# Lecturers' Union

*Statistics 141SL*
*University of California Los Angeles, CA 90029*

## ABSTRACT

The goal of this project was to assess 2007-2016 UCLA Lecturers' Union data and seek out the answers to Union's inquiries. The parameters in question were: the appointment percentage and length for the new hires plus the annual pay salary for the first year of hire. All three parameters were analyzed based on the gender and division. The main objective was to identify whether there is exists the gender discrimination problem among lecturers within UCLA College of Letters and Sciences.

The series of steps were implemented to attain necessary data and arrange it in format suitable for analysis. The data consisted of nine Excel files, one file per academic year starting in the Fall quarter of 2007. The files were combined in one uniformly formatted data set and transferred into R. The unique ids were assigned to each lecturer for confidentiality purpose. To obtain critical missing information, e.g. faculty gender, we employed a probabilistic approach which required accumulating enough name and gender data to make a prediction based solely on name. Divisions had to be obtained and added to the main dataset. A variety of data web scraping procedures were used to gather this data from an array of sources.

The methods used for the analysis combine evaluation of boxplots, histograms, violin plots and calculations of an effect size (Cohens d) for subgroups within the data. Subgroups were determined based on the division and academic year. Nearly all the subgroups showed a either a negligible or small effect size in terms of the difference among males and females.

While our results were satisfactory, we nevertheless feel key future step is to perform similar analysis for all the UC campuses.

# Introduction

The Lecturers' Union exists for lecturers to ensure their fair employment treatment by University. The data collected by the Union is not permitted to include gender, hence the Union never could perform the analysis which would identify whether the gender discrimination in regards to fair employment for lecturers takes place on UC campuses. We attempted this project to assist the Lecturers' Union in tackling this query.

# Research Questions

Is there a difference in appointment length for new hires? If so, does one exist in regards to gender?
In each division, is there a difference between appointment percentage for new hires, based off gender?
In each division, is there a difference in annual pay salary between men and women for the first year of hire?

# Data

## Primary

Our primary data source was the data collected and provided to us by the UCLA Lecturers Union. The data consisted of nine academic years (Fall 2007- Spring 2016) for the seven UC campuses. This data was presented in Excel format, each file represented academic year and consisted of three Excel sheets, one per quarter (Fall, Winter and Spring). Summer data was also available but not used in the analysis due to its substantially different lecturers appointments and pay arrangements. As we were primarily interested in investigating potential gender-based appointment length, percentage and annual pay inequality among newly hired lecturers within only UCLA campus, we set the non-pertinent data aside for potential future analysis and are limiting our data to UCLA employees who were in their fist appointment as a lecturer with an appointment percentage greater than 0%, resulting in a final tally of 1073 observations for analysis.

The most salient limitation of this data set is the lack of gender coding of the individual records, requiring us to rely on secondary resources to infer that data as best we can.

## Secondary - Gender

In order to identify the genders of the individual faculty members we had to rely on several external sources of data.

### California Baby Names 2009-2014

The California Health and Human Services Open Data website has many free data sets available for public use. We found one titled California Baby Names 2009-2014 (CB) https://chhs.data.ca.gov/Demographics/Most-Popular-Baby-Names-2009-2014/ypbh-uwxf and used it as one further resource for name/gender pairs. The (CB) data set contains 351,760 unique given name entries during that six year period. To avoid anomalies, we limited inclusion to names which occurred a minimum of three times, culling the (CB) data down to 66,124 observations.

### SSA Data

Additionally we used the collected data from Social Security Administration (SSA) registrations, years 1880-2015, found at: https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data. We combined the 136 text files into a single document using a shell script, then imported it into R. Using tapply we were able to get quick tallies for all names and both genders over all the years included. We again used this data to get probabilistic values for any matching names in our data set.

### Voter Registration

Finally we used a collection of voter registration files from Los Angeles County, and the states of Washington, Texas, Ohio, and Florida to compile a database of names and their associated gender probabilities.

### Composite Gender Probability

For each name in the data set we calculated a composite gender probability by weighting each source listed above by the number of instances where the name appeared in each data set. Additionally, for those lecturer's with both a given and middle name listed we weighted the middle name value inversely proportional to the certainty of the given name score. That is to say if a given name was 0.999 male the middle name would be weighted near 0, while if a given name was evenly split at 0.50 male the name would be scored entirely by the middle name.

