dataAnalysisOCS

OCS Team

5/22/2020

```
#retrieve the dataset from the survey and select the relevant columns
ocs <- read_csv("OCS_OG_dataset.csv")
```

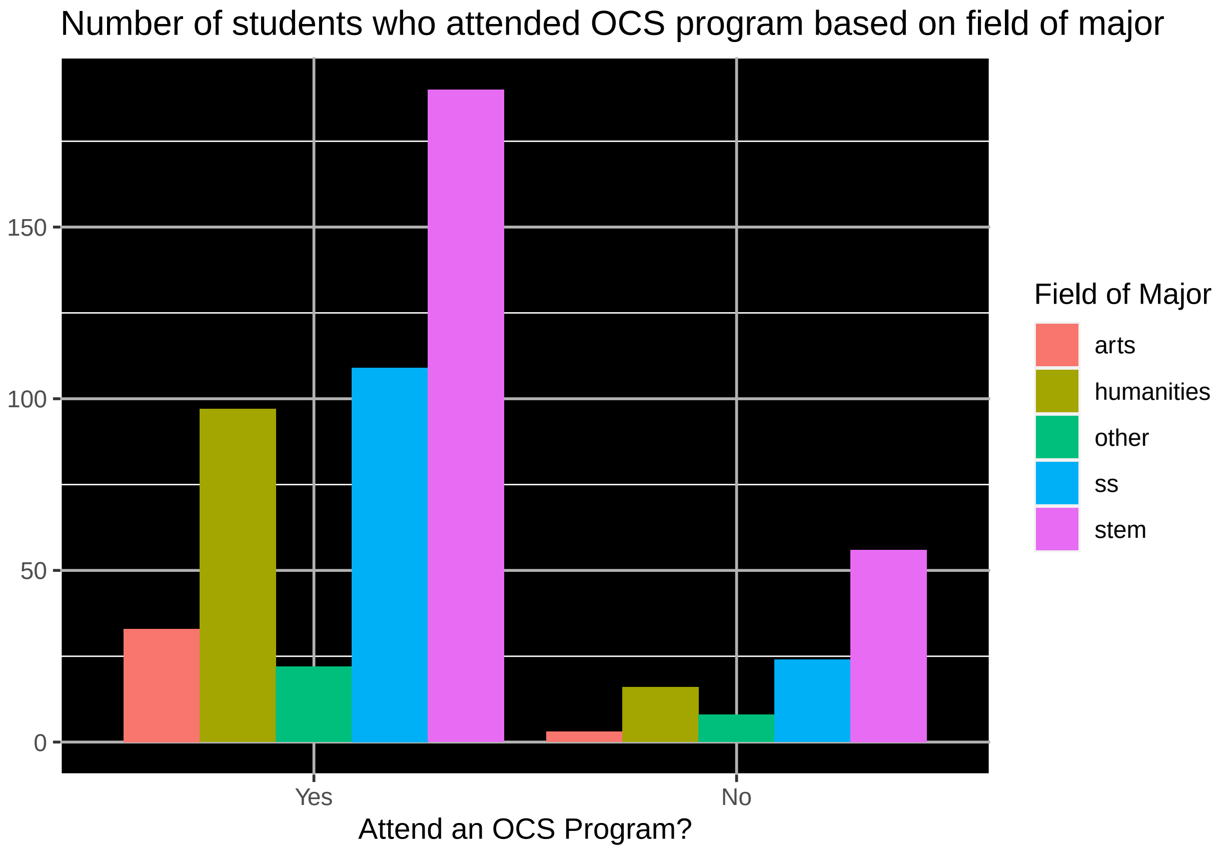
```
## Parsed with column specification:
## cols(
##   .default = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
ocs <- ocs %>%
  select(starts_with("Q"))

#rename column names
names(ocs) <- c("attend_OCS", "reason_not", "reason_not_text", "abroad_classYear", "europe", "reason_eu

#filtering out the first two rows
ocs <- ocs[3:nrow(ocs),]

#find out the percentage of varsity students at Carleton so we can normalize
#https://apps.carleton.edu/voice/?story_id=1836663&section_id=353600&issue_id=1836011
#70% of varsity students study abroad


#Factorize some columns

ocs$attend_OCS <- as.factor(ocs$attend_OCS) %>%
  recode_factor("1" = "Yes", "2" = "No")

ocs$varsity <- as.factor(ocs$varsity) %>%
  recode_factor("1" = "Yes", "2" = "No")

ocs$europe <- as.factor(ocs$europe) %>%
  recode_factor("1" = "Yes", "2" = "No")
```

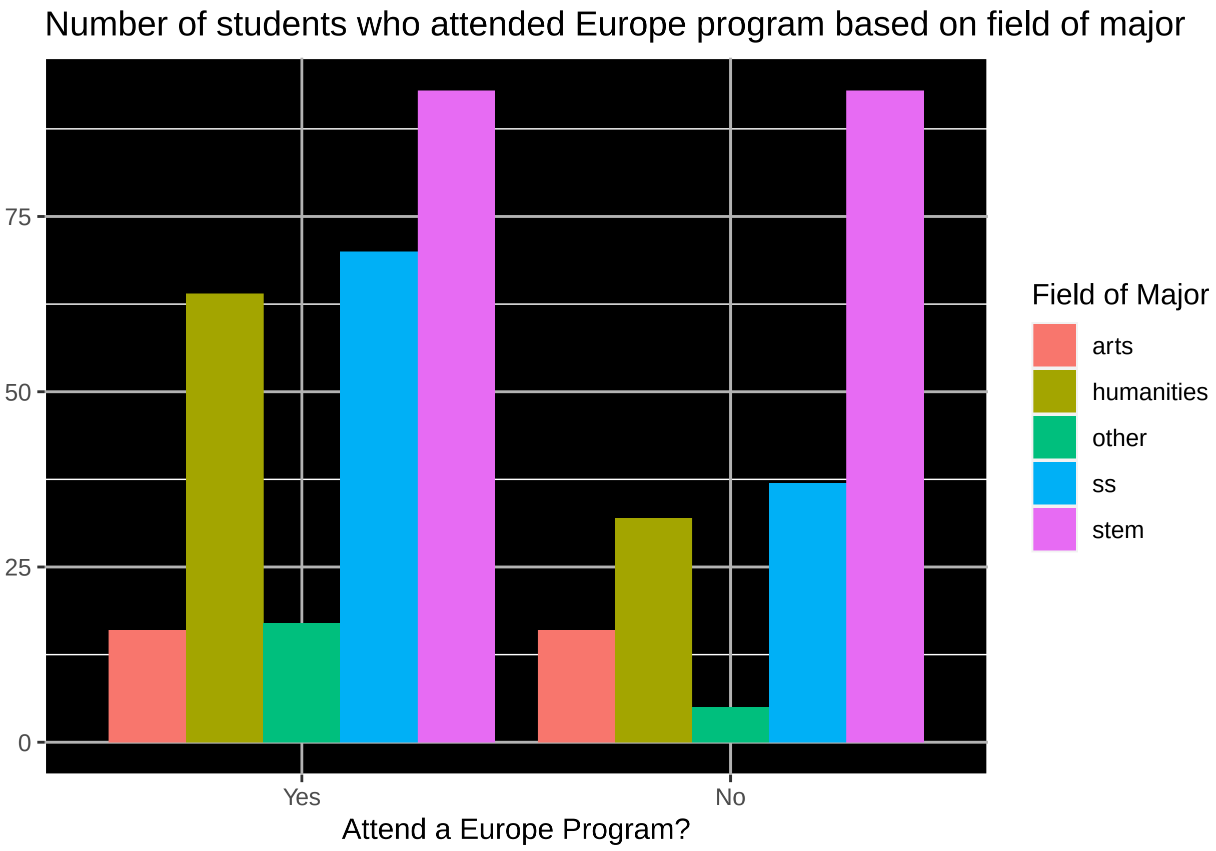
```
#recount double+ majors as separate rows

ocs$stem <- str_detect(ocs$major, pattern = zero_or_more(ALPHA) %R% "STEM")
ocs$humanities <- str_detect(ocs$major, pattern = zero_or_more(ALPHA) %R% "Humanities")
ocs$ss <- str_detect(ocs$major, pattern = zero_or_more(ALPHA) %R% "Social Sciences")
ocs$arts <- str_detect(ocs$major, pattern = zero_or_more(ALPHA) %R% "Arts")
ocs$other <- str_detect(ocs$major, pattern = zero_or_more(ALPHA) %R% "Other")

ocs_pivottedMajor <- pivot_longer(ocs, cols = stem:other,
                                   names_to = "noDoubleMajor",
                                   values_to = "majorTF") %>%
  filter(majorTF == "TRUE") %>%
  select(-majorTF)

#bar chart of majors vs study abroad
ocs_pivottedMajor %>%
  group_by(noDoubleMajor, attend_OCS) %>%
  summarize(count = n()) %>%
  ggplot(aes(x = attend_OCS, y = count, fill = noDoubleMajor)) +
  geom_bar(position = "dodge", stat = "identity") +
  labs(x = "Attend an OCS Program?",
       title = "Number of students who attended OCS program based on field of major",
       fill = "Field of Major") +
  theme(
    panel.background = element_rect(fill = "black",colour = "white",size = 0.5),
    panel.grid.major = element_line(colour = "grey70"),
    axis.title.y = element_blank())
```

```
## `summarise()` regrouping output by 'noDoubleMajor' (override with `.groups` argument)
```



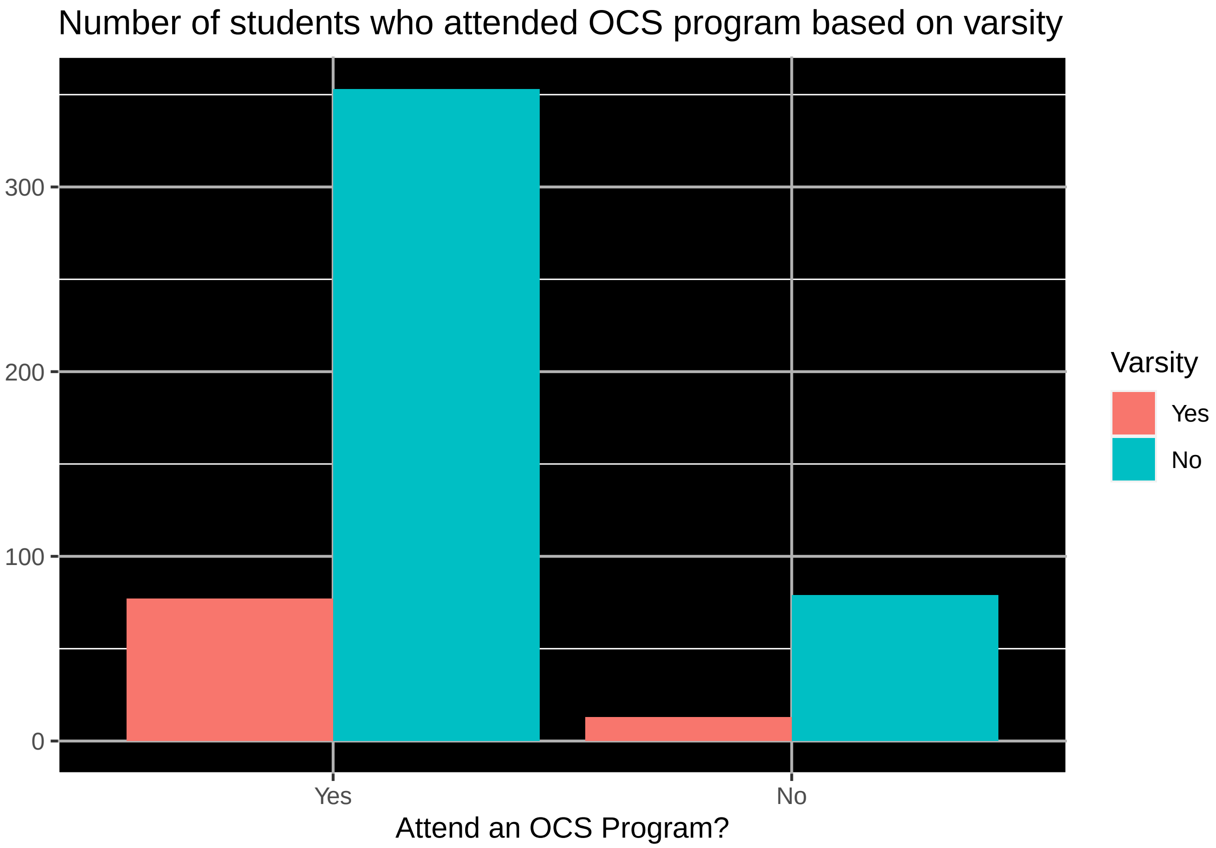
```
#bar chart of majors vs europe program

ocs_pivottedMajor %>%
  drop_na(europe) %>%
  group_by(noDoubleMajor, europe) %>%
  summarize(count = n()) %>%
  ggplot(aes(x = europe, y = count, fill = noDoubleMajor)) +
  geom_bar(position = "dodge", stat = "identity") +
  labs(x = "Attend a Europe Program?",
       title = "Number of students who attended Europe program based on field of major",
