ST 538 Project 1

Group 4

2024-04-18

```
r <- getOption("repos")
r["CRAN"] <- "https://cloud.r-project.org/"
options(repos=r)

if(!require(readxl)) {
    install.packages("readxl")
}
```

```
## Loading required package: readxl
```

```
if(!require(reader)) {
    install.packages("reader")
}
```

```
## Loading required package: reader
```

```
## Warning: package 'reader' was built under R version 4.3.3
```

```
## Loading required package: NCmisc
```

```
## Warning: package 'NCmisc' was built under R version 4.3.3
```

```
##
## Attaching package: 'reader'
```

```
## The following objects are masked from 'package:NCmisc':
##
##     cat.path, get.ext, rmv.ext
```

```
if(!require(car)) {
    install.packages("car")
}
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.3.3
```

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

```
## Warning: package 'carData' was built under R version 4.3.3
```

```
library(readxl)
library(readr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
```

```
## v lubridate 1.9.2     v tidyr     1.3.0
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(car)
```

#Data Import and Cleaning

```
temp <- tempfile()
download.file("https://www2.census.gov/programs-surveys/acs/data/pums/2022/5-Year/csv_por.zip",temp)
data <- read_csv(unz(temp, "psam_p41.csv"))
```

```
## Rows: 205072 Columns: 290
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (27): RT, SERIALNO, SPORDER, PUMA10, PUMA20, JWTRNS, SCHG, SCHL, ANC1P,...
## dbl (263): DIVISION, REGION, ST, ADJINC, PWGTP, AGEP, CIT, CITWP, COW, DDRS,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
unlink(temp)
```

```
#Select Needed Columns From Larger Dataset
Proj1 <- data %>%
          select(SOCP, WAGP, SEX, AGEP, SCHL, SCIENGP)
```

```
#Challenge- SOCP Code as character, despite integer, and no real info in code
#Select Only Teachers in SOCP Code:

Proj1 <- Proj1 %>%
  subset(SOCP %in% c("251000", "252010", "252020", "252030", "252050"))

Proj1 <- Proj1 %>%
  mutate(TEACH.TYPE = case_when(SOCP == "251000" ~ "Postsecondary",
                                SOCP == "252010" ~ "Preschool And Kindergarten",
                                SOCP == "252020" ~ "Elementary And Middle",
                                SOCP == "252030" ~ "Secondary",
                                SOCP == "252050" ~ "Special Ed"))

Proj1$TEACH.TYPE <- as.factor(Proj1$TEACH.TYPE)
str(Proj1)
```

```
## tibble [5,101 x 7] (S3: tbl_df/tbl/data.frame)
##  $ SOCP      : chr [1:5101] "251000" "251000" "251000" "251000" ...
##  $ WAGP      : num [1:5101] 0 6000 0 0 19000 6000 70000 73000 75000 45000 ...
##  $ SEX       : num [1:5101] 1 2 1 1 2 2 2 2 2 2 ...
##  $ AGEP      : num [1:5101] 36 21 36 38 31 21 36 45 40 64 ...
##  $ SCHL      : chr [1:5101] "21" "19" "21" "21" ...
##  $ SCIENGP   : num [1:5101] 1 NA 1 2 2 NA 2 1 2 2 ...
##  $ TEACH.TYPE: Factor w/ 5 levels "Elementary And Middle",..: 2 2 2 2 1 2 5 4 1 2 ...
```

```r
#Challenge- Sex as 1/2 rather than 0/1 factor
#Making Sex a flag 1=male)
Proj1$SEX[Proj1$SEX==2] <- 0
Proj1$SEX<- as.factor(Proj1$SEX)
```

```r
#CHALLENGE- Many Preschool Teachers Without Degrees. Need to make Degree Level Flag
Proj1<- Proj1 %>%
  mutate(DEGREE = case_when(SCHL <20 ~ "No Degree",
                 SCHL == 20 ~ "Associates",
                 SCHL == 21 ~ "Bachelors",
                 SCHL == 22 ~ "Masters",
                 SCHL == 23 ~ "Professional",
                 SCHL == 24 ~ "Doctorate"))
Proj1$DEGREE <- as.factor(Proj1$DEGREE)
```

