data used : from census_income.csv Lets dig deeper in our data to find, to examine significant [Or not] interactions among different features [variables].

Note: In the solution , discussion on whether tests results conform to visual evidence has been left for forum discussions

One

`Are capital.gain and capital.loss really different ?`

From a data perspective this question can be taken as , is the average difference between capital.gain and capital.loss significantly different from zero? Since in this we can calculate individual differences between capital.gain and capital.loss , we will be using paired t-test.

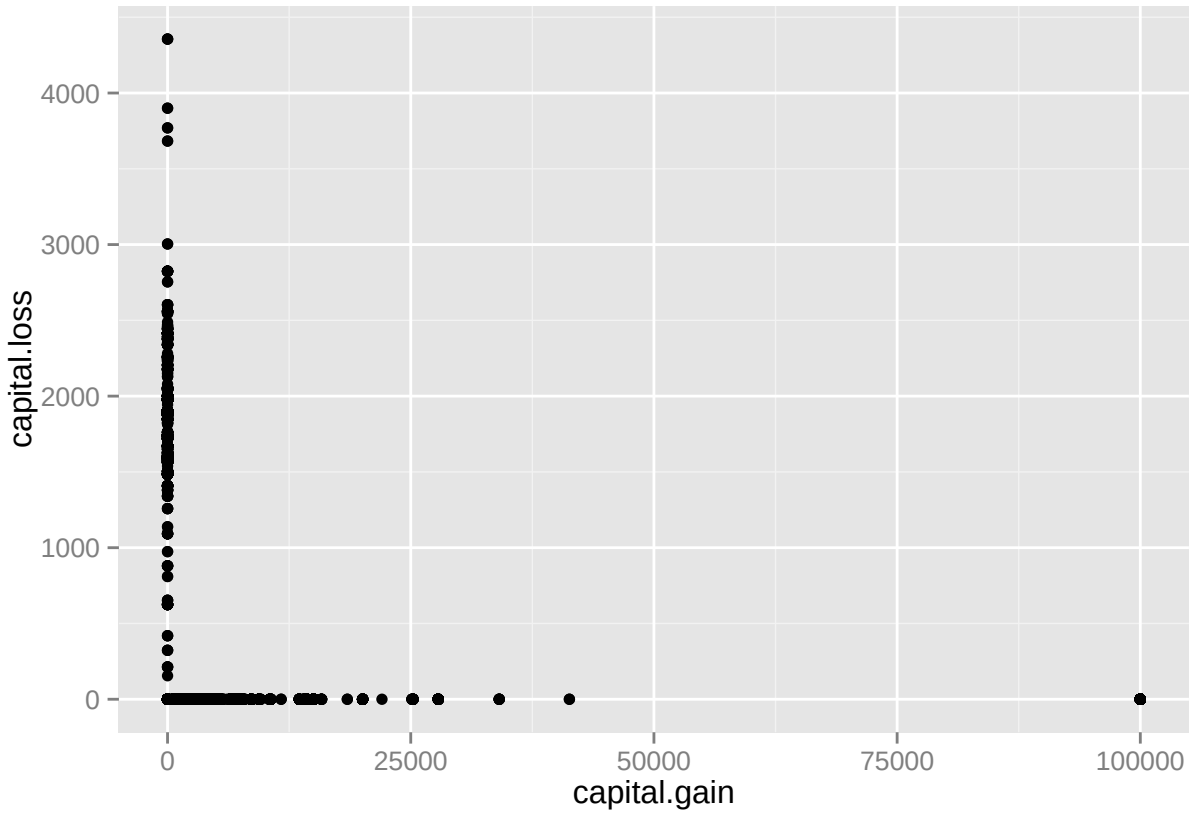
```
setwd("/Users/lalitsachan/Desktop/March onwards/CBAP with R/Data/")
# You'll have to chose path accroding to location of file in your machine

d=read.csv("census_income.csv",stringsAsFactors = F)
t.test(d$capital.loss,d$capital.gain,paired = T)
```

```
##
##  Paired t-test
##
## data:  d$capital.loss and d$capital.gain
## t = -24.12, df = 32560, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1070.8225  -909.8676
## sample estimates:
## mean of the differences
##                -990.345
```

very small p-value indicates that NULL hypothesis is not true. Also if you look at mean of the difference it is much far from 0.

```
library(ggplot2)
ggplot(d,aes(capital.gain,capital.loss))+geom_point()
```



Two

`Find out if hours.per.week are significantly different across genders , income levels.`

