Data Viz Final Project

Ari Kassin-Fuentes

2023-05-11

Introduction

The dataset consists of survey results given to UTEP students throughout the last 3 to 4 years regarding food security. The survey had many questions asking for demographic, academic and socioeconomic information. Also, it asks about academic performance factors, for example, ability to concentrate or degree completeness due to lack of food security.

My approach to this final project is to use each question as a guide to start exploring the dataset and produce stories and insightful visualizations that together, tackle each question.

Stories and Questions

Next, I will present each question with the story and visualizations I created to answer it.

Data Setup

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggrepel)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
library(leaflet)
library(leaflet.extras)
```

```
## Warning: package 'leaflet.extras' was built under R version 4.1.3
```

```
library(ggplot2)
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.1.3
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(pls)
```

```
##
## Attaching package: 'pls'
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```
library(WOCR)
library(pracma)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:pracma':
##
##     logit
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(DT)
library(KernSmooth)
```

```
## Warning: package 'KernSmooth' was built under R version 4.1.3
```

```
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
```

```
library(raster)
```

```
## Warning: package 'raster' was built under R version 4.1.3
```

```
## Loading required package: sp
```

```
##
```

```
## Attaching package: 'raster'

## The following object is masked from 'package:MASS':
##
##     select

## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(rgdal)
```

```
## Warning: package 'rgdal' was built under R version 4.1.3

## Please note that rgdal will be retired by the end of 2023,
## plan transition to sf/stars/terra functions using GDAL and PROJ
## at your earliest convenience.
##
## rgdal: version: 1.5-32, (SVN revision 1176)
## Geospatial Data Abstraction Library extensions to R successfully loaded
## Loaded GDAL runtime: GDAL 3.4.1, released 2021/12/27
## Path to GDAL shared files: C:/Users/kassinfuentesa/Documents/R/win-library/4.1/rgdal/gdal
## GDAL binary built with GEOS: TRUE
## Loaded PROJ runtime: Rel. 7.2.1, January 1st, 2021, [PJ_VERSION: 721]
## Path to PROJ shared files: C:/Users/kassinfuentesa/Documents/R/win-library/4.1/rgdal/proj
## PROJ CDN enabled: FALSE
## Linking to sp version:1.5-0
## To mute warnings of possible GDAL/OSR exportToProj4() degradation,
## use options("rgdal_show_exportToProj4_warnings"="none") before loading sp or rgdal.
```

```
library(ggforce)
```

```
## Warning: package 'ggforce' was built under R version 4.1.3
```

```
library(treemapify)
```

```
## Warning: package 'treemapify' was built under R version 4.1.3
```

```
library(ggmosaic)
```

```
## Warning: package 'ggmosaic' was built under R version 4.1.3

##
## Attaching package: 'ggmosaic'

## The following object is masked from 'package:raster':
##
##     mosaic
```

```
library(forcats)
library(rcompanion)
library(lsr)
```

```
## Warning: package 'lsr' was built under R version 4.1.3

##
## Attaching package: 'lsr'

## The following object is masked from 'package:pracma':
##
##     who
```

```
library(vcd)
```

```
## Loading required package: grid
```

```
##
## Attaching package: 'vcd'
```

```
## The following objects are masked from 'package:ggmosaic':
##
##     mosaic, spine
```

```
## The following object is masked from 'package:raster':
##
##     mosaic
```

```
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.1.3
```

```
##
## Attaching package: 'DescTools'
```

```
## The following object is masked from 'package:car':
##
##     Recode
```

```
## The following objects are masked from 'package:pracma':
##
##     Mode, Rank
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.2     v purrr   0.3.4
## v tidyr   1.1.3     v stringr 1.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x scales::col_factor() masks readr::col_factor()
## x purrr::cross()       masks pracma::cross()
## x purrr::discard()     masks scales::discard()
## x tidyr::extract()     masks raster::extract()
## x dplyr::filter()      masks stats::filter()
## x dplyr::lag()         masks stats::lag()
## x car::recode()        masks dplyr::recode()
## x raster::select()     masks MASS::select(), dplyr::select()
## x purrr::some()        masks car::some()
## x tidyr::who()         masks lsr::who(), pracma::who()
```

```
library(treemap)
```

```
## Warning: package 'treemap' was built under R version 4.1.3
```

```
options(scipen=999)  # turn off scientific notation like 1e+06

df <- read.csv("master.csv", header = TRUE, sep=",", na.strings = "")
dim(df)
```

```
## [1] 11763    59
```

```r
df2 = filter(df, Year != 2022)
dim(df2)
```

```
## [1] 10020    59
```

```r
df3 = filter(df2, FedAid != "UTEP's COVID CARES Act Fund")
dim(df3)
```

```
## [1] 9999   59
```

```r
table(df3$FedAid)
```

```
##
## Emergency Loan         Grants          Loans          Other    Scholarship
##            163           1689           3020              4            952
##     Work-study
##           4171
```

```r
table(df3$USDAcat)
```

```
##
##           Low FS Marginal/High FS               NA      Very Low FS
##             1918             3903             1385             2793
```

```r
# Additional data
extraDF <- read.csv("extra_questions_withID.csv", header = TRUE, sep=",", na.strings = "")
dim(extraDF)
```

```
## [1] 1743   57
```

```r
common_colnames <- intersect(names(df), names(extraDF))
full_data <- merge(df, extraDF, by=common_colnames)

