Final Project- Immunization Rates in California 7th Graders

Jane Huston

November 3, 2014

California has experienced high levels of Pertussis in recent years while simultaneously struggling with declining immunization rates. While immunization is required for school attendance, exemptions are allowed when medically contraindicated or when immunization contradicts the parent's personal or religious beliefs. Public health officials are concerned that rising rates of personal belief exemptions (PBEs) create a dangerous opportunity for infectious disease outbreaks.

My primary research objective is to determine if the proportion students submitting PBEs are different in public versus private schools in California. Secondary objectives are to understand the number and proportion of PBEs across schools of different sizes and different counties.

First, I imported the dataset. The data are immunization status of 7th graders in the state of California, presented at school-level and available at http://www.cdph.ca.gov/programs/immunize/pages/immunizationlevels.aspx. Data is restricted to those schools with 10 or more students enrolled in the 7th grade.

Students are considered up to date if they have received pertussis-containing booster shot on or after the 7th birthday. A permanent medical exemption (PME) is allowed upon presentation of a written statement from physician that immunization is not indicated due to medical circumstances. A personal belief exemption (PBE) occurs when a parent requests exemption from the immunization requirement for school entry because all or some immunizations are contrary to the parent's belief

```
setwd("/Users/janehuston/Documents/Programming in R/FinalProject")
iz1314<-read.csv("2013-2014CA7thGradeData.csv")
```

Next, I performed some transformations on the data to prepare for later analysis. I've removed unecessary columns, renamed a column, and omitted incomplete cases. I then created a categorical variable based on the size of the school, and computed the ratio of PBEs per 100 students for each school.

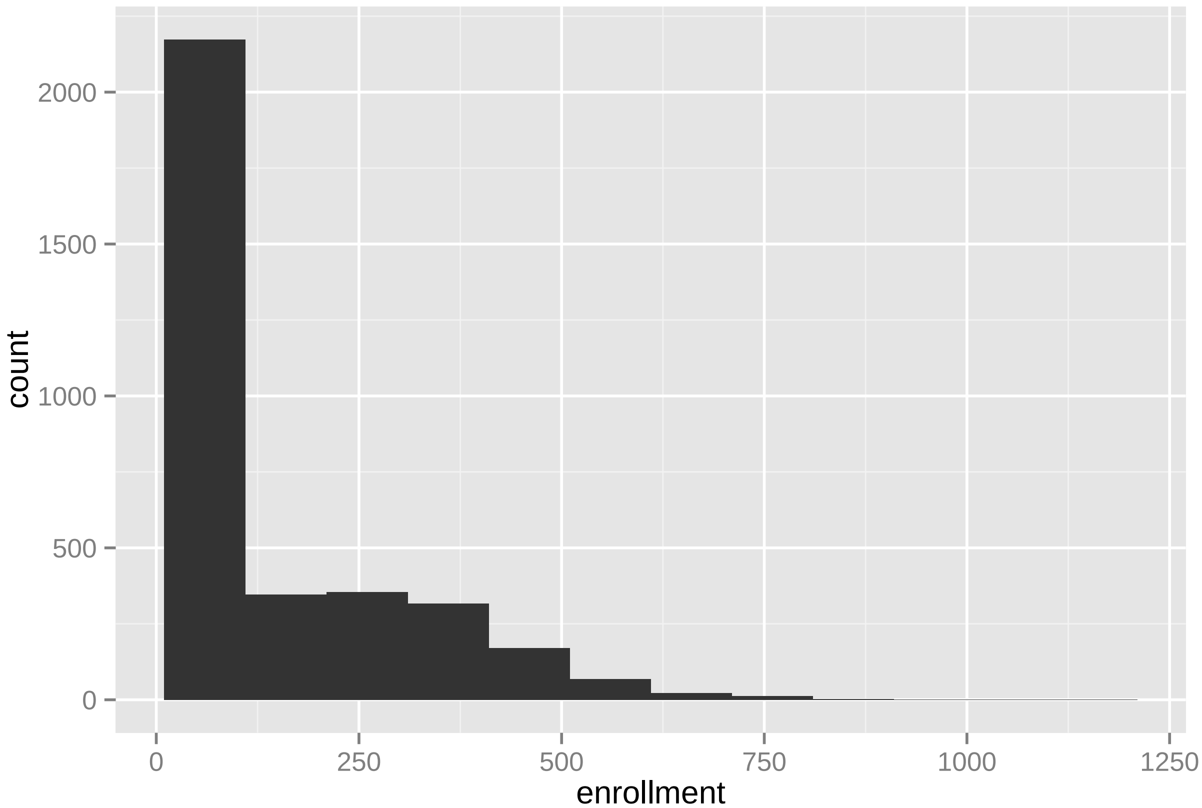
```
names(iz1314)
```

```
##  [1] "SCHOOL.CODE"           "COUNTY"
##  [3] "PUBLIC...PRIVATE"      "PUBLIC.SCHOOL.DISTRICT"
##  [5] "CITY"                  "SCHOOL.NAME"
##  [7] "ENROLLMENT"            "UTD_NUM"
##  [9] "UTD_PCT"               "PME_NUM"
## [11] "PME_PCT"               "PBE_NUM"
## [13] "PBE_PCT"               "REPORTED"
```

```
names(iz1314)<- tolower(names(iz1314))
newiz1314<- iz1314[-c(4, 9, 11, 13)] #remove unnecessary columns

names(newiz1314)[3]<- "public_private" #rename columns
newiz1314<- na.omit(newiz1314) #deal with missing values

#derive categorical variable
library(ggplot2)
ggplot(newiz1314, aes(enrollment))+ geom_histogram(binwidth=100, origin=10)
```




```
quantile(newiz1314$enrollment, c(.33, .66))
```

```
##   33%   66%
##  32.0 134.2
```

```
newiz1314<- within(newiz1314,{
                   size<- vector()
                   size[enrollment <= 32] <- "Small"
                   size[enrollment > 32 & enrollment<= 134] <- "Medium"
                   size[enrollment >134] <- "Large"
                   })
newiz1314$size <- factor(newiz1314$size)

#create a new variable of pberate
newiz1314$pberate<- round((newiz1314$pbe_num/newiz1314$enrollment)*100, 2)
```

