Statistical Consulting – Senior Project

Cary Hernandez

Advisor Dr. Gary Hughes| Spring Quarter 2015|

Table of Contents

Introduction and Overview

The past two quarters I have spent working as a statistical consultant alongside Dr. Gary Hughes, Dr. John Walker, Dr. Steve Rein, and Professor Heather Smith. In Winter Quarter 2015, Dr. Hughes and Professor Smith were the lead consultants in the Statistical Consulting Services at Cal Poly while Spring Quarter 2015 brought in Dr. Rein and Dr. Walker. The whole team and I combined have worked on a widespread array of projects from all of the colleges here at Cal Poly. We have seen a total of 64 clients spanning 28 different departments. Personally, I have worked with 15 individual clients over the past two quarters and conducted at least 30 meetings. These meetings ranged from short 45-minute discussions of design of experiment to extensive two and a half hour walkthroughs of various analyses.

The Statistical Consulting Services at Cal Poly is designed such that two professors from the department each quarter are the lead consultants. The database I developed was made not only to explore the client demographics but also to aid in the continuity of situations in which a client will need help over consecutive quarters. This issue arises very frequently in which a client will meet with a consultant for help in design of experiment one quarter and then the client will request advice the next quarter in analyzing their data appropriately. However, the consultants the following quarter will be different than those from the quarter prior, thus forcing the consultant to backtrack what has been discussed already. To help alleviate this problem, the database has recorded meeting notes with each client for every meeting so that when a client informs the services that he or she has sought

consulting help in the past, any consultant can look up in the database what has been talked about already with this specific client and with what professor from the department.

This report's goal is to describe a couple intriguing case studies of master theses I have helped with and to give an overview of the database I created. The database can be used to help discover client demographics, explore the amount of time spent consulting, and search for other interesting aspects about our cliental information.

Case Study 1 – Trevor Curry

One of the most interesting experiments and clients I was fortunate to collaborate with was Trevor Curry, a kinesiology masters student, who was testing the effectiveness of a weighted suit used during exercise for his thesis. I cannot state what the name is of the company whom designed the suit due to a disclosure agreement with the company and Curry. However, the experiment's goal was to search for statistical evidence that the suit increases an individual's aerobic capacity when compared to people who exercised without a suit.

The six-week exercise study was conducted on 25 students from California Polytechnic State University San Luis Obispo ages 18 – 30 years old, although 4 students dropped out of the study for various reasons. The weighted suit was approximately 13 lbs. and was randomly assigned to the subjects and the remaining subjects made up the control group who worked out for 6 weeks without the suit. The subjects were kept from extensive physical activity the month prior to the exercise phase and also they abstained from all supplements. The exercise phase consisted of training on a treadmill three days a week while connected to machines that tracked things such as heart rate and blood pressure. The individuals' Body Mass Index (BMI), body composition, and leg strength were all recorded both before and after the six-week exercise phase but the most important variable recorded was maximal aerobic capacity (VO2 max). A person's V02 max is a measure of how well the body can move oxygen in and out of the system during the hardest point of exercise, also known as maximum capacity.

The company and Trevor Curry wanted to see if the V02 max difference from before the exercise phase to post exercise phase was significantly larger for those subjects wearing the suit during training. Dr. Rein, Trevor Curry, and myself decided a multiple regression analysis would be most appropriate here in order to take into account whether they were assigned the suit, the subject's pre BMI, their pre leg strength, and pre lean mass. These were taken into account because we discussed the differences in ability to increase one's aerobic capacity at various fitness levels. For example, for someone who is extremely fit already, there is only so much of an increase in aerobic capacity possible, also known as plateauing, when compared to someone who is out of shape and starting to work out within this study.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	20.565593	4.965147	4.14	0.0014*
Group[No]	0.8656172	0.730304	1.19	0.2588
Pre BMI	-0.100319	0.281217	-0.36	0.7275
Pre Lean Mass (kg)	-0.122819	0.059032	-2.08	0.0596
Pre Leg Press (lbs)	-0.004256	0.010376	-0.41	0.6889

Figure 1. Parameter estimates in JMP for suit analysis of aerobic capacity.

After taking all of these variables into account and using VO2 max difference as the response, we ended up not getting significant results for the suit (See Figure 1). Even more so, the mean of VO2 max difference for suit group was smaller than the change in VO2 max for those without the suit. The data was telling the complete opposite story than what Curry wished to tell. Explaining this and the positive coefficient on the suit parameter estimate was a difficult, yet rewarding conversation with Trevor Curry. In order to first explain that the data was telling the opposite story that what he desired, we looked at a few exploratory data analysis

items such as histograms and boxplots of the differences for both groups – suit and control. To help the conversation I told him that one cannot ensure to get statistically significant results, especially with an end sample size of 18 subjects, and also there were many other items not taken into account. I mentioned if a follow-up study were conducted it would be helpful to think of a method to track eating habits so it could be taken into account.

The company was very set on getting some statistical backing for their product and there was much interest in a larger study to be conducted so I introduced the idea of a power analysis. I explained the differences between prospective and retrospective power analyses and when to use which type accordingly. Curry was set on using a retrospective power analysis utilizing the data on the 18 subjects as a basis, which is displayed in Figure 2. He performed it in JMP and wanted to see the sample size needed to see significant results given our data. This was also another challenging conversation to describe the Least Significant Number in JMP, also known as LSN.05, will be telling the wrong story again. After this talk, I suggested a prospective power analysis as something to think of moving forward with investigating the amount of differences between the weighted suit group and the control.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	LSV0.05	LSN0.05	AdjPower0.05
Intercept	20.565593	4.965147	4.14	0.0014*	10.81813	8.52	0.9166
Group[No]	0.8656172	0.730304	1.19	0.2588	1.591195	49.14	0.0668
Pre BMI	-0.100319	0.281217	-0.36	0.7275	0.612719	515.61	0.0500
Pre Lean Mass (kg)	-0.122819	0.059032	-2.08	0.0596	0.12862	18.26	0.3182
Pre Leg Press (lbs)	-0.004256	0.010376	-0.41	0.6889	0.022608	390.63	0.0500

Figure 2. Trevor Curry's retrospective power analysis in JMP to inspect LSN.05

This project's most difficult aspect was being able to learn how to translate some of the more complicated statistical concepts to people that are not statistically inclined in any fashion. Explaining the p-values for each of the parameters in the regression model and also the concept of power was very challenging at first but became easier to do and the conversations more effective as time went on.