dim(full_data)
```

```
## [1] 1743  115
```

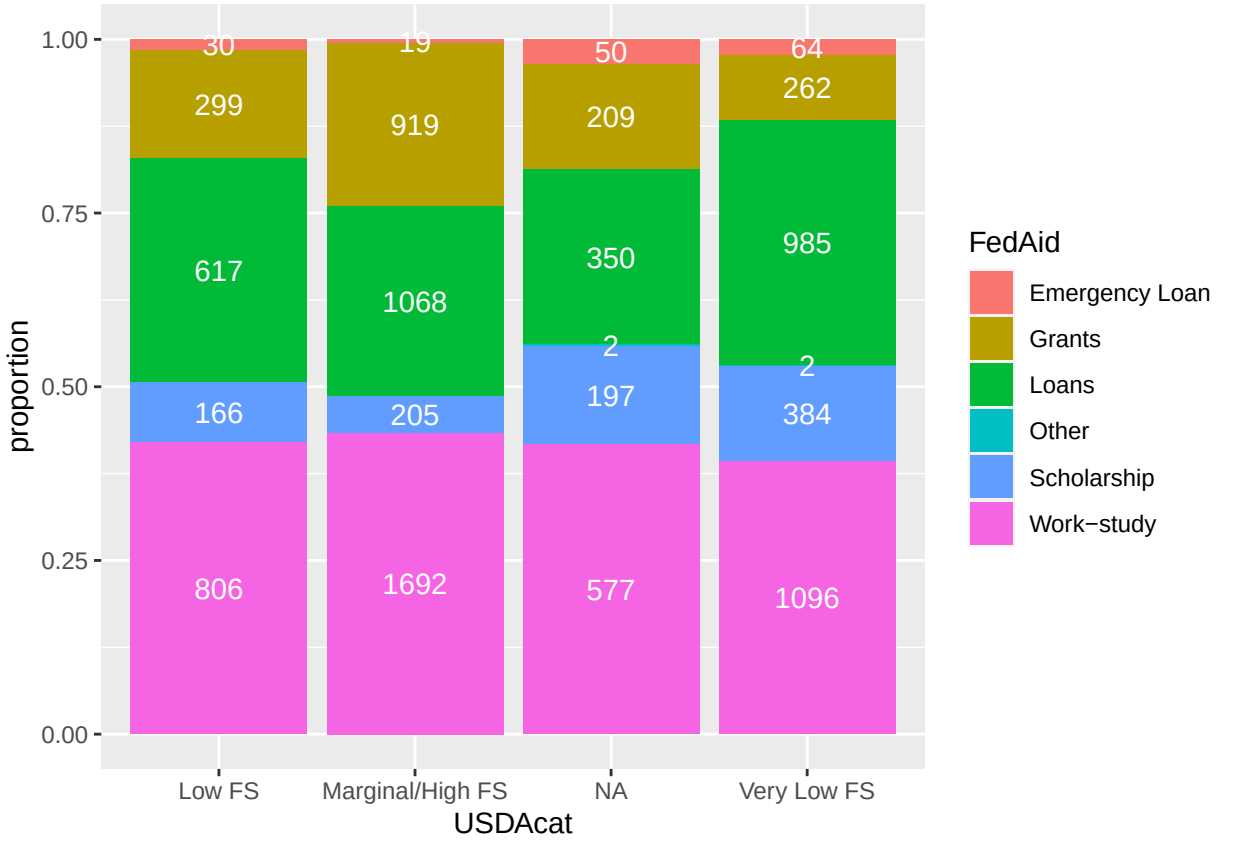
Question 1 - How is use of government federal aid/assistance associated with food insecurity as measured by the USDA index or categories?

The next chart plots USDAcat vs Fedaid variables. Notice that for Very Low FS and Low FS, there is a difference between Grants and Loans. I want to explore this further but before that, let's check other variables such as Income.

```r
# USDAcat vs FedAid
ggplot(df3, aes(x = USDAcat, fill = FedAid)) +
  geom_bar(position = "fill") + ylab("proportion") +
  stat_count(geom = "text",
             aes(label = stat(count)),
             position=position_fill(vjust=0.5), colour="white")
```

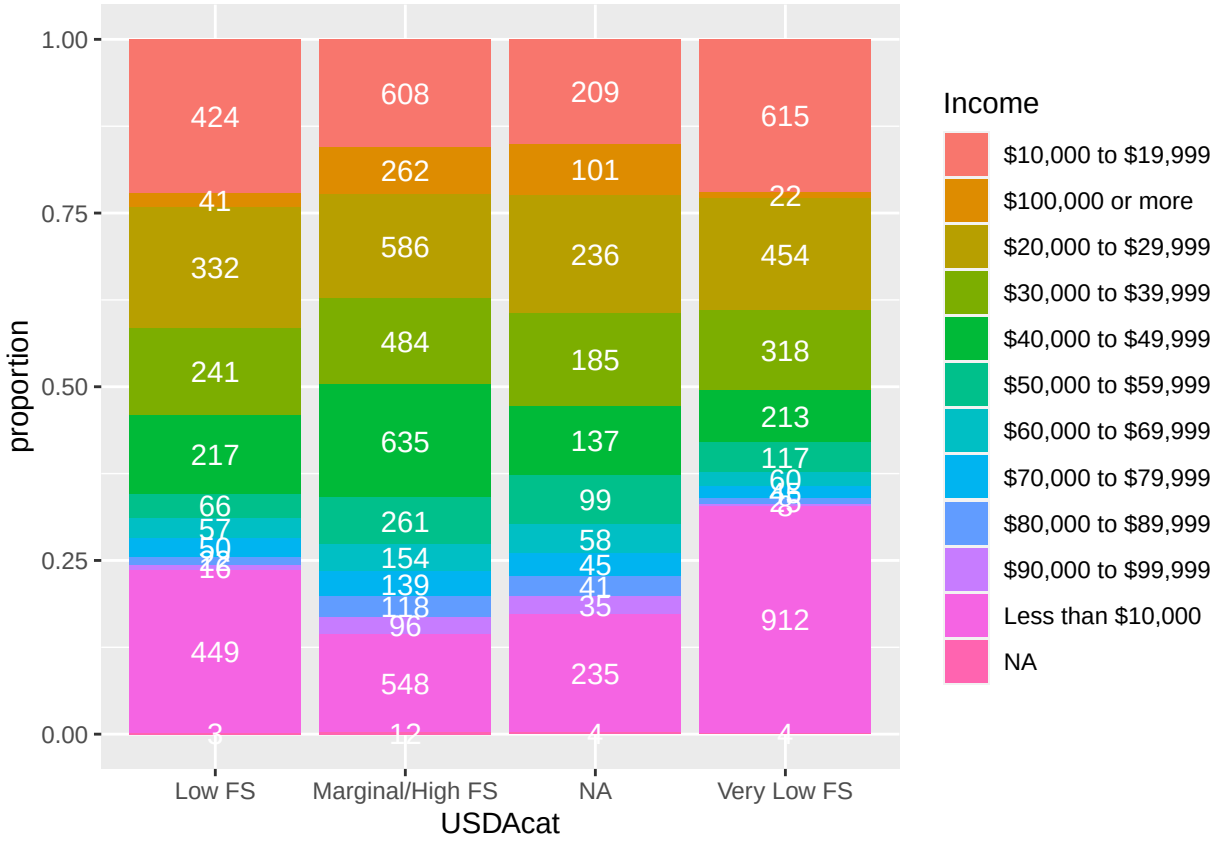
```
## Warning: `stat(count)` was deprecated in ggplot2 3.4.0.

## Warning: Please use `after_stat(count)` instead.
```



The next chart plots USDAcat vs Income variables. As expected, for Very Low FS and Low FS there is a lot of students with an income of less than $10,000 or between $10,000 and 19,999. This means that the students with the lowest income are the ones struggling.