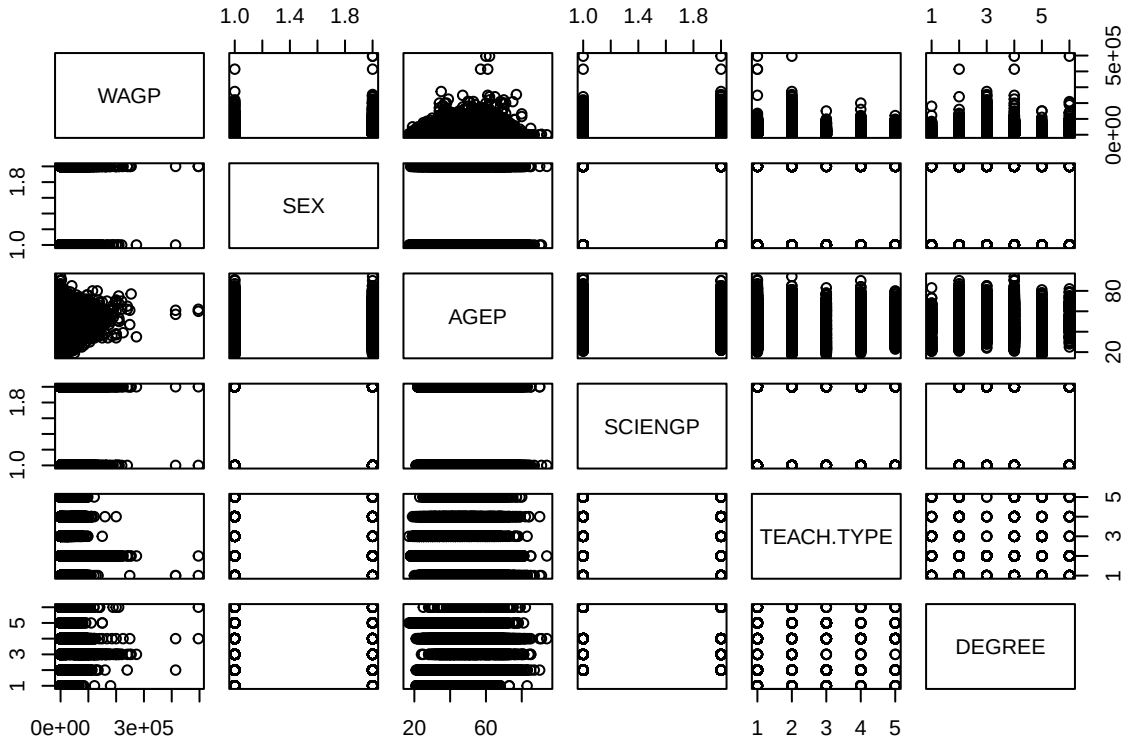
```r
#Challenge- SCIENGP as 1/2 rather than 0/1 factor
#Making SCIENGP a flag 1=STEM Degree)
Proj1$SCIENGP[Proj1$SCIENGP==2] <- 0
Proj1$SCIENGP<- as.factor(Proj1$SCIENGP)
```

```r
str(Proj1)
```

```
## tibble [5,101 x 8] (S3: tbl_df/tbl/data.frame)
##  $ SOCP      : chr [1:5101] "251000" "251000" "251000" "251000" ...
##  $ WAGP      : num [1:5101] 0 6000 0 0 19000 6000 70000 73000 75000 45000 ...
##  $ SEX       : Factor w/ 2 levels "0","1": 2 1 2 2 1 1 1 1 1 1 ...
##  $ AGEP      : num [1:5101] 36 21 36 38 31 21 36 45 40 64 ...
##  $ SCHL      : chr [1:5101] "21" "19" "21" "21" ...
##  $ SCIENGP   : Factor w/ 2 levels "0","1": 2 NA 2 1 1 NA 1 2 1 1 ...
##  $ TEACH.TYPE: Factor w/ 5 levels "Elementary And Middle",..: 2 2 2 2 1 2 5 4 1 2 ...
##  $ DEGREE    : Factor w/ 6 levels "Associates","Bachelors",..: 2 5 2 2 2 5 4 4 4 4 ...
```

