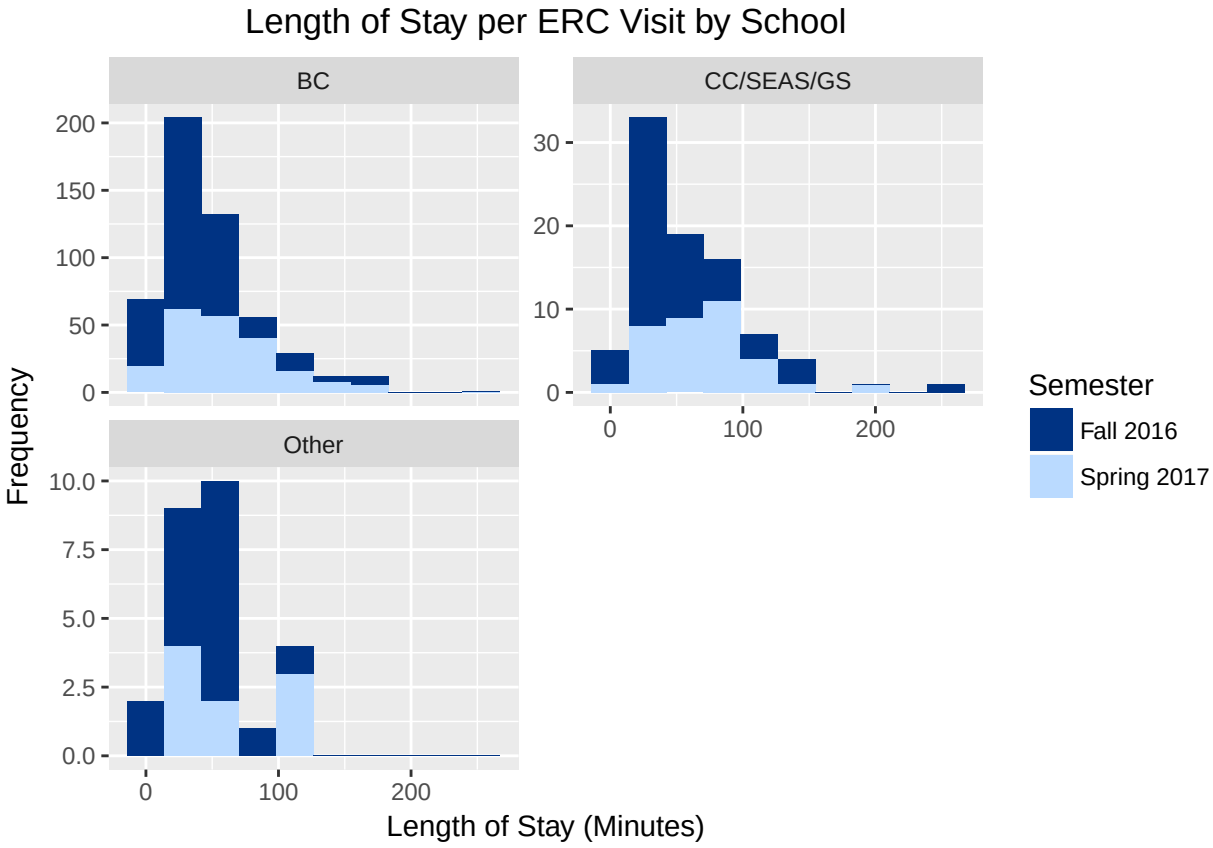
Excluding programs with fewer than ten observations, visitors tend to stay the longest when working in GIS and Python, averaging 73 and 77 minutes per visit, respectively. On the other hand, visitors using Excel and SPSS stay the shortest on average. For the ANOVA test, we reject the null hypothesis that students working in each program have the same mean length of stay at the $p < .001$ level. The pairwise t-tests show that most pairs of group means have a significant difference with some exceptions (see Table 15).