### Manual assignment

Originally, the data had 1073 people. Utilizing the probabilistic method explained above, 956 got classified to one gender with more than 90% probability. 100 got classified to one gender under a 90% probability. 17 ended up unclassified by the model. This was due to unique cultural names or names that the model has not seen. To maintain accuracy, the team manually searched the ones whom the model could not predict above 90% accuracy. Departments sites, posted pictures, Daily Bruin articles, Ratemyprofessor, Facebook, and other on-line resources were all cross-referenced to verify the gender and the results were manually hard-coded. This method although time-consuming helped maintain accuracy and reduce error.

### Concerns

Many of our secondary resources are likely to draw from the same primary sources resulting in very high correlations for gender predictions among the various methods. This concern is mitigated somewhat by the high confidence we have in these predictors and the added value they provide in 'genderizing' incrementally more names with each additional inclusion.

## Secondary - Division/Department

The other key data point missing from the Lecturers' Union data was the individuals division within the university. We used several university resources in attempt to get as complete data of this as possible.

### Campus Directory

The online campus directory http://www.directory.ucla.edu/search.php was instrumental in filling in many of the gaps in our division data. Searching by department codes which were gleaned from the PDF dept-contact-list-0710.pdf found on www.payroll.ucla.edu. we were able to scrape the directory in its entirety.

### CCLE

Using the public CCLE website https://ccle.ucla.edu/, we were able to acquire a complete list of all classes taught by all lecturers and professors from Spring 2012 to Winter 2017. This data is used to create the department-divisions dataset for the entire UCLA (including David Geffen School of Medicine, The Law School of Law, The Herb Alpert School of Music, etc.). We focused our analysis only on the College of Letters and Sciences resulting in 5833 number of observations for 9 years (Table 1).

### Manual assignment

The provided data had inconsistencies with the obtained department-division data. 30 departments and 30 divisions had to be manually reedited in order to have it match to the department-division data, resulting in 16898 'divisionized' observations out of 17762 total UCLA observations.

Table 1: Final Count-Divisions.

| Humanities | Life Science | Physical Science | Social Science |
|---|---|---|---|
| 3457 | 450 | 597 | 1329 |

# Study Variables

| | | |
|---|---|---|
| Appointment Percentage | Appt% | Percentage time lecturer is assigned to the appointment. |
| Length of First Appointment | dur | Number of days of the first lecturer appointment. |

Appointment percentage was given to us by the client, while length of first appointment was a variable we computed by calculating the difference in the start date and end date of the lecturer's first appointment.

# Methods Used

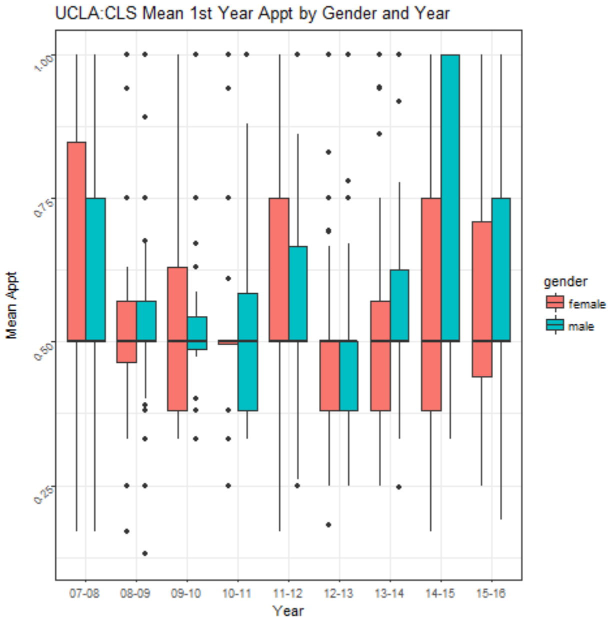
The methods used for the analysis combine evaluation of boxplots, histograms, violin plots and calculations of an effect size (Cohens d). Given that the data set was a population, Cohen's d seemed the most appropriate. Cohen's d is an effect size used to indicate the standardized difference between two means.

