NoahCollin_607_Project2

Noah Collin

Noah Collin 607

DataSet 1: Candy Data

This dataset was posted in the discussion boards on CUNY Blackboards. It was posted by Coffy Andrews-Guo. This link might not work, but you can see his post here: https://bbhosted.cuny.edu/webapps/discussionboard/do/message?action=list_messages&course_id=_2010110_1&nav=discussion_board&conf_id=_2342995_1&forum_id=_2997791_1&message_id=_53840343_1

Data is from: https://www.scq.ubc.ca/so-much-candy-data-seriously/

The survey that was actually used is here: https://www.scq.ubc.ca/wp-content/uploads/2017/10/candyhierarchysurvey2017.pdf

```
Sys.setenv("VROOM_CONNECTION_SIZE" = 131072 * 20)
rawCandy <-  read.csv("candyhierarchy2017.csv", encoding = "UCS-2LE")
```

```
?gsub
```

```
## starting httpd help server ... done
```

```
#names(rawCandy)
colnames(rawCandy) <-  gsub(("Q\\d+"),"", colnames(rawCandy))

colnames(rawCandy) <-  gsub(("\\.\\.+"),"", colnames(rawCandy))

colnames(rawCandy) <-  gsub(("\\."),"  ",
colnames(rawCandy))

colnames(rawCandy) <- str_trim(colnames(rawCandy), side = "left")

#drop columns that aren't candy questions... Still some questionable "candies" in here...
rawCandy <- rawCandy[(-c(112:110))]
#names(rawCandy)
```