	EXCEL	STATA	SPSS	R	GIS
STATA	0.00				
SPSS	1.00	0.16			
R	0.02	1.00	0.55		
GIS	0.00	0.03	0.00	0.00	
MATLAB	0.00	0.01	0.00	0.00	1.00

Table 15: p-values for pairwise t-tests with pooled SD and bonferroni adjustment

3.4 Length of Stay by School

Next we break down length of stay by school. We include "Other" schools in this analysis.



BC	48.20
CC/SEAS/GS	60.75
Other	51.61

Table 16: Average Length of Stay by School

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
school	2	12546.28	6273.14	4.11	0.0169
Residuals	650	993249.66	1528.08		

Table 17: ANOVA Results

	BC	CC/SEAS/GS
CC/SEAS/GS	0.01	
Other	1.00	0.83

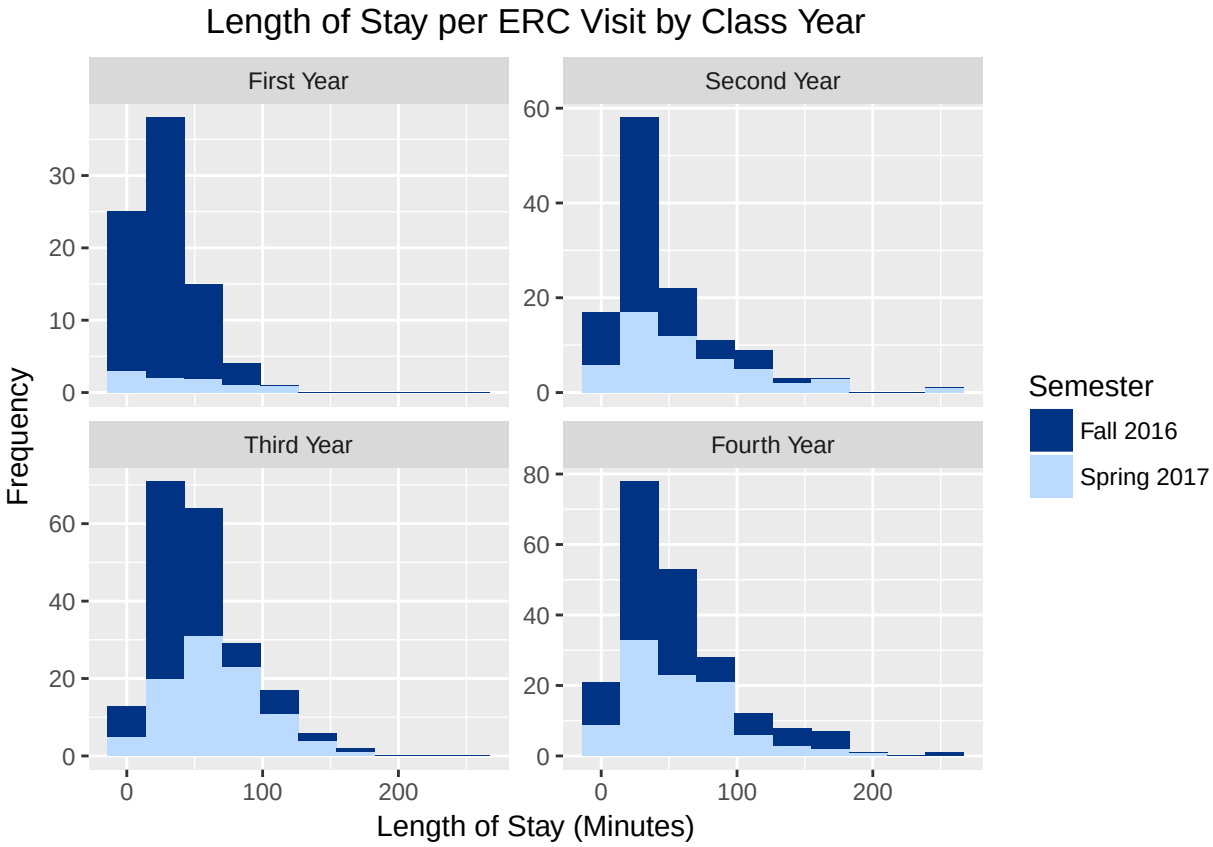
Table 18: p-values for pairwise t-tests with pooled SD and bonferroni adjustment

Findings:

There is a statistically significant difference between the length of stay of Barnard students and that of students from the other Columbia colleges at the $p < .01$ level. Based on the averages from the observed data, Barnard students stay for about 48 minutes, while Columbia students stay for over an hour.

