To: Ms. Cynthia E. Russell

            Department of Mathematics and Statistics, University of Houston-Downtown

            S705A, One Main Street, Houston TX 77002

From: Mai Ha

Subject: Transmittal Memo

Date: Oct. 30, 2018

This is a recommendation report prepared for Dr. Ryan D. Pepper, the Mathematics and

Statistics Department Chair of the University of Houston-Downtown. Please kindly forward this

document to Dr. Pepper.

Thank you for your help.

LIMITED RESOURCES TEACHING R PROGRAMMING AT THE UNIVERISTY OF HOUSTON-DOWNTOWN

Prepared for: Dr. Ryan D. Pepper

                         Professor and Department Chair

                         Department of Mathematics and Statistics

                         University of Houston-Downtown

Prepared by: Mai Ha

                         Student in Applied Statistics

                         University of Houston-Downtown

Abstract

The purpose of this document is to present the research done to find the most appropriate solution that can help retain, attract and help students with learning R Programming (R) at the University of Houston-Downtown (UHD). The problem with the current teaching of R at UHD is that there are too few resources available for students. In order to find the solution, a survey on students, interviews with professors, and multiple comparison analyses were conducted. These methods helped confirm that students have problems learning R and determine what the best solution is. The most appropriate and immediate action recommended is to create a reference guide that includes all necessary codes and their functions.

Table of Contents

List of Illustrations

Executive Summary

One of the most common tools in data analytics and statistics is R Programming. Despite its popularity and functionality, R is not being taught with enough resources at the University of Houston-Downtown. Many students learn R without a lot of experience in programming. As a result of the lack of resources and experience, several students in the Applied Statistics program struggle. Even though statistics and data analytics are growing fields, the statistics and data science programs offered at UHD do not appeal to students partly due to the difficulty of learning to program with R. Therefore, in order to retain and attract students to the program, the Department of Mathematics and Statistics must incorporate additional resources.

In order to find the correct resource needed, several methods of research were implemented. These methods are survey, interview, and comparison analysis. A survey was created and sent to UHD students who either have taken or is taking a class using R. Interviews were conducted with 2 professors in the Department of Mathematics and Statistics. These professors provided what resources they used, what students struggle with the most, and whether knowledge in R would help students become better statisticians and data analysts. Lastly, multiple comparison analyses were performed. These analyses are comparisons of multiple degree programs that teach R at UHD, between the improvements in the programs and the survey results, and between the University of Houston (UH) and UHD.

The survey yielded interesting results. Many students found the codes provided by their professors to be the most helpful. Several students struggled with syntax, vocabulary, knowing how a code works, and getting the necessary codes for assignments. Students agreed that having a manual for using R made specific for a class could have improved their initial

experience with R. Additionally, the professors mentioned during their interviews that students had trouble with basic tasks in R such as importing and exporting data and writing functions. Overall, the professors agreed that students did not have similar basic backgrounds in R. The first comparison analysis showed that the new and updated degrees either included more R classes or introduce R earlier in the degree plan. However, these changes did not address the problems mentioned in the survey responses from students. A comparison of the UH and UHD resources showed that UHD did not have as many resources as UH. The conclusion was that an additional resource must be created.

The recommended resource to be added to the new degrees is a reference guide. This guide would include all necessary codes and its function, which is similar to a reference guide provided by UH. The information needed for making this guide would be collected from students in the form of an assignment. Codes will be chosen by professors, and assembled by either a contractor or someone in the department. This guide could possibly cost from nothing to $320, depending on which option the department chooses.

Introduction

R is currently one of the most common statistical analysis tools. Since it is a free open-source programming language, not a lot of attention and resources at UHD are directed towards improving its accessibility for students. A comprehensive R manual is available on the R official website to accommodate learners, which has a section shown below. Problem arises because this user guide assumes that learners are familiar with programming vocabulary and provides overly detailed explanation on how each code works. Such a guide would only benefit those learners who are sufficiently familiar with programming and its vocabulary.