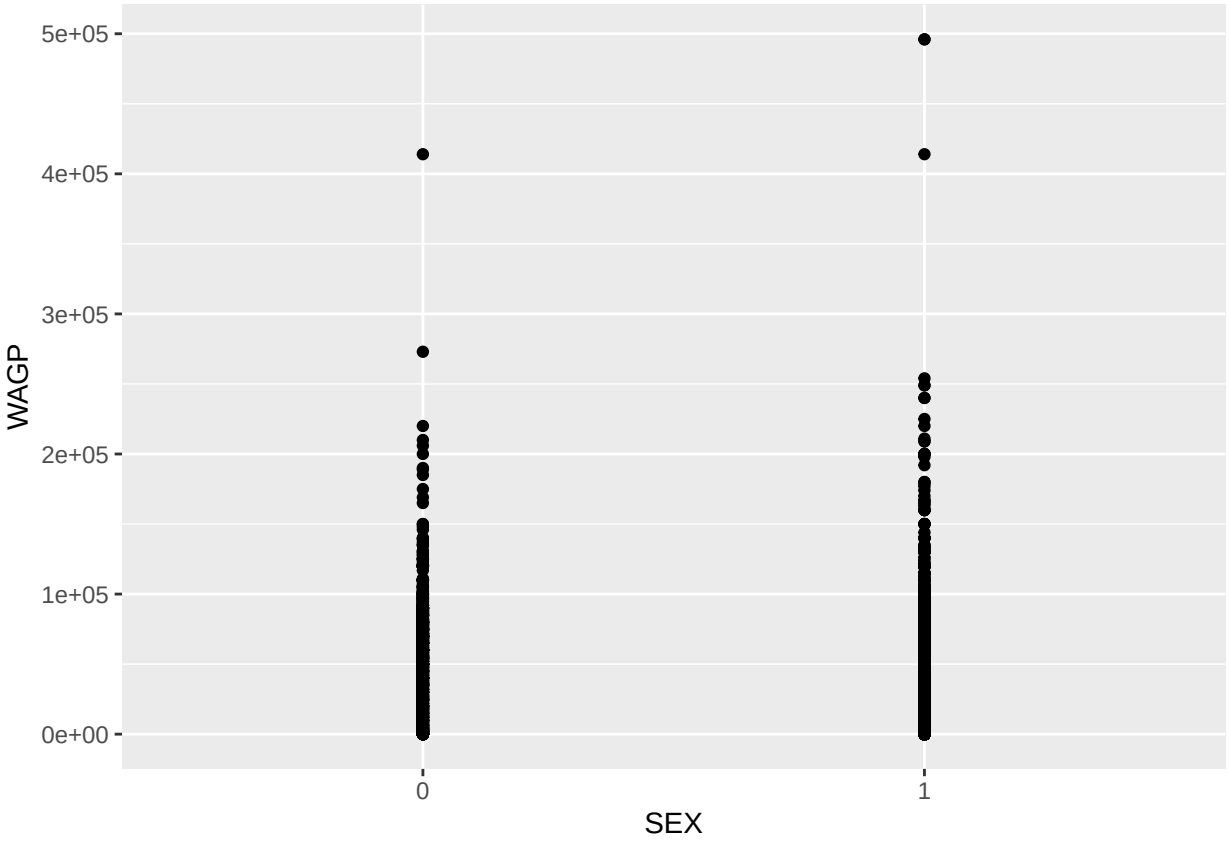
#Visualizations

```r
#Created Visualization on pairs
pairs(Proj1[,-c(1,5)])
```