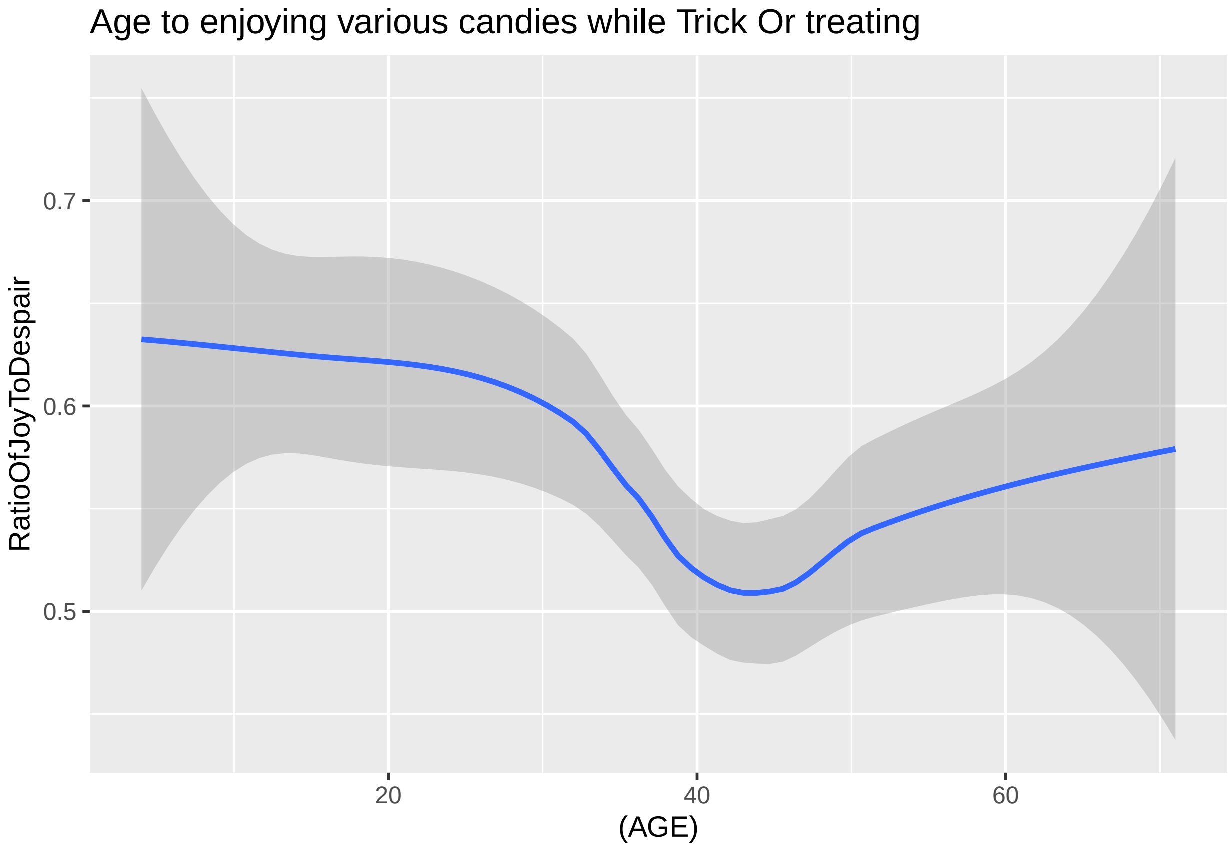
```
trickOrTreaters <- filter(rawCandy, rawCandy$"GOING OUT" == "Yes")

OldGrumps <- filter(rawCandy, rawCandy$"GOING OUT" == "No")

##TODO

trickOrTreaters <-  trickOrTreaters %>% rowwise() %>% mutate(SumJoys = sum(c_across(all_of(7:109)) == "
trickOrTreaters <-  trickOrTreaters %>% rowwise() %>% mutate(SumDespair = sum(c_across(all_of(7:109)) ==
trickOrTreaters <- trickOrTreaters %>% mutate(RatioOfJoyToDespair = SumJoys / sum(SumJoys ,SumDespair))


mostJoyfulTrickOrTreaters <-  arrange(trickOrTreaters, desc(trickOrTreaters$RatioOfJoyToDespair))

mostJoyfulTrickOrTreaters <- mostJoyfulTrickOrTreaters %>% transform()

#class(mostJoyfulTrickOrTreaters$AGE)



mostJoyfulTrickOrTreaters$AGE <-as.numeric( (mostJoyfulTrickOrTreaters$AGE) )
```

```
## Warning: NAs introduced by coercion
```

```
mostJoyfulTrickOrTreaters$AGE <- ifelse(mostJoyfulTrickOrTreaters$AGE < 90, mostJoyfulTrickOrTreaters$A


mostJoyfulTrickOrTreaters %>% ggplot(aes(x=(AGE), y = RatioOfJoyToDespair)) +
  geom_smooth() +
  ggtitle("Age to enjoying various candies while Trick Or treating")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 82 rows containing non-finite values (stat_smooth).
```



40 year olds seem to be the grumpiest trick-or-treaters. I think this graph partially shows what a silly dataset this is. I actually looked to see if there was a real relationship for children under 18, but there didn't seem to be one.

Dataset 2: Ficticious Financial Data

I made a fake dataset here:

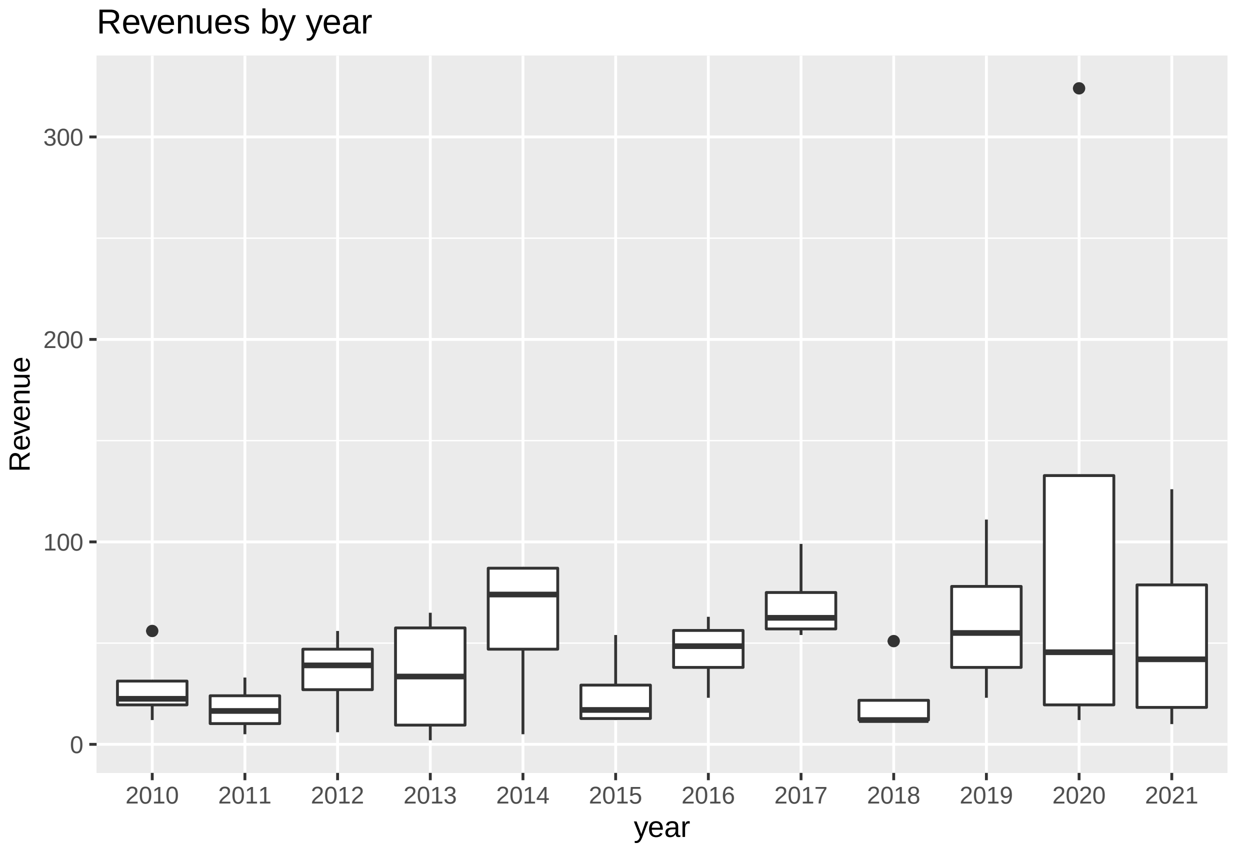
```
dumb1 <- tibble(Group = c(1,2,3,4,1,2,3,4,1,2,3,4), year = c(2010,2011,2012,2013, 2014,2015,2016,2017, 2
                Qtr.3 = c(56,12,34,65,87,13,63,58,12,23,22,63),
                Qtr.4 = c(22,33,44,55,87,12,43,54,51,67,12,126))
```