2.1 Vectors and assignment

R operates on named *data structures*. The simplest such structure is the numeric *vector*, which is a single entity consisting of an ordered collection of numbers. To set up a vector named x, say, consisting of five numbers, namely 10.4, 5.6, 3.1, 6.4 and 21.7, use the R command

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

This is an *assignment* statement using the *function* c() which in this context can take an arbitrary number of vector *arguments* and whose value is a vector got by concatenating its arguments end to end.[7]

A number occurring by itself in an expression is taken as a vector of length one.

Notice that the assignment operator ('<-'), which consists of the two characters '<' ("less than") and '-' ("minus") occurring strictly side-by-side and it 'points' to the object receiving the value of the expression. In most contexts the '=' operator can be used as an alternative.

Assignment can also be made using the function assign(). An equivalent way of making the same assignment as above is with:

```
> assign("x", c(10.4, 5.6, 3.1, 6.4, 21.7))
```

Figure 1 - R Manual section (R Core Team)

On the other hand, many college students attending introductory classes with R do not have much programming experience. Typically, professors will briefly discuss how to install, set up, and use R syntax the classroom. Since the classes are taught use R but are not teaching R specifically, professors must cover the required materials and thus, cannot spend much time showing students the basics of R. Thus, students who are not versed in programming languages generally have to rely on the R comprehensive manual, which is too advanced for most of them.

As a result, many Applied Statistics students at UHD had trouble learning R. Students who struggle in upper-level classes are usually those who do not have sufficient training in R from their pre-requisite courses. Consequently, those struggling students may question if they have chosen the right major, and some of them might change to a different degree or drop out of school.

Recently, an updated version of the Applied Statistics degree and a new degree called Data Science were created in an attempt to combat such a problem. The Data Science degree plan combines statistics, mathematics and computer programming, and includes more programming classes than the traditional Statistics program. Alternatively, the updated Applied Statistics degree has several positive changes such as requiring more R classes and teaching R earlier in the program.

Although R will be introduced early in these new degree plans, a true fix for the problem has not emerged. Students will still struggle with R syntax, vocabulary, and coding with the limited resources currently available. Thus, in order to improve retention rates for these new programs and train successful statisticians and data analysts for the future, a better, more permanent solution must transpire.

Methods

Method 1: Survey

I created an online survey for current UHD students who have taken at least one class using R. The survey aimed to determine what struggles students face in learning R, and

gathered students' thoughts on possible improvements. Survey responses came from students in both the Statistics and Data Science degree programs.

Method 2: Interview

All 3 of these questions were asked in the interviews:

1. What aspect of R do you find students struggle with the most?

2. What resources are you using to teach R?

3. Do you think knowing R will help students become better statisticians and data analyst?

I either emailed or met with faculties at UHD for their response to these questions. These faculties are Dr. Dexter Cahoy and Dr. Ha Nguyen, who are professors in the Department of Mathematics and Statistics. These interviews helped assess how well students perform and identified all available resources that are provided for students in courses teaching R at UHD.

Method 3: Comparison

I determined improvements in teaching with R by a comparison analysis between degree plans of the original Statistics program, the updated Statistics program, and the new Data Science program. Then, I analyzed which of the students' struggles and suggestions in the surveys were addressed by these improvements. The closest campus to UHD is UH, which also teaches using R despite not offering a statistics degree. Thus, I also compared resources utilized to teach R between UHD and UH for additional suggestions.