       fill = "Field of Major") +
  theme(
    panel.background = element_rect(fill = "black",colour = "white",size = 0.5),
    panel.grid.major = element_line(colour = "grey70"),
    axis.title.y = element_blank())
```

```
## `summarise()` regrouping output by 'noDoubleMajor' (override with `.groups` argument)
```


Number of students who attended Europe program based on field of major

```
#bar chart of varsity vs study abroad
ocs %>%
  drop_na(varsity) %>%
  group_by(attend_OCS, varsity) %>%
  summarize(count = n()) %>%
  ggplot(aes(x = attend_OCS, y = count, fill = varsity)) +
  geom_bar(position = "dodge", stat = "identity") +
  labs(x = "Attend an OCS Program?",
       title = "Number of students who attended OCS program based on varsity",
       fill = "Varsity") +
  theme(
    panel.background = element_rect(fill = "black",colour = "white",size = 0.5),
    panel.grid.major = element_line(colour = "grey70"),
    axis.title.y = element_blank())
```

```
## `summarise()` regrouping output by 'attend_OCS' (override with `.groups` argument)
```



```
## 1. Major vs. study abroad

#NULL HYPOTHESIS: No association between major and study abroad
#ALTERNATIVE: Association between major and study abroad

#major_abroad <- table(ocs_pivottedMajor$noDoubleMajor, ocs_pivottedMajor$attend_OCS)

#there is no evidence that there is association between major and studying abroad
#that is, there is no evidence that whether students study abroad or not is dependent on what field of
chisq.test(ocs_pivottedMajor$noDoubleMajor, ocs_pivottedMajor$attend_OCS)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ocs_pivottedMajor$noDoubleMajor and ocs_pivottedMajor$attend_OCS
## X-squared = 7.8052, df = 4, p-value = 0.09898
```

```
#confirm to see if they are independent using poisson regression
y <- c(33, 3, 97, 16, 22, 8, 109, 24, 190, 56)
major <- factor(c(rep("arts",2),rep("humanities",2), rep("other",2), rep("ss",2), rep("stem",2)))
studyAbroad <- factor(rep(c("yes","no"),5))
major_abroad_df <- data.frame(y, major, studyAbroad)
major_abroad_df
```

```
##      y      major studyAbroad
## 1   33       arts         yes
## 2    3       arts          no
## 3   97 humanities         yes
## 4   16 humanities          no
## 5   22      other         yes
## 6    8      other          no
## 7  109         ss         yes
## 8   24         ss          no
## 9  190       stem         yes
## 10  56       stem          no
```

```
major_abroad_glm <- glm(y ~ major + studyAbroad + major:studyAbroad, family=poisson, data=major_abroad_
summary(major_abroad_glm)
```

```