Case Study 2 – Natalie Rossington

Natalie Rossington, a biology masters student, was working on her thesis research that she carried out for approximately 6 months. Rossington was studying certain characteristics on flower's germination and flowering comparing two different species and multiple families within each species. The two different species of flowers she was testing were *jonesii* and *platyglossa* and their seeds were planted in grassy area, called grassland, and in a rocky, barren environment, named outcrop. She also planted *jonesii* in a third environment that was filled with invasive weeds so she can compare the *jonesii's* characteristics in the weeded environment to the more promising grassland.

She had three main project goals while she studied the germination, flowering, and viable seed counts of these plants. Her first goal was to see if there was a difference of the treatment factor of location and of the species when it came to the number of seeds that germinated after planting. Of the seeds that germinated, she wanted to discover whether there is a significant difference in the number that flowered when taking into account the species and treatment effect. Her second

project goal was to compare if there was a significant difference in number of viable seeds that were produced of the flowered plants depending on species and treatment. The last project goal investigated was seeing if there was a difference in survival time depending on treatment and species.

The first two project goals were pretty straightforward. Dr. Walker and I worked on this project extensively and decided upon using a general linear mixed model utilizing SAS's Proc Glimmix for the germination and flowering analyses. The mixed effect in these analyses was the nested mixed effect of the seed's family within species as each family corresponded to only a specific species, producing it's own subgroup. Dr. Walker and I also chose a general regression model for the viable seed count by treatment and species. We used the square root of viable count, as there was an issue with the fanning in the residual vs. predicted residual plot. The main issue came with the survival analysis of the flowered plants with the treatment and species factors.

While looking at the days survived data in the middle of a consulting meeting, I noticed a pattern - the time points were clumped together with gaps between the groups. I asked Natalie during the meeting a few simple probing questions to get her to explain how she recorded days survived which ended up revealing the explanation of the pattern I saw. She went every Saturday or sometimes Sunday to the sites where she planted these seeds to observe whether the plant had died since her last visit a week prior. Therefore, this was causing interval censoring. I had to explain this to her and inform her that we needed to investigate this further as I have never performed a mixed effect interval censored problem before. After a lot of

extensive research, we were unable to successfully include the nested effect of family and species but we did do a regression using the observed date of the flower's death as a response which did not produce significant results. We then tried a Cox proportional hazards model with the treatment and species as the explanatory variables, which resulted in significant results for treatment but not for species.

The main question within the survival analysis was whether the species survival length was different because intuition already claimed that the locations should be different. Although we did not get significant results, we performed analyses to the best of our capabilities trying to include as much as we could for Natalie Rossington's thesis.

What interested me the most working on this thesis was the amount of statistical applications applied to this one study. Dr. Walker, Natalie Rossington, and I performed a wide range of analyses all on the same dataset all answering different questions so that we could reveal more and more information and get a better idea of what the study can and cannot conclude.

Statistical Consulting Services Database Overview

The database was created to track cliental information. I created a Google Form for data input, which then outputs its responses to a Google Sheet. The reason for this was for the simple accessibility from anyone in the department with the provided password and also for it's ease of use in changing past responses. This capability of changing recorded responses was essential to those who needed to go back in and add information for a given meeting's notes field for more extensive documentation if new information arose from the client.

The Google Form, shown in Figure 3, was made so that each submission of the form will correspond to a unique meeting. This made it easier to find out what consultants were at each meeting, when the meetings took place, and the frequency of meetings per client. The variables that are tracked in this database are displayed in Table 1.

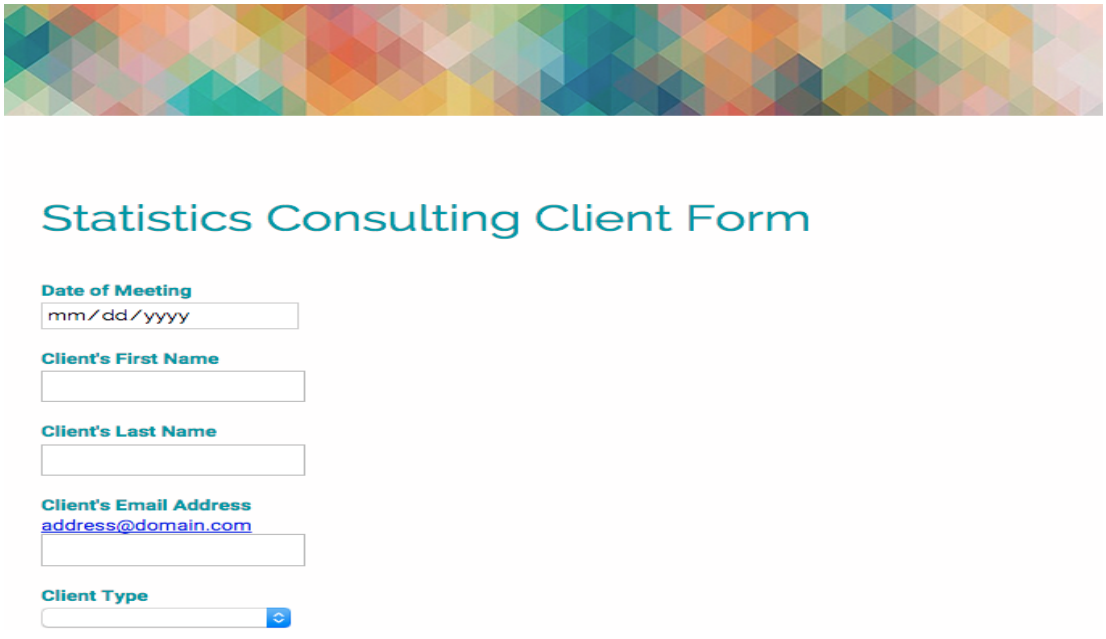
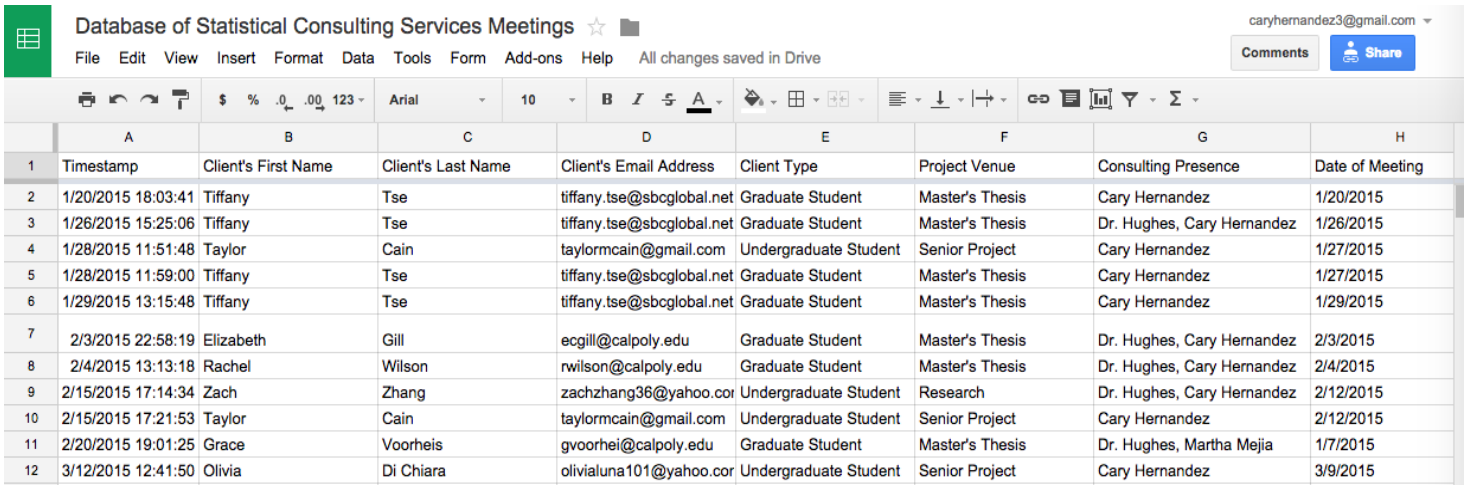