Results

Result 1: Survey

The online survey for UHD received 32 responses (See Appendix for the survey template and additional survey results). 17 students confirmed that they benefit most from codes provided by the professors (53%). Additionally, an equal number of students found the textbook and the R manual helpful (4 students or 12.5% for each resource). Of all the resources utilized in the classroom, the syllabus was helpful to only 3 people (9.4%).
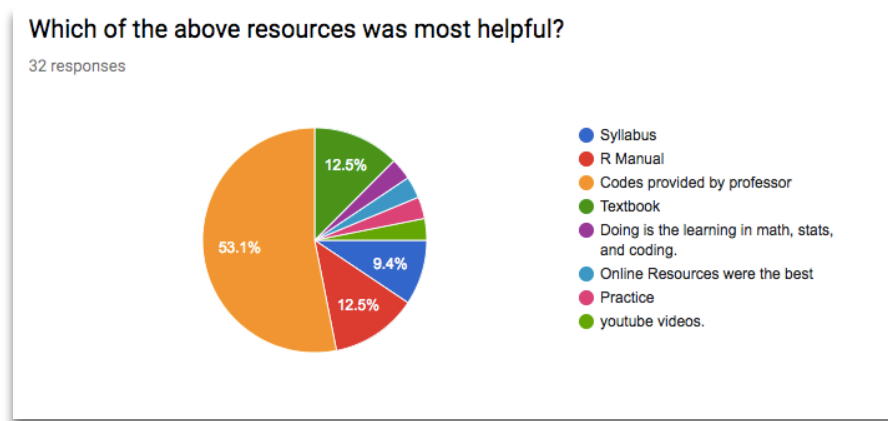
Figure 2 - Survey result

Students struggle the most with learning syntax, vocabulary, knowing when to use a command, and understanding the use of each command. Students also struggle due to not getting enough practice and not having prior coding experience. Additionally, a student mentioned that the textbook in their class does not provide all necessary codes for assignments.

Result 2: Interview

According to Dr. Cahoy, students struggle the most with data import and export, writing functions, and getting enough practice. He uses the textbook and provides codes for students in

this class. Dr. Nguyen utilizes 2 textbooks, the R manual, and also provides codes for his students. He mentioned that because a standard for basic R knowledge is not established, students who learned R before would still struggle to learn R in a different, higher-level class.

Result 3: Comparison

1. Original and updated Applied Statistics vs. Data Science degrees

Below are the requirements and electives of the Applied Statistics degree requirements and electives. According to Dr. Cahoy, the classes that utilize R in this degree plan are STAT 4397, STAT 4300, STAT 4306, STAT 4307, STAT 4309, and STAT 4310.

Mathematical Sciences Requirements (39-41 hours)

- MATH 2305 - Discrete Mathematical Structures (3 hours of this count in the common core)
- MATH 2401 - Calculus I
- MATH 2402 - Calculus II
- MATH 2307 - Linear Algebra
- MATH 3302 - Probability and Statistics
- STAT 3311 - Statistical Analysis for Natural Science
- STAT 4318 - SAS Fundamentals and Applications
- STAT 4397 - Statistical Computing
- STAT 4395 - Senior Project in Statistics (or a department approved W-course)

Figure 3 - Original Applied Statistics requirements (University of Houston-Downtown)

Of all the classes that utilize R in the Applied Statistics degree, only STAT 4397 is required. The rest of the classes that teaches R are electives (5 out of 7 classes) as shown in the figure below. It is important to note that a Statistics student must take at least 4 electives (12 hours). Thus, a typical Statistics student would take at least 3 classes using R during their course of study.

Figure 4 - Original Applied Statistics electives (University of Houston-Downtown)

Alternatively, below are the requirements for the Data Science degree. According Dr. Cahoy, the following classes utilize R in their curriculum: DATA 2401, STAT 3333, and STAT 4310. At first glance, the Data Science degree has a lot of more required classes that teaches using R compared to the Applied Statistics degree. One important difference between the original Statistics degree and the update Statistics and newly create Data Science degrees is that the newer degrees introduces students to R very early in their course of study. Specifically, the class DATA 2401 in the updated Statistics degree and the Data Science degree is a sophomore level class, whereas all classes teaching R in the original Applied Statistics degree are a senior level class (Cahoy).

