# Appendix A. Survey Questionnaire

**Maynooth Post-Grad Survey 2019**

**Student Number \_\_\_\_\_\_\_\_\_ Age \_\_\_ Gender** [M/F] **\_\_\_**

**Course** Applied Computer Science \_\_\_\_ Data Sc.\_\_\_\_ Software Eng/Dev. \_\_\_\_ Others(Name) \_\_\_\_\_\_\_\_

**Level** Masters \_\_\_\_ HDip \_\_\_\_

**No. Of Modules** [Current Semester] **\_\_\_\_ Lecture Attendance** [Weekly %] **\_\_\_**

**Hours spent per Week** [On Average]**:**

[Please write 0 where not relevant]

Lectures / Labs \_\_\_\_ Essays/ Assignments \_\_\_\_

Self-Study \_\_\_\_ Paid Job or Internship \_\_\_\_

Travel time to-from University \_\_\_\_

**How important were each of the following reasons in your decision to enter your study programme, rank in terms of importance.**

[1-Most Important; 4 - least Important]

|  |  |
| --- | --- |
| **Reason** | **RANK [1-4]** |
| **Interest in the subject** |  |
| **Reputation of Course/University** |  |
| **Professional/Career development** |  |
| **Recommendation** |  |

**P.T.O**

**Please rate your satisfaction with your current workload**

(Single choice Mandatory each row)

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **I Want Less** | **Adequate** | **I Want More** |
| **Current Curriculum** |  |  |  |
| **Jobs or Internship** |  |  |  |
| **Prep for Future Jobs/ Study** |  |  |  |
| **Overall Workload** |  |  |  |

**What do you plan to do within a year after finishing your current study programme?**

[Single choice ]

1 Paid Employment \_\_\_\_

2 Further Studies \_\_\_\_

3 Set-up my own business (self-employed) \_\_\_\_

4 Travel \_\_\_\_

**What kind of Initiatives the University should take for the betterment of the course, rank in terms of importance.**

[1-Most Important; 4 - least Important]

|  |  |
| --- | --- |
| **Topic** | **RANK [1-4]** |
| Dedicated Internship Program |  |
| Career Counselling (One-One) |  |
| Career Fair (Like GradIreland) |  |
| More Seminars / Visiting Lectures |  |

# Appendix B. SAS Source Code

/\*

\* Program: ST662 - Project - Postgrad Survey (3).sas

Team : Pratik Kumar Budhdeo (18250375), Suchismit Gupta (18250123),

Sean O’Riogain (18145426), Megan Weston (18145442).

Date : 24th April 2019

\*/

/\* Load the postgraduate student survey data \*/

proc import datafile="/home/seanoriogain200/ST662/Student\_Data.csv" out=survey replace dbms=CSV;

getnames=yes;

run;

/\*\* Comment: The Age variable is in character format which may indicate the presence of invalid

or missing values.\*/

/\* Display details of the dataset's structure etc. \*/

proc contents data=survey;

run;

/\*\* Comment: The number of variables (p=25) is pretty large relative to the number of

observations (n=54). So it might be a good idea to combine some of them and/or

to drop some of the less useful/interesting ones. \*/

/\* Let's take a high-level look at the contents of the dataset \*/

proc means data=survey min mean max nmiss;

run;

/\*\* Comments:

1. At first glance, we appear to have no missing data.

2. Age and Gender do not appear in the list (because they are character variables).

3. The ranges of Course, Level and Modules values seem to be correct.

4. There appears to be a wide range in Attendance values - possibly a good focus area.

5. Respondents appear to have a range of views on the number of lectures they have

to attend.

6. Some respondents are either not doing their assignments or can complete them easily

while other have to invest more effort (or are struggling).

7. There is a wide spread in Self Study values - with at least one respondent doing none

and at least one spending over half a working week.

8. There is also a wide spread in Job\_Intern values with at least one respondent

working half-time while others are not employed at all (the low mean value

suggests that the former category is likely to be outlier).

9. There is a reasonable spread in the Travel values indicating the extent of the

courses' catchment area, while the low mean value suggests that most respondents

live relatively close to the campus.

10. The value ranges multiple-choice Decision, Uni and Satif variable groupings

indicate that the relevant questions have been responded to correctly.

There may be scope for combining them somehow to reduce dataset dimensionality.

11. The value range of the Res\_Poststudy\_plan is as expected.

12. The variables reviewed under points 4, 5, 6 and 7 appear to be the best candidates

for further analysis (as response variables), followed by those under points

8 and 9.\*/

/\* Now let's take a closer look at the contents of the individual variables \*/

/\*\*\*\*\* Variable: Sl\_No \*\*\*\*/

/\* Check that this variable takes unique values \*/

proc freq data=survey nlevels;

run;

/\*\* Comment: This variable does take unique values because its NLEVELS value equals the number

of observations (54) provided earlier by the CONTENTS procedure.\*/

/\*\*\*\*\* Variable: Age \*\*\*\*/

proc freq data=survey;

tables age;

run;

/\*\* Comment: Most of the respondents are in their twenties, followed by a few students in their

thirties (8) and forties (3), with one outlier in their sixties. The NA value indicates

that Age values are missing for 3 respondents. (Some sensitivity to providing age details

is to be expected.) \*/

/\* Change values of NA to null (.) to highlight that missing data - and the variable's format to

numeric. \*/

data survey;

set survey;

if Age="NA" then Age = .;