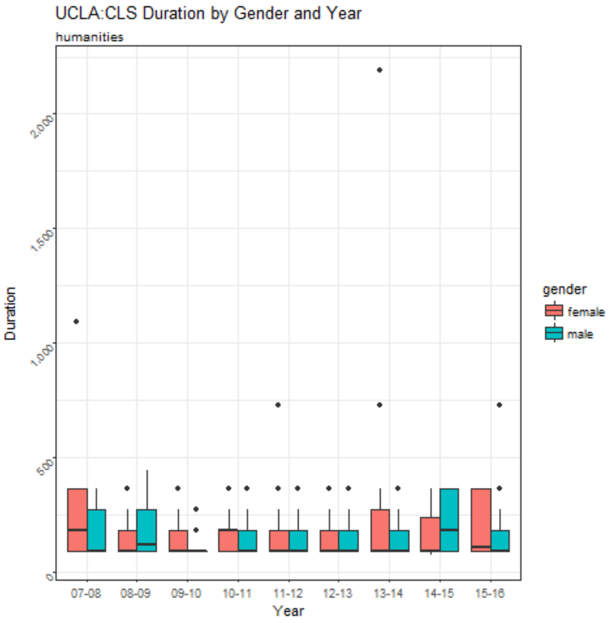
# Analysis

## Overall

Table 2: Counts of New Lecturers by Year

| | 07-08 | 08-09 | 09-10 | 10-11 | 11-12 | 12-13 | 13-14 | 14-15 | 15-16 |
|---|---|---|---|---|---|---|---|---|---|
| female | 139 | 63 | 25 | 39 | 60 | 56 | 65 | 78 | 58 |
| male | 106 | 57 | 35 | 35 | 49 | 53 | 45 | 53 | 57 |