Again the underlying questions are , is there a significant difference on an average in hours.per.week when genders are different [Female, Male] or when income levels are different? Since we can not calculate individual differences in this case , we will be using unpaired t-test. Also , we first need to test whether variances are equal across two groups in order to provide appropriate input to t.test function.

```
var.test(d$hours.per.week[d$sex==" Female"],d$hours.per.week[d$sex==" Male"])
```

```
##
##  F test to compare two variances
##
## data:  d$hours.per.week[d$sex == " Female"] and d$hours.per.week[d$sex == " Male"]
## F = 0.94975, num df = 10770, denom df = 21789, p-value = 0.002041
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9193215 0.9813577
## sample estimates:
## ratio of variances
##          0.9497464
```

```
var.test(d$hours.per.week[d$Y==" <=50K"],d$hours.per.week[d$Y!=" <=50K"])
```

```
##
```

```
##  F test to compare two variances
##
## data:  d$hours.per.week[d$Y == " <=50K"] and d$hours.per.week[d$Y != " <=50K"]
## F = 1.2512, num df = 24719, denom df = 7840, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.206880 1.296796
## sample estimates:
## ratio of variances
##           1.251243
```

This result indicates that variances are unequal in both the cases.

```
t.test(d$hours.per.week~d$sex,paired=F,var.equal=F)
```

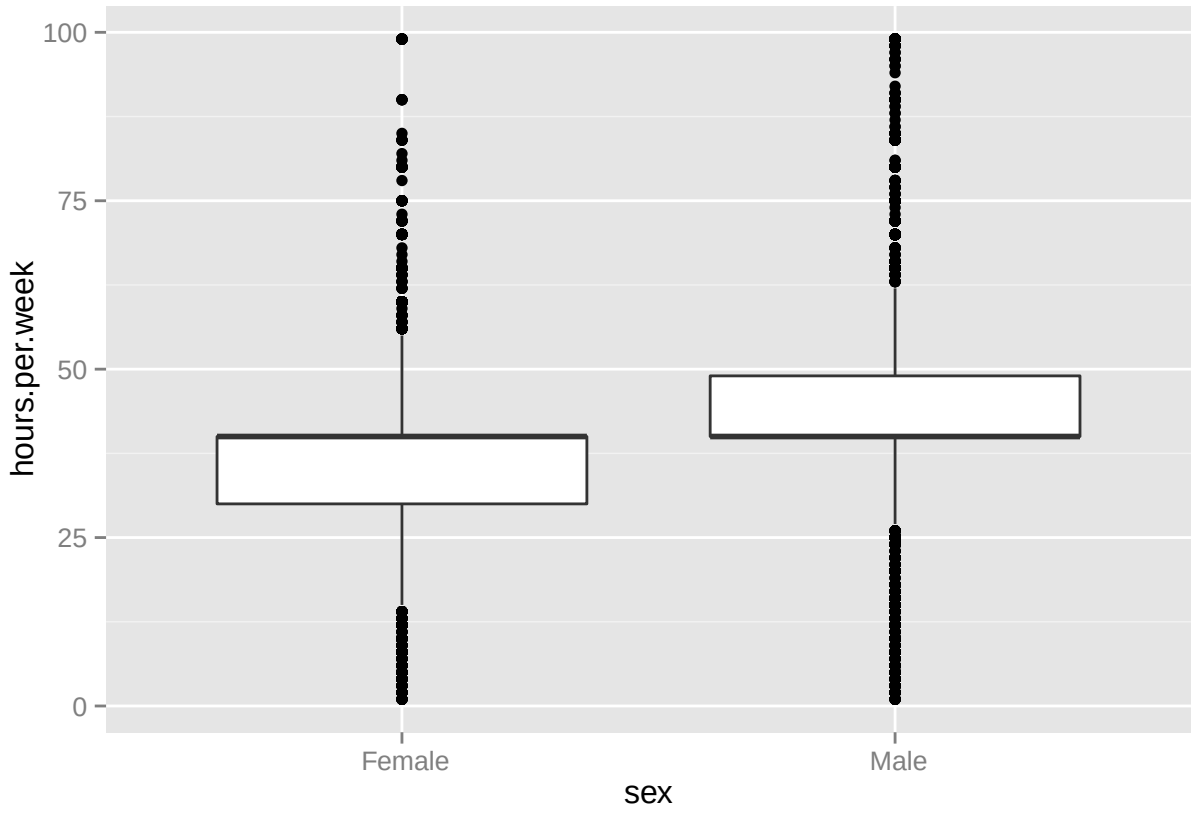
```
##
##  Welch Two Sample t-test
##
## data:  d$hours.per.week by d$sex
## t = -42.882, df = 21958, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.292787 -5.742664
## sample estimates:
## mean in group  Female   mean in group  Male
##              36.41036              42.42809
```

```
t.test(d$hours.per.week~d$Y,paired=F,var.equal=F)
```