Figure 5 - Data Science requirements (College of Sciences and Technology)

2. All degree improvements vs. Survey suggestions

The improvements on the original Statistics degree include more classes that teaches R (updated Statistics degree) and earlier introduction to R (both degrees). The students suggested that a manual made specific to their class and supplemental instruction sessions would be helpful (See figure below for students' thoughts). Although some of the needs of students are addressed by the new degrees' improvements, students' specific struggles such as syntax, leveling sessions and the necessary codes, are not addressed.
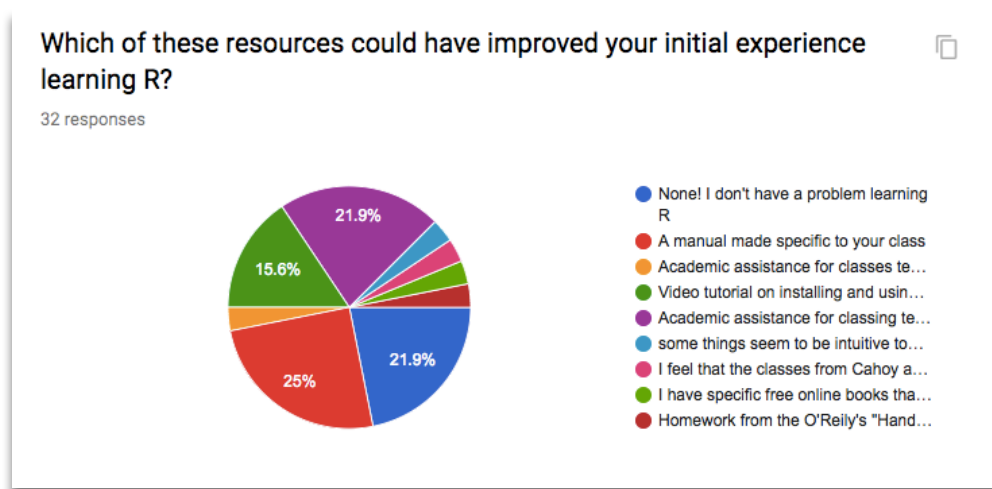
Figure 6 - Survey result 2

3. UHD vs. UH resources

The figure below is a screenshot of the available online resources for UH students who are taking MATH 2311 – Introduction to Probability and Statistics. From this screenshot, UH apparently has more resources than UHD. Particularly, in addition to the syllabus, R manual, and lecture notes that are provided by both institutions, UH also includes a tutoring service, videos using R, and a quick reference guide. Compared to UH, there are too few resources offered at UHD for teaching using R.

MATH 2311

Links to the Math 2311 Instructor's sites: Cathy Poliak   Thomas Weber   Matt Caputo

General Syllabus for Math 2311

Tutoring Schedule for Spring 2018

 LAUNCH Tutoring

Class information:

- **Homework Assignments**
- **How to Upload Homework in CourseWare/CASA**
- **Quizzes**
- **Videos Using TI-84**
- **Videos Using R**
- **Using R Studio:**   How to install R-Studio   R Download Site   R Studio Download Site   An Introduction to R (user manual)   Mosaic Reference
  Using R and Install MosaicData
  R-Studio Quick Reference Guide

- **Using the Library Virtual Commons to access R studio: Instructions**
- **Using the TI-83/84 Calculator:**   TI 83/84 Calculator Basics (videos)   MathBits   TI Guidebooks

Figure 7 - UH all resources statistics

Conclusion

Despite all the efforts towards improving the Statistics and Data Science degrees, many students still struggle with learning R. A more tangible solution must be developed in order to improve the quality of students' learning outcomes and the program's growth and retention rates.

Recommendation

Applied Statistics is a field of study that interpret information and analyze what happens around us. The Applied Statistics degree at UHD prepares students for areas requiring discrete mathematics and statistics (Career Development Center). The national average salary of a statistician is $85,160 (Department of Mathematics and Statistics). Data Science is an emerging discipline that combines mathematics, statistics, and computation to acquire actionable information from data. Students in the Data Science program at UHD will learn to collect, prepare, transform, and model data to make decisions (Career Development Center). The projects salary for data scientist is $127,760 as of September 27, 2017 (Department of Mathematics and Statistics). Since both of these degrees require students to work with R, it is essential to provide students with as many resources as possible to increase students' success and program's attractiveness.