##
## Call:
## glm(formula = y ~ major + studyAbroad + major:studyAbroad, family = poisson,
##     data = major_abroad_df)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         1.0986     0.5774   1.903 0.057060 .
## majorhumanities                     1.6740     0.6292   2.661 0.007798 **
## majorother                          0.9808     0.6770   1.449 0.147399
## majorss                             2.0794     0.6124   3.396 0.000684 ***
## majorstem                           2.9267     0.5926   4.939 7.86e-07 ***
## studyAbroadyes                      2.3979     0.6030   3.976 6.99e-05 ***
## majorhumanities:studyAbroadyes     -0.5958     0.6606  -0.902 0.367157
## majorother:studyAbroadyes          -1.3863     0.7308  -1.897 0.057839 .
## majorss:studyAbroadyes             -0.8846     0.6438  -1.374 0.169431
## majorstem:studyAbroadyes           -1.1762     0.6219  -1.891 0.058578 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance:  5.1452e+02  on 9  degrees of freedom
## Residual deviance: -1.5321e-14  on 0  degrees of freedom
## AIC: 72.747
##
## Number of Fisher Scoring iterations: 3
```

```
## 2. Varsity vs. Study abroad

#varsity_abroad <- table(ocs$varsity, ocs$attend_OCS)

#there is no evidence that there is association between studying abroad and varsity
chisq.test(ocs$varsity, ocs$attend_OCS)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  ocs$varsity and ocs$attend_OCS
## X-squared = 0.51596, df = 1, p-value = 0.4726
```

```
#confirm to see if they are independent using poisson regression
y <- c(77, 13, 353, 79)
varsity <- factor(c(rep("yes",2),rep("no",2)))
studyAbroad <- factor(rep(c("yes","no"),2))
varsity_abroad_df <- data.frame(y, varsity, studyAbroad)
varsity_abroad_df
```

```
##     y varsity studyAbroad
## 1  77     yes         yes
## 2  13     yes          no
## 3 353      no         yes
## 4  79      no          no
```

```
varsity_abroad_glm <- glm(y ~ varsity + studyAbroad + varsity:studyAbroad, family=poisson, data=varsity_
summary(varsity_abroad_glm)
```

```
## 
## Call:
## glm(formula = y ~ varsity + studyAbroad + varsity:studyAbroad, 
##     family = poisson, data = varsity_abroad_df)
## 
## Deviance Residuals: 
## [1]  0  0  0  0
## 
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                  4.3694     0.1125  38.837  < 2e-16 ***
## varsityyes                  -1.8045     0.2993  -6.029 1.65e-09 ***
## studyAbroadyes               1.4970     0.1245  12.028  < 2e-16 ***
## varsityyes:studyAbroadyes    0.2818     0.3247   0.868    0.385    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance:  4.8202e+02  on 3  degrees of freedom
## Residual deviance: -2.0428e-14  on 0  degrees of freedom
## AIC: 32.514
## 
## Number of Fisher Scoring iterations: 3
```

```
## 3. Major vs. Study Europe
europe_major_df <- ocs_pivottedMajor %>%
  drop_na(europe) %>%
  filter(attend_OCS == "Yes")

#major_europe <- table(major_df$noDoubleMajor, major_df$europe)

#EDA shows that 50% of arts and stem students study in Europe,
#whereas ~67% of humanities students study in europe and
#~65% of social science students study in Europe.