```
# USDAcat vs Income
ggplot(df3, aes(x = USDAcat, fill = Income)) +
  geom_bar(position = "fill") + ylab("proportion") +
  stat_count(geom = "text",
             aes(label = stat(count)),
             position=position_fill(vjust=0.5), colour="white")
```



For the next plot I want to focus on Very Low FS and Less than $10,000, hence I will filter the data to those values only. Now the following plot shows frequency by Academic level to see the spread of students within the filtered data. As we can see, surprisingly seniors and juniors are the 2 top categories.

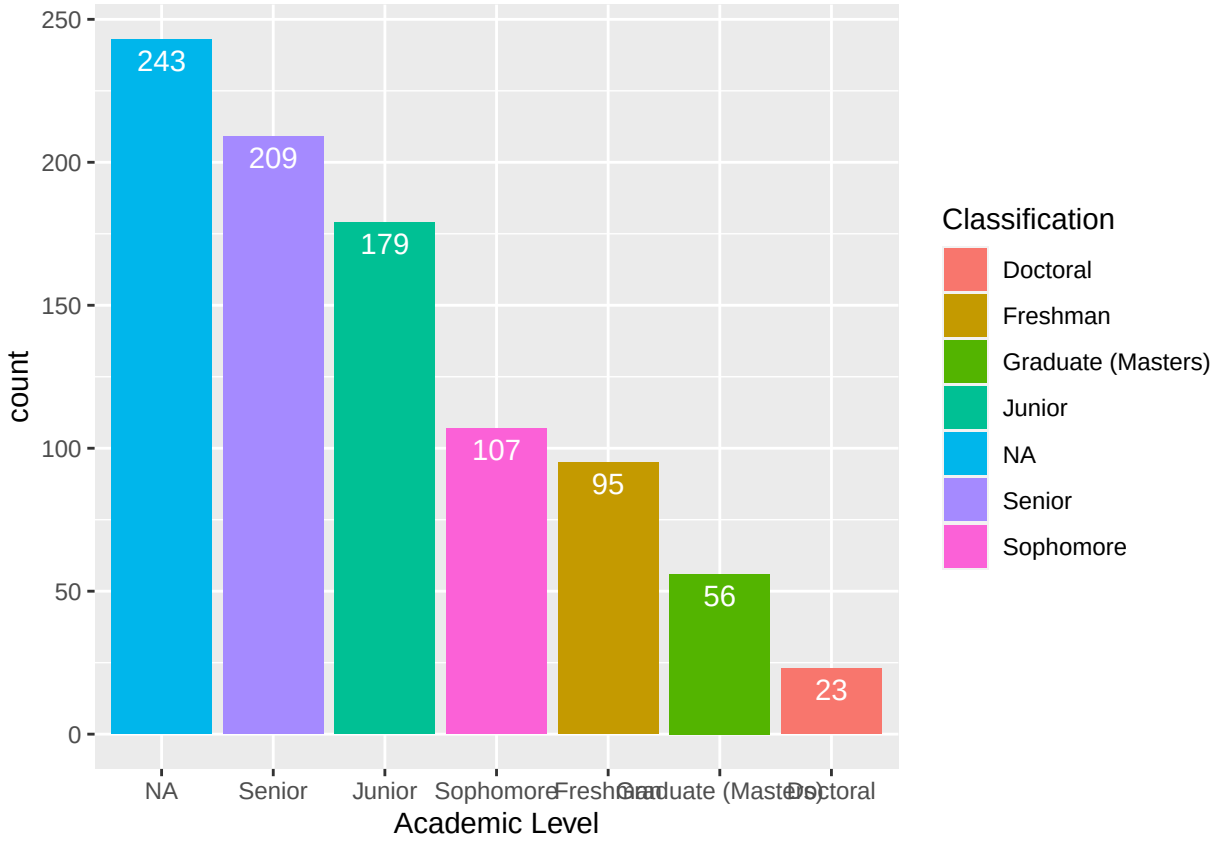
```
# Filter by Very Low FS and Less than 10,000
dfFiltered = filter(df3, USDAcat=="Very Low FS", Income =="Less than $10,000")
dim(dfFiltered)
```

```
## [1] 912  59
```

```
table(dfFiltered$USDAcat)
```

```
##
## Very Low FS
##         912
```

```
# Filtered by Very Low FS and Less than 10,000 then frequency by Academic Level
ggplot(dfFiltered, aes(x = fct_infreq(Classification), fill = Classification)) +
  geom_bar() +
  labs(x = "Academic Level") +
  stat_count(geom = "text",
             aes(label = stat(count)), vjust = 1.5,
             position=position_dodge(width = 0.9), colour="white")
```

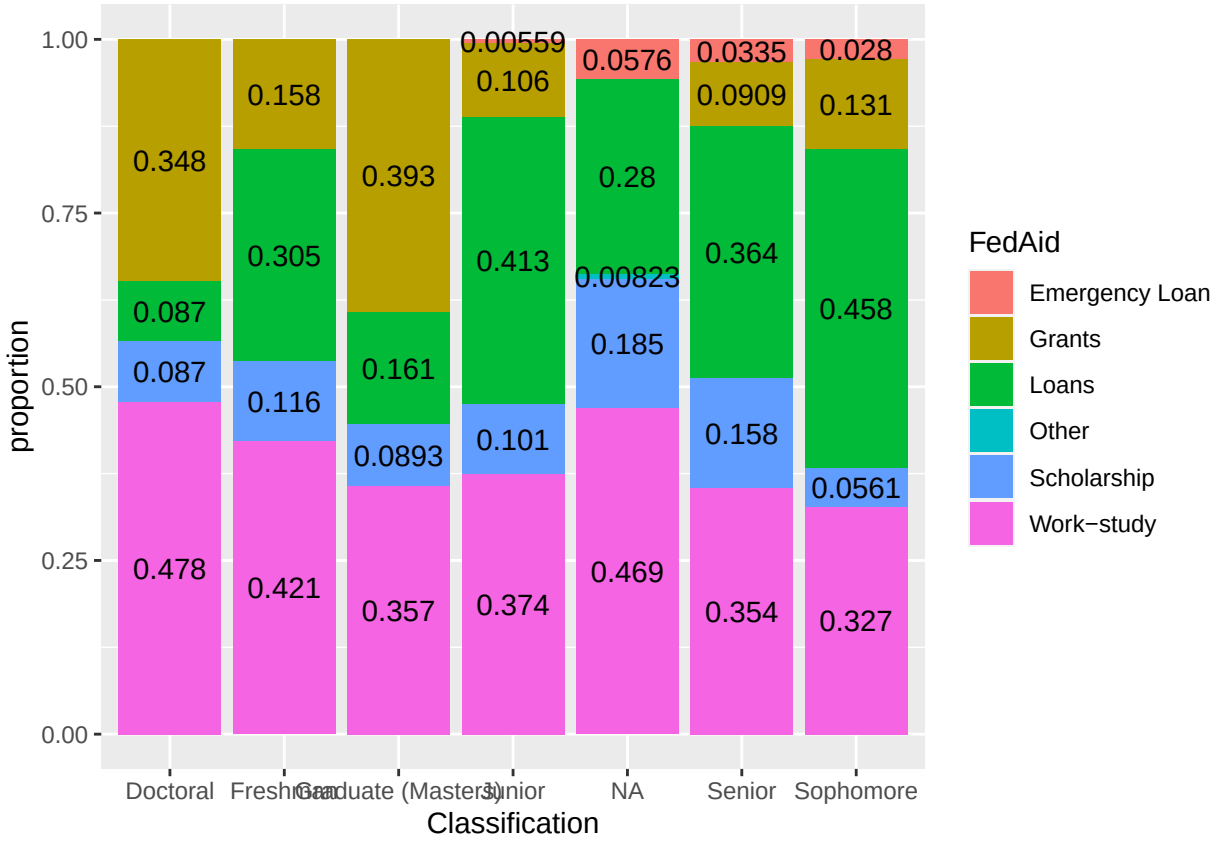


Finally, the next plot shows Academic level vs FedAid using the filtered data by Very Low FS and Less than $10,000. What I see is that Seniors and Juniors have way more Loans than Grants.