3.5 Length of Stay by Class Year

Finally, the same length of stay analysis is broken down by class year, with plots, ANOVA, and paired t-tests excluding "Other."



First Year	29.06
Second Year	49.06
Third Year	55.09
Fourth Year	55.80
Other	26.00

Table 19: Average Length of Stay by Class Year

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
year	3	48534.25	16178.08	10.73	0.0000
Residuals	614	926096.66	1508.30		

Table 20: ANOVA Results

	First Year	Second Year	Third Year
Second Year	0.00		
Third Year	0.00	1.00	
Fourth Year	0.00	0.76	1.00

Table 21: p-values for pairwise t-tests with pooled SD and bonferroni adjustment

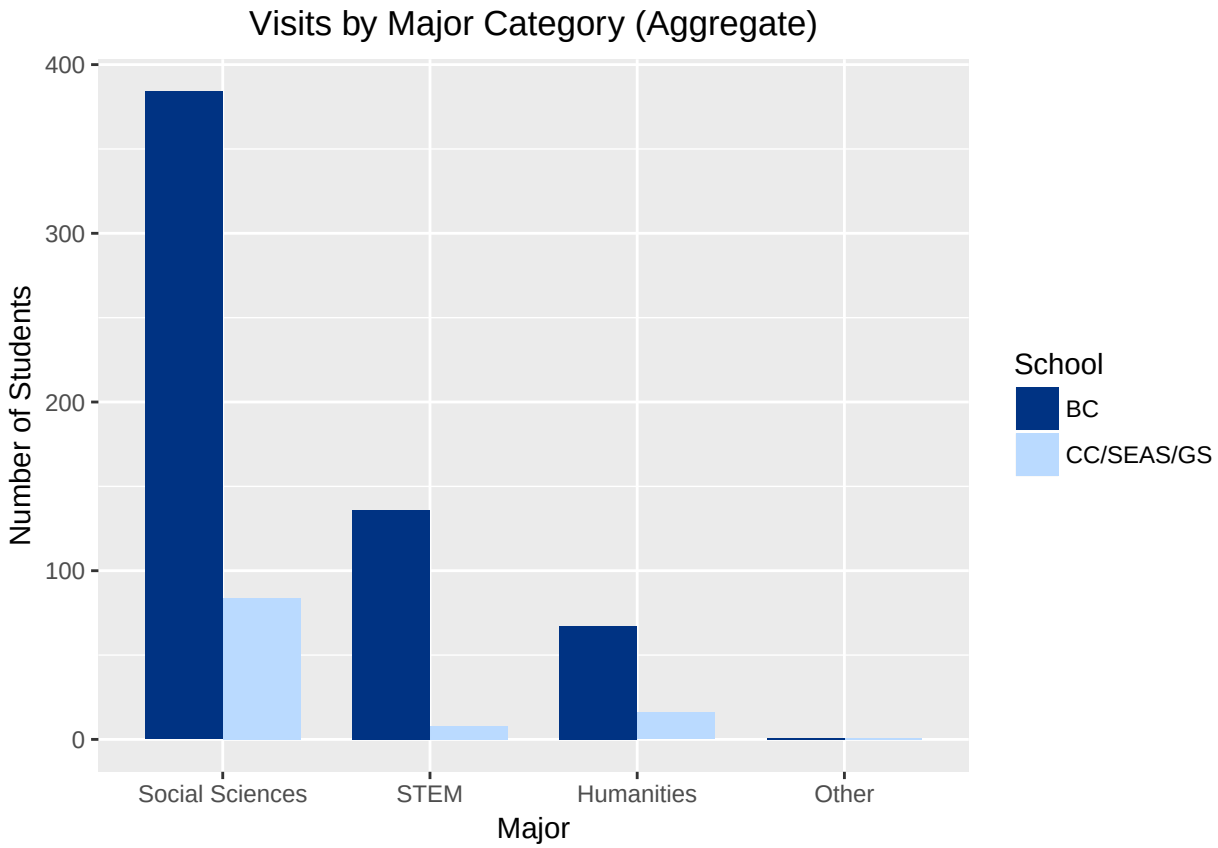
Findings:

Limiting our analysis to undergraduate first, second, third, and fourth years, we see a statistically significant difference in the length of stay of first year students and each of the other class years, all at the $p < .01$ level. In Fall 2016 and Spring 2017, the average length of stay for first years was under 30 minutes, while that of the other class years was at least 49 minutes.

4 Summary Statistics by School

In order to gain more insight into the types of students visiting the ERC from the different Columbia University colleges, we present plots for the number of visits from students of each school broken down by major category, software program/programming language, and class year. In section 4, we provide tables of the number of visitors by course. In descriptions of findings, we may use CC/SEAS/GS and Columbia interchangeably.

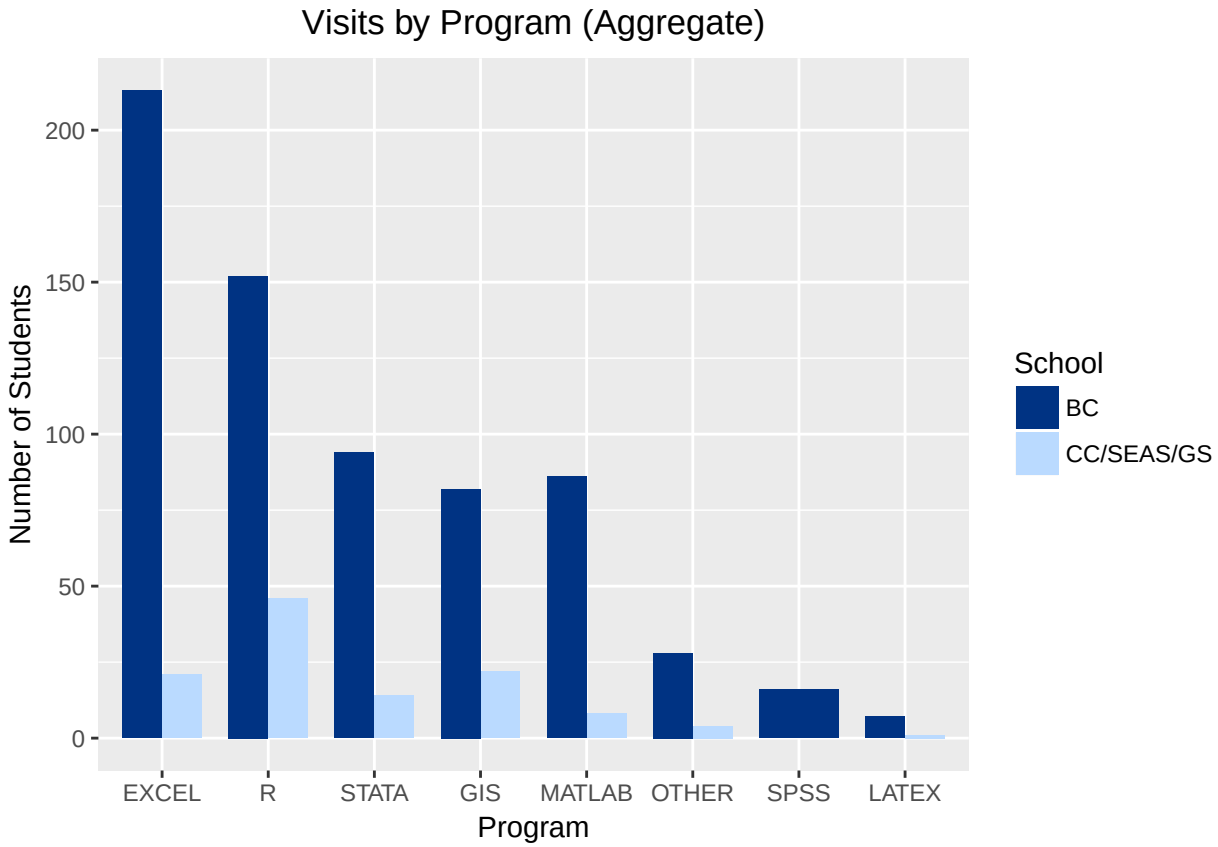
4.1 Major Category



Findings:

For both Barnard and CC/SEAS/GS, by far the most students visited the ERC from Social Science fields. However, for Columbia students, we saw more visitors from humanities and STEM majors, while the opposite was true for Barnard students.