#Chisq test shows that these differences are statistically significant.
#that is, humanities and social science students seem to prefer studying in Europe over the rest of the
```

```r
chisq.test(europe_major_df$noDoubleMajor, europe_major_df$europe)
```

```
##
##  Pearson's Chi-squared test
##
## data:  europe_major_df$noDoubleMajor and europe_major_df$europe
## X-squared = 14.655, df = 4, p-value = 0.005474
```

```r
#confirm to see if they are independent using poisson regression
y <- c(16, 16, 64, 32, 17, 4, 69, 37, 93, 91)
major <- factor(c(rep("arts",2),rep("humanities",2), rep("other",2), rep("ss",2), rep("stem",2)))
studyEurope <- factor(rep(c("yes","no"),5))
major_europe_df <- data.frame(y, major, studyEurope)
major_europe_df
```

```
##     y      major studyEurope
## 1  16       arts         yes
## 2  16       arts          no
## 3  64 humanities         yes
## 4  32 humanities          no
## 5  17      other         yes
## 6   4      other          no
## 7  69         ss         yes
## 8  37         ss          no
## 9  93       stem         yes
## 10 91       stem          no
```

```r
major_europe_glm <- glm(y ~ major + studyEurope + major:studyEurope, family=poisson, data=major_europe_
#other vs. study abroad seems significant...
summary(major_europe_glm)
```

```
##
## Call:
## glm(formula = y ~ major + studyEurope + major:studyEurope, family = poisson,
##     data = major_europe_df)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    2.773e+00  2.500e-01  11.090  < 2e-16 ***
## majorhumanities                6.931e-01  3.062e-01   2.264  0.02359 *
## majorother                    -1.386e+00  5.590e-01  -2.480  0.01314 *
## majorss                        8.383e-01  2.992e-01   2.802  0.00508 **
## majorstem                      1.738e+00  2.711e-01   6.412 1.43e-10 ***
## studyEuropeyes                -3.507e-15  3.536e-01   0.000  1.00000
## majorhumanities:studyEuropeyes  6.931e-01  4.146e-01   1.672  0.09454 .
## majorother:studyEuropeyes      1.447e+00  6.587e-01   2.197  0.02804 *
## majorss:studyEuropeyes         6.232e-01  4.081e-01   1.527  0.12672
## majorstem:studyEuropeyes       2.174e-02  3.831e-01   0.057  0.95474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance:  2.3404e+02  on 9  degrees of freedom
## Residual deviance: -1.4655e-14  on 0  degrees of freedom
## AIC: 72.747
##
## Number of Fisher Scoring iterations: 3
```

