ocsProject

Kunsang

4/15/2020

```
ocs <- read_excel("ocs_data.xlsx")
```

```
## New names:
## * `` -> ...12
```

```
unique(ocs$`Ethnic Group Code`)
```

```
## [1] "FW" "FU" "FT" "FB" "FH" "FI" "FA" "FN" "FP"
```

```
#what is FN?
for (row in 1:nrow(ocs)) {
    ethnic <- ocs[row, "Ethnic Group Code"]
    if(ethnic == 'FW') {
        ocs[row, "Ethnic Group Code"] <- 'White'
    } else if(ethnic == 'FU') {
        ocs[row, "Ethnic Group Code"] <- 'Unidentified'
    } else if(ethnic == 'FT') {
        ocs[row, "Ethnic Group Code"] <- 'Multiple'
    } else if(ethnic == 'FB') {
        ocs[row, "Ethnic Group Code"] <- 'Black'
    } else if(ethnic == 'FH') {
        ocs[row, "Ethnic Group Code"] <- 'Hispanic'
    } else if(ethnic == 'FI') {
        ocs[row, "Ethnic Group Code"] <- 'International'
    } else if(ethnic == 'FA') {
        ocs[row, "Ethnic Group Code"] <- 'Asian'
    } else if(ethnic == 'FP') {
        ocs[row, "Ethnic Group Code"] <- 'PacificIslander'
    }
}
```

```
#double major problem
#separate double major into two columns
ocs <- separate(ocs, 'Majors', paste("Major", 1:2, sep="_"),sep=",", extra = "drop")
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 3858 rows [1, 2,
## 3, 5, 6, 9, 10, 12, 13, 14, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, ...].
```

```
ocs$Major_1 <- as.factor(ocs$Major_1) #convert as factors
ocs$Major_2 <- as.factor(ocs$Major_2) #convert as factors

#same with minor