```
ggplot(dfFiltered, aes(x = Classification, fill = FedAid)) +
  geom_bar(position = "fill") + ylab("proportion") +
  geom_text(
    aes(label=signif(..count.. / tapply(..count.., ..x.., sum)[as.character(..x..)], digits=3)),
    stat="count",
    position=position_fill(vjust=0.5))
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.

## Warning: Please use `after_stat(count)` instead.
```



Overall story and conclusion: Data filtered by Very Low FS and Less than 10,000. Variables were Academic level and FedAid. Graduate and Doctoral students have majority of Grants instead of Loans, in contrast, seniors, juniors and sophomore have majority of loans and almost no grants. My conclusion is that for the group with the lowest income and less food security (poorest) it is counterproductive to loan, i.e. ask to pay back the money. It is best to give the money for free, i.e. grants. Maybe grants should go to income level instead of grades.

Question 2 - Does food insecurity (as measured by USDA index or categories) have a relationship with the items pertaining to concentration on school and degree progress/completion?

For this question, I merged the two datasets provided by the project. Something to note is that only records for 2022 had complete data, specifically, the index for food security. Because of this, the dataset will be reduced to 2022 data only.

First I wanted to check relationship between variables, in other words, check if the variables are dependent or independent amongst themselves. Using the following methods to check independence in categorical variables; Chi-squared test, (corrected) contingency coefficient and Cramer's V, we can state that since we get a p-value of less than the significance level of 0.05, we can reject the null hypothesis and conclude that the variables are, indeed, dependent.

Next, I plotted I raster graph between three variables - DiffConcentrate, DelaycompDegree and Index. I can see that students who answered Almost every day and Once a week have the greater index value, meaning, Very Low FS. I want to explore other variables next.

```
table(full_data$DelayComplDegree)
```

```
##
##                          NA                          No
```

```
##                          28                         1325
##          Yes, by 1 semester Yes, by 2 semesters or more
##                         180                          210
```

```
table(full_data$DiffConcentrate)
```

```
##
## About once a month  About once a week   Almost every day                 NA
##                493                385                317                 28
##              Never
##                520
```

```
table(full_data$DiffConcentrate, full_data$DelayComplDegree)
```

```
##
##                       NA  No Yes, by 1 semester Yes, by 2 semesters or more
##   About once a month   0 408                 46                          39
##   About once a week    0 259                 68                          58
##   Almost every day     0 171                 52                          94
##   NA                  28   0                  0                           0
##   Never                0 487                 14                          19
```

```
table(full_data$index)
```

```
##
##   0   1   2   3   4   5   6  NA
##   4  14  28  61 105 155 449 927
```

```
assocstats(xtabs(~full_data$DiffConcentrate + full_data$DelayComplDegree))
```

```
##                      X^2 df P(> X^2)
## Likelihood Ratio  514.47 12        0
## Pearson          1974.06 12        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.729
## Cramer's V        : 0.614
```

```
chisq.test(full_data$DiffConcentrate, full_data$DelayComplDegree)
```

```
## Warning in chisq.test(full_data$DiffConcentrate, full_data$DelayComplDegree):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  full_data$DiffConcentrate and full_data$DelayComplDegree
## X-squared = 1974.1, df = 12, p-value < 0.00000000000000022
```

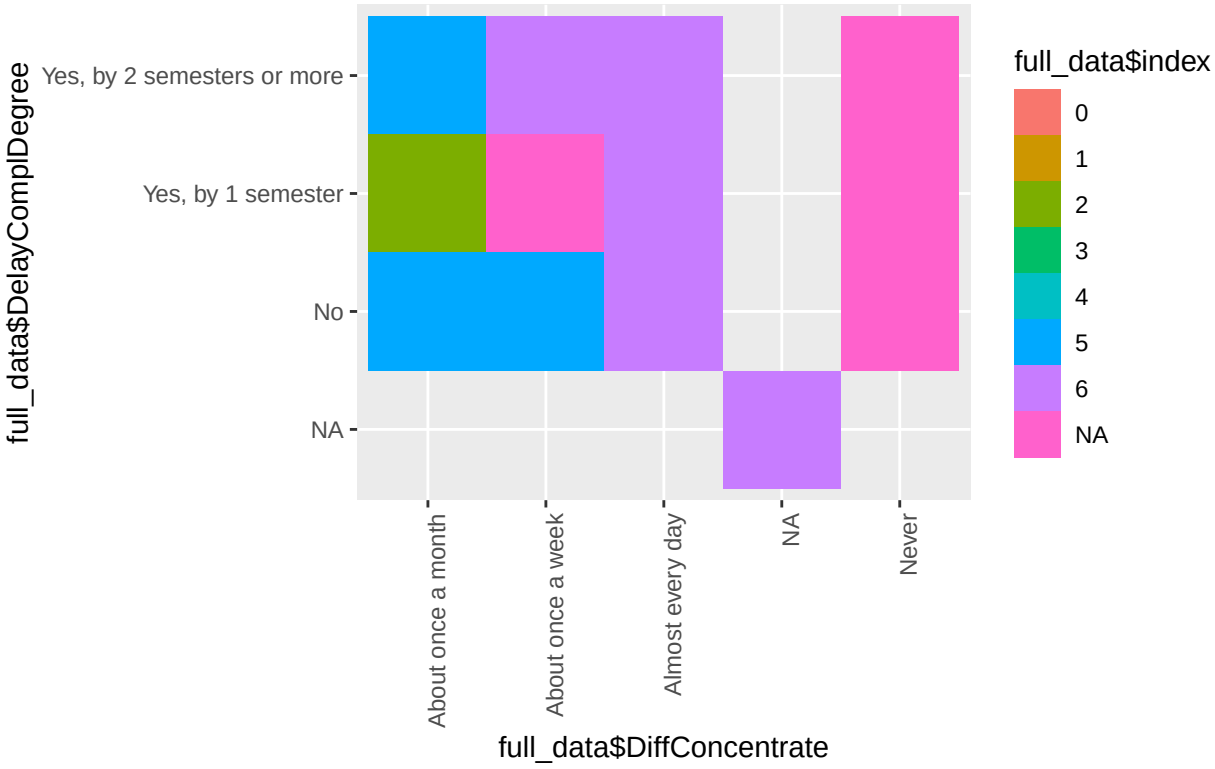
```
# Raster of the 3 variables
ggplot(full_data, aes(x=full_data$DiffConcentrate, y=full_data$DelayComplDegree, fill=full_data$index)) +
  geom_raster() +
  coord_fixed() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1))
```