```
#but other has no discerning feature -- not important factor to consider
other <- ocs_pivottedMajor %>%
  filter(noDoubleMajor == "other")

#the statistical significance seen earlier with chisq test might be because of the major factor other
#there is no evidence that there is association between major and studying in europe
#that is, there is no evidence that whether students study in europe or not is dependent on what field

#let's repeat without other factor


europe_major_df <- ocs_pivottedMajor %>%
  drop_na(europe) %>%
  filter(attend_OCS == "Yes", noDoubleMajor != "other")

#major_europe <- table(major_df$noDoubleMajor, major_df$europe)

chisq.test(europe_major_df$noDoubleMajor, europe_major_df$europe)
```

```
##
##  Pearson's Chi-squared test
##
## data:  europe_major_df$noDoubleMajor and europe_major_df$europe
## X-squared = 10.182, df = 3, p-value = 0.01708
```

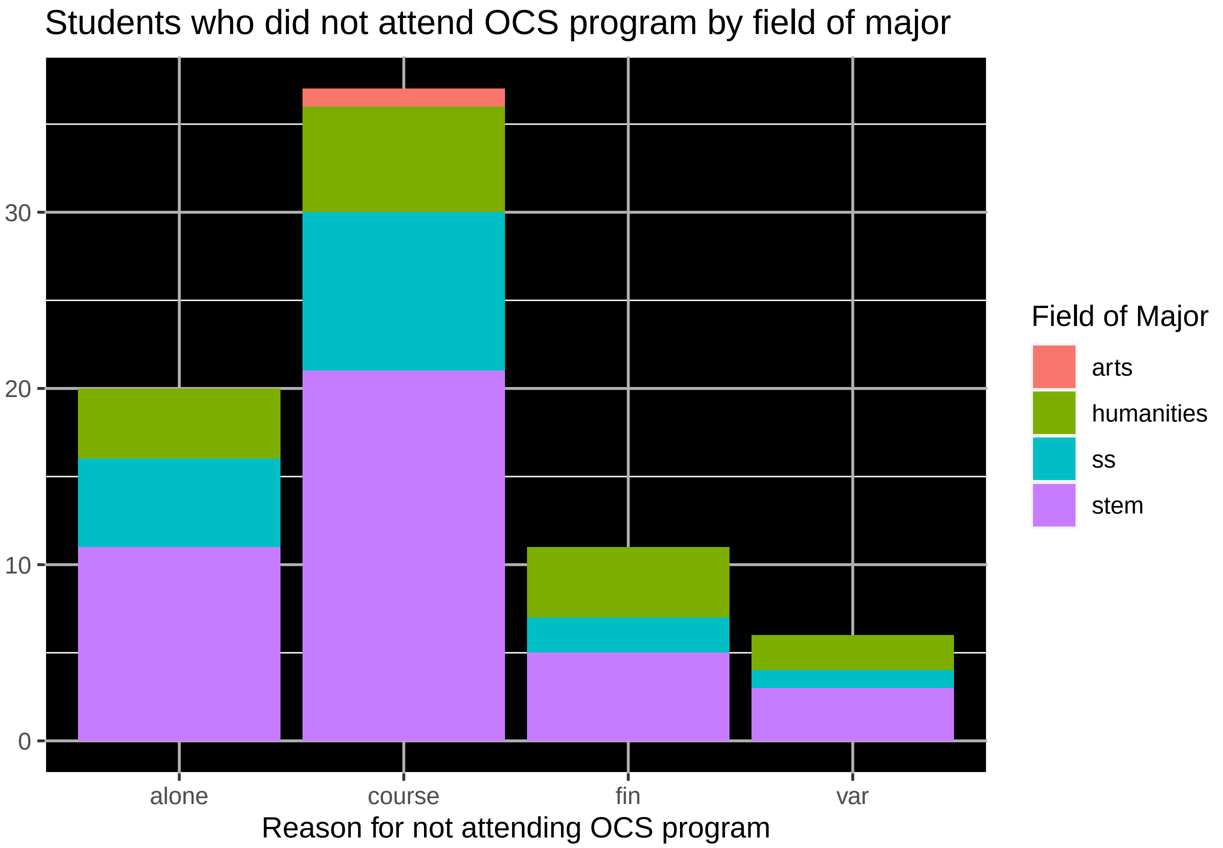
```
#confirm to see if they are independent using poisson regression
y <- c(16, 16, 64, 32, 69, 37, 93, 91)
major <- factor(c(rep("arts",2),rep("humanities",2), rep("ss",2), rep("stem",2)))
studyEurope <- factor(rep(c("yes","no"),4))
major_europe_df <- data.frame(y, major, studyEurope)
major_europe_df
```

```
##    y      major studyEurope
## 1 16       arts         yes
## 2 16       arts          no
## 3 64 humanities         yes
## 4 32 humanities          no
## 5 69         ss         yes
## 6 37         ss          no
## 7 93       stem         yes
## 8 91       stem          no
```

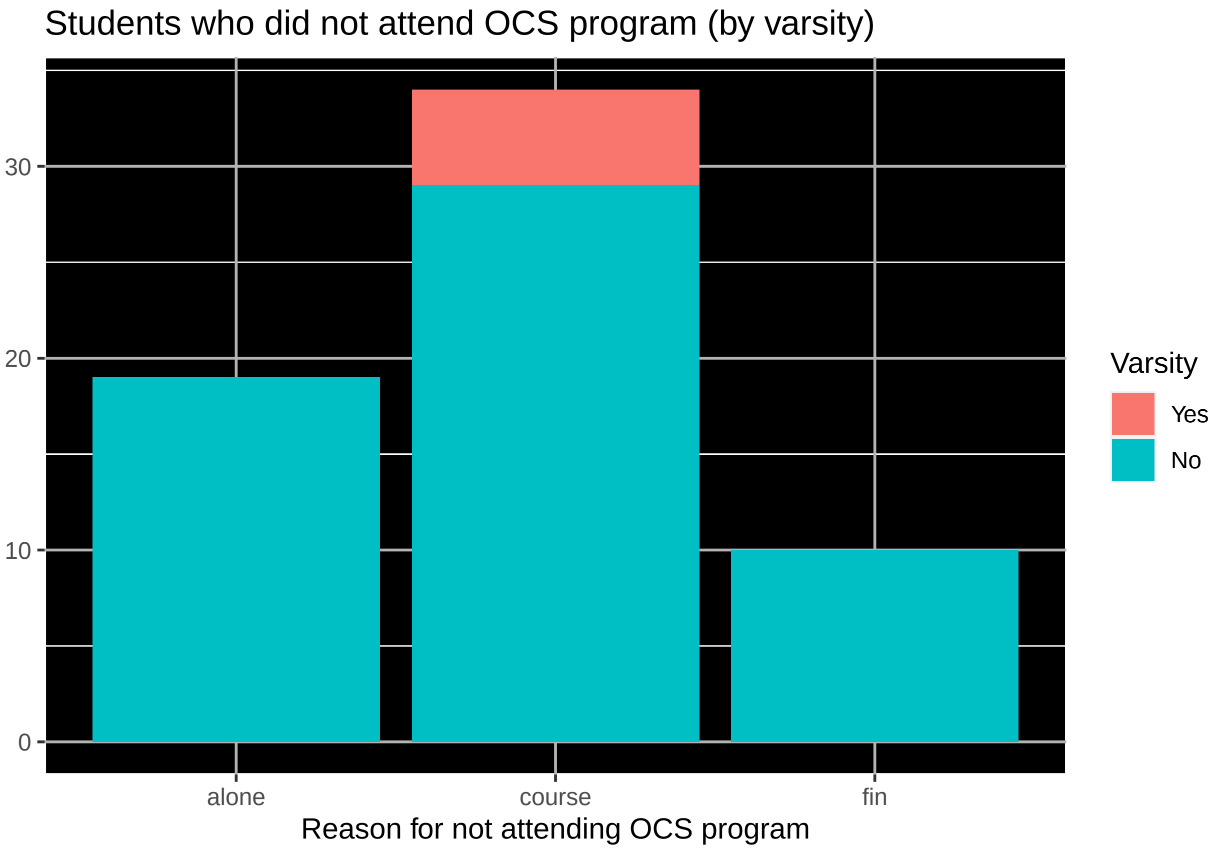
```
major_europe_glm <- glm(y ~ major + studyEurope + major:studyEurope, family=poisson, data=major_europe_

#this test tells us that whether students study in Europe or not is independent of their field of major
summary(major_europe_glm)
```

```
##
## Call:
## glm(formula = y ~ major + studyEurope + major:studyEurope, family = poisson,
##     data = major_europe_df)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      2.773e+00  2.500e-01  11.090  < 2e-16 ***