Now, I'd like some summary statistics to better understand the data. I want to know the counts for the different categorical variables, and the means and standard deviations for the countinuous variables. I'd also like to sum the total enrollment and total UTD, PME, and PBE numbers to calculate the statewide rates.

```
apply(newiz1314[ , c(3, 11)], 2, FUN=table) #counts for categorical variables
```

```
## $public_private
##
## PRIVATE  PUBLIC
##    1102    2369
```

```
##
## $size
##
##  Large Medium  Small
##   1180   1128   1163
```

```
apply(newiz1314[ , c(6, 7, 8, 9)], 2, mean) #mean for continuous variables
```

```
## enrollment    utd_num    pme_num    pbe_num
##    140.441    135.779      0.263      4.399
```

```
apply(newiz1314[ , c(6, 7, 8, 9)], 2, sd) #sd for continuous variables
```

```
## enrollment    utd_num    pme_num    pbe_num
##     158.63     155.88       2.51      11.17
```

```
#create a function to find statewide rates
state_rate<- function(x){
  sum(x)/sum(newiz1314$enrollment)
}

#apply function to variables of interest
attach(newiz1314)
state_rate(utd_num)
```

```
## [1] 0.9668
```

```
state_rate(pme_num)
```

```
## [1] 0.001873
```

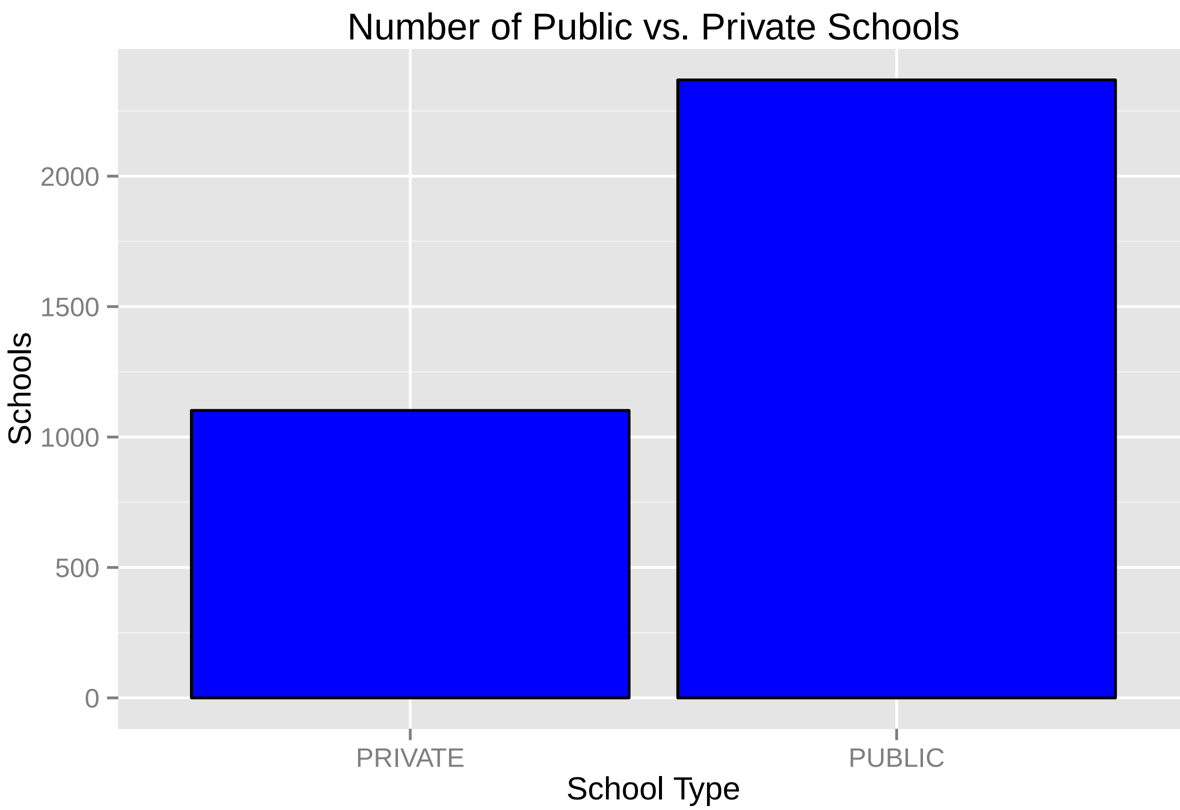
```
state_rate(pbe_num)
```

```
## [1] 0.03132
```