Gather and seperate

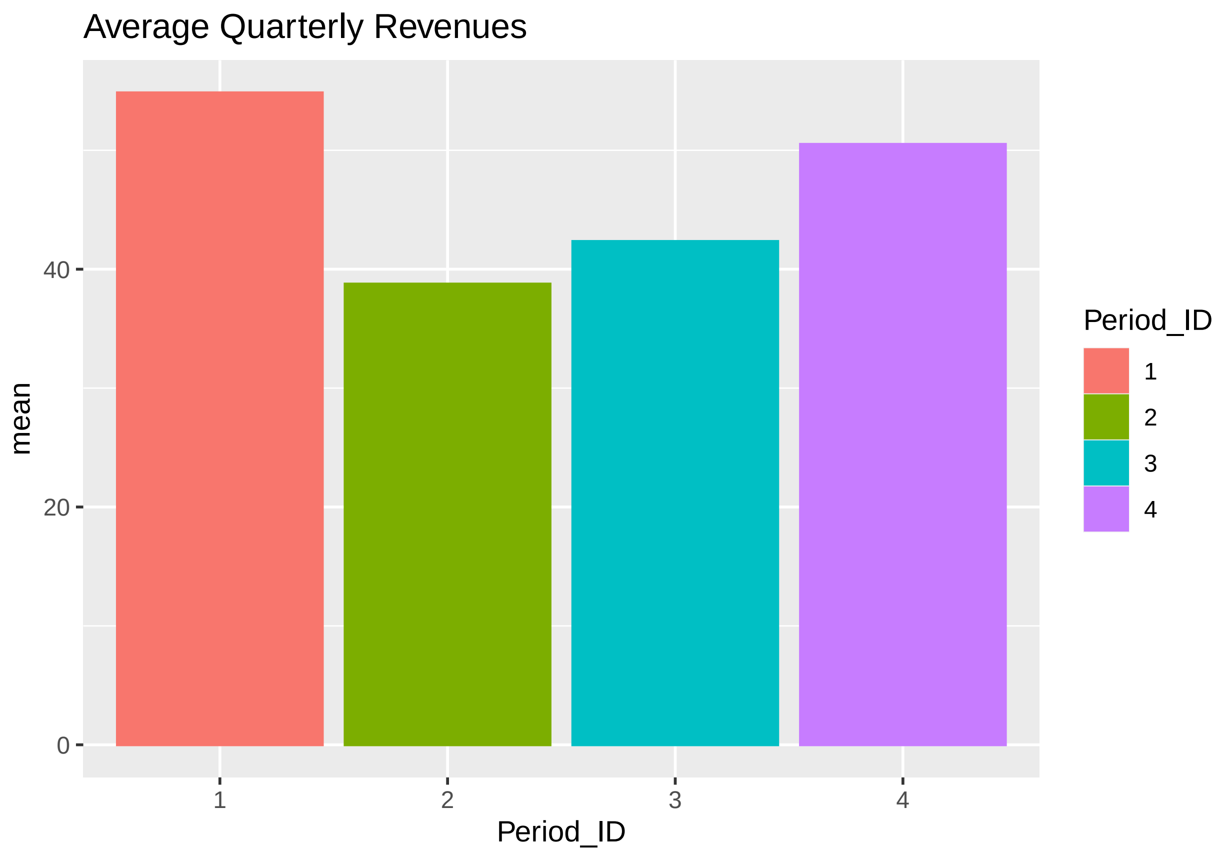
```
cleanTibble <- dumb1 %>% gather(Quarter, Revenue, Qtr.1:Qtr.4)

cleanTibble <-  cleanTibble %>% separate(Quarter, c("Time_Period", "Period_ID"), sep = "\\.")

cleanTibble$year <- as.factor(cleanTibble$year)
```

#Groupby

```
ggplot(cleanTibble, aes(x =year,y = Revenue) ) +
  geom_boxplot() +
  ggtitle("Revenues by year")
```



```
dumbGrouped <- cleanTibble %>% group_by(Period_ID) %>% summarise(mean = mean(Revenue) , n = n())

ggplot(dumbGrouped, aes(x=Period_ID, y = mean, colour = Period_ID, fill = Period_ID)) +
  geom_col() +
  ggtitle("Average Quarterly Revenues")
```



Dataset 3: Fictional Education Outcomes Dataset

I didn't make this dataset. Source of this fictional test data: https://www.kaggle.com/spscientist/students-performance-in-exams also: http://roycekimmons.com/tools/generated_data/exams