## majorhumanities                  6.931e-01  3.062e-01   2.264  0.02359 *
## majorss                          8.383e-01  2.992e-01   2.802  0.00508 **
## majorstem                        1.738e+00  2.711e-01   6.412 1.43e-10 ***
## studyEuropeyes                  -1.759e-15  3.536e-01   0.000  1.00000
## majorhumanities:studyEuropeyes   6.931e-01  4.146e-01   1.672  0.09454 .
## majorss:studyEuropeyes           6.232e-01  4.081e-01   1.527  0.12672
## majorstem:studyEuropeyes         2.174e-02  3.831e-01   0.057  0.95474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1.3990e+02  on 7  degrees of freedom
## Residual deviance: 1.1102e-15  on 0  degrees of freedom
## AIC: 60.8
##
## Number of Fisher Scoring iterations: 3
```

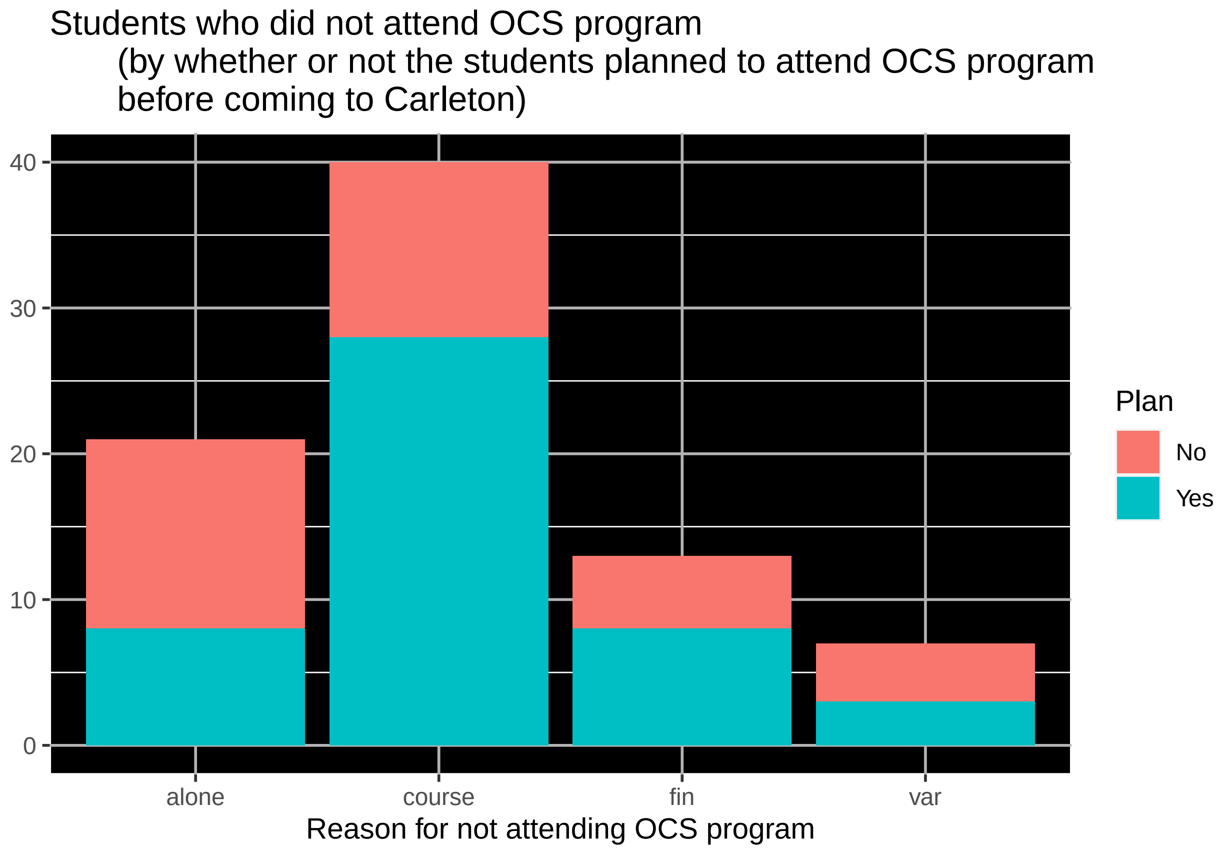
```r
#MORE EDA
#RELATIONSHIP BETWEEN MAJOR AND WHY NOT ABROAD

notAbroad <- ocs_pivottedMajor %>%
  filter(attend_OCS == "No") %>%
  select(reason_not, reason_not_text, varsity, ocs_before, noDoubleMajor)

notAbroad$fin <- str_detect(notAbroad$reason_not, pattern = zero_or_more(ALPHA) %R% "Financial")
notAbroad$course <- str_detect(notAbroad$reason_not, pattern = zero_or_more(ALPHA) %R% "Course")
notAbroad$alone <- str_detect(notAbroad$reason_not, pattern = zero_or_more(ALPHA) %R% "Alone")
notAbroad$var <- str_detect(notAbroad$reason_not, pattern = zero_or_more(ALPHA) %R% "varsity")

notAbroad <- pivot_longer(notAbroad, cols = fin:var,
                                  names_to = "reason_not_2.0",
                                  values_to = "reason_not_TF") %>%
  filter(reason_not_TF == "TRUE") %>%
  select(-reason_not_TF)

#bar chart of majors vs study abroad
notAbroad %>%
  filter(noDoubleMajor != "other") %>%
  ggplot() +
  geom_bar(aes(reason_not_2.0, fill = noDoubleMajor)) +
  labs(x = "Reason for not attending OCS program",
       title = "Students who did not attend OCS program by field of major",
       fill = "Field of Major") +
  theme(
    panel.background = element_rect(fill = "black",colour = "white",size = 0.5),