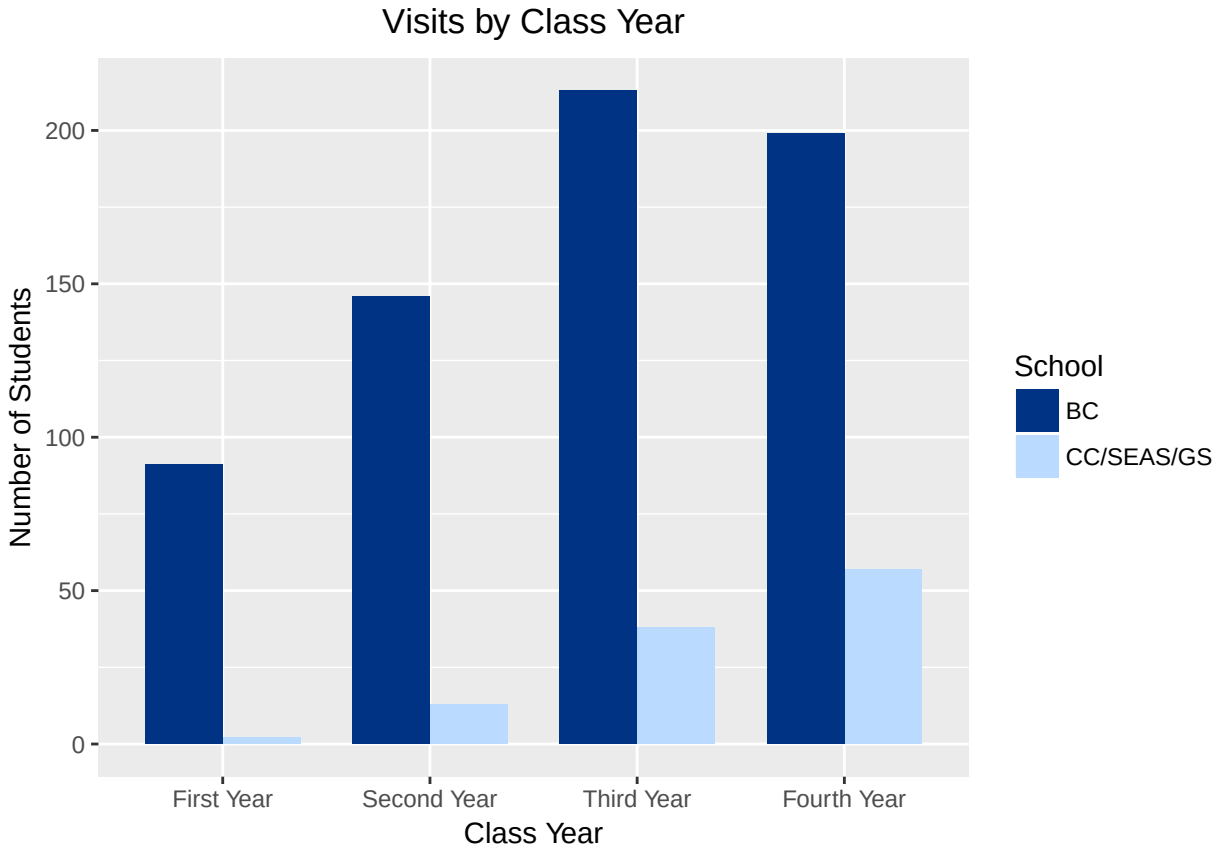
4.2 Program



Findings:

We received the most Columbia visitors using R and GIS (followed closely by Excel) and the most Banrard visitors using Excel and R.