Age\_Num = input(Age, 3.);

drop Age;

rename Age\_Num=Age;

run;

/\* The new Age variable is at the end of the dataset so we will now put it back to its original

position. To do this, a format statement is needed before the set statement. \*/

data survey;

format Sl\_No Age Gender Course Level Modules Attendance Lecture Essay\_Assign Self\_Study

Job\_Intern Travel Decision\_Interest Decision\_Reput\_Uni Decision\_Career\_Dev Decision\_Reccom

Satisf\_WL\_CC Satisf\_WL\_Job Satisf\_WL\_Future Satisf\_WL\_Overall Res\_Poststudy\_plan

Uni\_Init\_Itern Uni\_Init\_CC Uni\_Ini\_CF Uni\_Ini\_Semin;

set survey;

run;

/\*\*\*\*\* Variable: Gender \*\*\*\*/

proc freq data=survey;

tables gender;

run;

/\*\* Comments:

1. The expected value range (M, F) is achieved but some are in lowercase.

2. The male-female gender ratio is approximately 3:1 (40:14). \*/

/\* Ensure all values are in uppercase \*/

data survey;

set survey;

gender = upcase(gender);

run;

proc freq data=survey;

tables gender;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Plot Age and Gender Distribution\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*step 1 create dataset with percentage count of age and gender,

differentiate gender by sign change \*/

proc sql noprint;

create table left as select Gender,Age, -1\*count(Age) as population from survey where Gender="F"

group by Gender,Age;

create table right as select Gender,Age, count(Age) as population from survey where Gender="M"

group by Gender,Age;

quit;run;

data both;

set left right;

percentage = 100\*population/54;

run;

proc sql;

drop table left;

drop table right;

quit;run;

/\*step 2 hbarparm uses categorical groups so create more observations

to spread the distribution across the true age range\*/

data spread;

Gender="M";

population=.;

percentage=.;

do Age = 20 to 62;

output;

end;

run;

data age\_gender;

set both spread;

run;

proc sql;

drop table both;

drop table spread;

quit;run;

/\*step 3 creating the graph\*/

ods listing style=htmlblue gpath='/home/meganweston20190/ST662Lib/';

ods graphics/ width=6in height=4in imagename='age\_gender\_d' noborder;

proc sgplot data=age\_gender noborder;

styleattrs datacolors=(cxefafaf cxafafef) datacontrastcolors=(red blue);

hbarparm category=Age response=percentage / group=Gender groupdisplay=cluster grouporder=descending

clusterwidth=0.25 barwidth=0.1 baselineattrs=(thickness=0);

scatter y=Age x=percentage / group = Gender groupdisplay=cluster

markerattrs=(symbol=circlefilled size=18) filledoutlinedmarkers name='a' dataskin=sheen;

xaxis display=(noticks noline) grid values=(-8 -4 0 4 8 12)

valuesdisplay=("8" "4" "0" "4" "8" "12");

yaxis display=(noticks noline);

label percentage = "Percentage of total respondents (%)";

run;

/\* note: this method of plotting, inspired by lollipop plots in r is not defined in sas. So to create

horizontal lines, thin horizontal bars are used. This throws up warnings in Sas but after investigating

it is clear that the plot executed as expected and the warnings may be ignored\*/

/\*The resulting graph displays the age distribution of male respondents on the right and female on

the left. It is clear that there were more male respondents than female. Of those who provided their

ages, all females were under the age of 33, with a large propportion under 23. The most frequent male

ages were 25 and 26 but there was a greater spread up to higher ages.

Note: This plot was also panel segmented by course but the results were not of great interest so it

was omitted

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\* Variable: Course \*\*\*\*/

proc freq data=survey;

tables course;

run;

/\*\* Comment: The expected number of values (3) - and only those - are present. Course 2

(Data Science) is specified for 50% (27) of the respondents, which looks high. Further

investigation is required.\*/

/\*\*\*\*\* Variable: Level \*\*\*\*/

proc freq data=survey;

tables level;

run;

/\*\* Comment: Here, c. 46% (25) of the respondents specify H. Dip. (2) which conflicts with the

answer to previous question.\*/

/\* Let's take a closer look at Course and Level together....\*/

proc sql number;

select course, level, count(\*) as count

from survey

group by course, level

order by 1, 2;

quit;

/\*\* Comment: This tells us that the only course for where respondents represent both levels

(Masters (18) and H. Dip. (9)) is Data Science.

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Plot of Course and Level\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*step 1 for the rest of the graphs we will be grouping

by level and course categorically, so lets create a mirror

dataset with strings for these variables\*/

data survey\_g;

set survey;

if Course=1 then Course\_n = 'Applied Comp. Sci.';

else if Course=2 then Course\_n = 'Data Science';

else if Course=3 then Course\_n = 'Software Eng./Dev.';

drop Course;

rename Course\_n=Course;

if Level = 1 then Level\_n='Masters';

else if Level=2 then Level\_n = 'HDip';

drop Level;

rename Level\_n=Level;

run;

/\*step 2 Lets group the new dataset as before, counting

the population of Level and Course\*/

proc sql noprint;

create table level\_course as

select course, level, count(\*) as count

from survey\_g

group by course, level order by course,level;

quit;run;