From personal experience in the Applied Statistics program, I initially predicted that a simplified manual would be the solution. This manual would be shorter, thus, easier to navigate. Using such a manual, beginners would be able to forego unnecessary background information as presented in the comprehensive R manual. Problems such as syntax and having the correct code would be alleviated by this manual.

After my survey, interviews, and research, I realized there are multiple resources necessary to address all problems students face. However, creating and providing multiple resources will require a lot of capital that is not readily available. It is more sensible to recommend an immediate solution and advise on further research and future resources.

Therefore, I propose that the Department of Mathematics and Statistics create a reference guide.

Reference Guide Layout

As the name implies, a reference guide is a list of simplified items that are accessible for any given task. In the case of learning R, a reference guide would have all the basic codes, what they do, and how they function. For problems such as syntax and having the necessary codes, a reference guide is the most effective item. The figure below is the quick reference guide for UH students in MATH 2311. This reference guide only includes the basic codes for specific actions in R, such as "mean" and "standard deviation" of a data set. There are only a few pages in the reference guide, and extraneous information is nonexistent.

RStudio Quick Reference Guide — *Page 1 of 3*

The program may be used as a regular calculator.
- $+$ to add
- $-$ to subtract
- $*$ to multiply
- $/$ to divide
- $\wedge$ to raise to a power
- sqrt to square root; any other root, use a fractional exponent

To enter a data set
- c( )

The cursor will then appear inside the parenthesis and you'll enter the data set, separating each number with a comma. Lastly, hit enter.

Name a data set
- name=c( )

Mean of a data set
- mean(name of data set) or mean(enter the data set)

Median of a data set
- median(name of data set) or median(enter the data set)

Sort data
- sort(name of data set) or sort(enter the data set)

Variance of a data set
- var(name of data set) or var(enter the data set)

Standard Deviation of a data set
- sd(name of data set) or sd(enter the data set)

Five Number Summary
- fivenum(name of data set) or fivenum(enter data set)

Factorial
- factorial(number)

For permutations, use the factorial command.

Combination
- choose(n,r)

Binomial Distributions
- $P(X = k) = \text{dbinom}(k,n, p)$
- $P(X \le k) = \text{pbinom}(k,n,p)$
- $P(X > k) = 1 - \text{pbinom}(k,n,p)$

In the command, n = number of trials, k = number of successes and p = probability of success

Geometric Distributions
- $P(X = n) = \text{dgeom}(n - 1, p)$
- $P(X \le n) = \text{pgeom}(n - 1,p)$
- $P(X > n) = 1 - \text{pgeom}(n - 1,p)$

where n = nth trial and p = probability of success

Normal Distributions
- $P(X < b) = \text{pnorm}(b, \mu, \sigma)$
- $P(X > a) = 1 - \text{pnorm}(a, \mu, \sigma)$
- $P(a < X < b) = \text{pnorm}(b, \mu, \sigma) - \text{pnorm}(a, \mu, \sigma)$

If the random variable is the standard normal variable, then leave μ and σ blank.
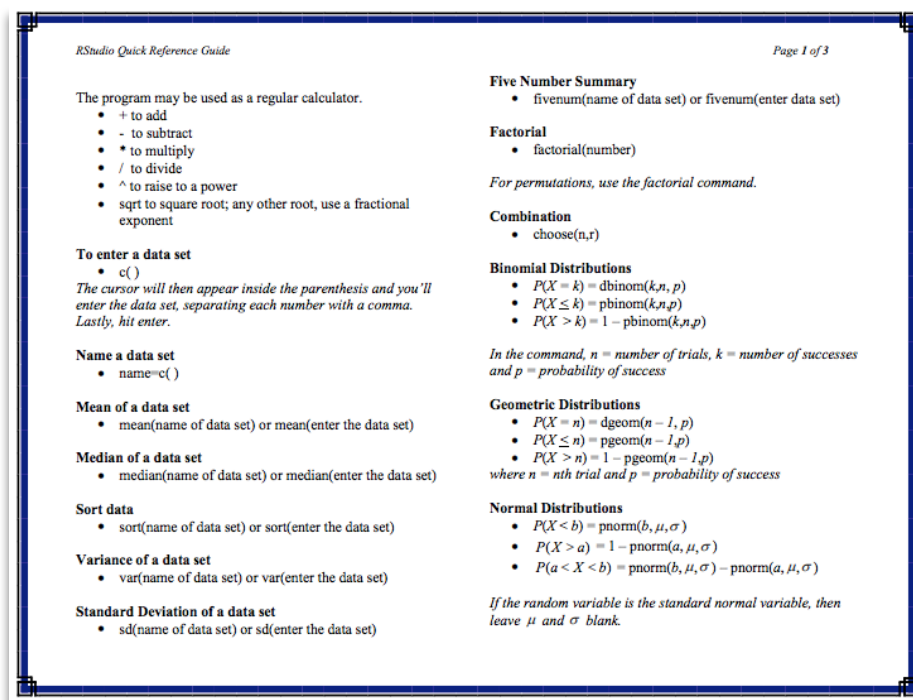
Figure 8 - UH Quick Reference Guide