```
## Warning: Use of `full_data$DiffConcentrate` is discouraged.
## i Use `DiffConcentrate` instead.
```

```
## Warning: Use of `full_data$DelayComplDegree` is discouraged.
```

```
## i Use `DelayComplDegree` instead.
```

```
## Warning: Use of `full_data$index` is discouraged.
## i Use `index` instead.
```



Index vs DiffConcentrate. Another visualization supporting that DifConcentrate and Index have a positive relationship, that is, trouble concentrating is related to food security. Next I want to explore Mental Health.

```
# index vs diffconcentrate or delay are independent also.
chisq.test(full_data$DiffConcentrate, full_data$index)
```

```
## Warning in chisq.test(full_data$DiffConcentrate, full_data$index): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  full_data$DiffConcentrate and full_data$index
## X-squared = 480.58, df = 28, p-value < 0.00000000000000022
```

```
chisq.test(full_data$index, full_data$DelayComplDegree)
```

```
## Warning in chisq.test(full_data$index, full_data$DelayComplDegree): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  full_data$index and full_data$DelayComplDegree
## X-squared = 147.42, df = 21, p-value < 0.00000000000000022
```

```
chisq.test(full_data$index, full_data$RateMentalHealth)
```

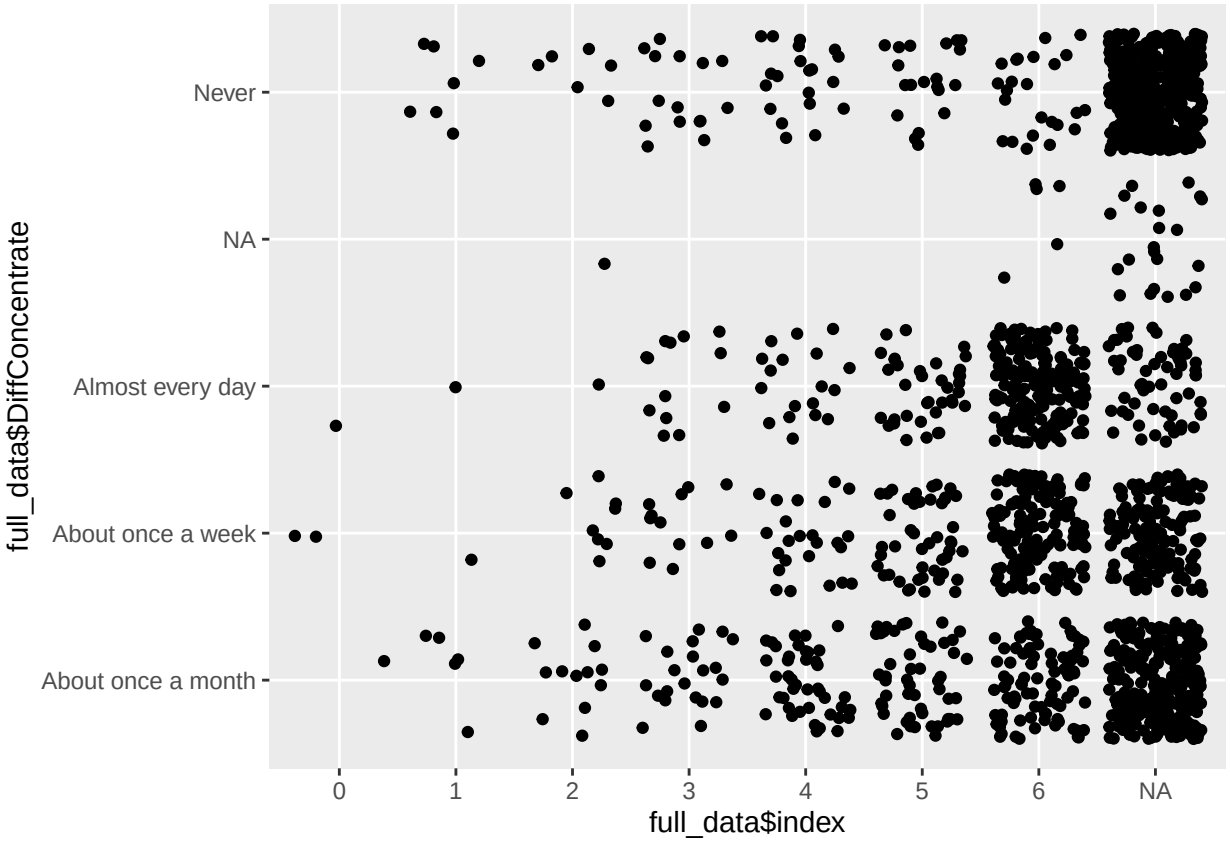
```
## Warning in chisq.test(full_data$index, full_data$RateMentalHealth): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  full_data$index and full_data$RateMentalHealth
## X-squared = 102.77, df = 35, p-value = 0.00000001389
```

```
# index vs diffconcentrate
ggplot(full_data, aes(x=full_data$index, y=full_data$DiffConcentrate)) +
  geom_jitter()
```

```
## Warning: Use of `full_data$index` is discouraged.
## i Use `index` instead.
```