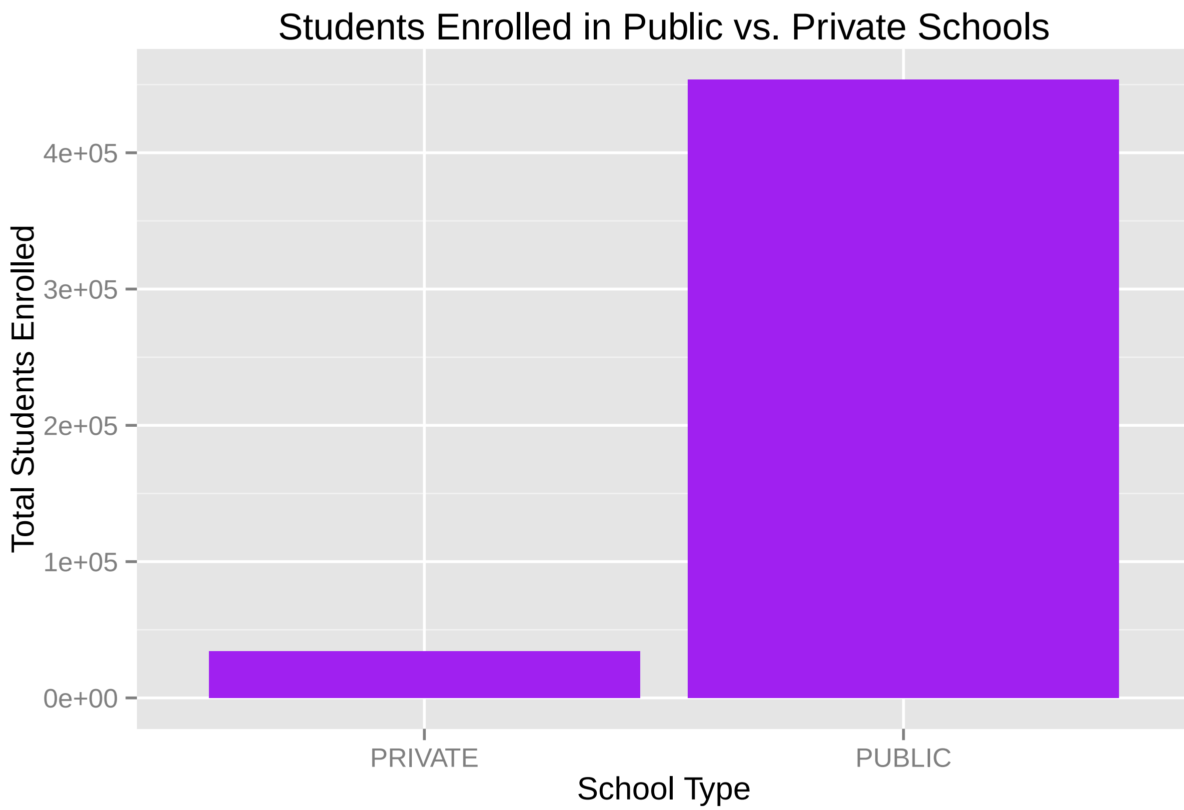
```
detach(newiz1314)
```

One interesting note, in understanding the landscape of the 7th grade in California: while there are about twice as many public schools as private schools, they enroll many more times the students.

```
p<- ggplot(newiz1314, aes(x=public_private))
p+ geom_bar(color="black", fill="blue")+
  ggtitle("Number of Public vs. Private Schools")+
  xlab("School Type")+
  ylab("Schools")
```




```
p+ geom_bar(aes(y=enrollment), stat="identity", fill="purple")+
  ggtitle("Students Enrolled in Public vs. Private Schools")+
  xlab("School Type")+
  ylab("Total Students Enrolled")
```



Public vs. Private Schools

First, let's compare the statewide rates for public vs. private schools.

```
df<- aggregate(newiz1314[ ,c(6, 9)], by=list(Type=newiz1314$public_private), FUN=sum)
df$rate<- df$pbe_num/ df$enrollment
print(df)
```

```
##      Type enrollment pbe_num    rate
## 1 PRIVATE      34041    1607 0.04721
## 2  PUBLIC     453431   13663 0.03013
```

The statistical test to compare 2 independent propotions in R is the prop.test( ) function, which takes 2 vectors. I'll take them from my df dataframe.

```
prop.test(df$pbe_num, df$enrollment)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  df$pbe_num out of df$enrollment
## X-squared = 303.7, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.01475 0.01940
## sample estimates:
##  prop 1  prop 2
## 0.04721 0.03013
```

This tells me private schools have a higher rate of PBEs than public schools across the state as a whole.

Let's drill down and compare the average school PBE rate between public and private schools. I'll need to do a Wilcoxon rank-sum test (the non-parametric version of a t-test) to determine if the rates of PBEs are truly different between public and private schools.

```
aggregate(newiz1314$pberate, by=list(Type=newiz1314$public_private), mean)
```

```
##      Type     x
## 1 PRIVATE 5.619
## 2  PUBLIC 5.398
```

```
wilcox.test(pberate~public_private, newiz1314)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pberate by public_private
## W = 1127534, p-value = 3.501e-11
## alternative hypothesis: true location shift is not equal to 0
```

The Wilcoxon rank sum test returns a p-value under 0.05, meaning we can reject the null hypothesis that there is no difference between the two groups, and accept the alternative hyptothesis that there is a statistically significant difference in PBE rates between public and private schools.

I want to make a boxplot of PBE rate for public versus private schools, but I know my data is pretty skewed, so I'm going to first remove the outliers that are 3 or more standard deviations from the mean. This should make my boxplot easier to read though it's not exactly good practice for analysis.

```
mean(newiz1314$pberate)+ 3*sd(newiz1314$pberate) #remove obs over this value
```