/\*Step 3 create the plot\*/

ods listing style=htmlblue gpath='/home/meganweston20190/ST662Lib/';

ods graphics/ width=5in height=4in imagename='level\_course\_count' noborder;

proc sgplot data=level\_course noborder;

styleattrs datacolors=(Yellow Indigo);

vbarparm category=Course response=count/ group=Level groupdisplay=cluster

datalabel=count datalabelattrs=(size=12)

clusterwidth=0.9 barwidth=0.9 outlineattrs=(color=gray77) baselineattrs=(thickness=0);

yaxis display=(novalues noticks noline) label="Respondent count";

run;

/\*The resulting plot displays what was previously shown in the frequency table, Data Science had

the most respondents, and was the only course with both HDip and Masters students. The second

highest response was from Applied Computer Science, with 16 HDip respondents.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\* Variable: Modules \*\*\*\*/

proc freq data=survey;

tables modules;

run;

/\*\*\*\*\* Variable: Attendance \*\*\*\*/

proc freq data=survey;

tables Attendance;

run;

/\*\* Comment: Most of the respondents (40) appear to be taking 5 modules while the remainder are

taking either 4 modules (5) or 6 modules (9). \*/

/\* Let's do a consistency check by course and level...\*/

proc sql number;

select course, level, modules, count(\*) as count

from survey

group by course, level, modules

order by 1, 2, 3;

quit;

/\*\* Comment: Respondents from Courses 1 and 2 (Computer Science and Data Science) have a common

view of the number of modules in accordance with their levels, while those

from Course 3 (Software Development) at Level 1 (Masters) are inconsistent

with 5 of them claiming to take 4 modules while the remaining 6 claim to

take 5 modules. Some imputation may be required here.\*/

/\*\*\*\*\* Variable: Attendance \*\*\*\*/

proc freq data=survey;

tables Attendance;

run;

/\*Comment: The majority of the respondents (36/54=67%) claim to have attended 90%+ of

their lectures with an additional 14 (26%) claiming to have attended more

than 50%, leaving 4 (7%) who claim to have attended 50% or less. We need to

check if course satisfaction is able to explain the low attenders.\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Plot of attendance by Course\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*step 1 create a boxplot grouped by course\*/

ods listing style=htmlblue gpath='/home/meganweston20190/ST662Lib/';

ods graphics/ width=5in height=4in imagename='attendance\_course\_box' noborder;

proc sgplot data=survey\_g noborder;

styleattrs datacolors=(Yellow GreenYellow Indigo) datacontrastcolors=(dab) datasymbols=(diamondfilled);

vbox Attendance/ group=Course boxwidth=0.7;

yaxis label="Attendance (%)" ;

run;

/\* This plot shows the distribution of attandence across the three courses. It appears that Applied

Computer Science has the largest spread in attendance values and also the lowest mean attendance.

Software Eng./Dev. have the highest mean attendance and Data Science lies slightly lower, with a low

outlier at 20% attendance.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\* Variable: Lecture \*\*\*\*/

proc freq data=survey;

tables Lecture;

run;

/\*\* Comment: The respondents submit widely differing estimates of the number of lectures they

have to attend per week during semester 2.\*/

/\* Let's take a closer look at how this breaks down by Course and Level....\*/

proc sql number;

select course, level, lecture, count(\*) as count

from survey

group by course, level, lecture

order by 1, 2, 3;

quit;

/\*Comment: For Course 1 Level 1 (H. Dip. in Computer Science), most (9/16) of the respondents

estimate that they spend 16 hours at lectures/labs while the remainder

have different estimates ranging from 8 to 14 hours.

For Course 2 Level 1 (Masters in Data Science), estimates vary widely between 3 to

12 hours.

For Course 2 Level 1 (H. Dip. in Data Science) estimates are more consistent,

ranging between 10 and 12 hours with the majority (5/9) going with 11 hours.

For Course 3 Level 1 (Masters in Software Development) estimates range between

10 to 16 hours with most (6/11) submitting estimates in the 14-16 hour range.

Does this stem from the low attendance rate or a misinterpretation of the question?

/\*\*\*\*\* Variable: Essay\_Assign \*\*\*\*/

proc freq data=survey;

tables Essay\_Assign;

run;

/\*\*Comment: There is a wide range of values here too - ranging from 2 to 24 hours per week.

The highest number of respondents with the same estimate is 11 with 10 hours.

Assuming that all respondents have been submitting their assignments on time,

this seems to suggest that some are finding the assignments very easy while

others are finding them very difficult. This may be an indicator of mixed

ability levels across the courses.\*/

/\* Let's take a closer look by Course and Level....\*/

proc sql number;

select course, level, essay\_assign, count(\*) as count

from survey

group by course, level, essay\_assign

order by 1, 2, 3;

/\*\*Comment: For Course 1 Level 2 (H. Dip. in Computer Science), the responses range from 2 to 20

hours per week, with 10 hours being the most common (6/16) while the remaining

10 responses are split equally having values above and below 10 hours.

For Course 2 Level 1 (Masters in Data Science), the responses range from 3 to 20

hours per week, 15 hours being the most common (4/18) while the remaining 14

responses are mostly less than 15 hours.

For Course 2 Level 2 (H. Dip. in Data Science), the responses range from 4 to 22

hours. 2 of the values in that range (8 and 22) were specified by 2 respondents

each while all other values are unique to individual respondents.