4.3 Class Year



Findings:

While the general trend of more visitors for each increasing class year was true for both schools, there were slightly more Barnard third years than fourth years in the observed data set.

4.4 Courses

SOCIAL RESEARCH METHODS	27
ENVIRONMENTAL ECONOMICS	9
PROGRAMMING FOR BEHAVIORAL SCIENCES	7
EMPIRICAL DEVELOPMENT ECONOMICS	6
QUANTITATIVE POLITICAL RESEARCH	6
INTRO GIS METHODS	5

Table 22: Number of Visitors by Course, CC/SEAS/GS

GENERAL CHEMISTRY LAB	71
PROGRAMMING FOR BEHAVIORAL SCIENCES	54
INTRO ECON	30
SOCIAL RESEARCH METHODS	27
PSYCH STATS	25
ENVIRONMENTAL ECONOMICS	21
APPLIED STATISTICAL COMPUTING	19
PUBLIC OPINION	19
THESIS	19
ECONOMETRICS	15
INDEPENDENT STUDY	15
DEVELOPMENT ECONOMICS	14
AMERICAN POLITICAL PARTIES	12
GIS METHODS	12
EMPIRICAL DEVELOPMENT ECONOMICS	11
AMERICAN ELECTIONS	10
PHYSICS LAB	10
INTRO GIS METHODS	9
HAPPINESS ECONOMICS SEMINAR	7
APPLIED LINEAR REGRESSION ANALYSIS	6
PERSONALITY PSYCH LAB	5

Table 23: Number of Visitors by Course, BC

5 Tracking Repeat Visitors

5.1 Method

We are interested in determining whether the same students keep returning to the ERC. Our sign-in sheets include fields for name and UNI; but, because ERC fellows input the information into the Qualtrics survey based on handwritten surveys filled out by the visitors, there are many typos. In this analysis, we consider visitors to be a match if either the name OR the UNI matches. A more full-proof approach would be for a person to manually fix the name and UNI typos by cross-checking the information with the Columbia University directory.

A student ID is assigned to each unique student as determined by the above approach. Tables 20 and 21 display the number of students who have visited the ERC once, twice, three times and so on for both the aggregate data set and for each semester. We also determine the number of unique visitors to the ERC based on the student ID.

Finally, we create a new dummy variable called manyVisits, which takes the value 1 if the student in the given observation has visited the ERC multiple times (i.e. at least twice) and 0 otherwise. In the next section, we determine the variables that have a significant association with manyVisits in order to gain insight into why a student may or may not return to the ERC.

5.2 Results

The number of unique visitors in...
-The aggregate data set was 390
- Fall 2016 was 262
- Spring 2017 was 152

No. Visits	Frequency
1	226
2	74
3	37
4	20
5	9
6	6
7	4
8	5
9	3
10	1
11	2
12	2

Table 24: Frequency of Number of Visits per Student

No. Visits	Fall 2016	Spring 2017
1	155	88
2	58	23
3	26	17
4	11	6
5	4	3
6	4	2
7	2	4
8	1	3
9	0	2
10	0	1
11	0	2

Table 25: Frequency of Number of Visits per Student by Semester

6 Variables Associated with Return Visitors

In order to determine the variables that are associated with repeat visitors, we first create contingency tables that display the frequencies of observations that fall into each combination of categories. We also present the row percentages – that is, the percent of visitors who are one time visitors and the percent who are repeat visitors in a given category of a variable.

Next, we perform a chi-squared test on each of the variables of interest and manyVisits to determine if there is a statistically significant association.

To assess the signifcance of individual categories, we look at the standardized residuals of the chi-squared test. Each standardized residual is a z-score, so if the value lies outside of ± 1.96, it is significant at $p < .05$.

Finally, for 2x2 contingency tables, we use the odds ratio to calculate effect size, reporting both the odds ratio estimate and the 95% confidence interval.

Note that the following analysis is based on the entire data set (i.e. all visits), not just unique visitors.