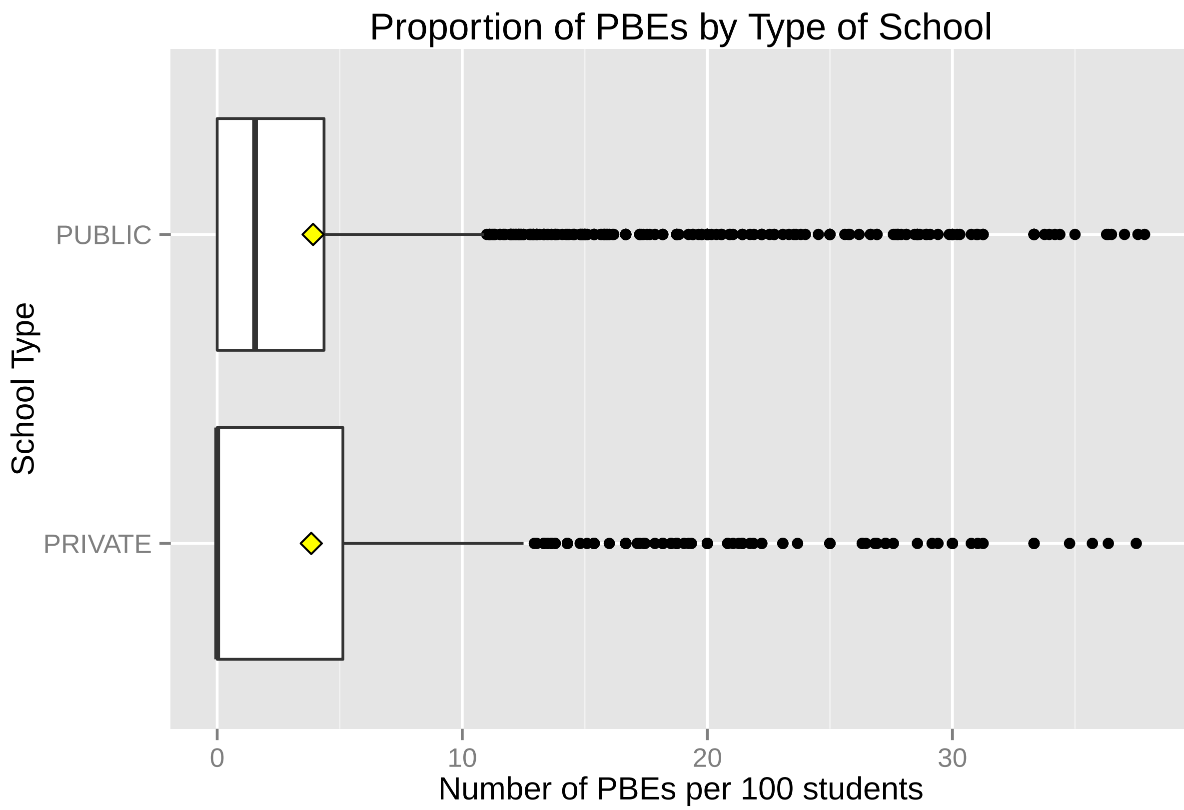
```
## [1] 38.05
```

```
length(which(newiz1314$pberate>=38.05)) #double check how many rows are dropped
```

```
## [1] 114
```

```
no.outliers<- subset(newiz1314, pberate<=38.05, select=c(school.code, public_private, pberate))

ggplot(no.outliers, aes(public_private, pberate))+
  geom_boxplot()+
  coord_flip()+
  stat_summary(fun.y="mean", geom="point", shape=23, size=3, fill="yellow")+
  ggtitle("Proportion of PBEs by Type of School")+
  xlab("School Type")+
  ylab("Number of PBEs per 100 students")
```



Size

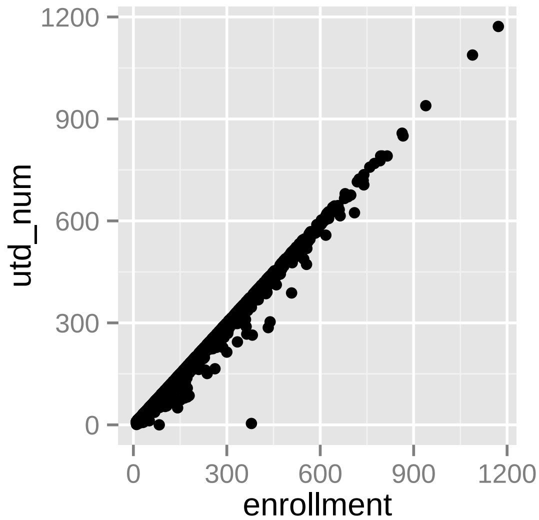
Let's look at the question of size. Do schools of different enrollment sizes report different rates of PBEs?

```
cor(newiz1314[ ,c(6, 7, 8, 9)])
```