For Course 3 Level 1 (Masters in Software Development), the response range from 1

20 hours, with 12 hours being the most common (3/11). Apart from 5 hours, which

was specified by 2 respondents, all of the other values were unique to individual

respondents.\*/

/\*\*\*\*\* Variable: Self\_Study \*\*\*\*/

proc freq data=survey;

tables Self\_Study;

run;

/\*\*Comment: The values of this variable range from 0 to 22 hours per week, with the most common

(9/54) value being 10 hours. 7 respondents claim to spend less than 3 hours per

week on self study while another 7 claim to spend 18 hours or more.

Again, this may be an indicator of mixed abilities in the sample population.\*/

/\*\*\*\*\* Variable: Job\_Intern \*\*\*\*/

proc freq data=survey;

tables Job\_Intern;

run;

/\*\*Comment: 38 of the respondents (or just over 70%) are full-time students with the remainder

engaged in some type of part-time activity ranging from as little as 3 hours

per week to as much as 20 hours (3 respondents).\*/

/\*\*\*\*\* Variable: Travel \*\*\*\*/

proc freq data=survey;

tables Travel;

run;

/\*\*Comment: Values for this variable range from 2 hours to 12 hours per week, with over 70% of respondents

specifying a value of 5 hours or less.

These values demonstrate that virtually all of the respondents live within a

relatively small distance from the university.

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*Plot of time spent on Activities per Course \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*With so many variables there are many ways to plot this data to examine different relationships.

Four plots will be made here, but not all will make it to the final report\*/

/\*step 1 : Transform from wide to long format\*/

data timer(keep=Sl\_No Course time1 time2 time3 time4 time5);

set survey\_g(rename=(Lecture=time1

Essay\_Assign=time2

Self\_Study=time3

Job\_Intern=time4

Travel=time5));

run;

proc transpose data=timer

out=timer\_t (rename=(col1=total \_name\_=time));

var time1-time5;

by Sl\_No Course;

run;

data timer\_t;

set timer\_t;

length Activity $ 30;

if time="time1" then Activity="Lectures/Labs";

if time="time2" then Activity="Essays/Assignments";

if time="time3" then Activity="Self-Study";

if time="time4" then Activity="Paid Job/Internship";

if time="time5" then Activity="Travel to-from University";

run;

proc sql;

drop table timer;

quit;run;

/\*step 2: summarize mean time spent per course per activity\*/

proc sql noprint;

create table time\_m as select

Course,Activity, round(avg(total),0.01) as Mean\_time

from timer\_t group by Course,Activity order by Course, Mean\_time desc;

quit; run;

/\*\*\*\*\*\*\*\*Plot 1: Panelled bar chart of mean time per course\*\*\*\*\*\*\*\*\*\*\*\*\*/

ods listing style=htmlblue gpath='/home/meganweston20190/ST662Lib/';

ods graphics/ width=7in height=4in imagename='panel\_time\_mean' noborder;

proc sgpanel data=time\_m noautolegend;

styleattrs datacolors=(Yellow GreenYellow Green BlueViolet Indigo) datacontrastcolors=(dab);

panelby Activity/ novarname;

vbarparm category=Course response=Mean\_time / group = Activity groupdisplay=cluster

barwidth=0.9 outlineattrs=(color=gray77) datalabel=Mean\_time datalabelattrs=(size=12) ;

rowaxis label="Mean Time Spent (hrs/week)" display=(novalues noticks);

run;

/\*The main point of this plot is to see the difference in mean time spent by each

course on each activity. Overall, Applied Computer Science spend the most time in

lecture or labs and studying while Data Science have spent the most time on essays or

assignments.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*Plot 2: Clustered bar chart of mean time by course and activity\*\*\*\*\*\*\*\*\*\*\*/

ods listing style=htmlblue gpath='/home/meganweston20190/ST662Lib/';

ods graphics/ width=7in height=4in imagename='cluster\_time\_mean' noborder;

proc sgplot data=time\_m noborder;

styleattrs datacolors=(Yellow GreenYellow Green BlueViolet Indigo) datacontrastcolors=(dab);

vbarparm category=Course response=Mean\_Time / group = Activity groupdisplay=cluster

barwidth=0.9 outlineattrs=(color=gray77) datalabel=Mean\_time datalabelattrs=(size=10);

yaxis label="Mean Time Spent (hrs/week)" display=(novalues noticks noline);

run;

/\*This plot is just another way of visualising the same data. From it we can easily pick out

the most and least time consuming activities for each course and the colours allow us to compare

by course.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*Plot 3: Denisty panel of time spent per activity per course\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Mean is not always the best measure of center so lets look at the distribution of

answers in each group\*/

ods listing style=htmlblue gpath='/home/meganweston20190/ST662Lib/';

ods graphics/ width=3in height=7in imagename='panel\_time\_density' noborder;

proc sgpanel data=timer\_t;

styleattrs datacontrastcolors=(Yellow GreenYellow Indigo);

panelby Activity/ onepanel columns=1 uniscale=column novarname;

density total / group = Course;

colaxis min=0 label="Time Spent (hrs/week)";

run;

/\*This plot is informative because we see the spread of responses. The Essay/Assignment

plot is similar for each course with Data Science pulled slightly higher up the time

axis. Time spent studying or in Lectures or Labs is noticeably lower for the vast majority of