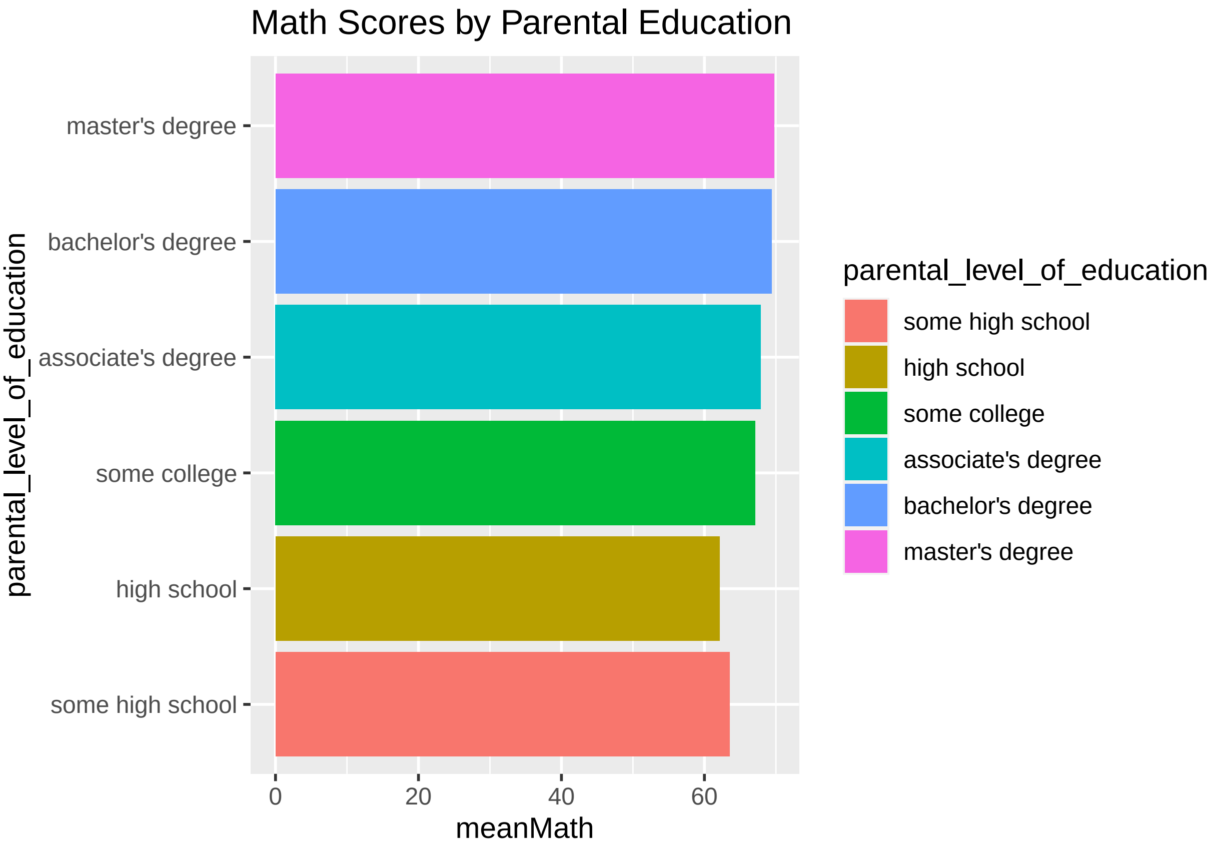
```
rawTests <- read_csv("StudentsPerformance.csv")
```

```
## Rows: 1000 Columns: 8
```