ocs <- separate(ocs, 'Minors', paste("Minor", 1:2, sep="_"),sep=",", extra = "drop")
```

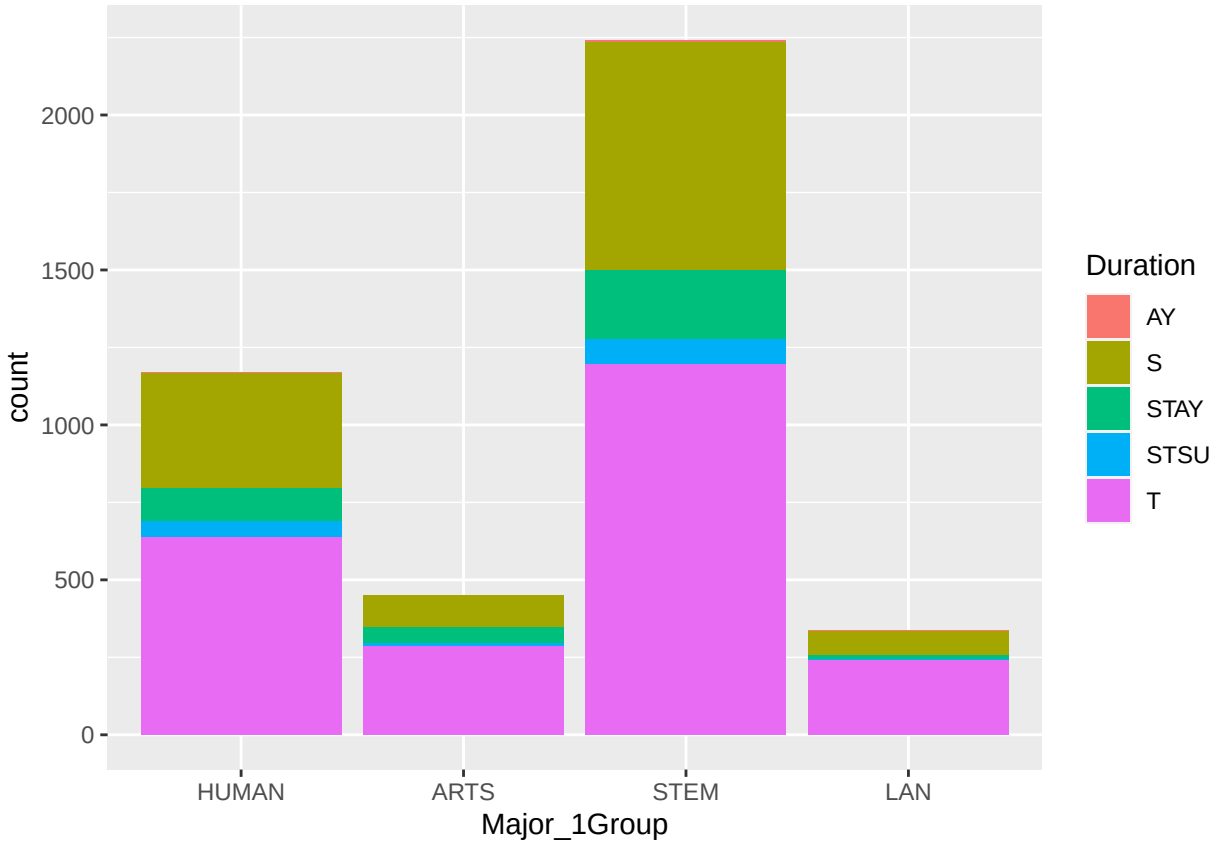
```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 1054 rows [5, 6,
```

```
## 9, 12, 17, 22, 35, 40, 45, 46, 48, 56, 71, 73, 75, 82, 83, 84, 86, 87, ...].
```

```
ocs$Minor_1 <- as.factor(ocs$Minor_1) #convert as factors
ocs$Minor_2 <- as.factor(ocs$Minor_2) #convert as factors

#create a new column(group)
ocs <- mutate(ocs, Major_1Group = fct_collapse(Major_1, ARTS=c("ARTH","ARTS","CAMS","DANCE","MUSC","THEA
 STEM=c("BIOL","CGSC","CHEM","COGSC","CS","ECON","ENST","GEOL","MATH","MATS","PHYS","PSYC","STAT"),
HUMAN=c("AFAM","AFST","AMEST","AMST","ASST","CLAS","CLSS","HIST","LING","LTAM","PHIL","POSC","POSI","PP
 LAN=c("CHINA","CLLN","ENGL","FRST","JALLI","JLALI","RUSS","SPAN","GERM"),
OTHER=c("SPECL","UNDC")))
```

```
ocs %>%
  drop_na(Major_1Group, Duration) %>%
  filter(Major_1Group != "OTHER") %>%
  ggplot() +
  geom_bar(aes(Major_1Group, fill = Duration))
```




```
ocs %>%
  filter(Duration == "S" | Duration == "T") %>%
  filter(Major_1Group != "OTHER") %>%
  group_by(Major_1Group, Duration) %>%
  summarize(count = n())
```

```
## `summarise()` regrouping output by 'Major_1Group' (override with `.groups` argument)
```

```
## # A tibble: 8 x 3
## # Groups:   Major_1Group [4]
##   Major_1Group Duration count
```

```
