Capstone DataWrangling and EDA

## The dataset

By law (California Education Code Section 60800), all public local educational agencies in California are required to administer the Physical Fitness Test (PFT) annually to all students in grades five, seven, and nine. The main goal of the test is to help students in starting life-long habits of regular physical activity. The test has six parts that show a level of fitness that offer a degree of defense against diseases that come from inactivity.During the month of February, March, April, or May, the governing board of each school maintaining grades five, seven, or nine shall administer to each student in those grades the PFT.

## The goal

The goal of this project is to determine if, given certain characteristics of the student (age,race, gender,school type,socio economic information) and school (county, district) can the fitness level of the student be predicted. Furthermore, this project aims to explore if there are any trends or patterns in types of schools and school districts/counties and fitness levels and if age plays a part in the fitness level of a student.

## Cleaning the data

The data was in a xls format, it was converted to csv format. Rows that had no relevant student data at all were removed. Aggregation data present in the original dataset was also removed.

# Read data and replace \*\* strings with NA  
fitness <- read.csv(file="C:/Users/s/Documents/data.csv", header=TRUE, sep=",",na.strings = "\*\*",stringsAsFactors=FALSE)  
  
head(fitness)

## Level\_Number Report\_Number Table\_Number Line\_Number CO DIST SCHL  
## 1 1 0 1 4 1 61150 132225  
## 2 1 0 1 5 1 61150 132225  
## 3 1 0 1 6 1 61150 132225  
## 4 1 0 1 1 1 61150 6090351  
## 5 1 0 1 2 1 61150 6090351  
## 6 1 0 1 3 1 61150 6090351  
## Line\_Text NoStud5 NoHFZ5 Perc5a Perc5b Perc5c  
## 1 Trunk Extension Strength 0 0 0.0 0.0 0.0  
## 2 Upper Body Strength 0 0 0.0 0.0 0.0  
## 3 Flexibility 0 0 0.0 0.0 0.0  
## 4 Aerobic Capacity 78 42 53.8 38.5 7.7  
## 5 Body Composition 78 49 62.8 16.7 20.5  
## 6 Abdominal Strength 78 71 91.0 9.0 0.0  
## NoStud7 NoHFZ7 Perc7a Perc7b Perc7c NoStud9 NoHFZ9 Perc9a Perc9b Perc9c  
## 1 0 0 0 0 0 721 674 93.5 6.5 0  
## 2 0 0 0 0 0 721 641 88.9 11.1 0  
## 3 0 0 0 0 0 721 682 94.6 5.4 0  
## 4 0 0 0 0 0 0 0 0.0 0.0 0  
## 5 0 0 0 0 0 0 0 0.0 0.0 0  
## 6 0 0 0 0 0 0 0 0.0 0.0 0  
## ChrtNum  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0

# If there are no students tested at all make the column values as NA  
fitness$NoStud5[fitness$NoStud5 == 0] <- NA  
  
fitness$NoStud7[fitness$NoStud7 == 0] <- NA  
  
fitness$NoStud9[fitness$NoStud9 == 0] <- NA  
  
#clean up rows that have no data for number of students tested in grade 5, 7 and 9  
fitness <- fitness[!(is.na(fitness$NoStud5)) | !(is.na(fitness$NoStud7))| !(is.na(fitness$NoStud9)),]  
  
#Clean up rows that dont have data about the excercies  
fitness <- fitness[!grepl("fitness", fitness$Line\_Text),]  
  
#Clean up any aggregartion in the data set  
fitness <- fitness[!grepl("Total", fitness$Line\_Text),]  
  
#Clean up aggregation with the Report Number in the data set  
fitness <- fitness[!grepl("0", fitness$Report\_Number),]  
  
  
  
#clean up the percentage columns  
  
fitness$Perc5b[is.na(fitness$NoStud5) & fitness$Perc5a ==0] <- NA   
  
fitness$Perc5b[is.na(fitness$NoStud5) & fitness$Perc5b ==0] <- NA   
  
fitness$Perc7b[is.na(fitness$NoStud7) & fitness$Perc7b ==0] <- NA   
  
fitness$Perc9b[is.na(fitness$NoStud9) & fitness$Perc9b ==0] <- NA

## Creating new columns

Based on the data present in the Report\_Number column ethnicity,sex. socio economic based information can be obtained. That information is then stored in seperate columns to be used for pattern detection.