6.1 Major Category

All Major Categories

majorCat	manyVisits 0	1	Total
Social Sciences			
N	122	328	450
Row(%)	27.1111%	72.8889%	68.1818%
STEM			
N	32	107	139
Row(%)	23.0216%	76.9784%	21.0606%
Humanities			
N	27	44	71
Row(%)	38.0282%	61.9718%	10.7576%
Total	181	479	660

	0	1
Social Sciences	-0.13	0.08
STEM	-0.99	0.61
Humanities	1.71	-1.05

Table 27: Standard Residuals

Pearson's Chi-squared Test:

X-squared = 5.3869765, df = 2, p-value = 0.0676446

Findings:

-Of the three major categories, STEM had the most repeat visitors (~77% of the observations), followed by the social sciences (~73%), and finally the humanities (~62%). The chi-squared test is not significant at the $p < .05$ level, so we cannot reject the null hypothesis that major category and manyVisits are independent (no association).

STEM vs Humanities

We run the same analysis as above, limiting the major categories to the humanities and STEM.

	manyVisits		
majorCat	0	1	Total
Humanities			
N	27	44	71
Row(%)	38.0282%	61.9718%	33.8095%
STEM			
N	32	107	139
Row(%)	23.0216%	76.9784%	66.1905%
Total	59	151	210

	0	1
Humanities	1.58	-0.99
STEM	-1.13	0.71

Table 29: Standard Residuals

Pearson's Chi-squared Test:

X-squared = 5.2387455, df = 1, p-value = 0.0220892

Fisher's Exact Test for Count Data (two-sided):

p-value = 0.0243104

95% Confidence Interval: 1.0478749, 3.9893931

Odds ratio estimate: 2.0444306

Findings:

-Here, we do find a statitically signifcant relationship ($p < .05$) between manyVisits and major category (STEM or Humanities), although none of the standardized residuals are statistically significant.

-Based on the odds ratio, the odds of a student being a repeat visitor to the ERC were 2.04 (1.05, 3.99) times as high for STEM majors than for humanities majors.

6.2 Program

program1	manyVisits 0	1	Total
EXCEL			
N	79	157	236
Chi-square	2.7899	1.0714	
Row(%)	33.4746%	66.5254%	
			32.4176%
STATA			
N	28	57	85
Chi-square	0.8264	0.3174	
Row(%)	32.9412%	67.0588%	
			11.6758%
SPSS			
N	9	5	14
Chi-square	6.7361	2.5869	
Row(%)	64.2857%	35.7143%	
			1.9231%
R			
N	49	151	200
Chi-square	0.7601	0.2919	
Row(%)	24.5000%	75.5000%	
			27.4725%
GIS			
N	32	69	101
Chi-square	0.5639	0.2166	
Row(%)	31.6832%	68.3168%	
			13.8736%
MATLAB			
N	5	87	92
Chi-square	16.5068	6.3391	
Row(%)	5.4348%	94.5652%	
			12.6374%
Total	202	526	728

	0	1
EXCEL	1.67	-1.04
STATA	0.91	-0.56
SPSS	2.60	-1.61
R	-0.87	0.54
GIS	0.75	-0.47
MATLAB	-4.06	2.52

Table 31: Standard Residuals

Pearson's Chi-squared Test:

X-squared = 39.0063944, df = 5, p-value = 2.367809×10^{-7}

Findings:

-The p-value of the chi-squared test is significant at the $p < .001$ level, so we reject the null hypothesis of independence and conclude that there is a statistically significant relationship between program and return visitors.

-From the proportion table, we see that the major category with the largest proportion of repeat visitors was MATLAB, followed by R, GIS, Stata, Excel, and finally SPSS.

-Looking at standard residuals:

for students working in SPSS, significantly more people than expected did not return ($p < .01$).

for students working in MATLAB, significantly fewer people than expected did not return (p<.001) and significantly more people than expected did return ($p < .05$).