##   <fct>        <chr>    <int>
## 1 HUMAN        S          370
## 2 HUMAN        T          637
## 3 ARTS         S          102
## 4 ARTS         T          284
## 5 STEM         S          735
## 6 STEM         T         1196
## 7 LAN          S           76
## 8 LAN          T          240
```

```
t <- ocs %>%
  filter(Duration == "S" | Duration == "T") %>%
  filter(Major_1Group != "OTHER")

table(t$Major_1Group, t$Duration)
```

```
##
##            S    T
##   HUMAN  370  637
##   ARTS   102  284
##   STEM   735 1196
##   LAN     76  240
##   OTHER    0    0
```

```
y <- c(370, 637, 102, 284, 735, 1196, 76, 240)
major <- factor(c(rep("humanities",2),rep("arts",2), rep("stem",2), rep("language",2)))
duration <- factor(rep(c("Semester","Term"),4))
major_duration_df <- data.frame(y, major, duration)
major_duration_df
```

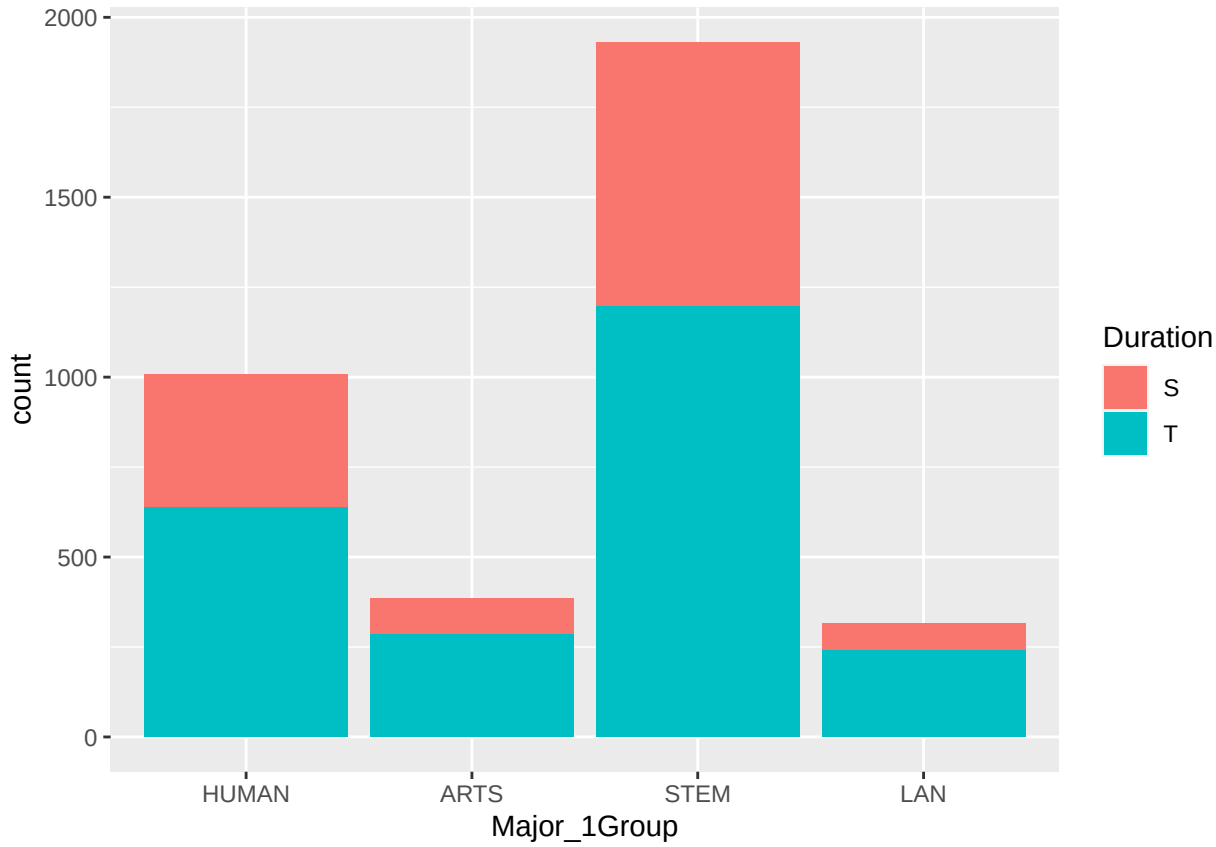
```
##      y      major duration
## 1  370 humanities Semester
## 2  637 humanities     Term
## 3  102       arts Semester
## 4  284       arts     Term
## 5  735       stem Semester
## 6 1196       stem     Term
## 7   76   language Semester
## 8  240   language     Term
```

```
major_duration_glm <- glm(y ~ major + duration + major:duration, family=poisson, data=major_duration_df)
summary(major_duration_glm)
```

```
##
## Call:
## glm(formula = y ~ major + duration + major:duration, family = poisson,
##     data = major_duration_df)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  4.62497    0.09901  46.710  < 2e-16 ***