```
## Warning: Use of `full_data$DiffConcentrate` is discouraged.
## i Use `DiffConcentrate` instead.
```



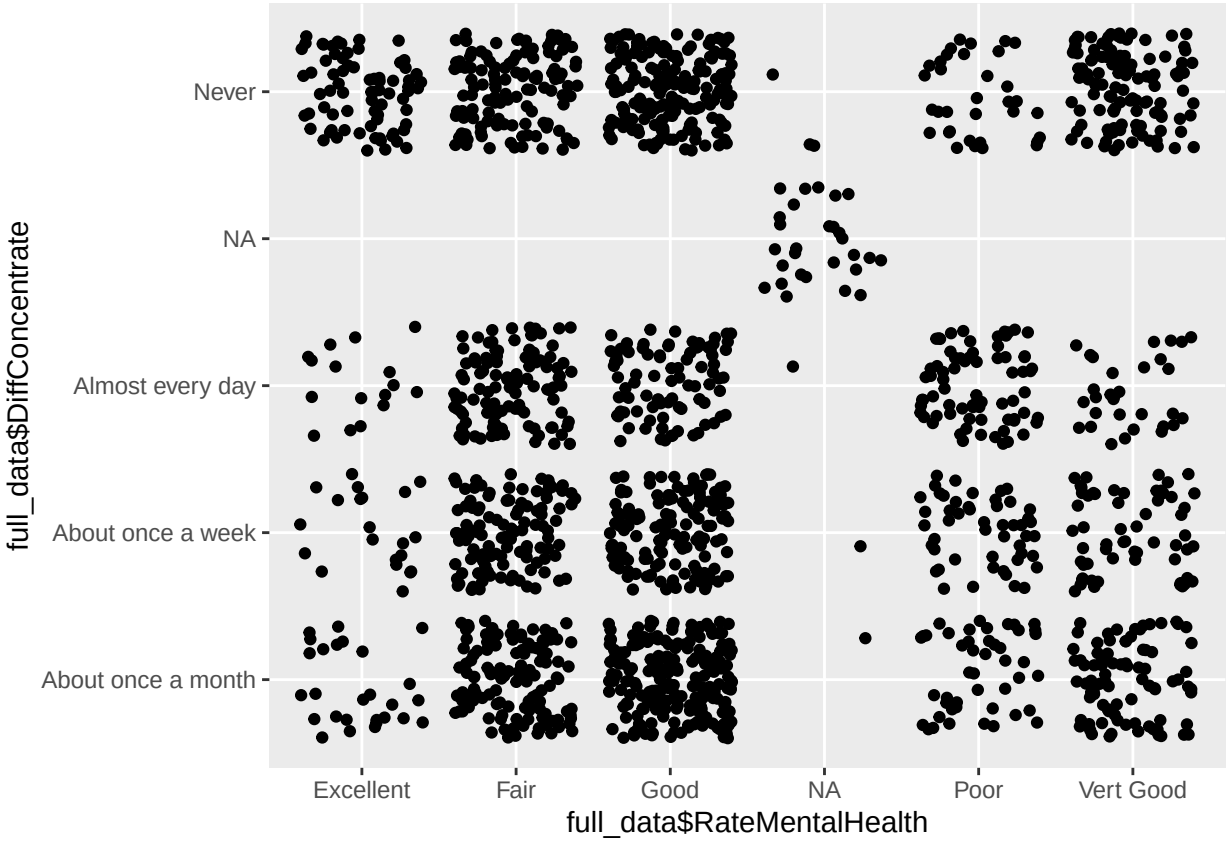
This is an interesting variable relationship between DiffConcentrate and MentalHealth. I can't tell that difficulty to concentrate produces the lowest mental health value (Poor) but if we compare it with the next lowest value, i.e., Fair then there is definitively a relationship. This is expected behavior as one would not think mental health is bad if we can't concentrate but instead could think is just Fair. Additionally, there are more concentration of values in Almost every day and Once a week.

```
# MentalHealth vs diffconcentrate
ggplot(full_data, aes(x=full_data$RateMentalHealth, y=full_data$DiffConcentrate)) +
```

```
  geom_jitter()
```

```
## Warning: Use of `full_data$RateMentalHealth` is discouraged.
## i Use `RateMentalHealth` instead.
```

```
## Warning: Use of `full_data$DiffConcentrate` is discouraged.
## i Use `DiffConcentrate` instead.
```



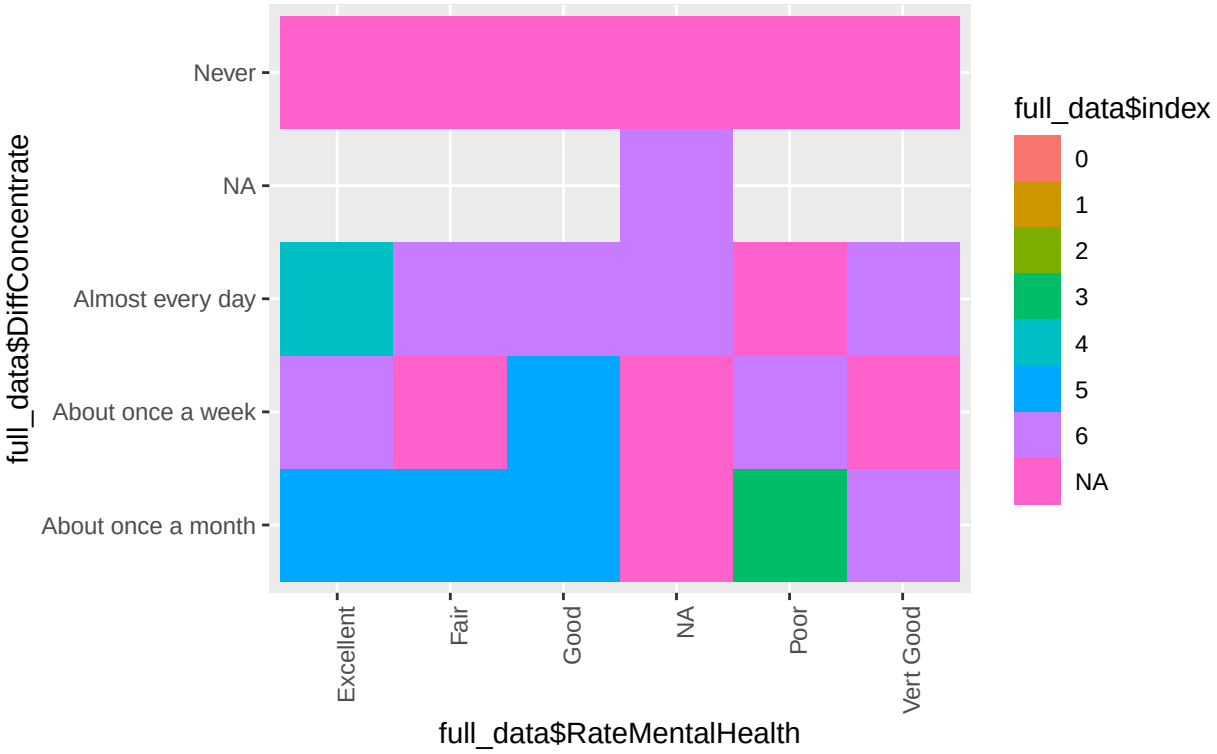
Finally, a raster plot of index vs diffconcentrate vs MentalHealth.
Using only year 2022, Plotting index with difficult to concentrate and delay completing a degree tells me that people with low FS have difficulty concentrating everyday. I want to know if this affects their mental health. It does at a certain degree as the majority of students with highest level of difficulty to concentrate (almost everyday) said their mental health is Fair. Not the extreme as in poor but fair. So this might be an indicator that food security could be an important factor in mental health but not too acute.