    panel.grid.major = element_line(colour = "grey70"),
    axis.title.y = element_blank())
```



```r
#bar chart of varsity vs study abroad
notAbroad %>%
  drop_na(varsity) %>%
  filter(noDoubleMajor != "other") %>%
  ggplot() +
  geom_bar(aes(reason_not_2.0, fill = varsity)) +
  labs(x = "Reason for not attending OCS program",
       title = "Students who did not attend OCS program (by varsity)",
       fill = "Varsity") +
  theme(
    panel.background = element_rect(fill = "black",colour = "white",size = 0.5),
    panel.grid.major = element_line(colour = "grey70"),
    axis.title.y = element_blank())
```

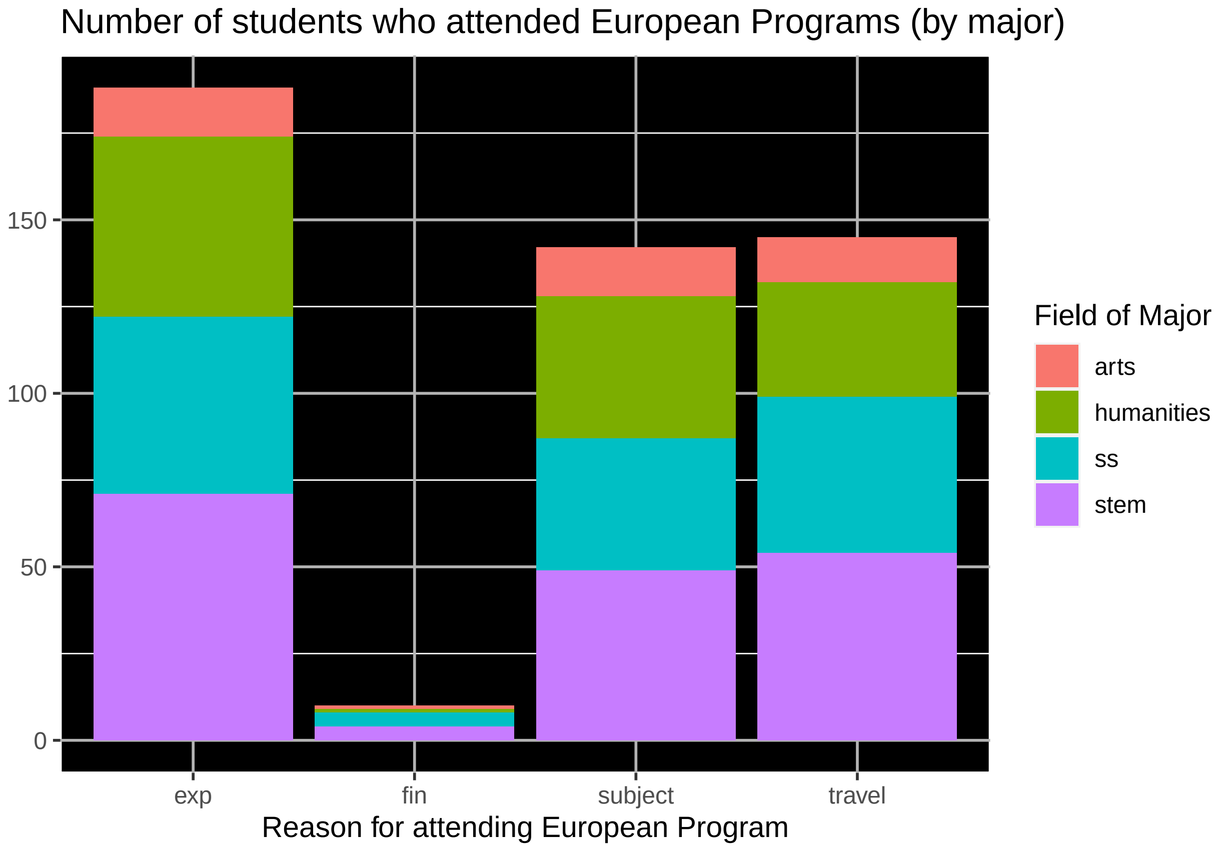


```
#bar chart of ocs_before vs study abroad
notAbroad %>%
  drop_na(ocs_before) %>%
  ggplot() +
  geom_bar(aes(reason_not_2.0, fill = ocs_before)) +
  labs(x = "Reason for not attending OCS program",
       title = "Students who did not attend OCS program
       (by whether or not the students planned to attend OCS program
       before coming to Carleton)",
       fill = "Plan") +
  theme(
    panel.background = element_rect(fill = "black",colour = "white",size = 0.5),
    panel.grid.major = element_line(colour = "grey70"),
    axis.title.y = element_blank())
```

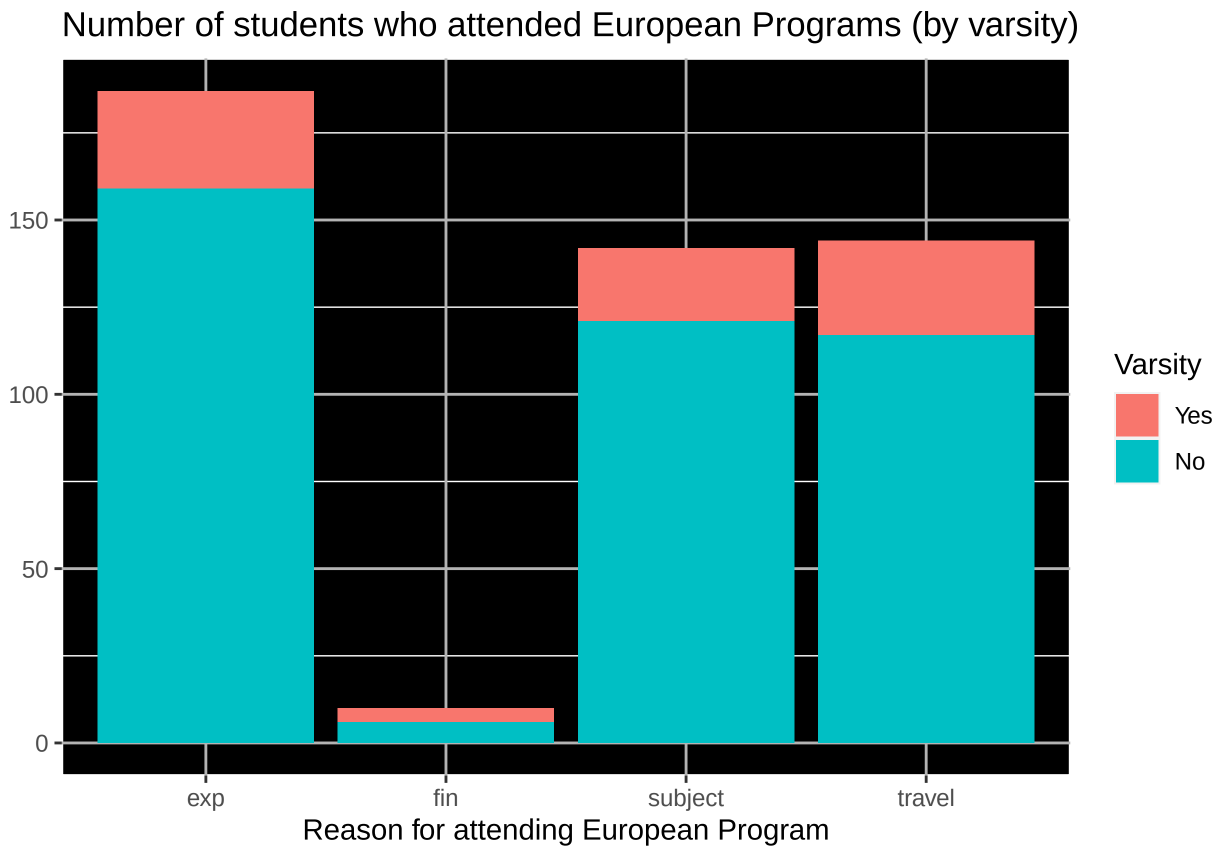


```
#INSERT REASON_NOT_TEXT VISUALIZATION FROM QUALTRICS AFTER THIS
```

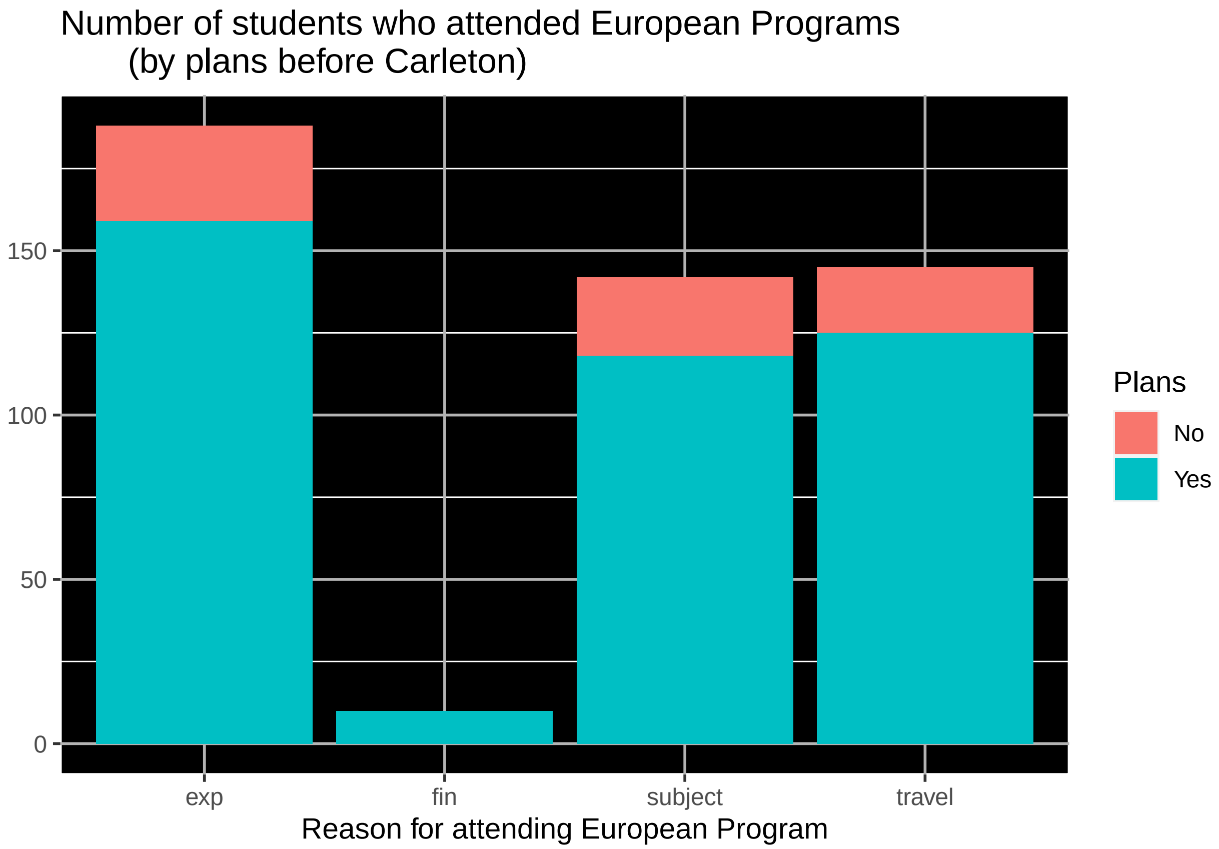
```
#RELATIONSHIP BETWEEN MAJOR AND WHY EUROPE

majorEurope <- ocs_pivottedMajor %>%
  filter(europe == "Yes") %>%
  select(reason_europe, reason_europe_text, varsity, ocs_before, noDoubleMajor)

majorEurope$exp <- str_detect(majorEurope$reason_europe, pattern = zero_or_more(ALPHA) %R% "Explore")
majorEurope$subject <- str_detect(majorEurope$reason_europe, pattern = zero_or_more(ALPHA) %R% "Subject
majorEurope$fin <- str_detect(majorEurope$reason_europe, pattern = zero_or_more(ALPHA) %R% "Financial")
majorEurope$travel <- str_detect(majorEurope$reason_europe, pattern = zero_or_more(ALPHA) %R% "Travel")