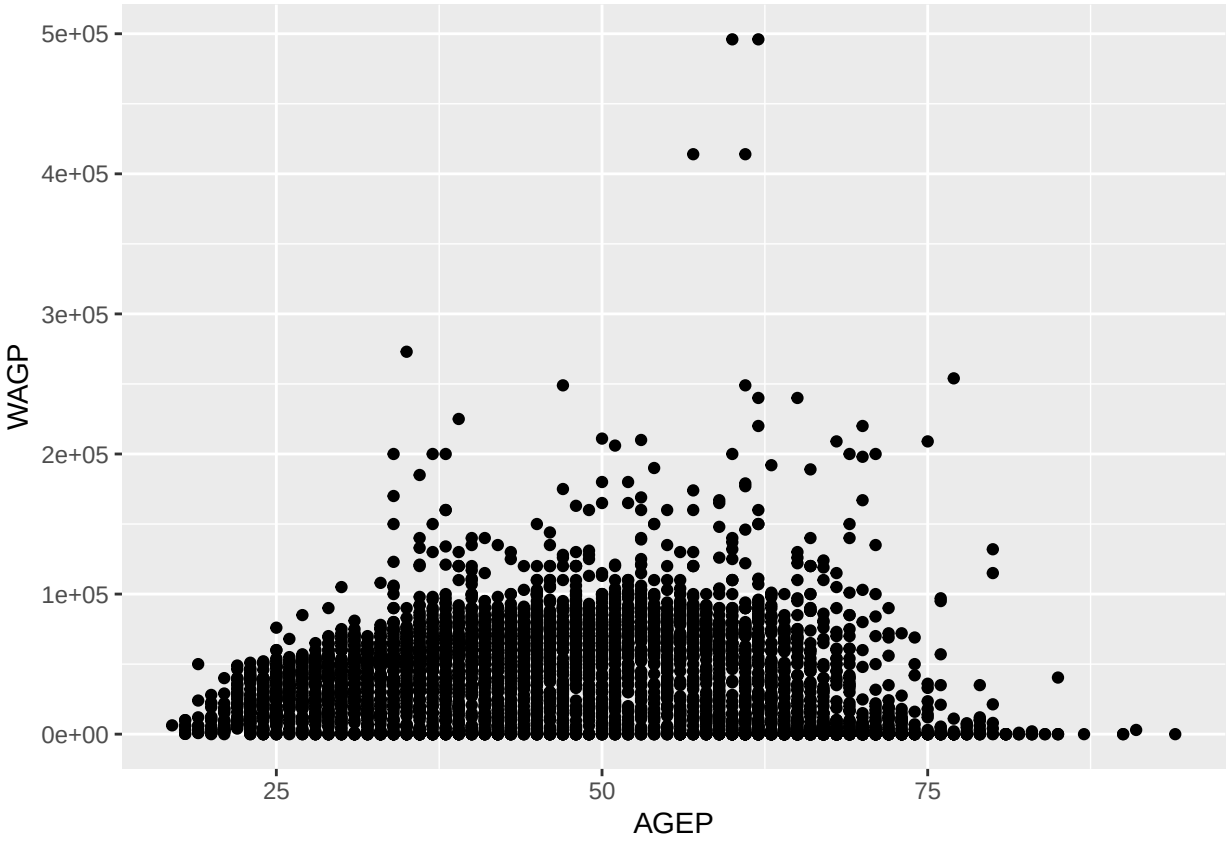



```
qplot(SEX,WAGP, data=Proj1)