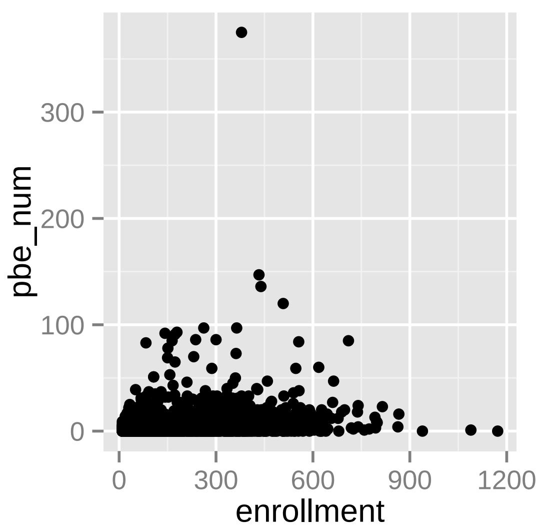
```
##             enrollment utd_num pme_num pbe_num
## enrollment    1.00000  0.9975 0.08242 0.26350
## utd_num       0.99747  1.0000 0.06560 0.19602
## pme_num       0.08242  0.0656 1.00000 0.03031
## pbe_num       0.26350  0.1960 0.03031 1.00000
```

I'd expect enrollment to correlate with the number of UTDp– more students means more up-to-date students. PMEs don't seem to correlate with anything– they are relatively rare (only 900 in the state). But, surprisingly, enrollment and the number of PBEs filed are not strongly correlated. Logically, a bigger school should mean more PBEs but the correlation coefficient indicates otherwise.

```
ggplot(newiz1314, aes(enrollment, utd_num))+ geom_point()
```




```
ggplot(newiz1314, aes(enrollment, pbe_num))+geom_point()
```



It might be easier to see a difference using the ordinal variable "size":

```
library(plyr)
ddply(newiz1314, "size", summarize,
      UTD= sum(utd_num)/sum(enrollment),
      PME= sum(pme_num)/sum(enrollment),
      PBE= sum(pbe_num)/sum(enrollment))
```

```
##     size    UTD      PME     PBE
## 1  Large 0.9733 0.001521 0.02515
## 2 Medium 0.9496 0.003262 0.04718
## 3  Small 0.9145 0.003245 0.08227
```

Small schools have higher rates of PBEs that large schools. However, I suspect that private schools are generally smaller than public schools, meaning the public/private status could be a potential confounder for any difference in PBE rates across sizes of schools.

```
attach(newiz1314)
aggregate(enrollment, by=list(public_private), FUN=mean, na.rm=TRUE)
```

```
##   Group.1      x
## 1 PRIVATE  30.89
## 2  PUBLIC 191.40
```

```
table(size, public_private)
```

```
##         public_private
## size     PRIVATE PUBLIC
##   Large        6   1174
##   Medium     349    779
##   Small      747    416
```

```
detach(newiz1314)
```

This confirms my suspicion that private schools are on average smaller than public schools.

Geography

My final question was to try to understand how the different counties compare in their total rates of up-to-date students, students with PMEs, and students with PBEs?