## majorhumanities              1.28853    0.11183  11.522  < 2e-16 ***
## majorlanguage               -0.29424    0.15153  -1.942  0.05216 .
```

```
## majorstem                      1.97490    0.10566  18.691  < 2e-16 ***
## durationTerm                   1.02400    0.11543   8.871  < 2e-16 ***
## majorhumanities:durationTerm -0.48073    0.13266  -3.624  0.00029 ***
## majorlanguage:durationTerm    0.12590    0.17507   0.719  0.47204
## majorstem:durationTerm       -0.53713    0.12459  -4.311 1.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2.1405e+03  on 7  degrees of freedom
## Residual deviance: 9.5479e-15  on 0  degrees of freedom
## AIC: 76.851
##
## Number of Fisher Scoring iterations: 2
```

```
ocs %>%
  filter(Duration == "S" | Duration == "T") %>%
  filter(Major_1Group != "OTHER") %>%
  group_by(Major_1Group, Duration) %>%
  ggplot() +
  geom_bar(aes(Major_1Group, fill = Duration))
```




```
unique(ocs$Term)
```

```
##  [1] "15/WI" "14/FA" "13/FA" "13/WI" "17/SU" "19/WI" "17/FA" "12/WI" "11/FA"
## [10] "15/FA" "18/WI" "20/WI" "16/SP" "17/WI" "12/FA" "18/FA" "13/SP" "11/WI"
## [19] "10/FA" "12/SP" "14/SP" "15/SP" "19/SP" "19/SU" "16/FA" "19/FA" "16/WI"
```

```
## [28] "12/SU" "12/WS" "11/SU" "14/WI" "16/WS" "18/SP" "11/SP" "15/SU" "17/SP"
## [37] "14/SU" "16/SU" "10/SU" "13/SU" "13/AY" "20/WS" "18/SU" "11/AY" "10/AY"
## [46] "17/AY" "18/WS" "11/WS" "18/AY" "14/WS" "12/AY" "15/AY" "15/WS" "14/AY"
## [55] "19/WS" "16/AY" "17/FW" "19/AY" "17/WS" "19/FW" "11/FW"
```

```
ocs$winter <- str_detect(ocs$Term, pattern = zero_or_more(ALPHA) %R% "WI")
ocs$fall <- str_detect(ocs$Term, pattern = zero_or_more(ALPHA) %R% "FA")
ocs$summer <- str_detect(ocs$Term, pattern = zero_or_more(ALPHA) %R% "SU")
ocs$winterspring <- str_detect(ocs$Term, pattern = zero_or_more(ALPHA) %R% "WS")
ocs$spring <- str_detect(ocs$Term, pattern = zero_or_more(ALPHA) %R% "SP")
ocs$allyear <- str_detect(ocs$Term, pattern = zero_or_more(ALPHA) %R% "AY")
ocs$fallwinter <- str_detect(ocs$Term, pattern = zero_or_more(ALPHA) %R% "FW")

ocs_pivot <- pivot_longer(ocs, cols = winter:fallwinter,
                                  names_to = "term_2.0",
                                  values_to = "termTF") %>%
  filter(termTF == "TRUE") %>%
  select(-termTF)


ocs_pivot %>%
  filter(Duration == "S" | Duration == "T") %>%
  filter(Major_1Group != "OTHER") %>%
  filter(term_2.0 == "fall" | term_2.0 == "winter" | term_2.0 == "spring") %>%
  group_by(Major_1Group, Duration, term_2.0) %>%
  summarize(count = n())
```

```
## `summarise()` regrouping output by 'Major_1Group', 'Duration' (override with `.groups` argument)
```

```
## # A tibble: 20 x 4
## # Groups:   Major_1Group, Duration [8]
##    Major_1Group Duration term_2.0 count
##    <fct>        <chr>    <chr>    <int>
##  1 HUMAN        S        fall       356
##  2 HUMAN        S        spring       2
##  3 HUMAN        S        winter       5
##  4 HUMAN        T        fall        94
##  5 HUMAN        T        spring     324
##  6 HUMAN        T        winter     171
##  7 ARTS         S        fall       100
##  8 ARTS         T        fall        22
##  9 ARTS         T        spring     103
## 10 ARTS         T        winter     141
## 11 STEM         S        fall       699
## 12 STEM         S        spring       5
## 13 STEM         S        winter      21
## 14 STEM         T        fall       256
## 15 STEM         T        spring     225
## 16 STEM         T        winter     511
## 17 LAN          S        fall        73
## 18 LAN          T        fall        39
## 19 LAN          T        spring     107
## 20 LAN          T        winter      57
```

```
ocs_pivot %>%
  filter(Duration == "S" | Duration == "T") %>%
```

```
  filter(Major_1Group != "OTHER") %>%
  filter(term_2.0 == "fall" | term_2.0 == "winter" | term_2.0 == "spring") %>%
  group_by(Major_1Group, Duration, term_2.0) %>%
  ggplot() +
  geom_bar(aes(Major_1Group, fill = Duration)) +
  facet_wrap(term_2.0~.)
```