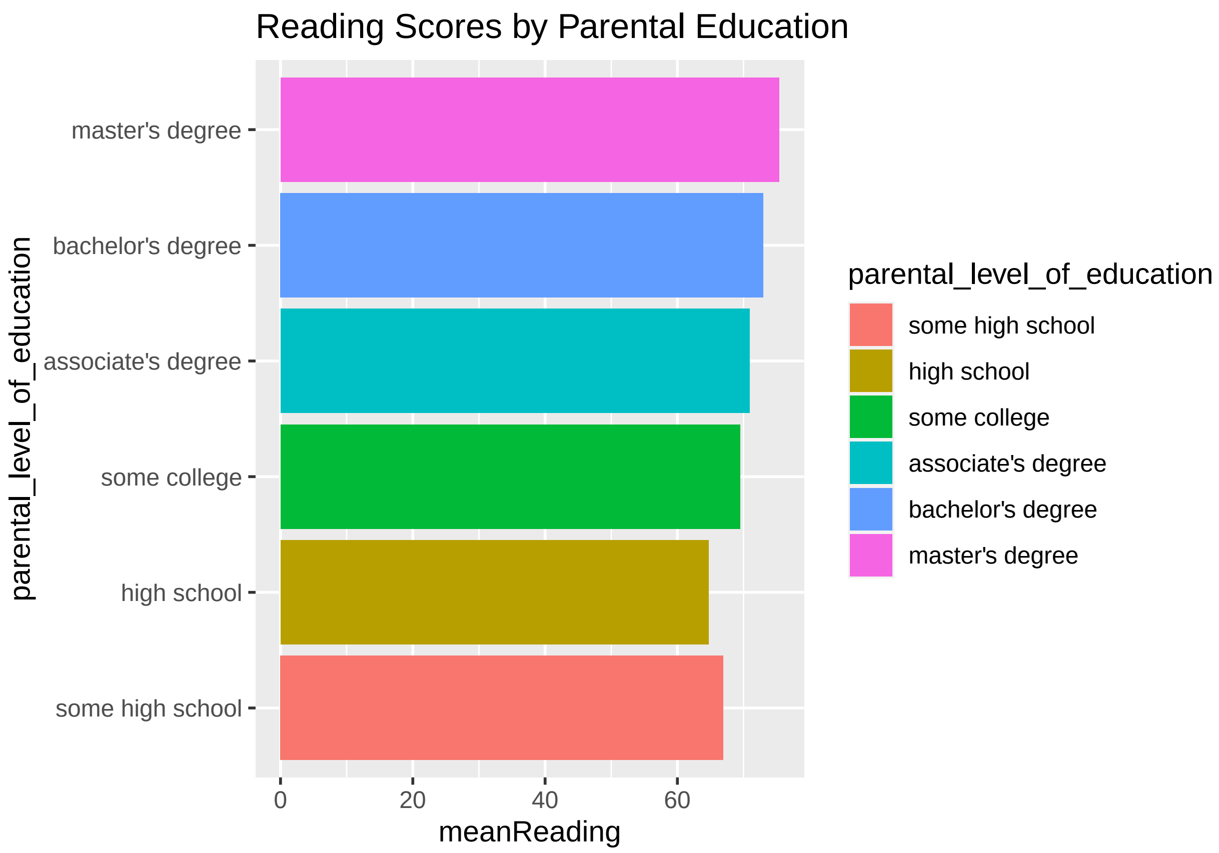
```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): gender, race/ethnicity, parental level of education, lunch, test pr...
## dbl (3): math score, reading score, writing score
```

```
##
## 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.
```