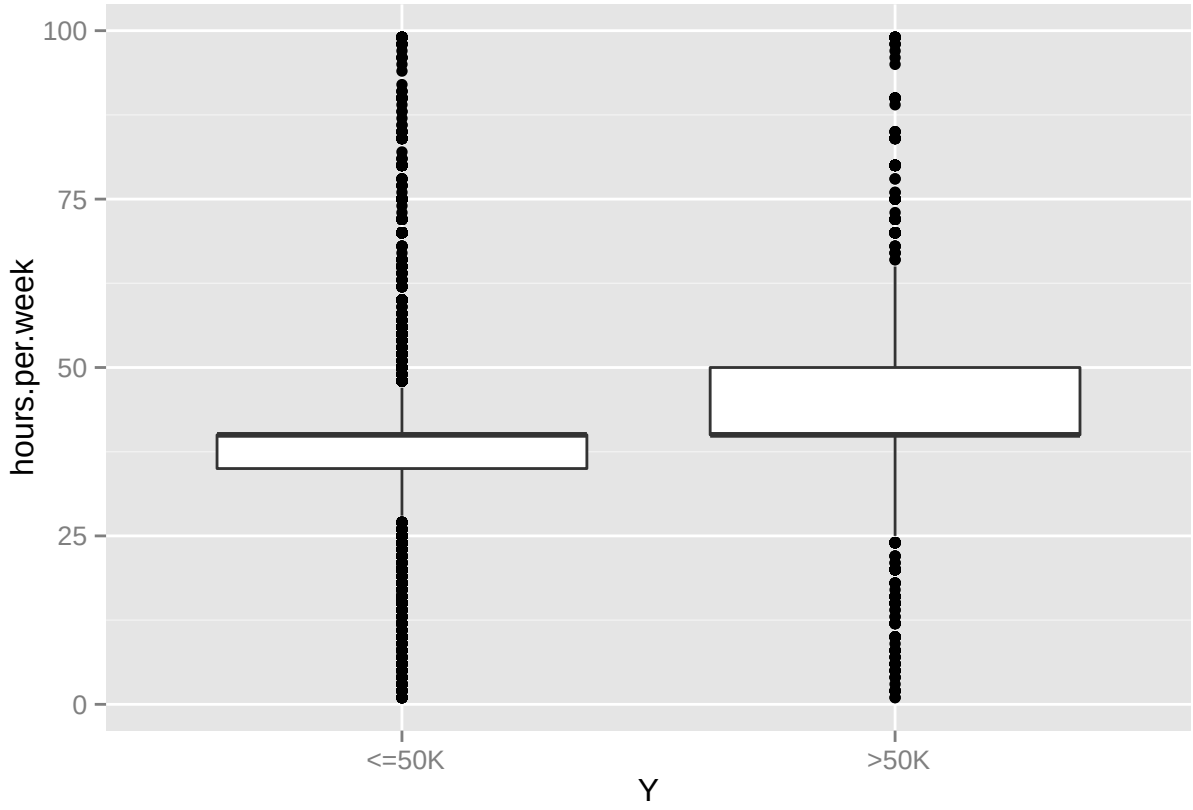
```
##
##  Welch Two Sample t-test
##
## data:  d$hours.per.week by d$Y
## t = -45.123, df = 14570, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.920943 -6.344690
## sample estimates:
## mean in group  <=50K  mean in group  >50K
##             38.84021             45.47303
```

Again in both the cases , p-values indicate that Hourse-Per-Week differ based on sex or income level.

```
ggplot(d,aes(x=sex,y=hours.per.week))+geom_boxplot()
```




```
ggplot(d,aes(x=Y,y=hours.per.week))+geom_boxplot()
```



Three

Intuitively it seems that education levels should be different across workclasses , examine whether this is true using your data. Also report which classes have education levels siginificantly higher/lower in comparison to others ; using an additional test. [variables involved : education.num and workclass].