6.3 School

	manyVisits		
school	0	1	Total
CC/SEAS/GS			
N	54	53	107
Row(%)	50.4673%	49.5327%	14.2477%
BC			
N	153	491	644
Row(%)	23.7578%	76.2422%	85.7523%
Total	207	544	751

	0	1
CC/SEAS/GS	4.51	-2.78
BC	-1.84	1.13

Table 33: Standard Residuals

Pearson's Chi-squared Test:

X-squared = 32.7848115, df = 1, p-value = 1.0294608×10^{-8}

Fisher's Exact Test for Count Data (two-sided):

p-value = 7.317833×10^{-8}

95% Confidence Interval: $2.0975901, 5.0843827$

Odds ratio estimate: 3.2635204

Findings:

-The chi-squared test is significant at the $p < .001$ level, so manyVisits and school are not independent. Based on the standard residuals, significantly more CC/SEAS/GS students did not return and significantly fewer of them did return.

-Based on the odds ratio, the odds of the student being a repeat visitor were 3.26 (2.10, 5.08) times higher if that student was from Barnard than if he/she was from Columbia ($p < .001$).

6.4 Workshop

workshop	manyVisits 0	1	Total
No			
N	126	258	384
Row(%)	32.8125%	67.1875%	50.3937%
Yes			
N	92	286	378
Row(%)	24.3386%	75.6614%	49.6063%
Total	218	544	762

	0	1
No	1.54	-0.97
Yes	-1.55	0.98

Table 35: Standard Residuals

Pearson's Chi-squared Test:

X-squared = 6.6970999, df = 1, p-value = 0.009657

Fisher's Exact Test for Count Data (two-sided):

p-value = 0.0103491

95% Confidence Interval: $1.0923531, 2.1127927$

Odds ratio estimate: 1.5173398

Findings:

-There is a statistically significant association between return visits and whether or not the visitor attended an ERC workshop ($p < .01$), but none of the standard residuals are significant.

-Based on the odds ratio, the odds of a student being a repeat visitor were 1.52 (1.09, 2.11) times higher if he/she attended a workshop (p<.05).

6.5 Class Year

year	manyVisits 0	1	Total
First Year			
N	27	66	93
Row(%)	29.0323%	70.9677%	12.4332%
Second Year			
N	48	104	152
Row(%)	31.5789%	68.4211%	20.3209%
Third Year			
N	59	188	247
Row(%)	23.8866%	76.1134%	33.0214%
Fourth Year			
N	71	185	256
Row(%)	27.7344%	72.2656%	34.2246%
Total	205	543	748

	0	1
First Year	0.30	-0.18
Second Year	0.98	-0.60
Third Year	-1.06	0.65
Fourth Year	0.10	-0.06

Table 37: Standard Residuals

Pearson's Chi-squared Test:

X-squared = 3.0055993, df = 3, p-value = 0.3907627

Findings:

-We found no statistically signifcant relationship between class year and return visitors at the $p < .05$ level.

7 Conclusions

Rather than summarize the findings again, we present ways in which the ERC can improve its services based on the data.

1) How can we reach out to more first year students outside of General Chemistry? Perhaps we can encourage first year classes in social science and STEM fields to include at least one empirical assignment and have professors refer their students to the ERC for assistance.

2) Since we saw a drop in the number of R users from Fall 2016 to Spring 2017, it might be useful to consider how we can continue to push students, particularly in the social sciences, to do their data analysis assignments and projects in R rather than STATA.

3) Very few students come into the ERC seeking help with research design and finding data. We may want to encourage students, particularly those writing theses, to come to us at the beginning of their research processes, so we can assist them in these areas.

4) Most of our visitors hear about the ERC though classes and professors, so we should continue to urge professors and TA's to mention us as a resource.

5) In terms of scheduling fellows, we should have the most people on staff early in the week based on our peak traffic times. We should also make sure to have adequate staff around 1pm and 4pm on most days of the week. Sunday night hours were very popular when they were offered in Fall 2016, so we might consider scheduling fellows at this time, as well.

6) We may want to find ways to encourage our Columbia visitors and students studying the humanities to come back for assistance, since the odds of these groups returning to the ERC is low based on the observed data.

7) Finally, we should continue to offer workshops for courses (and possibly apart from specific courses), since the odds of students visiting us multiple times are higher for workshop-attendees than non-workshop attendees.