Figure 3. A portion of the Google Form to collect cliental information and meeting notes.

Recorded Variables in Database and Descriptions	
Date of Meeting	The date the meeting occurred
First Name	The client's first name
Last Name	The client's last name
Client's Email Address	The email address for contact information purposes
Client Type	Whether the client was an undergraduate, graduate, faculty professor, or a member of Cal Poly's staff
Project Venue	Publication type of project (i.e. master's thesis, research)
Consulting Presence	Which consultants attended the meeting
Project Description	Description or title of project
Meeting Notes	Documentation of what was covered during the meeting
Preparation Time	Time spent preparing for the meeting
Duration	Time spent during the consulting meeting
Associated College	The client's associated college
Associated Department	The client's associated department

Table 1. List of variables recorded in database.



Database of Statistical Consulting Services Meetings

caryhernandez3@gmail.com

File Edit View Insert Format Data Tools Form Add-ons Help All changes saved in Drive

Comments Share

	A	B	C	D	E	F	G	H
1	Timestamp	Client's First Name	Client's Last Name	Client's Email Address	Client Type	Project Venue	Consulting Presence	Date of Meeting
2	1/20/2015 18:03:41	Tiffany	Tse	tiffany.tse@sbcglobal.net	Graduate Student	Master's Thesis	Cary Hernandez	1/20/2015
3	1/26/2015 15:25:06	Tiffany	Tse	tiffany.tse@sbcglobal.net	Graduate Student	Master's Thesis	Dr. Hughes, Cary Hernandez	1/26/2015
4	1/28/2015 11:51:48	Taylor	Cain	taylormcain@gmail.com	Undergraduate Student	Senior Project	Cary Hernandez	1/27/2015
5	1/28/2015 11:59:00	Tiffany	Tse	tiffany.tse@sbcglobal.net	Graduate Student	Master's Thesis	Cary Hernandez	1/27/2015
6	1/29/2015 13:15:48	Tiffany	Tse	tiffany.tse@sbcglobal.net	Graduate Student	Master's Thesis	Cary Hernandez	1/29/2015
7	2/3/2015 22:58:19	Elizabeth	Gill	ecgill@calpoly.edu	Graduate Student	Master's Thesis	Dr. Hughes, Cary Hernandez	2/3/2015
8	2/4/2015 13:13:18	Rachel	Wilson	rwilson@calpoly.edu	Graduate Student	Master's Thesis	Dr. Hughes, Cary Hernandez	2/4/2015
9	2/15/2015 17:14:34	Zach	Zhang	zachzhang36@yahoo.cor	Undergraduate Student	Research	Dr. Hughes, Cary Hernandez	2/12/2015
10	2/15/2015 17:21:53	Taylor	Cain	taylormcain@gmail.com	Undergraduate Student	Senior Project	Cary Hernandez	2/12/2015
11	2/20/2015 19:01:25	Grace	Voorheis	gvoorhei@calpoly.edu	Graduate Student	Master's Thesis	Dr. Hughes, Martha Mejia	1/7/2015
12	3/12/2015 12:41:50	Olivia	Di Chiara	olivialuna101@yahoo.cor	Undergraduate Student	Senior Project	Cary Hernandez	3/9/2015

Figure 4. Preview of Database in Google Sheet form with each observation being a meeting.

Figure 4 is a snapshot of the recorded responses from the Google Form. This makes it easy to change any one person's information due to newly acquired information or spelling errors and typos. There are two important things to note. First, are the data used for this project was data pulled from the end of Spring Quarter Week 8 totaling 19 weeks of data from winter and spring quarters combined. Second, the database is not complete with every single meeting accounted for and recorded. This is due to potential forgotten meetings and also missing recorded responses in Spring Quarter.

The dataset was then exported as an Excel file from the Google Sheet and inputted into SAS Studio 9.0. Within SAS, there was a little bit of data cleaning and a new variable of Client's Name was created as a concatenation of First Name and Last Name and sorted the dataset by this new variable which was performed to subset the dataset. After sorting the dataset by name, I deleted all meetings that were anything other than the initial meeting with a client so that we are left with one observation per client. This made exploratory analysis of cliental demographics

possible. I then transferred this SAS dataset into R to do some visual exploratory graphics using ggplot2.

The first thing I looked at was the client's type to see whether the Statistical Consulting Services was spending more time with undergraduates, graduate students, faculty professors, or staff members. As can be seen in Figure 3, a large majority of our time in the consulting services is spent with graduate students working on their theses as they make up about 56% of our clients by number these past two quarters. The undergraduates make up 21 of our total 64 clients and there were seven faculty members and one client that was from the Cal Poly staff.