```

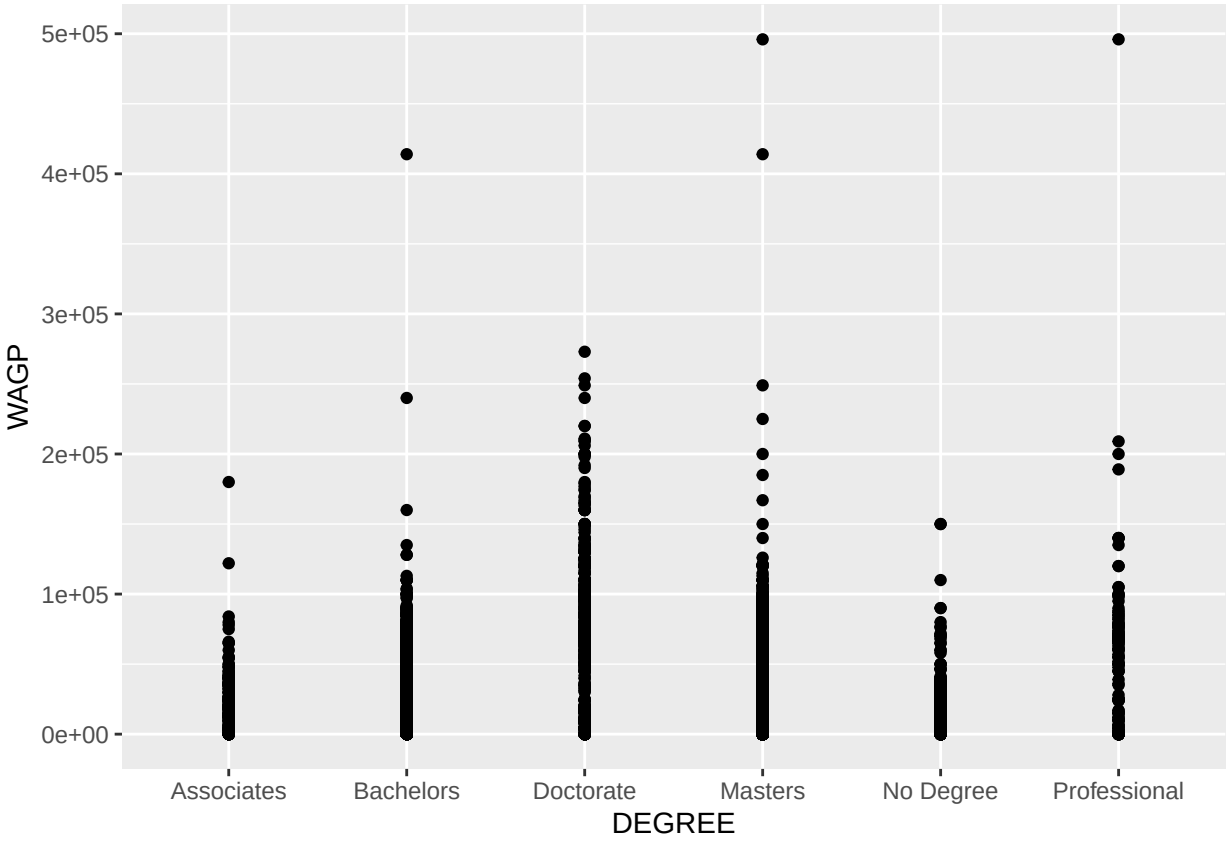
```