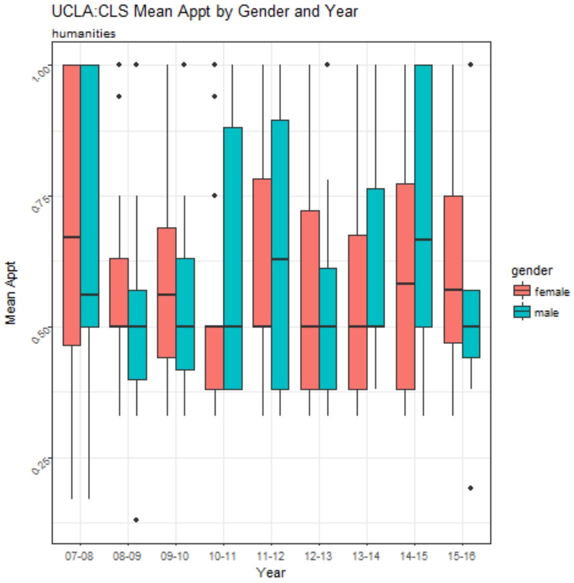


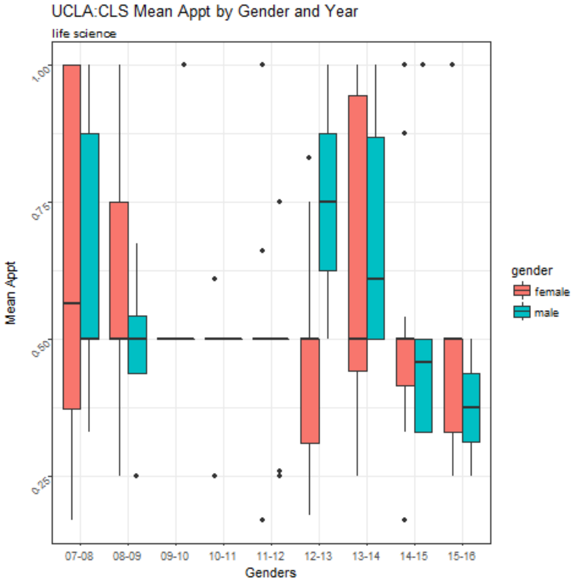


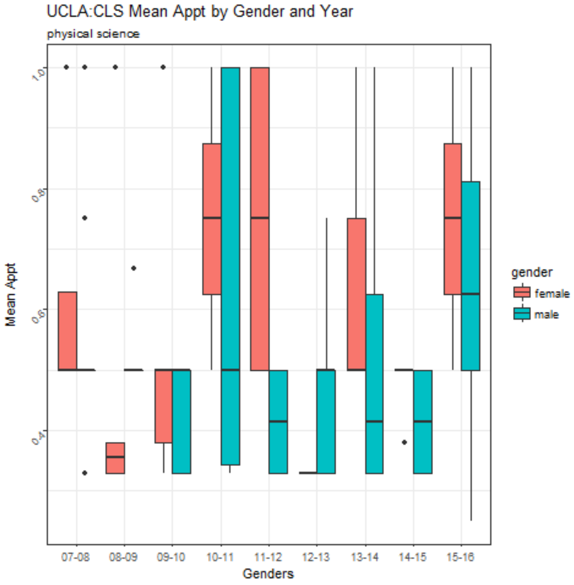
## Mean Appointment

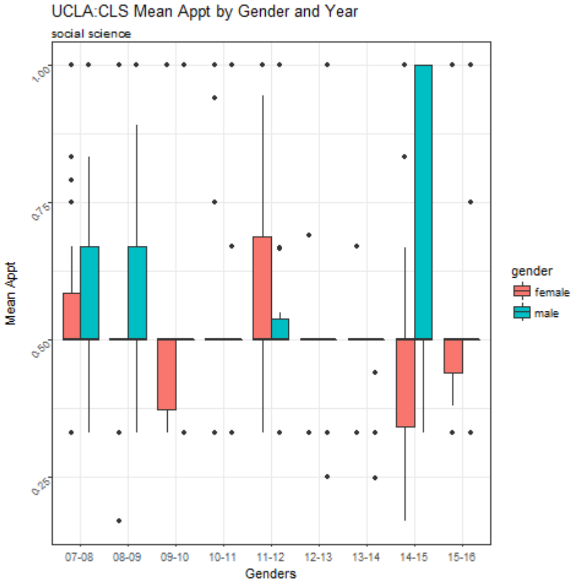
Table 3: Counts of New Lecturers by Year and Division

| | | 07-08 | 08-09 | 09-10 | 10-11 | 11-12 | 12-13 | 13-14 | 14-15 | 15-16 |
|---|---|---|---|---|---|---|---|---|---|---|
| Humanities | female | 73 | 25 | 11 | 17 | 32 | 23 | 38 | 36 | 31 |
| | male | 29 | 27 | 14 | 9 | 20 | 27 | 19 | 20 | 23 |
| Life Science | female | 14 | 7 | 3 | 5 | 10 | 12 | 5 | 15 | 9 |
| | male | 6 | 4 | 5 | 1 | 9 | 2 | 6 | 8 | 2 |
| Physical Sciences | female | 9 | 6 | 5 | 2 | 4 | 1 | 3 | 5 | 3 |
| | male | 17 | 5 | 7 | 10 | 6 | 14 | 4 | 6 | 14 |
| Social Sciences | female | 43 | 25 | 6 | 15 | 14 | 20 | 19 | 22 | 15 |
| | male | 54 | 21 | 9 | 15 | 14 | 10 | 16 | 19 | 18 |