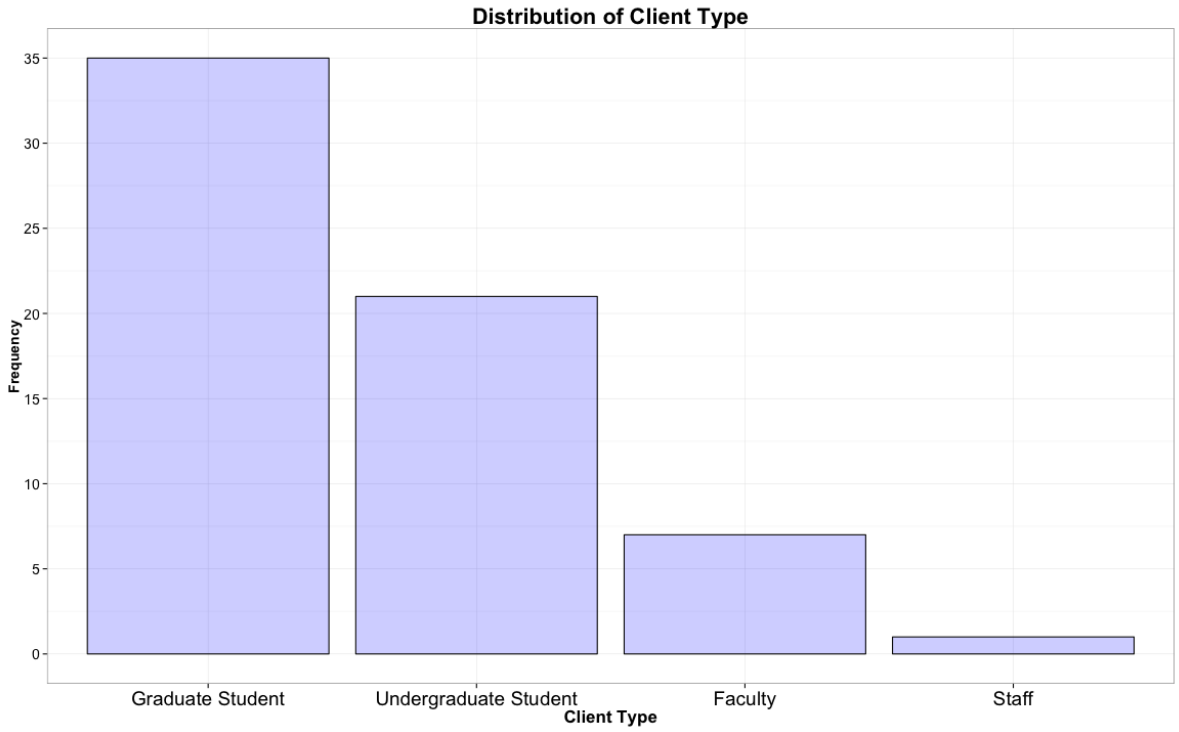


Figure 5. Bar-plot of distribution of Client Type.

I next inspected the client's associated college to see which college of Cal Poly requests the most amount of statistical advice through our services. Figure 4 shows that the College of Agriculture, Food & Environmental Sciences (CAFES) is the college that we consult with the most followed by the College of Science and Mathematics (COSAM). The majority of the COSAM clients are biology majors while a lot of the CAFES clients are from Dairy Science, Animal Science, and Fruit Science departments. This means consultants with a background in agricultural studies may be more capable with these clients.

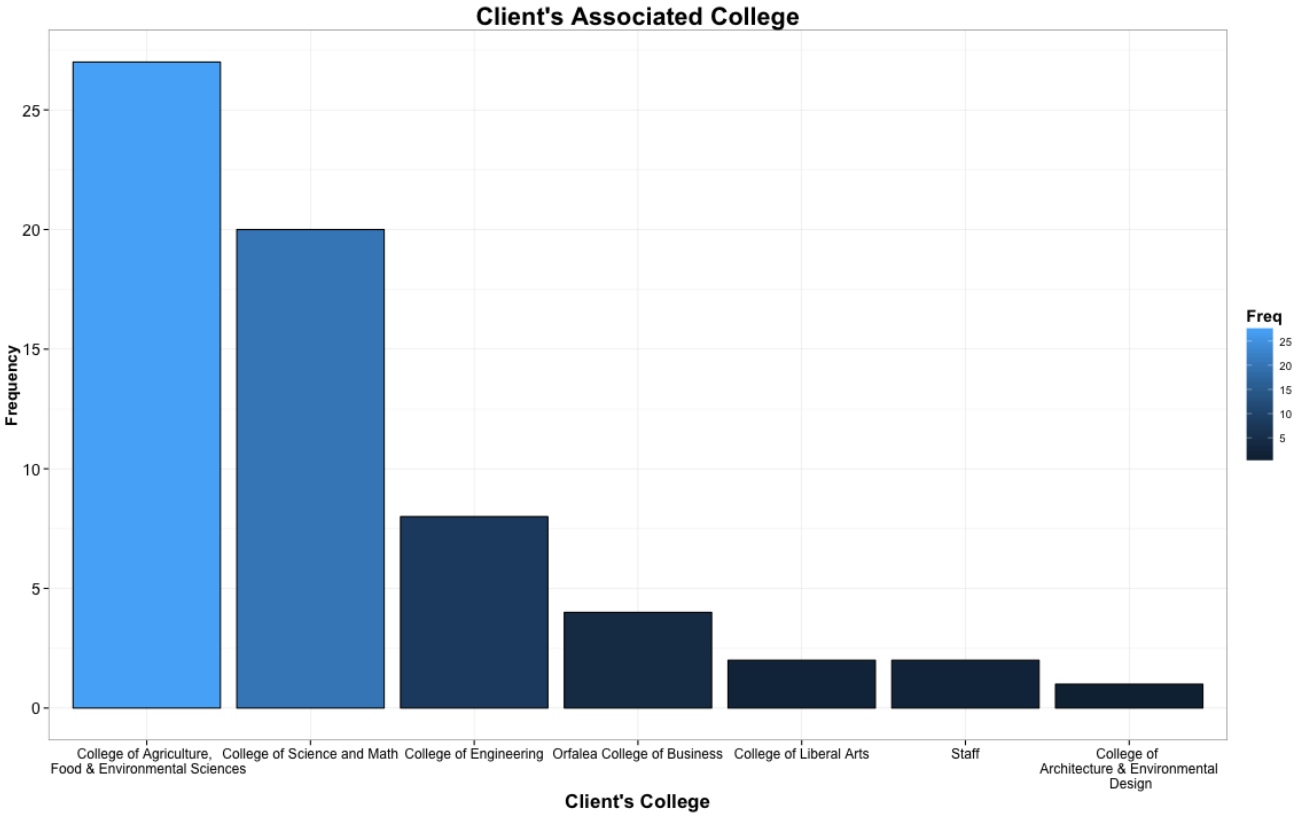


Figure 6. Bar-plot of frequency of Client's College.