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



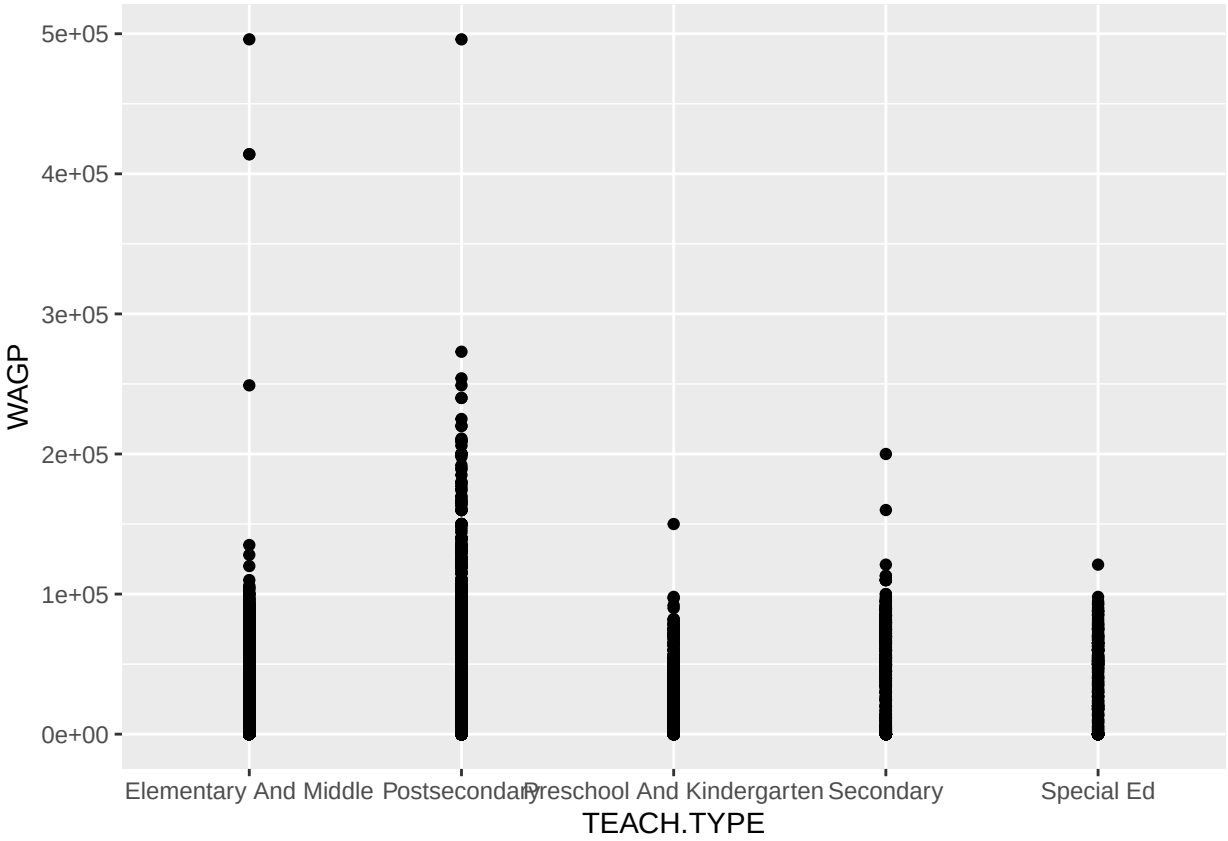