Data Science students. Most Applied Computer Science Students spend little to no time

at jobs or internships.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*Plot 4: Boxplot panel of time spent per activity per course\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

ods listing style=htmlblue gpath='/home/meganweston20190/ST662Lib/';

ods graphics/ width=3in height=7in imagename='panel\_time\_box' noborder;

proc sgpanel data=timer\_t;

styleattrs datacolors=(Yellow GreenYellow Indigo) datacontrastcolors=(dab) datasymbols=(diamondfilled);

panelby Activity/ onepanel columns=1 uniscale=column novarname;

vbox total / group = Course boxwidth=0.7;

rowaxis min=0 label="Time Spent (hrs/week)";

run;

/\*\*\*\*\* Variables: Decision\_Interest Decision\_Reput\_Uni Decision\_Career\_Dev Decision\_Reccom \*\*\*\*/

proc sql number;

select Decision\_Interest,Decision\_Reput\_Uni,Decision\_Career\_Dev,Decision\_Reccom,

count(\*) as count

from survey

group by Decision\_Interest,Decision\_Reput\_Uni,Decision\_Career\_Dev,Decision\_Reccom

order by 5 desc;

quit;

/\*\* Comment: The results of the query above show that 2-3-1-4 (21) and 1-3-2-4 (14) are the two

most popular rankings, accounting for 65% of the respondents.

Those results also reveal the presence of invalid ranking combinations

(e.g. 4-3-4-1) which will need to be addressed.\*/

/\* Identify all such anomalies \*/

proc sql number;

select Decision\_Interest as r1,

Decision\_Reput\_Uni as r2,

Decision\_Career\_Dev as r3,

Decision\_Reccom as r4

from survey

where r1=r2 or r1=r3 or r1=r4 or r2=r3 or r2=r4 or r3=r4;

quit;

/\* Identify all such anomalies \*/

proc sql number;

select Sl\_No, Decision\_Interest as r1,

Decision\_Reput\_Uni as r2,

Decision\_Career\_Dev as r3,

Decision\_Reccom as r4

from survey

where r1=r2 or r1=r3 or r1=r4 or r2=r3 or r2=r4 or r3=r4;

quit;

/\* The following 2 anomalies were found: 3-2-2-2 and 4-3-4-1.

/\*\* Comment: Here is the imputation approach we are going to take to resolve the anomalies

above:

1. Assume the ranking numbers up to where the (first) invalid number in the anomalous

ranking series are correct.

2. Find the most frequent, valid ranking sequence that begins with the correct

numbers and replace the invalid part of the anomalous ranking sequence with the

equivalent part of the most frequent, valid one.

3. If there is more than one such valid ranking with the same frequency, use the

first one.

4. If there is no such valid ranking sequence, replace only the invalid number(s) to

make the ranking sequence valid.\*/

/\* Perform imputation steps 1 & 2 \*/

proc sql number;

select Decision\_Interest as r1,

Decision\_Reput\_Uni as r2,

Decision\_Career\_Dev as r3,

Decision\_Reccom as r4,

count(\*) as count

from survey

where (r1=3 and r2=2) or (r1=4 and r2=3)

group by r1, r2, r3, r4

order by 5 desc;

quit;

/\* In accordance with imputation steps 2 to 4 above, 3-2-2-2 becomes 3-2-1-4 and 4-3-4-1

becomes 4-3-2-1 \*/

data survey;

set survey;

if Sl\_No=39 then do; /\* Resolve 3-2-2-2 \*/

Decision\_Career\_Dev = 1;

Decision\_Reccom = 4;

end;

else if Sl\_No=40 then do; /\* Resolve 4-3-4-1 \*/

Decision\_Career\_Dev = 2;

end;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Plot for ranked Reason of Joining the Course\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*These results were generally unchanged from course to course\*/

/\*step 1: vbox requires a transformation from wide to long\*/

data ranker(keep=Sl\_No Course decision1 decision2 decision3 decision4);

set survey\_g(rename=(Decision\_Interest=decision1

Decision\_Reput\_Uni=decision2

Decision\_Career\_Dev=decision3

Decision\_Reccom=decision4));

run;

proc transpose data=ranker

out=rank\_t (rename=(col1=rank \_name\_=decision));

var decision1-decision4;

by Sl\_No Course;

run;

data rank\_t;

set rank\_t;

length Reason $ 30;

if decision="decision1" then Reason="Interest in the subject";

if decision="decision2" then Reason="Reputation of Course/University";

if decision="decision3" then Reason="Professional/Career development ";

if decision="decision4" then Reason="Recommendation";

rank=-rank;

run;

proc sql;

drop table ranker;

create table rank\_t\_order as

select Sl\_No, Reason, rank, avg(rank) as mean\_rank from rank\_t group by Reason order by mean\_rank desc;

quit; run;

/\*step 2: create plot\*/

ods listing style=htmlblue gpath='/home/meganweston20190/ST662Lib/';

ods graphics/ width=7in height=4in imagename='reason\_box' noborder;

proc sgplot data=rank\_t\_order noborder;

styleattrs datacolors=(Yellow GreenYellow BlueViolet Indigo) datacontrastcolors=(dab)

datasymbols=(diamondfilled);

vbox rank/ group = Reason boxwidth=0.7;

yaxis values=(-1 -2 -3 -4) valuesdisplay=("1" "2" "3" "4") label="Least to Most Important";

run;

/\* This plot illustrates that the most important factor for students choosing their course

was professional career development, closely followed by interest in the subject. The

reputation of the course or university ranked third and recommendations of others ranked fourth

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\* Variables: Satisf\_WL\_CC Satisf\_WL\_Job Satisf\_WL\_Future Satisf\_WL\_Overall \*\*\*\*/

proc sql number;

select Satisf\_WL\_CC,Satisf\_WL\_Job,Satisf\_WL\_Future,Satisf\_WL\_Overall,

count(\*) as count

from survey

group by Satisf\_WL\_CC,Satisf\_WL\_Job,Satisf\_WL\_Future,Satisf\_WL\_Overall

order by 5 desc;

quit;

/\*\* Comment: Here, the most frequent (10/54=19%) response is 2-2-2-2, while the top 4

account for 46% of the responses (10+6+5+4=25/54).