```
# index vs diffconcentrate vs MentalHealth
ggplot(full_data, aes(x=full_data$RateMentalHealth, y=full_data$DiffConcentrate, fill=full_data$index))
  geom_raster() +
  coord_fixed() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1))
```

```
## Warning: Use of `full_data$RateMentalHealth` is discouraged.
## i Use `RateMentalHealth` instead.
```

```
## Warning: Use of `full_data$DiffConcentrate` is discouraged.
## i Use `DiffConcentrate` instead.
```

```
## Warning: Use of `full_data$index` is discouraged.
## i Use `index` instead.
```

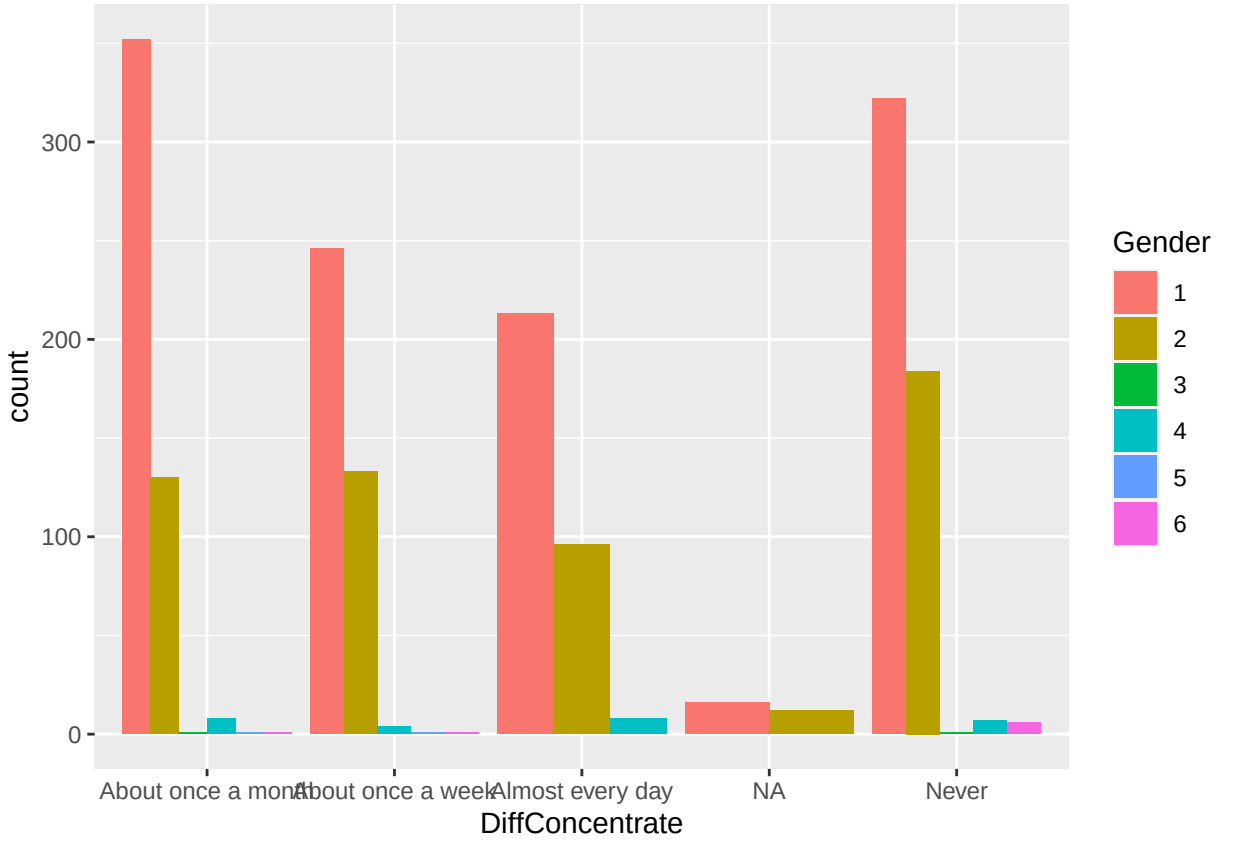


Question 3 - Are there gender or ethnicity differences in the items pertaining to concentration on school and degree progress/completion?

For the third and final question, I will focus on difficulty to concentrate and gender. I want to see if there is something interesting I can find with those variables. Reminder, this is 2022 data only.

Step one is to create a plot with DiffCouncentrate and group it by Gender. The most importnant remark here is that females outnumbered the next gender (male) almost 2 to 1, hence, I will focus on learning what is going on with this gender value.

```
ggplot(full_data, aes(fill=Gender, x=DiffConcentrate)) +
  geom_bar(position="dodge")
```



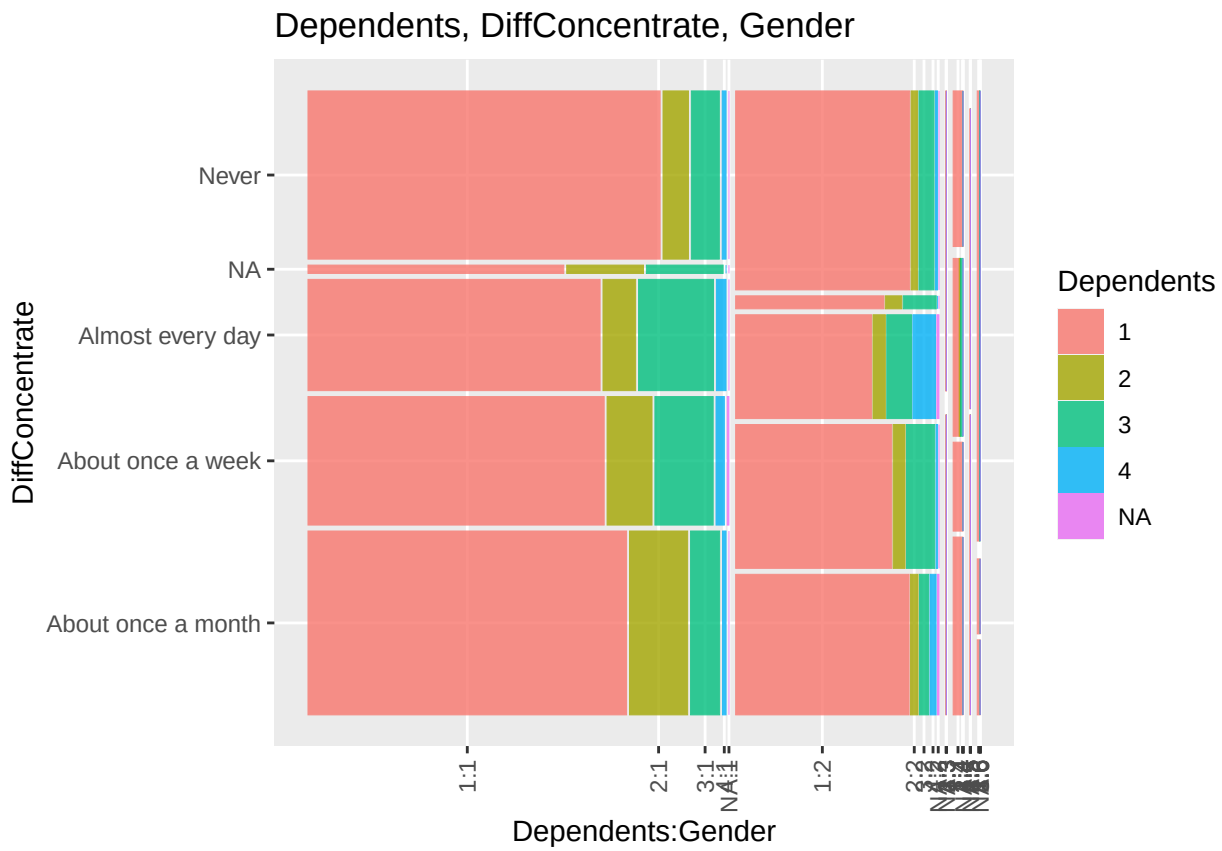