```
qplot(AGEP,WAGP, data=Proj1)
```



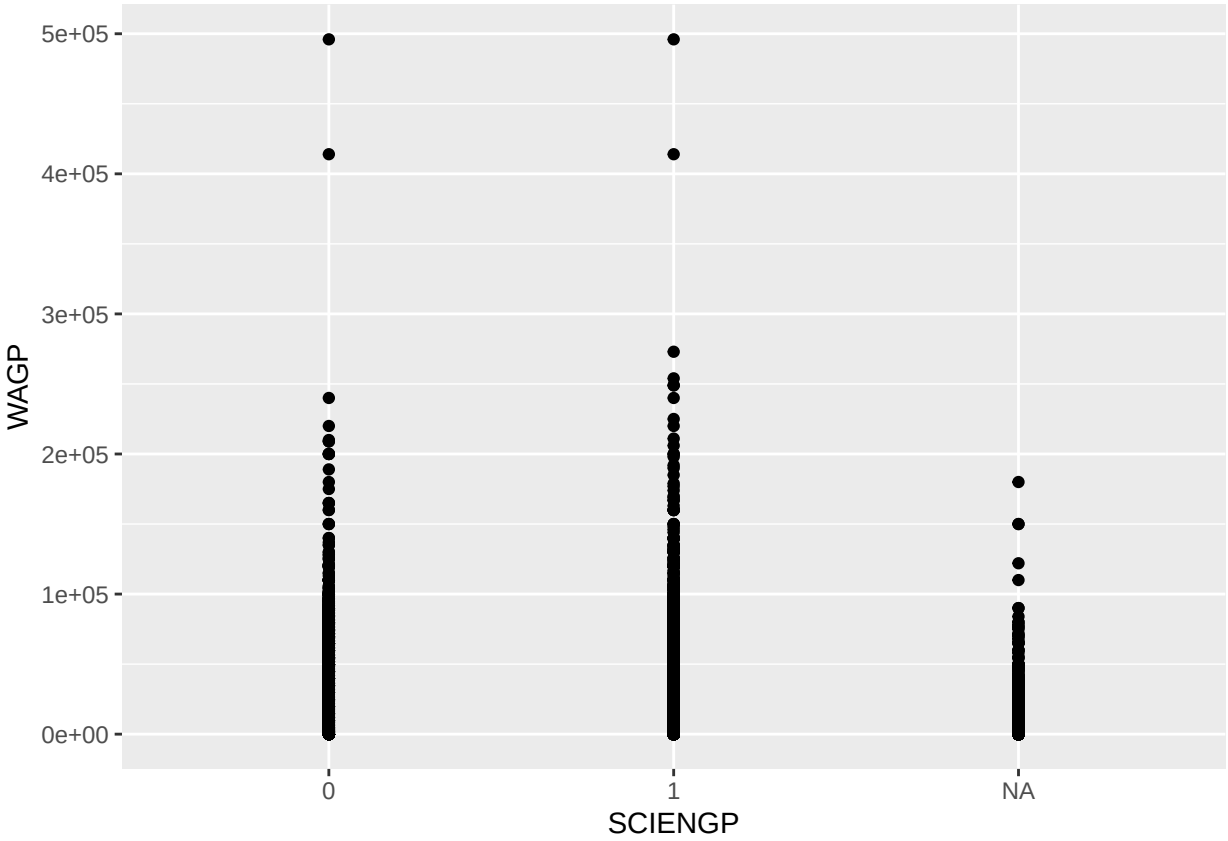

```
qplot(DEGREE,WAGP, data=Proj1)
```




```
qplot(TEACH.TYPE,WAGP, data=Proj1)
```




```
qplot(SCIENGP,WAGP, data=Proj1)
```




```
#Create Full Model:

ProjMod1 <- lm(WAGP ~ SEX + AGEP + TEACH.TYPE + DEGREE, data=Proj1)
summary(ProjMod1)
```

```
##
## Call:
## lm(formula = WAGP ~ SEX + AGEP + TEACH.TYPE + DEGREE, data = Proj1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83126 -23007  -1214  20119 447759
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          40606.87    3268.57  12.423  < 2e-16 ***
## SEX1                                  8392.76    1092.04   7.685 1.82e-14 ***
## AGEP                                  -360.47      33.73 -10.687  < 2e-16 ***
## TEACH.TYPEPostsecondary              -4229.73    1389.28  -3.045  0.00234 **
## TEACH.TYPEPreschool And Kindergarten -5339.25    1850.87  -2.885  0.00393 **
## TEACH.TYPESecondary                   2348.86    1413.52   1.662  0.09663 .
## TEACH.TYPESpecial Ed                   893.84    2152.13   0.415  0.67792
## DEGREEBachelors                       4921.13    2906.18   1.693  0.09045 .
## DEGREEDoctorate                      50611.82    3295.19  15.359  < 2e-16 ***
## DEGREEMasters                        20869.39    2858.67   7.300 3.31e-13 ***
## DEGREENo Degree                      -6830.88    3177.60  -2.150  0.03163 *
```

```
## DEGREEProfessional                   28182.22    4054.48   6.951 4.09e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33990 on 5089 degrees of freedom
## Multiple R-squared:  0.173,  Adjusted R-squared:  0.1712
## F-statistic: 96.78 on 11 and 5089 DF,  p-value: < 2.2e-16
```

```
vif_model <- vif(ProjMod1)
```