The most popular choice represents those who are satisfied with all aspects of their

course (although it may indicate a lack of thought in some cases). \*/

proc sql number;

select Satisf\_WL\_CC as r1,Satisf\_WL\_Job as r2,Satisf\_WL\_Future as r3,

Satisf\_WL\_Overall as r4

from survey

where r1 in(1,3) and r1 = r2 and r1 = r3 and r1 = r4;

quit;

/\* Now let's focus on the workload aspect of the courses....\*/

proc sql number;

select s.Course, s.Level, s.Satisf\_WL\_Overall, count(\*) as count,

round(count(\*)/X.students\*100) as percentage

from survey s join

(select Course, Level, count(\*) as students

from survey

group by Course, Level) X

on s.Course=X.Course and s.Level=X.Level

group by s.Course, s.Level, s.Satisf\_WL\_Overall;

quit;

/\*\* Comment: The results of the previous query show that the Data Science courses (both Masters

and H. Dip.) have the highest overall satisfaction rate (both 78%) with their

respective workloads, with nobody at H. Dip. level considering their workload

is too light.

The H. Dip in Computer Science students had the next highest satisfaction rating

(67%), while the Masters in Software Development has the lowest satisfaction

rate (27%) in this category.\*/

/\*\* Comment: We need check if there is a link between workload satisfaction and lecture

attendance.\*/

/\* Now let's look for any anomalous responses...\*/

proc sql number;

select Satisf\_WL\_CC as r1,Satisf\_WL\_Job as r2,Satisf\_WL\_Future as r3,Satisf\_WL\_Overall as r4

from survey

where r1 not in(1,2,3) or r2 not in(1,2,3) or r3 not in(1,2,3) or r4 not in(1,2,3);

quit;

/\* No anomalies were found. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Plot of satisfaction with current workload per course\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*step 1: transpose and summarise as before\*/

data satis(keep=Sl\_No Course s1 s2 s3 s4);

set survey\_g(rename=(Satisf\_WL\_CC=s1

Satisf\_WL\_Job =s2

Satisf\_WL\_Future=s3

Satisf\_WL\_Overall=s4));

run;

proc transpose data=satis

out=satis\_t (rename=(col1=res \_name\_=type));

var s1-s4;

by Sl\_No Course;

run;

data satis\_t;

set satis\_t;

length Workload $ 30;

length Response $ 30;

if type="s1" then Workload="Current Curriculum";

if type="s2" then Workload="Jobs or Internship";

if type="s3" then Workload="Prep for Future Jobs/ Study";

if type="s4" then Workload="Overall";

if res=1 then Response = "I Want Less";

if res=2 then Response = "Adequate";

if res=3 then Response = "I Want More";

run;

proc sql;

drop table satis;

create table workload\_summary as select

s.Course, s.Workload, s.Response, count(\*) as count, 100\*count(\*)/x.students as percentage

from satis\_t s join

(select Course, Workload, count(\*) as students

from satis\_t group by Course, Workload) x

on s.Course=x.Course and s.Workload = x.Workload

group by s.Course, s.Workload, s.Response;

quit;run;

/\*step 2: create the plot\*/

ods listing style=htmlblue gpath='/home/meganweston20190/ST662Lib/';

ods graphics/ width=7in height=4in imagename='panel\_workload' noborder;

proc sgpanel data=workload\_summary;

styleattrs datacolors=(Yellow Red Green) datacontrastcolors=(dab);

panelby Workload/ onepanel rows=2 novarname;

hbarparm category=Course response=percentage / group = Response groupdisplay=stack

barwidth=0.9 outlineattrs=(color=gray77);

rowaxis label="Course";

colaxis label="Respondents (%)" values=(0 50 100);

run;

/\*These stacked traffic light coloured plots display the level of satisfaction with

aspects of students workload broken down by course. Interestingly, Software Engineering

and Development who we know spend most of their time in lectures and labs are generally

unsatisfied with their curriculum and overall workload and most eager to spend their

time on jobs or job preparation. Majority of Applied Computer Science Students find

the time spent on job preparation to be adequate.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\* Variable: Res\_Poststudy\_plan \*\*\*\*/

proc sql number;

select Res\_Poststudy\_plan, count(\*) as count

from survey

group by Res\_Poststudy\_plan

order by 2 desc;

quit;

/\*\* Comment: The results of the previous query show that the vast majority (43/54=80%) intend

to seek paid employment after graduating, while 7 of them (13%) plan to go on to

further education and the remainder are intent on setting up their own business. Nobody

is planning to travel. \*/

/\* Now let's take a look at the breakdown by course: \*/

proc sql number;

select s.Course, s.Level, s.Res\_Poststudy\_plan, count(\*) as count,

round(count(\*)/X.students\*100) as percentage

from survey s join

(select Course, Level, count(\*) as students

from survey

group by Course, Level) X

on s.Course=X.Course and s.Level=X.Level

group by s.Course, s.Level, s.Res\_Poststudy\_plan;

quit;

/\*\* Comment: The results of the previous query show that all respondents from the H. Dip. in

Data Science intend to seek paid employment, followed by 78% of the Masters in

Data Science, 75% for the H. Dip.in Computer Science and 73% for the Masters in

Software Development. All of the prospective entrepreneurs are in the Computer

Science courses with 18% of the Masters in Software Development and 13% for the

H. Dip in Computer Science. At 22%, the Masters in Data Analytics had the

highest percentage that plan to go on to further education.