majorEurope$lang <- str_detect(majorEurope$reason_europe, pattern = zero_or_more(ALPHA) %R% "Language")

majorEurope <- pivot_longer(majorEurope, cols = exp:lang,
                                  names_to = "reason_europe_2.0",
                                  values_to = "reason_europe_TF") %>%
  filter(reason_europe_TF == "TRUE") %>%
  select(-reason_europe_TF)

#bar chart of majors vs studying in Europe

majorEurope %>%
  filter(noDoubleMajor != "other") %>%
  ggplot() +
  geom_bar(aes(reason_europe_2.0, fill = noDoubleMajor)) +
  labs(x = "Reason for attending European Program",
       title = "Number of students who attended European Programs (by major)",
       fill = "Field of Major") +
  theme(
    panel.background = element_rect(fill = "black",colour = "white",size = 0.5),
    panel.grid.major = element_line(colour = "grey70"),
    axis.title.y = element_blank())
```



```
#bar chart of varsity vs studying in Europe

majorEurope %>%
  drop_na(varsity) %>%
  filter(noDoubleMajor != "other") %>%
  ggplot() +
  geom_bar(aes(reason_europe_2.0, fill = varsity)) +
  labs(x = "Reason for attending European Program",
       title = "Number of students who attended European Programs (by varsity)",
       fill = "Varsity") +
  theme(
    panel.background = element_rect(fill = "black",colour = "white",size = 0.5),
    panel.grid.major = element_line(colour = "grey70"),
    axis.title.y = element_blank())
```



```
#
majorEurope %>%
  drop_na(ocs_before) %>%
  filter(noDoubleMajor != "other") %>%
  ggplot() +
  geom_bar(aes(reason_europe_2.0, fill = ocs_before)) +
  labs(x = "Reason for attending European Program",
       title = "Number of students who attended European Programs
       (by plans before Carleton)",
       fill = "Plans") +
  theme(
    panel.background = element_rect(fill = "black",colour = "white",size = 0.5),
    panel.grid.major = element_line(colour = "grey70"),
    axis.title.y = element_blank())
```



```
#INSERT REASON_EUROPE_TEXT VISUALIZATION FROM QUALTRICS AFTER THIS
```