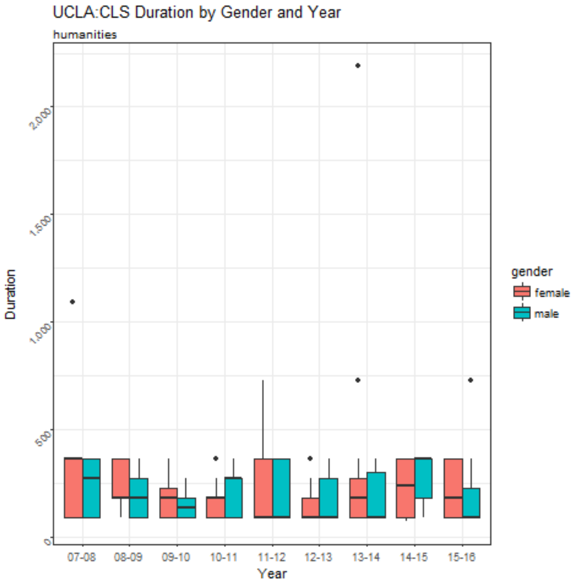




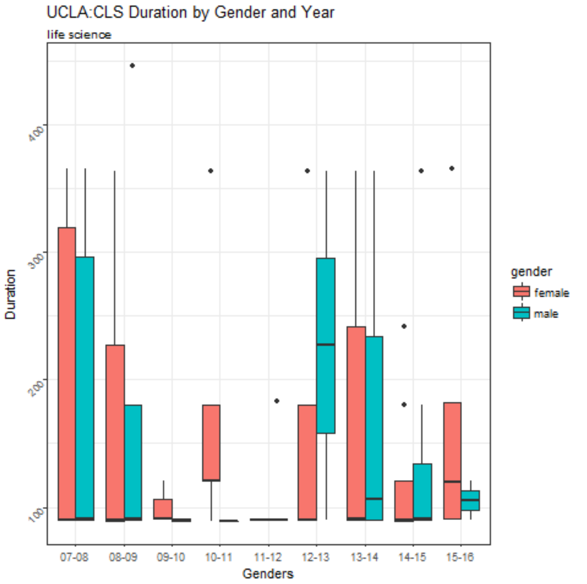
## Duration

Table 4: Counts of New Lecturers by Year and Division

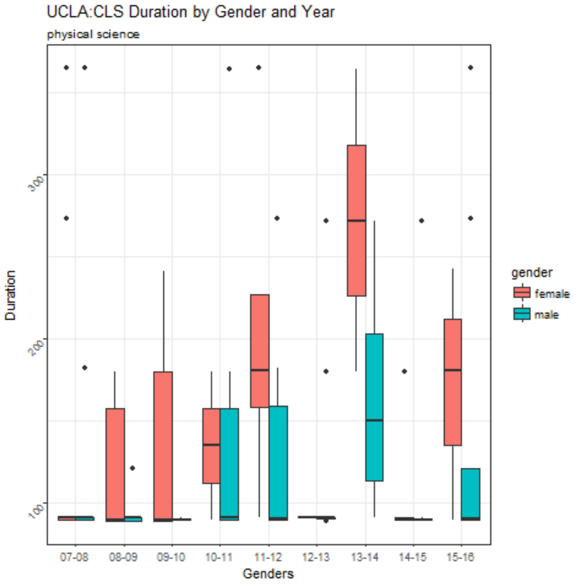
| | | 07-08 | 08-09 | 09-10 | 10-11 | 11-12 | 12-13 | 13-14 | 14-15 | 15-16 |
|---|---|---|---|---|---|---|---|---|---|---|
| Humanities | female | 73 | 25 | 11 | 17 | 32 | 23 | 38 | 36 | 31 |
| | male | 29 | 27 | 14 | 9 | 20 | 27 | 19 | 20 | 23 |
| Life Science | female | 14 | 7 | 3 | 5 | 10 | 12 | 5 | 15 | 9 |
| | male | 6 | 4 | 5 | 1 | 9 | 2 | 6 | 8 | 2 |
| Physical Sciences | female | 9 | 6 | 5 | 2 | 4 | 1 | 3 | 5 | 3 |
| | male | 17 | 5 | 7 | 10 | 6 | 14 | 4 | 6 | 14 |
| Social Sciences | female | 43 | 25 | 6 | 15 | 14 | 20 | 19 | 22 | 15 |
| | male | 54 | 21 | 9 | 15 | 14 | 10 | 16 | 19 | 18 |


UCLA:CLS Duration by Gender and Year

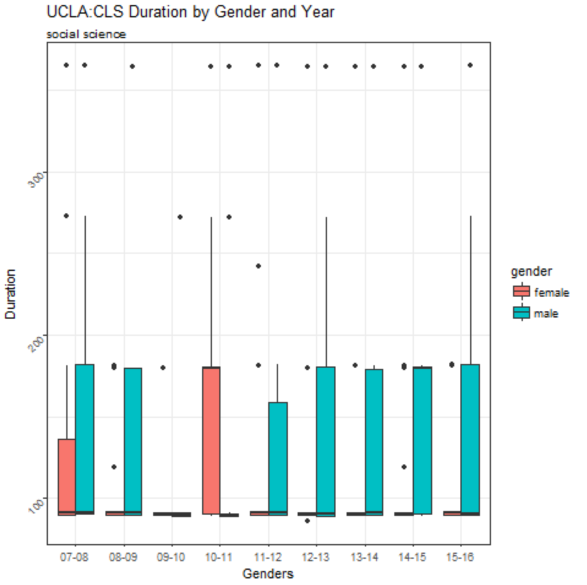
humanities


UCLA:CLS Duration by Gender and Year

life science


UCLA:CLS Duration by Gender and Year

physical science


UCLA:CLS Duration by Gender and Year

social science

## Cohen's d

Calculating the Cohen's d statistic we get negligible effects in Humanities, Life Sciences and Social Sciences. Physical Science division Cohen's d = 0.28 which is a small effect. Therefore, the difference in average appointment percentage between men and women lecturers is minimal in all four divisions.