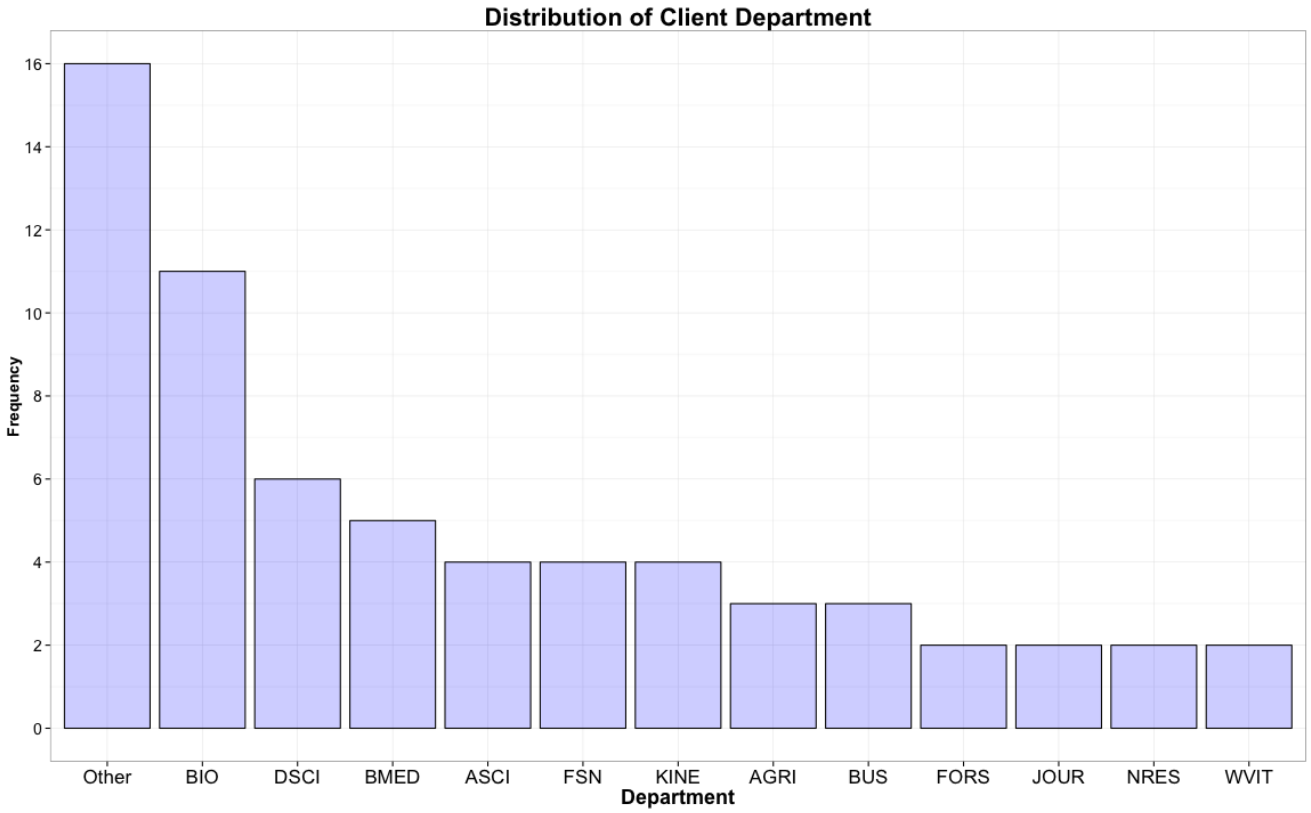


Figure 7. Distribution of Client's Associated Department. Note: Other category combines departments with frequency of one.

The graph in Figure 7 displays the client's associated department and their frequencies. The department that consists of the most clients is Biology from the College of Science and Mathematics. These projects spanned a wide range of topics, from seeing if a specific gene was associated with a certain disease symptom to helping with experimental design of testing if kangaroo rats from different habitats went quicker through veldt grass. The second most frequent department is Dairy Science followed by Biomedical Engineering. This bar-plot is eye opening because it is possible to see which departments have the most statistical applications within their field of study.