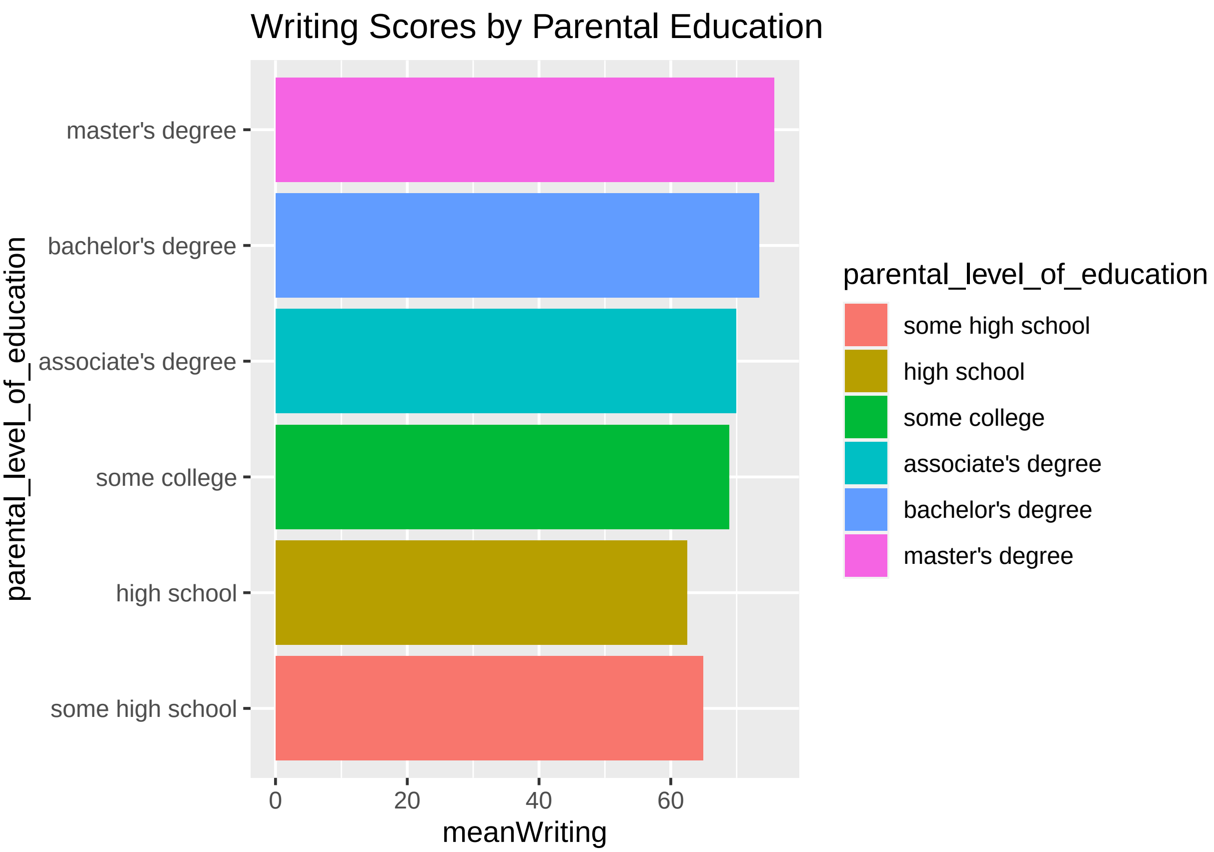
```
colnames(rawTests) <-  colnames(rawTests) <-  gsub("\\W","_", colnames(rawTests))
#names(rawTests)
#unique(rawTests$"parental_level_of_education")

rawTests$parental_level_of_education <- factor(rawTests$parental_level_of_education, levels = c("some hi
parentalEducationGrouped <- rawTests %>% group_by(parental_level_of_education) %>%
  summarise(meanMath = mean(math_score), meanWriting = mean(writing_score), meanReading = mean(reading_

ggplot(parentalEducationGrouped, aes(x = parental_level_of_education, y=meanMath, fill = parental_level
  geom_col() +
  coord_flip() +
  ggtitle("Math Scores by Parental Education")
```



```
ggplot(parentalEducationGrouped, aes(x = parental_level_of_education, y=meanReading, fill = parental_lev
  geom_col() +
  coord_flip() +
  ggtitle("Reading Scores by Parental Education")
```



```
ggplot(parentalEducationGrouped, aes(x = parental_level_of_education, y=meanWriting, fill = parental_l
geom_col() +
coord_flip() +
ggtitle("Writing Scores by Parental Education")
```



Interestingly, students whose parents have “some high school” seem to outperform those who have a high school diploma in the three R’s.