Table 5: Cohen's d Estimate

| Humanities | Life Science | Physical Science | Social Science |
|---|---|---|---|
| 0.01 | -0.08 | 0.28 | -0.19 |

## Findings

The following table shows the percentage appointment time for all new hires.
Contextual Knowledge: New Hires are given *Core Benefits* if their appointment time is above 43.5%, and they are given *Full Benefits* if their appointment time is above 50%.

Table 6: Appointment Percentage for New Hires

| 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|---|---|---|---|---|---|---|---|---|---|
| 12 | 16 | 196 | 31 | 504 | 51 | 62 | 20 | 7 | 174 |

# Conclusions

Statistically, there was no gender discrimination in regards to appointment percentage for first year hires. Comparably, no findings were obtained of any statistically significant difference for gender bias in the divisions.

Despite a few outliers, appointment length seemed to be pretty standard. If anything, possibly favorable to faculty.

# Shortcomings

### Data

A researcher is only as good as its data. In this project, the team had to spend time creating a model for gender classification, manually searching for genders, and manually classifying departments into divisions. While we are grateful for the given data, moving forward further data collection by the university can help enrich the analysis of future study. Additional data collection on the university's part for divisions and genders for its lecturers can save time, money, man-hours, and increase the quality of any future statistical analysis.

# Recommendations

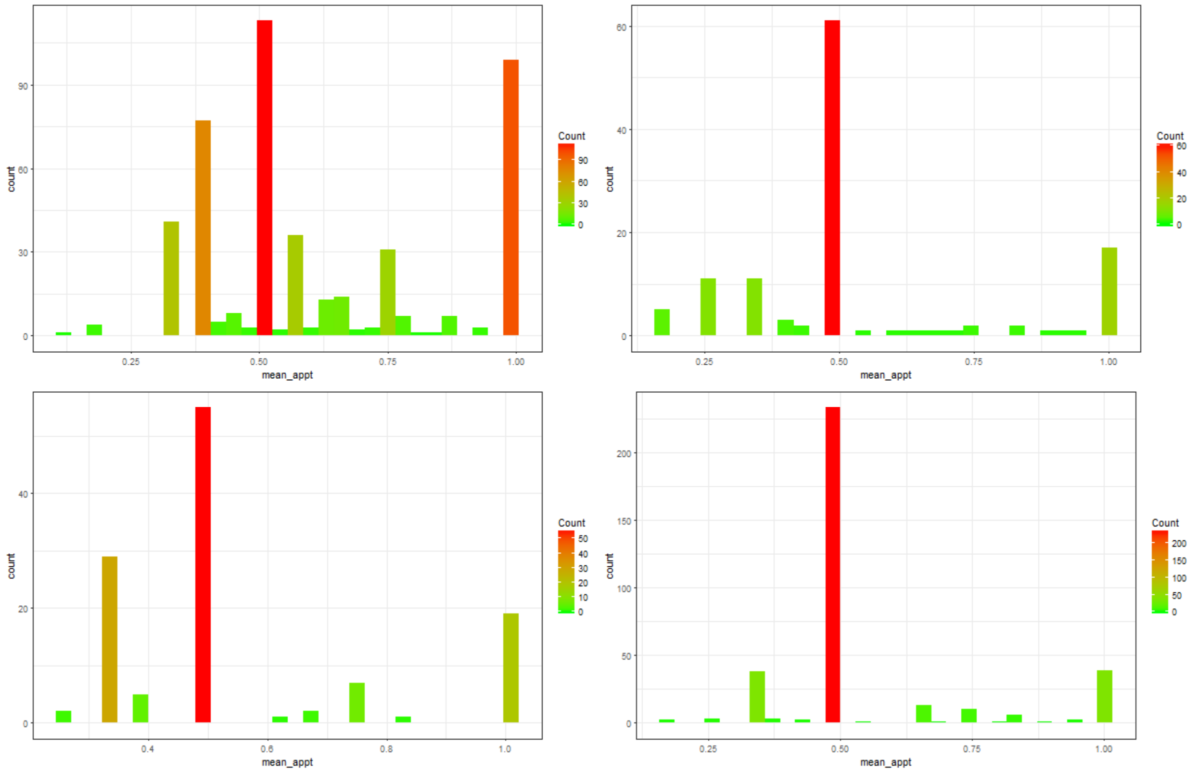
## Additional Tests

Further tests can be conducted to help answer the questions posed. To help distinguish if there is a difference in gender pay and appointment percentage. One could possibly use the Mann-Whitney U test (Wilcoxon rank-sum test) as a method to compare means. The Wilcoxon Rank-Sum Test is a non-parametric method that makes fewer assumptions, more robust, and is applicable to non-quantitative data. Furthermore, the Mann-Whitney U test does not require the assumption of normality. In conjunction with the Cohen's d test performed, the test can provide further insight and help validate our conclusion.