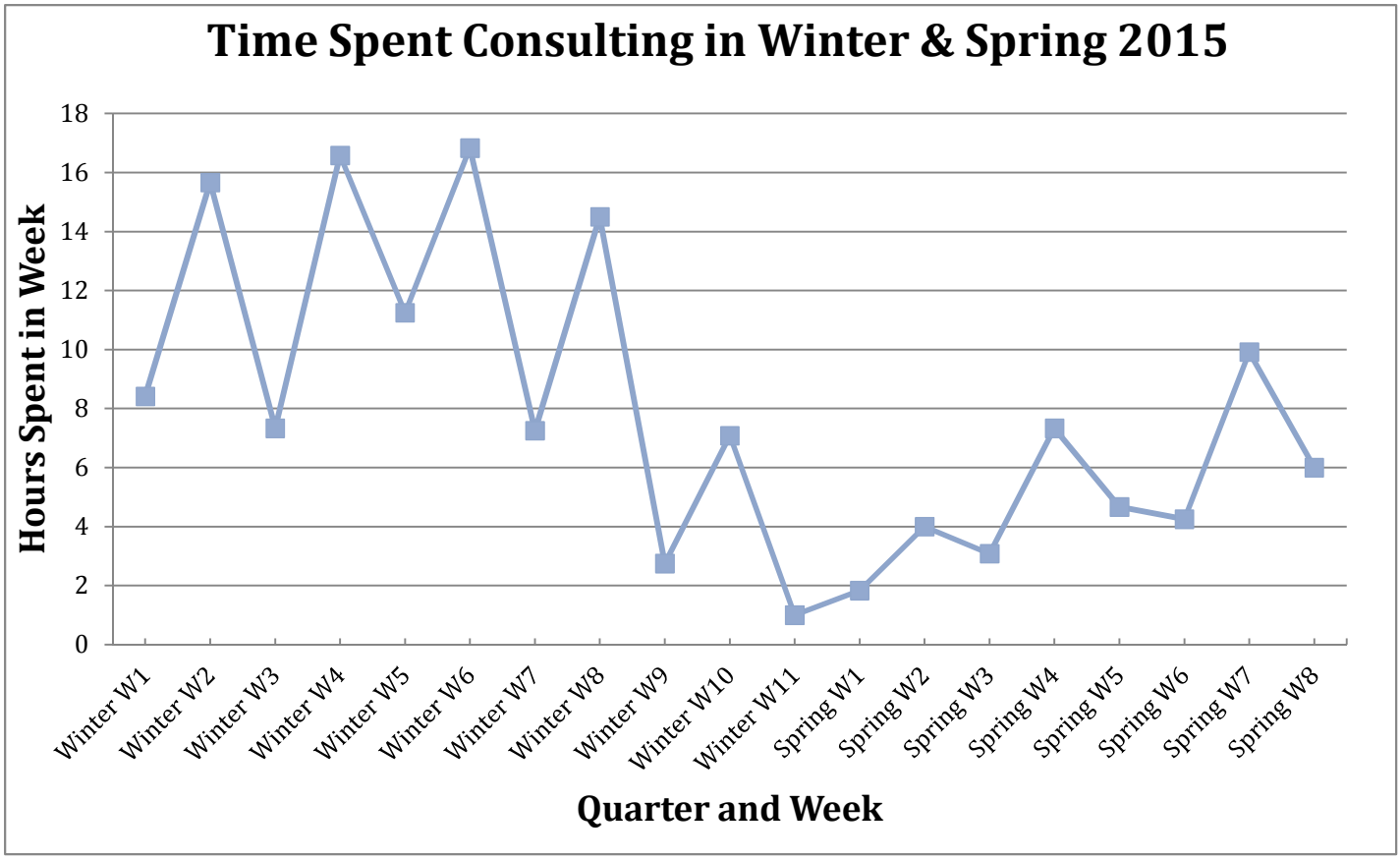


Figure 8. Time Progression of Hours Spent Consulting in Winter and Spring 2015

I also wanted to explore how much time, in hours, the other consultants and myself were spending on a weekly basis throughout the quarters (See Figure 8). Something to note especially here that I already mentioned is that we have a missing data records for Spring Quarter from one of the consultants. This can explain some of the decrease in time spent during Winter Quarter when compared to Spring 2015. It is also possible to observe that the number of hours spent during Winter Quarter fluctuated around 12 hours/week until Week 9. These fluctuations could be attributed to most follow-up meetings being scheduled about a week and a half apart from each other.

There is an immense decrease from approximately 14 hours spent in Winter Week 8 down to 3 hours in Week 9. I believe this is due to the fact that consultants' time availability at the end of a quarter becomes more limited as does availability for an undergraduate and graduate student whom is preparing for finals and projects.

Conclusion and Takeaways

During my first quarter at Cal Poly San Luis Obispo as a statistics major I sat in on a presentation from Professor Smith about her time as a statistical consultant. From that moment on, I was set on becoming a statistical consultant after graduation, wanting to develop the ability to work on a wide variety of projects with many different applications. After these past four years and especially these last two quarters as a statistical consultant, I have confirmed my goal to become a statistical consultant after the academia life.

The idea of traveling, meeting new people, and working on new challenges and tasks every project is extremely interesting to me. I believe that is what keeps life exciting - not knowing what is around the corner and always being on one's toes. These past two quarters I have been able to experience a little taste of this line of work and I cannot imagine a better, more enjoyable senior project for me.

From what I worked on, there are many various stages of statistical consulting advice that is required within the Statistical Consulting Services. There is a great deal of design of experiments projects but also a great number of analysis selection and interpretation requests. Even more so, there was a tremendous work load to be tackled. With the incomplete time spent data in the database, it is difficult to fully comprehend the amount of work that is being executed. However, it is still possible to get a roughly accurate idea of how much is spent when looking at the time spent in Winter 2015. Personally, I remember spending about 10 hours with one specific client in two consecutive weeks due to the client's time sensitivity of their finished thesis.

There is a lot of work required from the Statistical Consulting Services and the addition of one or two more upperclassmen statistics students would be very useful. The addition will not only alleviate some of the work but also give the opportunity for not one but two statistics students to experience what I did – assuming a senior statistics student will be aiding the services already. The experience of utilizing and applying statistical knowledge to various fields of applications is very valuable and extremely attractive to employers and I would highly suggest it to any statistics student.

Potential Next Steps

There are four prominent things that I would like to add into the database if I had more time for this project. These four items would make the analysis of the consulting database more efficient, accurate, and take more things into account while not making the process too burdensome.

First, I would like to somehow account for the time spent doing analysis from a client who has emailed a question over. If the consultant and client communicate efficiently via email and there is some analysis done on the consultant's end without an upcoming meeting, this time spent doing analysis and communicating with the client may never be tracked. This has happened a few times these past two quarters and including such data could add to the comprehensiveness of the database.

Not all clients who came to the Statistical Consulting Services were by themselves. A few times a client came representing a group's senior project or a couple of students would come at the same time. During the past two quarters, to insert these types of projects we would just agree on one individual's name to put into the database although we would be advising multiple people. I would enjoy attempting to take this into account somehow but still having all people linked to the corresponding project.

The way the database was reduced so that each observation was a unique client was using each client's full name as the identification. If there were two individuals with the same name, this would have deleted the second client by just coincidently having the same name – although I checked for this in the current dataset. As the dataset grows, we might want to consider using a combination of

random characters and digits or adding number(s) to the end of the client's full name to help better identify different clients with the same name. Searching by email may also work, assuming the addresses are input into the database in a consistent manner.

Finally, there were some issues with exporting the Google Sheet into Excel and having it turn into nice, clean dates. I attempted to do an Excel to SAS date conversion but nothing seemed to work. Therefore, my time series plot of hours spent consulting as the quarters progressed was conducted in Excel and not SAS or R. If I had more time I would spend more effort getting this issue resolved for whoever does further analysis on the database.