```
table(full_data$Gender)
```

```
##
##    1    2    3    4    5    6
## 1149  555    2   27    2    8
```

I want to use Dependent variable to see if Females with dependents have more issues concentrating or not. This is important because with more data, one can see if single mothers or mothers in general have more issues concentrating than male.

Using Gender, DiffConcentrate and Dependent variables I plotted a mosaic chart. I can see that women having difficulty concentrating are the ones with 2 or 1 dependents. More prominent when they have 2 dependents.

```
ggplot(data = full_data) +
  geom_mosaic(aes(x=product(Dependents, DiffConcentrate, Gender),
                  fill = Dependents)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = .5)) +
  labs(y="DiffConcentrate", x="Dependents:Gender", title = "Dependents, DiffConcentrate, Gender")
```

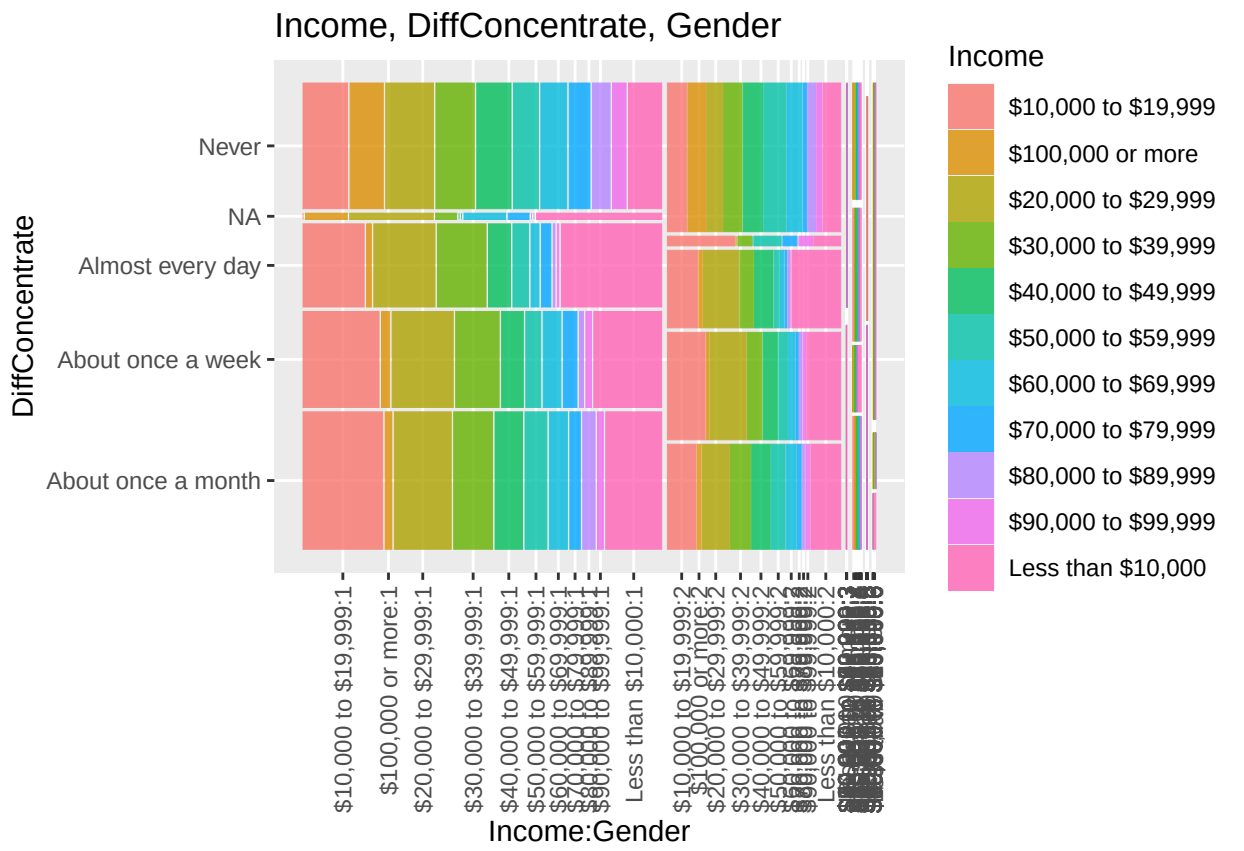


Finally, I want to see if Income plays a part in the story. I used Gender, DiffConcentrate and Income to create a mosaic plot.

The largest part of the mosaic (value = 1) is the female section. I can see that lowest income females have difficulty concentrating.

My conclusion is that females are twice as likely to have problems concentrating in school and the reason is two fold, a relationship between number of dependents and low income. This is consistent with what we observe in other social circumstances so I would recommend having more help and support to females with the above characteristics so the can concentrate in school. This could be related to food insecurity and deteriorating mental health as seen in the previous questions.

```
ggplot(data = full_data) +
  geom_mosaic(aes(x=product(Income, DiffConcentrate, Gender),
                  fill = Income)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = .5)) +
  labs(y="DiffConcentrate", x="Income:Gender", title = "Income, DiffConcentrate, Gender")
```



Conclusion

In general, I can say that low food security is more predominant in Seniors and Juniors which have difficulty to concentrate due to the fact that they don't have good income. Additionally, they use loans instead of grants which might be more of a burden that help.

Moreover, students having difficulty concentrating due to low food insecurity communicate a fair level of mental health. This is beyond hunger, meaning that this is a very important topic to tackle.

Furthermore, females are the gender most affected by low food insecurity and difficulty concentrating in school. The most affected group within women are low income and having dependents. This provides a very specific student group to target for immediate help.