Because UHD offers more classes with R, the UHD reference guide would have a few more pages than the UH reference guide. Like the UH reference guide, the UHD reference guide would only include the codes and its function. Additionally, an extra page on the reference guide could include all basic vocabulary with definitions. Thus, the UHD reference guide would combat all students' problems with syntax, vocabulary, and codes.

Creating Reference Guide

For the semester that this recommendation is implemented, professors would ask students to submit the codes and vocabulary they think are the most common and useful for their classes. The layout of these submissions should be similar to the UH reference guide, where the codes are attached until their functions. All vocabulary could be attached at the end for the beginning of the students' submissions. The professors would then mark which of the submitted codes are most appropriate for students to use.

For compilation, an outside contractor or possibly a student could be hired. Depending on the specifics of the contract and how many hours the process would take, the cost might be between $30 (1hr) to $320 (8hrs) based on basic copyediting rates (Editorial Freelancers Association). Alternatively, if someone from the department is assigned to complete this task, it could possibly cost the department nothing.

Recommendation for Further Improvements

In addition to syntax, vocabulary, and codes, other problems students face are getting more practice and gaining coding experience. These problems could be addressed with on-campus tutors and academic assistance services such as Supplemental Instruction (SI), the

Center for Mathematics and Statistics Support (C4MS$_2$) and the Collaborative Learning

Community Center (CLCC).

SI is a higher-education academic support program using peer-led sessions for students

that is proven to improve academic skills, grades, and student retention (Skoglund, Wall and

Kiene). C4MS$_2$ is a tutoring service that aim to develop students' ability to understand and apply

mathematical knowledge in math and statistics homework (University College). CLCC is a

tutoring and mentoring program for mathematics and computer science operated by the

Scholars Academy of UHD (College of Sciences and Technology). Although these services are

available on campus, they do not specifically offer help on R. If at least one of these services

could integrate R into their program, students would benefit immensely in their academic

careers.

Finally, due to the scope and length of this report, the above recommendation is not

comprehensive and could be improved with additional research. All data collected during the

construction of this report may be reused by anyone who intend to improve the problems

mentioned in the report.

Glossary

UH – University of Houston

UHD – University of Houston-Downtown

R – R Programming, a free programming language for statistical computing and graphics

Works Cited

Cahoy, Dexter. "New B.S. Applied Statistics Degree Requirements." Houston: University of
Houston-Downtown, 27 October 2018. Print.

Career Development Center. "Applied Statistics - What can I do with this major?" University of
Houston-Downtown, 29 10 2018. Print.

—. "Data Science - What can I do with this major?" University of Houston-Downtown, 29 10
2018. Print.

College of Sciences and Technology. *Bachelor of Science in Data Science*. 2 10 2018. University
of Houston-Downtown. 18 10 2018.
<https://www.uhd.edu/academics/sciences/mathematics-statistics/Pages/MS-BS-
DataScience.aspx>.

—. *UHD Scholars Academy - Tutors*. 14 4 2017. 29 10 2018.
<https://www.uhd.edu/academics/sciences/scholars/Pages/scholars-tutors.aspx>.