Underlying questions here is whether average education.num is significantly different across various levels of workclass. Since workclass has more than two levels we can not use simple unpaired t-test , we will be using ANOVA. To further assess which of the classes are different each other , we will be doing bon-ferroni test.

```
summary(aov(education.num~workclass,data=d))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## workclass       8   8393  1049.1   164.9 <2e-16 ***
## Residuals   32552 207118     6.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

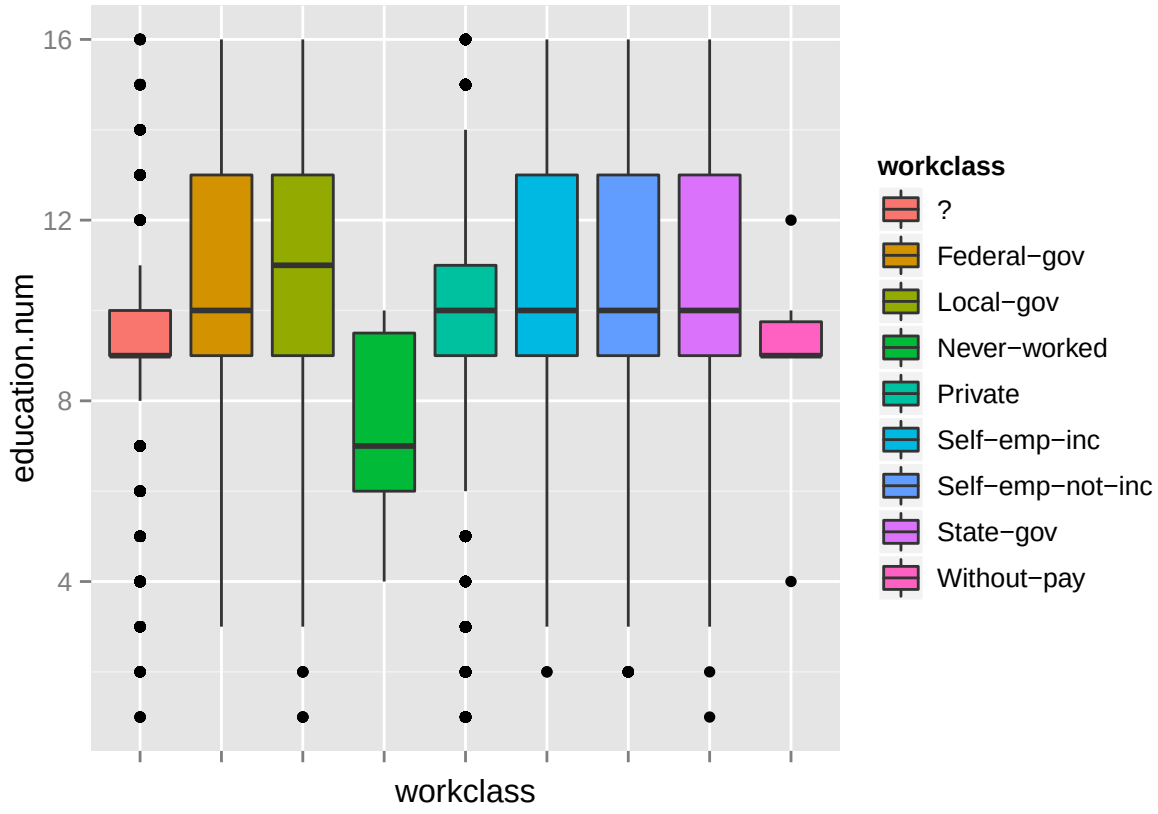
p-value for this indicates that there exist atleast one workclass for which education num is different than others. Now to find out which one it is, we'll do bon-ferroni test, you get a cross table of p-values for pair wise comparison of differences.:

```
pairwise.t.test(d$education.num, d$workclass, p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  d$education.num and d$workclass
##
##                    ?       Federal-gov  Local-gov  Never-worked  Private
##  Federal-gov       < 2e-16 -            -          -             -
##  Local-gov         < 2e-16 1.0000       -          -             -
##  Never-worked      1.0000  0.0076       0.0056     -             -
##  Private           < 2e-16 < 2e-16      < 2e-16    0.3657        -
##  Self-emp-inc      < 2e-16 1.0000       1.0000     0.0038        < 2e-16
##  Self-emp-not-inc  < 2e-16 1.9e-13      < 2e-16    0.1219        1.9e-09
##  State-gov         < 2e-16 0.0065       0.0065     0.0013        < 2e-16
##  Without-pay       1.0000  0.1831       0.1288     1.0000        1.0000
##                     Self-emp-inc  Self-emp-not-inc  State-gov
##  Federal-gov       -             -                 -
##  Local-gov         -             -                 -
##  Never-worked      -             -                 -
##  Private           -             -                 -
##  Self-emp-inc      -             -                 -
##  Self-emp-not-inc  < 2e-16       -                 -
##  State-gov         0.7330        < 2e-16           -
##  Without-pay       0.0838        1.0000            0.0243
##
## P value adjustment method: bonferroni
```

You can see p-value associated with Federal-gov and State-gov is fairly low which indicates the difference in education num in those two classes. Where as p-value corresponding to Federal-gov and Local-gov is 1, which means there is practically no difference in education num in those two groups. You can similarly check for other groups too.