#Create a new column for storing Ethnicity based information  
fitness$Ethnicity[fitness$Report\_Number == 3] <- "African American"  
fitness$Ethnicity[fitness$Report\_Number == 4] <- "Native Indian"  
fitness$Ethnicity[fitness$Report\_Number == 5] <- "Asian"  
fitness$Ethnicity[fitness$Report\_Number == 6] <- "Filipino"  
fitness$Ethnicity[fitness$Report\_Number == 7] <- "Hispanic"  
fitness$Ethnicity[fitness$Report\_Number == 8] <- "Pacific Islander"  
fitness$Ethnicity[fitness$Report\_Number == 9] <- "White"  
fitness$Ethnicity[fitness$Report\_Number == 10] <- "Mixed Race"  
  
#Create a new column for storing sex information  
fitness$Sex[fitness$Report\_Number == 1]<- "Female"  
fitness$Sex[fitness$Report\_Number == 2]<- "Male"  
  
#Create a new column for storing Economic information  
fitness$EconomicallyDisadvantaged[fitness$Report\_Number == 11]<- "Yes"  
fitness$EconomicallyDisadvantaged[fitness$Report\_Number == 12]<- "No"  
  
summary(fitness)

## Level\_Number Report\_Number Table\_Number Line\_Number   
## Min. :1.000 Min. : 1.000 Min. :1 Min. :1.0   
## 1st Qu.:1.000 1st Qu.: 3.000 1st Qu.:1 1st Qu.:2.0   
## Median :1.000 Median : 7.000 Median :1 Median :3.5   
## Mean :1.173 Mean : 6.613 Mean :1 Mean :3.5   
## 3rd Qu.:1.000 3rd Qu.:11.000 3rd Qu.:1 3rd Qu.:5.0   
## Max. :4.000 Max. :13.000 Max. :1 Max. :6.0   
##   
## CO DIST SCHL Line\_Text   
## Min. : 0.00 Min. : 0 Min. : 0 Length:574800   
## 1st Qu.:19.00 1st Qu.:64733 1st Qu.: 116970 Class :character   
## Median :30.00 Median :67124 Median :6009955 Mode :character   
## Mean :28.86 Mean :65427 Mean :3734906   
## 3rd Qu.:38.00 3rd Qu.:69468 3rd Qu.:6048144   
## Max. :99.00 Max. :77032 Max. :9010745   
##   
## NoStud5 NoHFZ5 Perc5a Perc5b   
## Min. : 1.00 Min. : 0.00 Min. : 0.00 Min. : 0.0   
## 1st Qu.: 3.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 12.1   
## Median : 17.00 Median : 6.00 Median : 29.50 Median : 22.4   
## Mean : 89.47 Mean : 53.93 Mean : 36.02 Mean : 26.2   
## 3rd Qu.: 46.00 3rd Qu.: 28.00 3rd Qu.: 72.30 3rd Qu.: 36.7   
## Max. :278870.00 Max. :232432.00 Max. :100.00 Max. :100.0   
## NA's :200862 NA's :156240 NA's :156240 NA's :357102   
## Perc5c NoStud7 NoHFZ7 Perc7a   
## Min. : 0.00 Min. : 1 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 0.00 1st Qu.: 4 1st Qu.: 0.00 1st Qu.: 0.00   
## Median : 0.00 Median : 17 Median : 0.00 Median : 0.00   
## Mean : 2.37 Mean : 147 Mean : 48.05 Mean : 19.14   
## 3rd Qu.: 0.00 3rd Qu.: 73 3rd Qu.: 7.00 3rd Qu.: 37.60   
## Max. :92.30 Max. :258185 Max. :221837.00 Max. :100.00   
## NA's :156240 NA's :356412 NA's :91302 NA's :91302   
## Perc7b Perc7c NoStud9 NoHFZ9   
## Min. : 0.0 Min. : 0.00 Min. : 1.0 Min. : 0.00   
## 1st Qu.: 10.1 1st Qu.: 0.00 1st Qu.: 4.0 1st Qu.: 0.00   
## Median : 18.9 Median : 0.00 Median : 21.0 Median : 0.00   
## Mean : 22.3 Mean : 1.28 Mean : 208.3 Mean : 45.84   
## 3rd Qu.: 30.3 3rd Qu.: 0.00 3rd Qu.: 115.0 3rd Qu.: 0.00   
## Max. :100.0 Max. :100.00 Max. :250451.0 Max. :221393.00   
## NA's :447714 NA's :91302 NA's :423414 NA's :59844   
## Perc9a Perc9b Perc9c ChrtNum   
## Min. : 0.00 Min. : 0.0 Min. : 0.00 Length:574800   
## 1st Qu.: 0.00 1st Qu.: 9.4 1st Qu.: 0.00 Class :character   
## Median : 0.00 Median : 17.7 Median : 0.00 Mode :character   
## Mean : 13.06 Mean : 21.1 Mean : 0.97   
## 3rd Qu.: 0.00 3rd Qu.: 27.7 3rd Qu.: 0.00   
## Max. :100.00 Max. :100.0 Max. :100.00   
## NA's :59844 NA's :483258 NA's :59844   
## Ethnicity Sex EconomicallyDisadvantaged  
## Length:574800 Length:574800 Length:574800   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##