Department of Mathematics and Statistics. "Applied Statistics." UHD College of Sciences and
Technology, 29 10 2018. Print.

—. "Data Science." College of Sciences and Technology, 29 10 2018. Print.

Editorial Freelancers Association. *Editorial Freelancers Association*. 2018. Editorial Freelancers
Association. 27 10 2018. <https://www.the-efa.org/rates/>.

R Core Team. *An Introduction to R*. 02 07 2018. 29 10 2018. <https://cran.r-
project.org/doc/manuals/R-intro.html#Index-vectors>.

Skoglund, Kirk, Timothy Wall and David Kiene. "Impact of Supplemental Instruction

Participation on College Freshman Retention." *Learning Assistance Review (TLAR)* 23.1

(2018): 115.

University College. *Center for Mathematics and Statistics Support*. 5 10 2018. University of

Houston-Downtown. 29 10 2018. <https://www.uhd.edu/academics/university-

college/mss/Pages/default.aspx>.

University of Houston-Downtown. *Bachelor of Science in Applied Statistics*. 2018. 20 10 2018.

<http://catalog.uhd.edu/preview_program.php?catoid=9&poid=2348&returnto=736>.

Appendices

Appendix A: Survey Template

Did you learn R before taking a course at UHD? *
○ Yes
○ No

What is your most recent class using R programming at UHD? (Can be your current class) *
○ DATA 2401 - Data Science I
○ DATA 3401 - Data Science II
○ STAT 4390 - Statistical Inference
○ STAT 4397 - Statistical Computing
○ Other...

Is the above class the first class you have taken at UHD that teaches R? *
○ Yes
○ No

Which resource(s) were utilized when you FIRST STARTED learning R at UHD? *
└ Syllabus (has guide on how to install R)
└ R Manual (on R website)
└ Codes provided by professor
└ Textbook
└ Other...

Which of the above resources was most helpful? *
○ Syllabus
○ R Manual
○ Codes provided by professor
○ Textbook
○ Other...

Please describe difficulties learning R, if any. *
Long answer text

Which of these resources could have improved your initial experience learning R? *
○ None! I don't have a problem learning R
○ A manual made specific to your class
○ Academic assistance for classes teachi...
○ Video tutorial on installing and using R f...
○ Other...

Do you think knowing R well will help you become a better statistician / data analyst? *
○ Yes
○ No
○ I don't know

Who directed you to this survey? *
○ Dr. Abbott
○ Dr. Cahoy
○ Dr. King
○ Dr. Nguyen
○ Dr. Redl
○ Supplemental Instruction
○ Other...

Extra: Any other suggestions / comments?
Long answer text

Figure 9 - Survey template

Appendix B: Additional survey results

Did you learn R before taking a course at UHD?
32 responses

- Yes
- No

87.5%
12.5%

What is your most recent class using R programming at UHD? (Can be your current class)
32 responses

- DATA 2401 - Data Science I
- DATA 3401 - Data Science II
- STAT 4390 - Statistical Inference
- STAT 4397 - Statistical Computing
- math 3302
- STAT 4300

9.4%
18.8%
65.6%

Is the above class the first class you have taken at UHD that teaches R?
32 responses

- Yes
- No

25%
75%

Which resource(s) were utilized when you FIRST STARTED learning R at UHD?
32 responses

Resource	Count
Syllabus (has guide on how to install R)	15 (46.9%)
R Manual (on R website)	10 (31.3%)
Codes provided by professor	26 (81.3%)
Textbook	17 (53.1%)
Google, StackOverflow, etc stack-exchange, all of the FREE introdu...	1 (3.1%)
Many different tutorials online	1 (3.1%)
Practice	1 (3.1%)

Do you think knowing R well will help you become a better statistician / data analyst?
32 responses