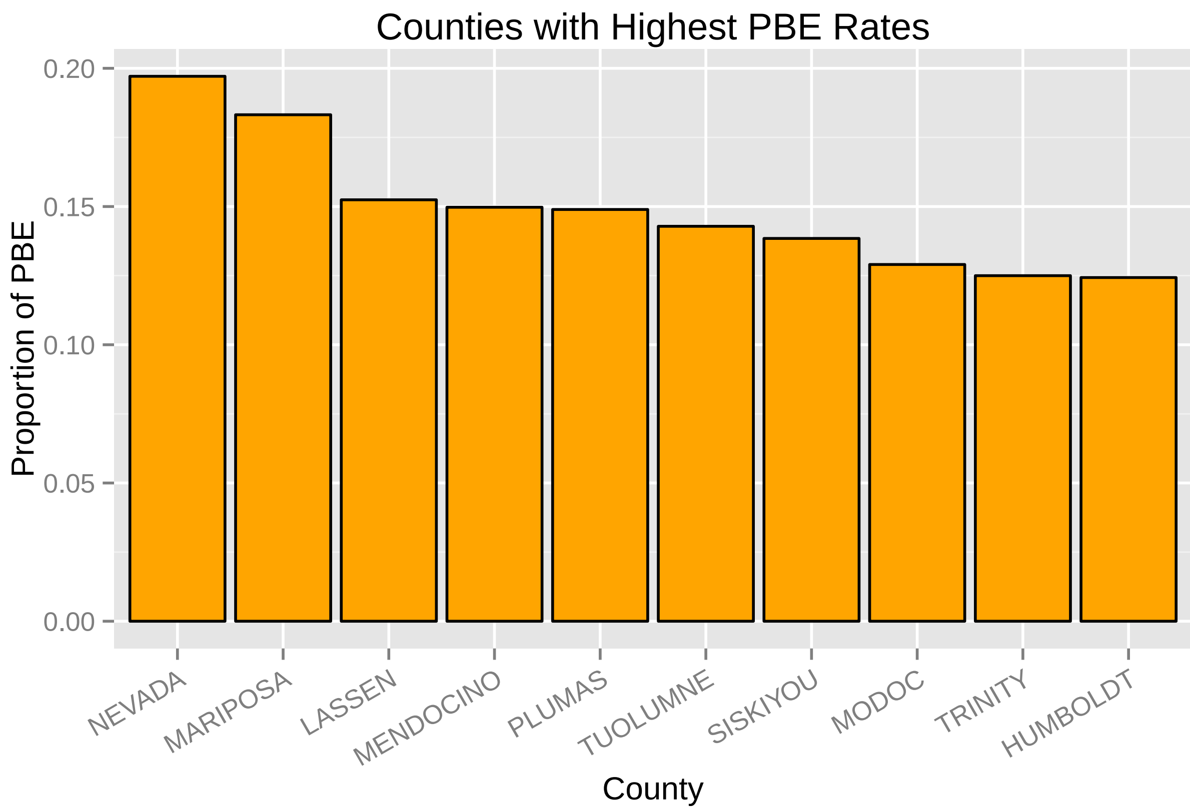
```
cstats<- ddply(newiz1314, "county", summarize,
      UTD= sum(utd_num)/sum(enrollment),
      PME= sum(pme_num)/sum(enrollment),
      PBE= sum(pbe_num)/sum(enrollment))
```

I'd like to create a bar graph showing the PBE rates of the 10 counties with the highest rates.

```
top10<- head(cstats[order(-cstats$PBE), ], 10)
print(top10)
```

```
##       county    UTD      PME    PBE
## 28    NEVADA 0.8029 0.000000 0.1971
## 21  MARIPOSA 0.8092 0.007634 0.1832
## 17    LASSEN 0.8476 0.000000 0.1524
## 22 MENDOCINO 0.8472 0.003077 0.1497
## 31    PLUMAS 0.8511 0.000000 0.1489
## 54  TUOLUMNE 0.8571 0.000000 0.1429
## 46  SISKIYOU 0.8590 0.002564 0.1385
## 24     MODOC 0.8387 0.032258 0.1290
## 52   TRINITY 0.8636 0.011364 0.1250
## 11  HUMBOLDT 0.8733 0.002331 0.1243
```

```
cnames<- top10[ ,1]
ggplot(top10, aes(county, PBE))+
  geom_bar(stat="identity", color="black", fill="orange")+
  ggtitle("Counties with Highest PBE Rates")+
  xlab("County")+
  ylab("Proportion of PBE")+
  scale_x_discrete(limits=cnames)+
  theme(axis.text.x= element_text(angle=30, hjust=1, vjust=1))
```


Counties with Highest PBE Rates
Proportion of PBE
0.20
0.15
0.10
0.05
0.00
NEVADA
MARIPOSA
LASSEN
MENDOCINO
PLUMAS
TUOLUMNE
SISKIYOU
MODOC
TRINITY
HUMBOLDT
County