## Healthy students in grades Vs Ethnicity

library(dplyr)

##   
## 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(tidyr)  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.4

healthystudents = fitness %>% group\_by(Ethnicity) %>% summarise(percentage\_healthy\_5th\_grade = mean(Perc5a,na.rm=T), percentage\_healthy\_7th\_grade = mean(Perc7a,na.rm=T), percentage\_healthy\_9th\_grade = mean(Perc9a,na.rm=T))  
View(healthystudents)  
  
  
library(reshape2)

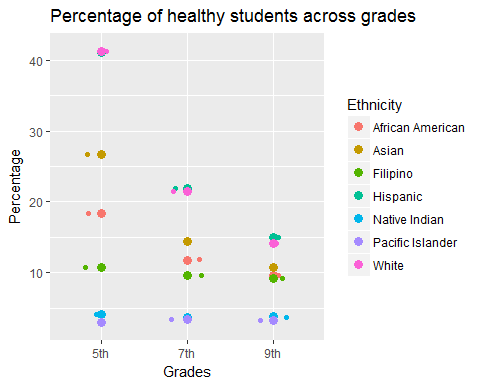
## Warning: package 'reshape2' was built under R version 3.4.4

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths

#when you melt essentially you create only one column with the value  
#and one column with the variable i.e. your x,y,z   
healthystudents <- melt(healthystudents, id.vars='Ethnicity')  
healthystudents <- healthystudents[!(is.na(healthystudents$Ethnicity)),]  
  
healthyplot <- ggplot(healthystudents, aes(x= variable, y=value)) +geom\_jitter(aes(color = Ethnicity, fill = Ethnicity),position=position\_jitter(0.2))+ geom\_dotplot(aes(color = Ethnicity, fill = Ethnicity),binaxis = "y", stackdir = "center")  
  
healthyplot <- healthyplot +ylab("Percentage")+xlab("Grades") + scale\_x\_discrete(breaks=c("percentage\_healthy\_5th\_grade","percentage\_healthy\_7th\_grade","percentage\_healthy\_9th\_grade"),  
 labels=c("5th", "7th", "9th"))  
  
healthyplot <- healthyplot +ggtitle("Percentage of healthy students across grades")  
  
  
plot(healthyplot)

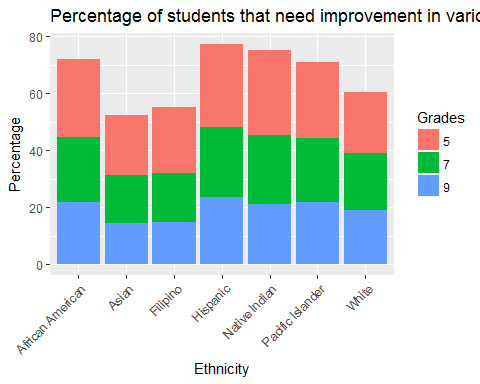
## `stat\_bindot()` using `bins = 30`. Pick better value with `binwidth`.