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Plot of future plans broken down by course \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc sql noprint;

create table plans as

select s.Course, s.Level, s.Res\_Poststudy\_plan, count(\*) as count,

round(count(\*)/X.students\*100,0.1) as percentage,

catx(", ", s.Course, s.Level) as Degree

from survey\_g s join

(select Course, Level, count(\*) as students

from survey\_g

group by Course, Level) X

on s.Course=X.Course and s.Level=X.Level

group by s.Course, s.Level, s.Res\_Poststudy\_plan;

quit;run;

data plans;

set plans;

length Plan $30;

if Res\_Poststudy\_plan=1 then Plan = "Paid Employment";

if Res\_Poststudy\_plan=2 then Plan = "Further Studies";

if Res\_Poststudy\_plan=3 then Plan = "Set up own business";

if Res\_Poststudy\_plan=4 then Plan = "Travel";

run;

ods listing style=htmlblue gpath='/home/meganweston20190/ST662Lib/';

ods graphics/ width=8in height=4in imagename='bar\_future' noborder;

proc sgplot data=plans noborder;

styleattrs datacolors=(GreenYellow Yellow Indigo) datacontrastcolors=(dab);

vbarparm category=Degree response=percentage/ seglabel seglabelattrs=(size=13 color=dab)

group = Plan groupdisplay=stack barwidth=0.95 clusterwidth=0 nooutline;

xaxis display=(noline);

yaxis display=(noline novalues noticks) values=(0 50 100) label = "Respondent (%)";

run;

/\*This plot shows the future plans of students per degree course. It echoes what we

already knew, that all Data Science HDips plan on going on to paid employment and the

Data Science Masters has the largest proportion of students planning to study further.

18.2% of Software Engineering or Development respondents plan on setting up their own

business

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\* Variables: Uni\_Init\_Itern Uni\_Init\_CC Uni\_Ini\_CF Uni\_Ini\_Semin \*\*\*\*/

proc sql number;

select Uni\_Init\_Itern,Uni\_Init\_CC,Uni\_Ini\_CF,Uni\_Ini\_Semin,

count(\*) as count

from survey

group by Uni\_Init\_Itern,Uni\_Init\_CC,Uni\_Ini\_CF,Uni\_Ini\_Semin

order by 5 desc;

quit;

/\*\* Comment: Here, the most frequent (16/54=30%) ranking sequence is 1-2-3-4, while the top 4

account for 67% of the responses (16+8+6+6=36/54).

/\* We will now apply the same anomaly detection and imputation approach here as was done for the

Decision\_\* ranking variable group above.\*/

/\* Detect any anomalies \*/

proc sql number;

select Sl\_No, Uni\_Init\_Itern as r1,

Uni\_Init\_CC as r2,

Uni\_Ini\_CF as r3,

Uni\_Ini\_Semin as r4

from survey

where r1=r2 or r1=r3 or r1=r4 or r2=r3 or r2=r4 or r3=r4;

quit;

/\* Only one anomaly was detected: 1-2-3-1 which we will now replace with 1-2-3-4 in accordance

with the selected imputation approach. \*/

data survey;

set survey;

if Sl\_No = 25 then Uni\_Ini\_Semin = 4;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Plot for suggested initiatives grouped by course\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*step 1: transpose and summarise\*/

data ranker(keep= Sl\_No Course decision1 decision2 decision3 decision4);

set survey\_g(rename=(Uni\_Init\_Itern=decision1

Uni\_Init\_CC=decision2

Uni\_Ini\_CF=decision3

Uni\_Ini\_Semin=decision4));

run;

proc transpose data=ranker

out=rank\_tr (rename=(col1=rank \_name\_=decision));

var decision1-decision4;

by Sl\_No Course;

run;

data rank\_tr;

set rank\_tr;

length Initiative $ 40;

if decision="decision1" then Initiative="Dedicated Internship Program";

if decision="decision2" then Initiative="Career Counselling (One-One)";

if decision="decision3" then Initiative="Career Fair (Like GradIreland)";

if decision="decision4" then Initiative="More Seminars / Visiting Lectures";

rank=-rank;

run;

proc sql;

drop table ranker;

create table rank\_tr\_order as

select Sl\_No, Course,Initiative, rank, avg(rank) as mean\_rank from rank\_tr

group by Course,Initiative order by mean\_rank desc;

quit; run;

/\*step 2: create the graph\*/

ods listing style=htmlblue gpath='/home/meganweston20190/ST662Lib/';

ods graphics/ width=8in height=4in imagename='panel\_initiative\_box' noborder;

proc sgpanel data=rank\_tr\_order;

styleattrs datacolors=(Yellow GreenYellow Blue Indigo) datacontrastcolors=(dab)

datasymbols=(diamondfilled);

panelby Course/ onepanel rows=1 uniscale=column novarname;

vbox rank/ group = Initiative boxwidth=0.8;

rowaxis grid values=(-1 -2 -3 -4) valuesdisplay=("1" "2" "3" "4") label="Least to Most Important";

run;