Appendix

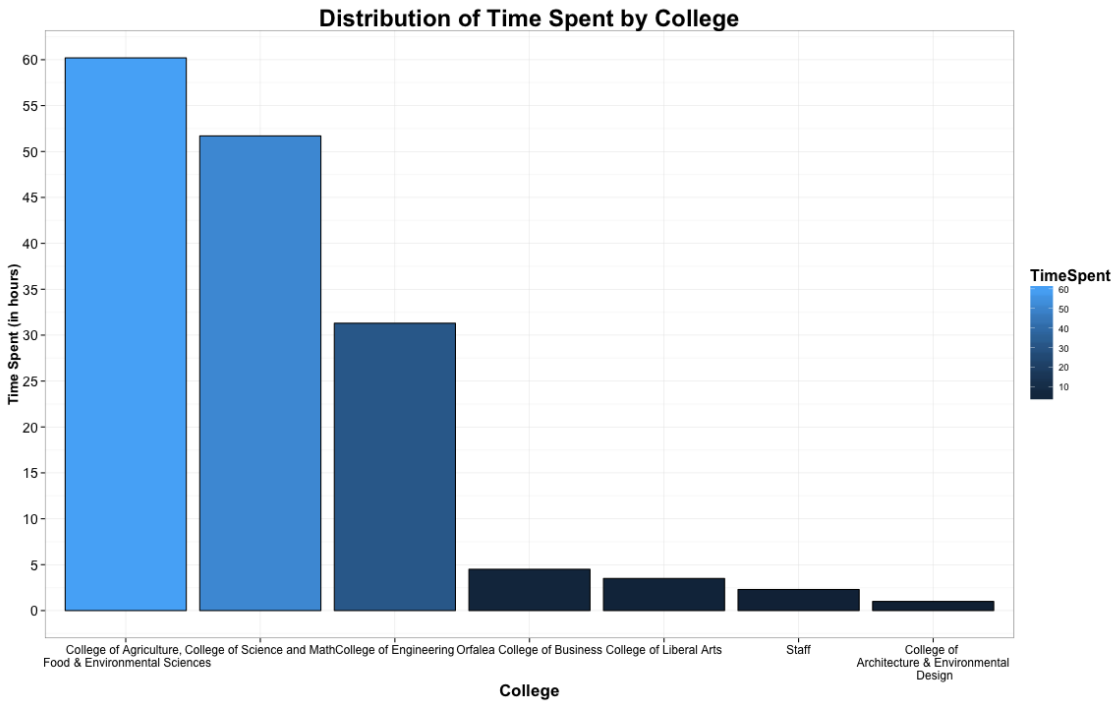


Figure 9. Distribution of time spent in hours by associated college.

The graph in Figure 9 corresponds with the Distribution of Clients by Associated College Figure 4. This graph however shows the extent of the amount of time spent within each college. I made this graph in case there were fewer clients in a specific college were requiring more time than more clients in a different college in other words, to inspect the possibility of fewer clients but more extensive and complicated projects.

Code for SAS:


```
1
2
3  libname fold '/folders/myfolders/';
4  /* Below just corrects all the College of Science and Math to be spelled correctly. Also concatenates the
5  first name and last name with each other
6  -- Maybe someone typed in a few college names without clicking the selection*/
7
8  data fold.ConsultwName; set fold.CResp;
9
10 if Client_s_Associated_College = 'College of Science and Mathematics' then
11     Client_s_Associated_College = 'College of Science and Math';
12 ClientName = CATX(' ',Client_s_First_Name,Client_s_Last_Name);
13
14 drop Client_s_First_Name Client_s_Last_Name;
15 run;
16
17 proc print data=fold.ConsultwName;
18 run;
19
20 * Shows a table of all meetings by frequency of college;
21
22 proc freq data=fold.ConsultwName;
23 tables Client_s_Associated_College;
24 run;
25 * Sorts the dataset by client name to get to subsetting the dataset to one observation per client
26  - This is done in order to get accurate client demographics distrubtions and avoid double counting;
27 proc sort data=fold.ConsultwName out=fold.ConsultSort;
28 by ClientName;
29 run;
30
31
32 * This below code gives us the resulting dataset of one observation per client.
33     NOTE: Provost and Student Affairs Clients Associated College is changed to Staff;
34 data fold.ConsultSort2; set fold.ConsultSort;
35 by ClientName;
36
37 if Client_s_Associated_College = 'Provost' then Client_s_Associated_College = 'Staff';
38 if Client_s_Associated_College = 'Student Affairs' then Client_s_Associated_College = 'Staff';
39 if Client_s_Associated_College = '' then delete;
40 if first.ClientName then output;
41 run;
42
43 proc print data=fold.ConsultSort2;
44 run;
45
46 * Gives us a frequency table by College - Each client only counted once;
47 proc freq data=fold.ConsultSort2;
48 tables Client_s_Associated_College;
49 run;
50
51 * Makes the 'Other' category for Frequency by Department Plot;
52 data fold.consult3; set fold.consultsort2;
53 if Client_s_Associated_Department = 'Provost' then Client_s_Associated_Department = 'Other';
54 if Client_s_Associated_Department = 'University Housing' then Client_s_Associated_Department = 'Other';
55 if Client_s_Associated_Department = 'PHYS' then Client_s_Associated_Department = 'Other';
56 if Client_s_Associated_Department = 'MATE' then Client_s_Associated_Department = 'Other';
57 if Client_s_Associated_Department = 'SS' then Client_s_Associated_Department = 'Other';
58 if Client_s_Associated_Department = 'NR' then Client_s_Associated_Department = 'Other';
59 if Client_s_Associated_Department = 'LA' then Client_s_Associated_Department = 'Other';
60 if Client_s_Associated_Department = 'IME' then Client_s_Associated_Department = 'Other';
61 if Client_s_Associated_Department = 'IT' then Client_s_Associated_Department = 'Other';
62 if Client_s_Associated_Department = 'EE' then Client_s_Associated_Department = 'Other';
63 if Client_s_Associated_Department = 'EDUC' then Client_s_Associated_Department = 'Other';
64 if Client_s_Associated_Department = 'CSC' then Client_s_Associated_Department = 'Other';
65 if Client_s_Associated_Department = 'CPE' then Client_s_Associated_Department = 'Other';
66 if Client_s_Associated_Department = 'CHEM' then Client_s_Associated_Department = 'Other';
67 if Client_s_Associated_Department = 'AEPS' then Client_s_Associated_Department = 'Other';
68 if Client_s_Associated_Department = 'HCS' then Client_s_Associated_Department = 'Other';
69 run;
70
71 proc print data=fold.consultsort2; run;
72
73 /* The below two datasets are subsets of the database to find number of unique clients I have met with (15)
74 and number of meetings I have reported (30)
75     Remember: These numbers are the minimum value as every consultant is not 100% accurate with
76 inputting meetings into the database as it is so new. */
```

```
77
78 data fold.consult_caryattend; set fold.consultsort2;
```



```
79 if index(Consulting_Presence, 'Cary Hernandez') ge 1 then output;
80 run;
81
82
83 data fold.consult_carymeet; set fold.ConsultSort;
84 if index(Consulting_Presence, 'Cary Hernandez') ge 1 then output;
85 run;
```