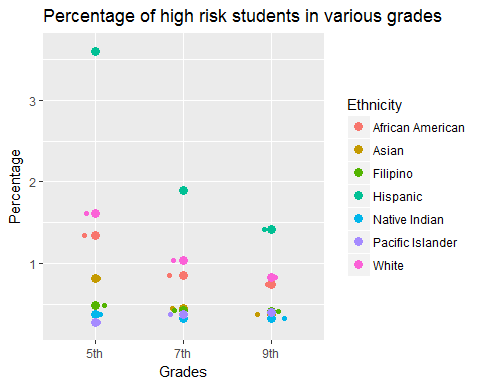
As shown in the plot above, certain ethnicities have a higher percentage of healthy students than others and the ratios of the various healthy students seem to stay consistent across the grades.For example ,Pacific Islanders seem to have the lowest percentage across all grades and White students seem to be the healthiest across all grades.

## Needs Improvement students across various ethnicities

#Needs Improv plot  
needsImprov = fitness %>% group\_by(Ethnicity) %>% summarise('5' = mean(Perc5b,na.rm=T), '7' = mean(Perc7b,na.rm=T), '9' = mean(Perc9b,na.rm=T))  
View(needsImprov)  
  
  
#when you melt essentially you create only one column with the value  
#and one column with the variable i.e. your x,y,z   
needsImprov <- melt(needsImprov, id.vars='Ethnicity')  
needsImprov <- needsImprov[!(is.na(needsImprov$Ethnicity)),]  
  
#ggplot it.   
  
needsImprovPlot <- ggplot(needsImprov, aes(x=Ethnicity, y=value, fill=variable)) + geom\_bar(stat='identity') + theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
needsImprovPlot <- needsImprovPlot +ylab("Percentage")+ labs(fill="Grades")  
  
needsImprovPlot <- needsImprovPlot +ggtitle("Percentage of students that need improvement in various grades")  
  
  
plot(needsImprovPlot)

 As shown in the ggplot above in the Needs Improvement plot, some ethnicities perform must better than others across grades. We can also see that students in Grade 5 need more improvement across all ethinicities than students in Grade 7 and Grade 9. Hispanic students need more improvement compared to other ethnicities.

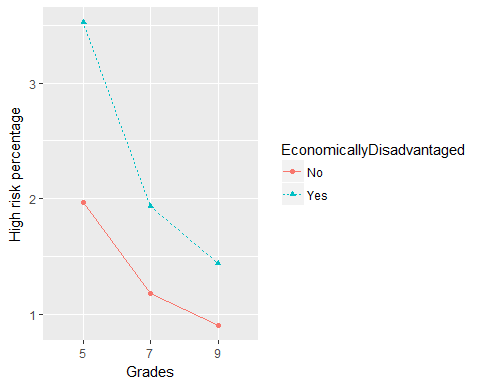
## `stat\_bindot()` using `bins = 30`. Pick better value with `binwidth`.



As shown in the plot above more students in the 5th grade are at high risk than students in grades 7 and 9. Hispanic students are in the higher risk percentges across all grades. Native Indian and Pacific Islander students have the lowest risk across all grades.

## Economics and High health risk students

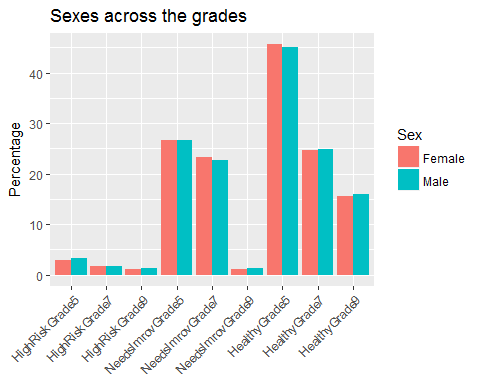
economic = fitness %>% group\_by(EconomicallyDisadvantaged) %>% summarise(percentage\_highrisk\_5th\_grade = mean(Perc5c,na.rm=T), percentage\_highrisk\_7th\_grade = mean(Perc7c,na.rm=T), percentage\_highrisk\_9th\_grade = mean(Perc9c,na.rm=T))  
View(healthystudents)  
  
library(reshape2)  
  
#when you melt essentially you create only one column with the value  
#and one column with the variable i.e. your x,y,z   
economic <- melt(economic, id.vars='EconomicallyDisadvantaged')  
economic <- economic[!(is.na(economic$EconomicallyDisadvantaged)),]  
View(economic)  
#ggplot it.  
library(ggplot2)  
  
economicPlot <- ggplot(economic, aes(x=variable, y=value, group=EconomicallyDisadvantaged)) +  
 geom\_line(aes(linetype=EconomicallyDisadvantaged , color = EconomicallyDisadvantaged))+  
 geom\_point(aes(shape=EconomicallyDisadvantaged, color = EconomicallyDisadvantaged))  
  
economicPlot <- economicPlot + xlab("Grades")  
  
economicPlot <- economicPlot +ylab("High risk percentage") + scale\_x\_discrete(breaks=c("percentage\_highrisk\_5th\_grade","percentage\_highrisk\_7th\_grade","percentage\_highrisk\_9th\_grade"),  
 labels=c("5", "7", "9"))  
  
  
plot(economicPlot)