```
ggplot(d,aes(x=workclass,y=education.num,fill=workclass))+
  geom_boxplot()+theme(axis.text.x=element_blank())
```



Four

`Find if income levels are affected by race.`

Since both of them are categorical variables , we will be doing a chi-sq test.

```
chisq.test(d$race,d$Y)
```

```
##
##  Pearson's Chi-squared test
##
## data:  d$race and d$Y
## X-squared = 330.92, df = 4, p-value < 2.2e-16
```

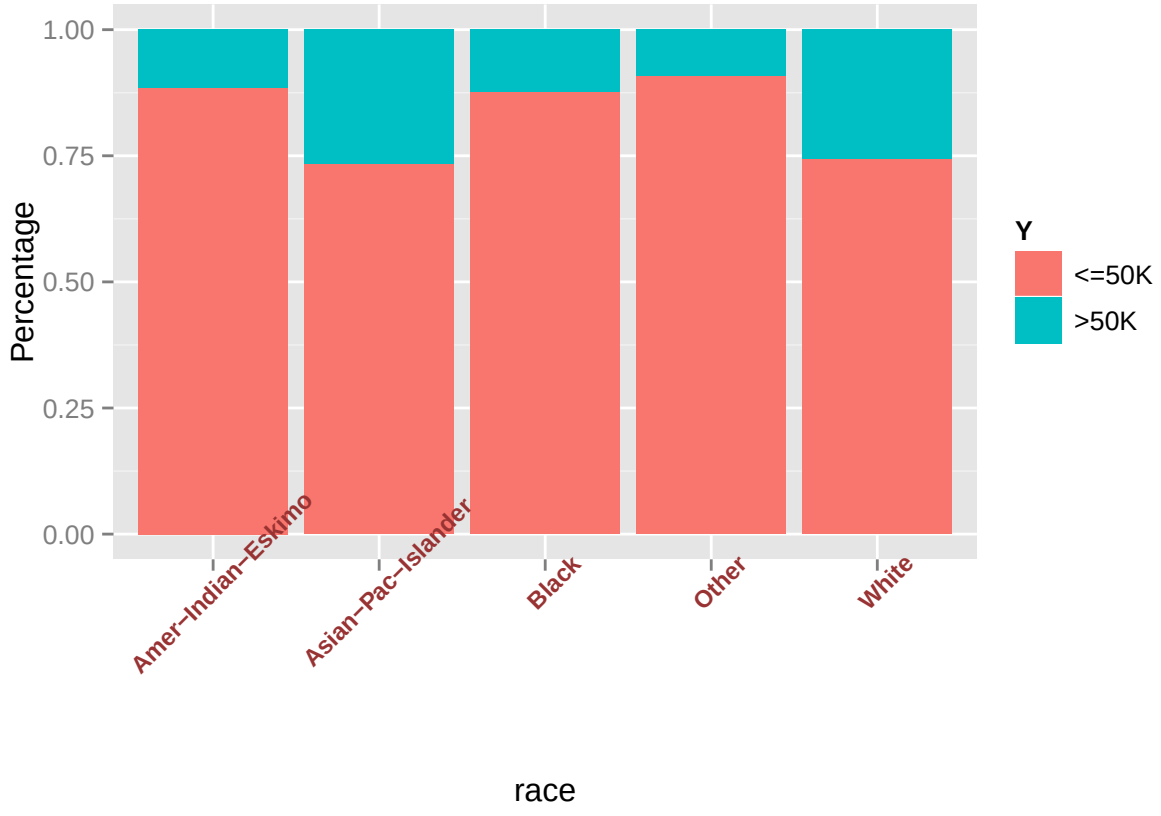
p-value indicates that race does affect income levels significantly. You can look at this proportion cross table also which will give an idea which race fare similarly in the data.

```
round(prop.table(table(d$race,d$Y),1),2)
```

```
##
##                        <=50K  >50K
```

```
##    Amer-Indian-Eskimo   0.88  0.12
##    Asian-Pac-Islander   0.73  0.27
##    Black                0.88  0.12
##    Other                0.91  0.09
##    White                0.74  0.26
```

```
ggplot(d,aes(x=race,fill=Y))+
  geom_bar(position = "fill")+
  ylab("Percentage")+
  theme(axis.text.x = element_text(face="bold", color="#993333",size=9, angle=45))
```



Note : Discussion on visual evidence is left for forum discussions