Code for R:

```
# setwd("~/Desktop/Spring 2015/Consulting")

## Imports the dataset into R for graphics
Consulting <- read.csv("~/Desktop/SPRING 2015/Consulting/Senior
Project/consultsort2.dat")
head(Consulting)

library(ggplot2)


# Gets a Barplot of Frequencies by Each College Sorted Descending.


freq <- table(Consulting$Client_s_Associated_College)

freq <- as.data.frame(table(Consulting$Client_s_Associated_College))

freq

freq[,1] <- c("College of Agriculture, \n Food & Environmental
Sciences",
              "College of \n Architecture & Environmental \n Design",
              "College of Engineering",
              "College of Liberal Arts",
              "College of Science and Math",
              "Orfalea College of Business",
              "Staff")

freq$Var1 <- with(freq, factor(freq$Var1, levels=freq[order(-Freq),
]$Var1))

## CLIENT COLLEGE DISTRIBUTION


college <- ggplot(data=freq, aes(x = Var1, y=Freq)) +
  geom_bar(stat="identity",breaks=seq(20, 50, by = 2), col="black",
           aes(fill=Freq)) + theme_bw() +
scale_y_continuous(breaks=seq(0,50,5))

college + labs(title="Client's Associated College") + xlab("Client's
College") + ylab("Frequency") + theme(title =element_text(size=18,
face='bold'),

axis.text.x = element_text(size=12),

axis.title.x = element_text(size=17, face="bold"),

axis.text.y = element_text(size=14),
```

```
axis.title.y = element_text(size=14, face="bold"))

### TIME SPENT ---- # Outputs the total time spent in meetings and
during prep time
ConsultFull <- read.csv("~/Desktop/SPRING 2015/Consulting/Senior
Project/consultsort.dat")



HoursSpent_PrepDuration <-
(sum(ConsultFull$Preparation_Time_minutes_in_dig) +

sum(ConsultFull$Duration_minutes_in_digits_))/60
HoursSpent_PrepDuration


freqTime <- table(ConsultFull$Client_s_Associated_College)

freqTime <- as.data.frame(table(Consulting$Client_s_Associated_College))
# 155.75 Hours spent

collegetime <- ggplot(data=ConsultFull, aes(x = , y=Freq)) +
  geom_bar(stat="identity",breaks=seq(20, 50, by = 2), col="black",
           aes(fill=Freq)) + theme_bw()

college + labs(title="Client Associated College") + xlab("Client's
College") + ylab("Frequency") + theme(title =element_text(size=18,
face='bold'),

axis.text.x = element_text(size=10),

axis.title.x = element_text(size=17, face="bold"),

axis.text.y = element_text(size=14),

axis.title.y = element_text(size=14, face="bold"))



### CLIENT TYPE ordered by descending
## Outputs a barplot of Client Type - Graudate, Undergraduate, Faculty
## Shows a lot of our time is helping students with their theses
freqType <- table(Consulting$Client_Type)

freqType <- as.data.frame(table(Consulting$Client_Type))

freqType$Var1 <- with(freqType, factor(freqType$Var1,
levels=freqType[order(-Freq), ]$Var1))
```

```
Type <- ggplot(data=freqType, aes(x= Var1, y=Freq)) +
              geom_bar(stat="identity",breaks=seq(20, 50, by = 2),
                 col="black",
                 fill="blue",
                 alpha = .2) +
            labs(title="Distribution of Client Type") +
            theme_bw() + xlab("Client Type") + ylab("Frequency") +
scale_y_continuous(breaks=seq(0,50,5))

Type + theme(title =element_text(size=18, face='bold'),
             axis.text.x = element_text(size=19),
             axis.title.x = element_text(size=17, face="bold"),
             axis.text.y = element_text(size=14),
             axis.title.y = element_text(size=14, face="bold"))

### DEPARTMENT/MAJOR DISTRIBUTION
# Inputs dataset with the frequency of 1 dept's deleted
ConsultDept <- read.csv("~/Desktop/SPRING 2015/Consulting/Senior
Project/consult3.dat")

ConsultDept$Client_s_Associated_Department

freqDept <- table(ConsultDept$Client_s_Associated_Department)

freqDept <-
as.data.frame(table(ConsultDept$Client_s_Associated_Department))

freqDept$Var1 <- with(freqDept, factor(freqDept$Var1,
levels=freqDept[order(-Freq), ]$Var1))

## Will need to adjust the max y axis once more data is filled in.
## Will also need to decide the other category once more data is
inputted.
Dept <- ggplot(data=freqDept, aes(x= Var1, y=Freq)) +
  geom_bar(stat="identity",breaks=seq(20, 50, by = 2),
                 col="black",
                 fill="blue",
                 alpha = .2) +
  labs(title="Distribution of Client Department") +
  theme_bw() + xlab("Department") + ylab("Frequency") +
scale_y_continuous(breaks=seq(0,50,2))

Dept  + theme(title =element_text(size=18, face='bold'),
              axis.text.x = element_text(size=17),
               axis.title.x = element_text(size=18, face="bold"),
               axis.text.y = element_text(size=14),
              axis.title.y = element_text(size=14, face="bold"))
```

```
freqDept

# Other category in this plot is Departments with frequency of 1

# Checks levels of the Departments to see if any are wrong -
levels(freqDept$Var1)


# Department by Department Time spent
# College of Agriculture, Food & Environmental Sciences 1720 or 60.2 hrs
# College of Architecture & Environmental Design 1 hour
# College of Engineering 530 31.3 hrs
# College of Liberal Arts 3.5 Hrs
# COSAM 51.7 hrs
# Business 4.5Hrs
# Staff 2.33
deptdept <- data.frame(College=c("College of Agriculture, \n Food &
Environmental Sciences",
                                 "College of \n Architecture &
Environmental \n Design",
                                 "College of Engineering",
                                 "College of Liberal Arts",
                                 "College of Science and Math",
                                "Orfalea College of Business",
                                "Staff"),
TimeSpent=c(60.2,1,31.3,3.5,51.7,4.5,2.3))
deptdept



deptdept$College <- with(deptdept, factor(deptdept$College,
levels=deptdept[order(-TimeSpent), ]$College))

## Will need to adjust the max y axis once more data is filled in.
## Will also need to decide the other category once more data is
inputted.
DeptT <- ggplot(data=deptdept, aes(x= College, y=TimeSpent)) +
  geom_bar(stat="identity",breaks=seq(20, 50, by = 2),
           col="black",
           fill="blue",
           alpha = .2) +
  labs(title="Distribution of Time Spent by College") +
  theme_bw() + xlab("College") + ylab("Time Spent (in hours)") +
scale_y_continuous(breaks=seq(0,80,5))


DeptT2 <- ggplot(data=deptdept, aes(x= College, y=TimeSpent)) +
  geom_bar(stat="identity",breaks=seq(20, 50, by = 2),
           col="black",
           aes(fill=TimeSpent)) + theme_bw() +
```

```
  labs(title="Distribution of Time Spent by College") + xlab("College")
+ ylab("Time Spent (in hours)") + scale_y_continuous(breaks=seq(0,80,5))

DeptT2  + theme(title =element_text(size=20, face='bold'),
                axis.text.x = element_text(size=11),
              axis.title.x = element_text(size=17, face="bold"),
              axis.text.y = element_text(size=14),
              axis.title.y = element_text(size=14, face="bold"))
```