As shown in the ggplot above Economically disadvantaged students are more at health high risk than students that are not. Also, students at smaller grade are more suspectible to this issue.The percentage difference is greatest in 5th and least in 9th grade.

## Sex and Health across the grades

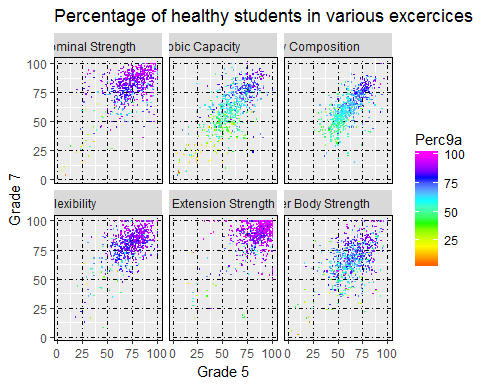
sex = fitness %>% group\_by(Sex) %>% summarise(HighRiskGrade5= mean(Perc5c,na.rm=T), HighRiskGrade7 = mean(Perc7c,na.rm=T), HighRiskGrade9= mean(Perc9c,na.rm=T),NeedsImrovGrade5= mean(Perc5b,na.rm=T), NeedsImrovGrade7 = mean(Perc7b,na.rm=T), NeedsImrovGrade9 = mean(Perc9c,na.rm=T), HealthyGrade5= mean(Perc5a,na.rm=T), HealthyGrade7= mean(Perc7a,na.rm=T), HealthyGrade9= mean(Perc9a,na.rm=T))  
  
sex <- sex[!(is.na(sex$Sex)),]  
  
sex <- melt(sex, id.vars='Sex')  
  
  
sexplot <- ggplot(sex, aes(x= variable, y=value,fill=Sex))   
sexplot <- sexplot + geom\_bar(stat="identity", position=position\_dodge()) + theme(axis.title.x=element\_blank())  
  
sexplot <- sexplot +ylab("Percentage") + theme(axis.text.x = element\_text(angle = 45, hjust = 1))   
  
sexplot <- sexplot +ggtitle("Sexes across the grades")  
  
plot(sexplot)



As shown in the plot above the healthy and the needs improvement percentages seem to be comparable across the sexes in all grades.In the high risk students male students seem to be at a slightly higher risk than female students.

## Comparison of the various excercies across healthy students

#Only use data that is non zero for this plot  
  
fitness <- fitness[!(is.na(fitness$Perc5a)),]  
fitness <- fitness[!(is.na(fitness$Perc7a)),]  
fitness <- fitness[!(is.na(fitness$Perc9a)),]  
  
fitness<-fitness[!(fitness$Perc5a==0),]  
fitness<-fitness[!(fitness$Perc7a==0),]  
fitness<-fitness[!(fitness$Perc9a==0),]  
  
variousExcercices <- ggplot()+ geom\_tile(data = fitness, aes(x=Perc5a, y=Perc7a,   
 colour=Perc9a),na.rm=T,alpha=1/5)+  
 facet\_wrap(~Line\_Text,ncol=3)+  
 scale\_colour\_gradientn(colours=rainbow(6))+  
 theme(panel.grid.major=element\_line(linetype=4,colour='Black'))+  
 theme(panel.border=element\_rect(linetype='solid',fill=NA))   
  
variousExcercices <- variousExcercices +ylab("Grade 7")  
  
variousExcercices <- variousExcercices +xlab("Grade 5")  
  
variousExcercices <- variousExcercices + labs(fill="Grade 9")  
  
variousExcercices <- variousExcercices +ggtitle("Percentage of healthy students in various excercices")  
  
plot(variousExcercices)



The plot above shows how the healthy students across various grades performed in the different kinds of excercise tests that were administered to them. Students are performing less well in the Aerobic capacity test than the Abdominal Strength test.