- Yes
- No
- I don't know

84.4%
12.5%

Who directed you to this survey?
32 responses

- Dr. Abbott
- Dr. Cahoy
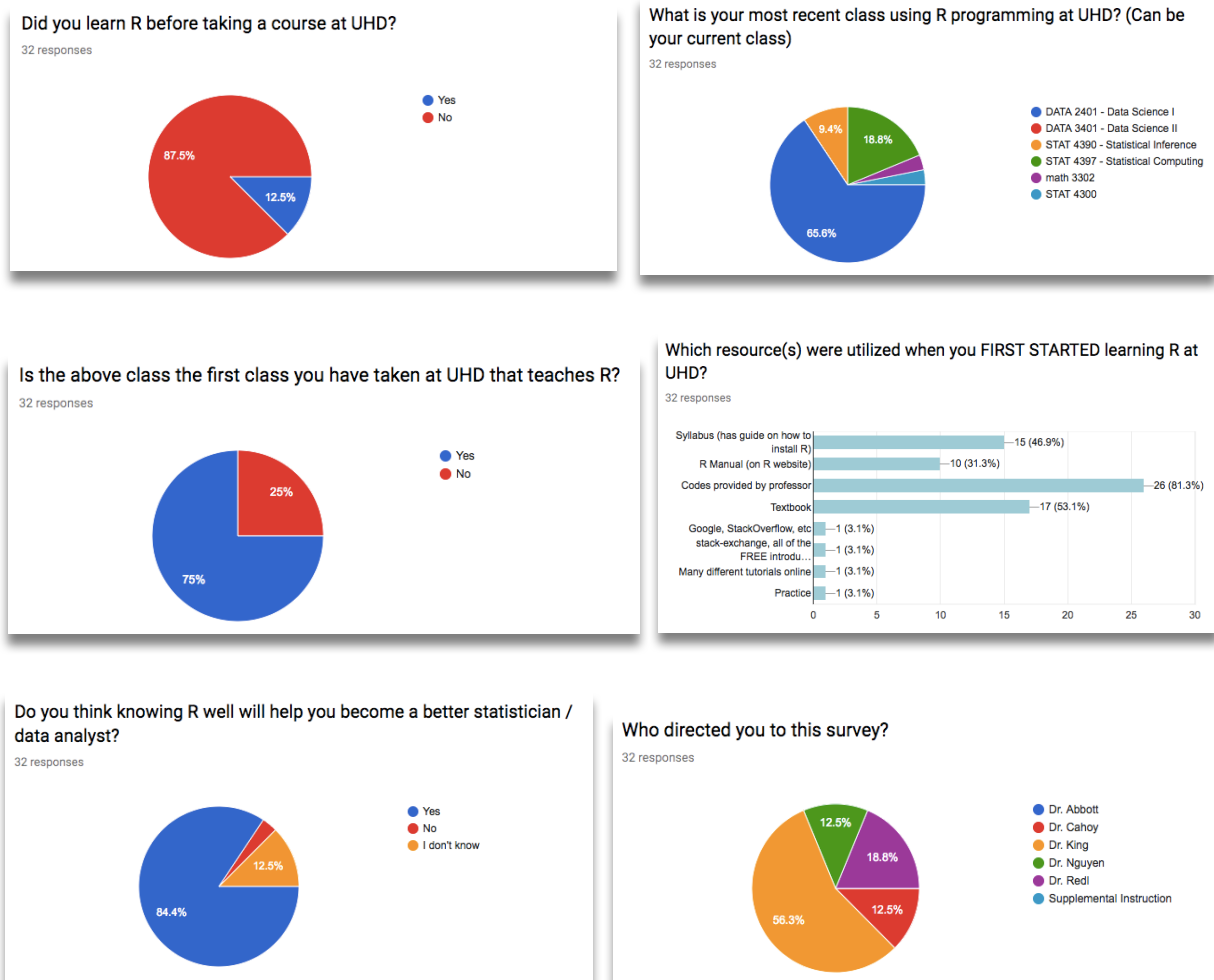- Dr. King
- Dr. Nguyen
- Dr. Redl
- Supplemental Instruction

12.5%
18.8%
12.5%
56.3%

Figure 10 - All additional survey results