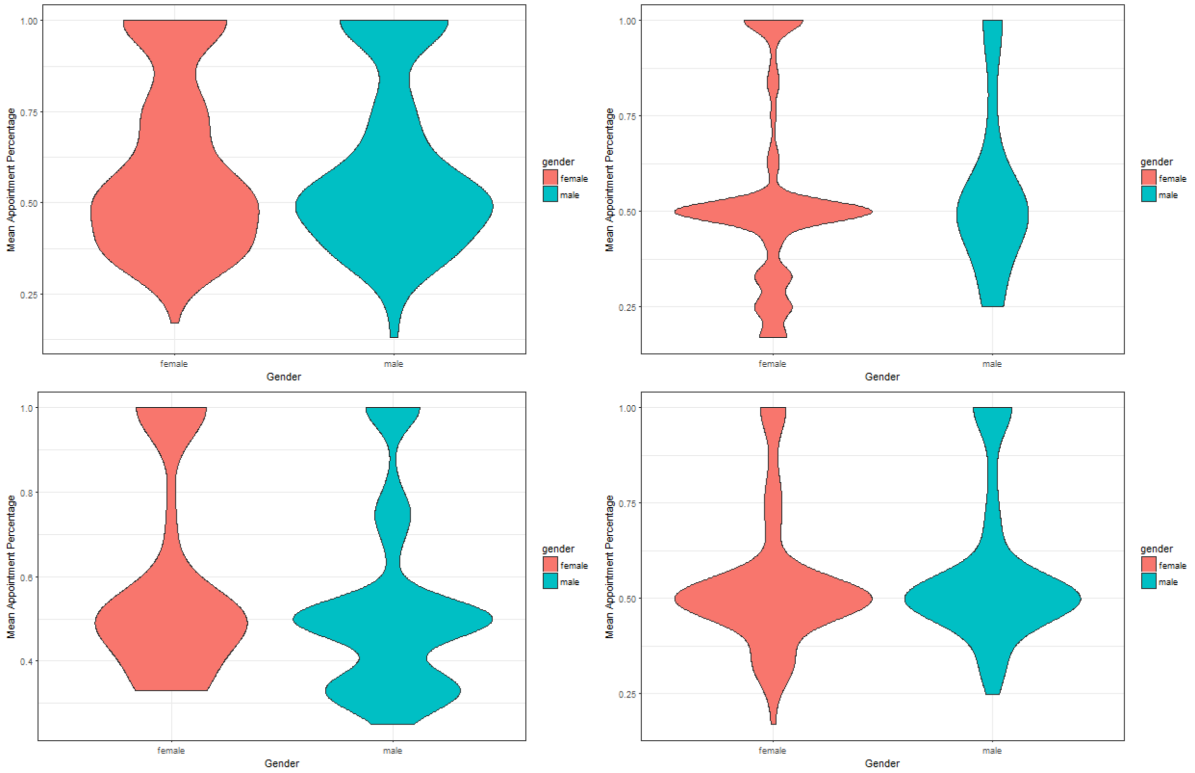
# Appendix

| Code Book | | |
|---|---|---|
| Variable Name | Variable Code | Variable Meaning |
| uid | | Unique identifier |
| gender | male/female | Gender |
| div | | Division |
| dept | | Department Name |
| year | | Academic Year |
| cto | | Title Code |
| hire | | Hire Date |
| start | | Appointment Begin |
| end | | Appointment End |
| dur | | Appointment Length |
| mean_appt | | Avrg. Appointment Percentage |
| p_male | | Probabilistic Score of the Male Name |
| p_female | | Probabilistic Score of the Female Name |

# Histograms



# Violin Plots


Mean Appointment Percentage
female
male
Gender
gender
female
male
Mean Appointment Percentage
female
male
Gender
gender
female
male
Mean Appointment Percentage
female
male
Gender
gender
female
male
Mean Appointment Percentage
female
male
Gender
gender
female
male