/\*This plot shows each course's ranked answers about the importance of certain initiatives. Everyone

agrees that a dedicated internship program and career counselling are of most importance, but unlike

the others, the average Data Science student ranks seminars above career fairs

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\* Modelling \*\*\*\*/

/\*\* Logistic Regression \*\*/

data survey1;

set survey;

if Age > 0;

Age1 = Age\*1;

drop Age;

run;

data survey1;

set survey1;

rename Age1 = Age;

run;

proc logistic data = survey1 outest= param;

class Satisf\_WL\_Overall (ref = "1") Gender Level ;

model Satisf\_WL\_Overall = Age Course Level Gender Job\_Intern/ link = glogit ;

run;

proc transpose data = param;

run;

proc print noobs;

run;

/\*\* Random Forest \*\*/

PROC HPFOREST data=survey maxtrees=50;

target Satisf\_WL\_CC/level=nominal;

input Attendance Lecture Essay\_Assign Self\_Study Job\_Intern Travel /level=interval;

input Gender Age Course Level / level=nominal;

run;

PROC HPFOREST data=survey maxtrees=50;

target Satisf\_WL\_Job/level=nominal;

input Attendance Lecture Essay\_Assign Self\_Study Job\_Intern Travel /level=interval;

input Gender Age Course Level / level=nominal;

run;

PROC HPFOREST data=survey maxtrees=50;

target Satisf\_WL\_Future/level=nominal;

input Attendance Lecture Essay\_Assign Self\_Study Job\_Intern Travel /level=interval;

input Gender Age Course Level / level=nominal;

run;

PROC HPFOREST data=survey maxtrees=50 ;

target Satisf\_WL\_Overall/level=ordinal;

input Attendance Lecture Essay\_Assign Self\_Study Job\_Intern Travel /level=interval;

input Gender Age Course Level / level=nominal;

run;

/\* Frequency of ranked variables \*/

proc freq data=survey;

tables Decision\_Interest Decision\_Career\_Dev Decision\_Reput\_Uni Decision\_Career\_Dev / ;

tables Uni\_Init\_Itern Uni\_Init\_CC Uni\_Ini\_CF Uni\_Ini\_Semin/ ;

run;

/\*

74% Of student who filled survey are male and 26% are female

50.00% from Data science, 29.63% from Computer sci, 20.37% from Applied computer science

53.70% for Master and 46.30% in HDip

\*/

proc freq data=survey;

tables Res\_Poststudy\_plan / ;

run;

/\* Plan to do within a year after finishing your current study programme: -

79.6% of students said they will look for Paid Employment

12.96% for Further Studies

7.41 are planning for set-up my business or self-employment

\*/

PROC sgscatter data=survey;

matrix Satisf\_WL\_CC Lecture Essay\_Assign Self\_Study Job\_Intern Travel ;

RUN;

proc sort data=survey;

by Course Level;

run;

proc means data=survey mean;

var Lecture Essay\_Assign Self\_Study Job\_Intern Travel ;

output out=temp mean(Lecture)=Lecture mean(Essay\_Assign)=Essay\_Assign mean(Self\_Study)=Self\_Study

mean(Job\_Intern)=Job\_Intern mean(Travel)=Travel;

run;

data temp;

set temp;

drop \_freq\_ \_type\_;

run;

/\* Histogram \*/

title "Histograms for Curiculum satisfaction";

proc sgpanel data = survey ;

panelby Course / columns = 3 rows = 1 ;

histogram Satisf\_WL\_CC / binwidth=0.5 datalabel=Percent fill fillattrs=(color= grey) ;

label Satisf\_WL\_CC = "1=I Want Less ; 2=Adequate ; 3=I Want More ";

run;

title "Histograms for Jobs satisfaction";

proc sgpanel data = survey ;

panelby Course / columns = 3 rows = 1 ;

histogram Satisf\_WL\_Job / binwidth=0.5 datalabel=Percent fill fillattrs=(color= yellow) ;

label Satisf\_WL\_Job = "1=I Want Less ; 2=Adequate ; 3=I Want More ";

run;

title "Histograms for Prep for Future Jobs satisfaction";

proc sgpanel data = survey ;

panelby Course / columns = 3 rows = 1 ;

histogram Satisf\_WL\_Future / datalabel=Percent fill fillattrs=(color= green) ;

label Satisf\_WL\_Future = "1=I Want Less ; 2=Adequate ; 3=I Want More ";

run;

title "Histograms for Overall Workload";

proc sgpanel data = survey ;

panelby Course / columns = 3 rows = 1 ;

histogram Satisf\_WL\_Overall / binwidth=0.5 datalabel=Percent fill fillattrs=(color= black) ;

label Satisf\_WL\_Overall = "1=I Want Less ; 2=Adequate ; 3=I Want More ";

run;

title "Histograms for Overall Workload by gender";

proc sgpanel data = survey ;

panelby Gender / columns = 2 rows = 1 ;

histogram Satisf\_WL\_Overall / binwidth=0.5 datalabel=Percent fill ;

label Satisf\_WL\_Overall = "1=I Want Less ; 2=Adequate ; 3=I Want More ";

run;