Ankit Midterm 1(R Programming)

```
# Imporitng the dataset for box plot of each region from Home Price Dataset
central<-read.csv('Centralzone.csv')
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec =
## dec, : embedded nul(s) found in input
```

```
west<-read.csv('West.csv')
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec =
## dec, : embedded nul(s) found in input
```

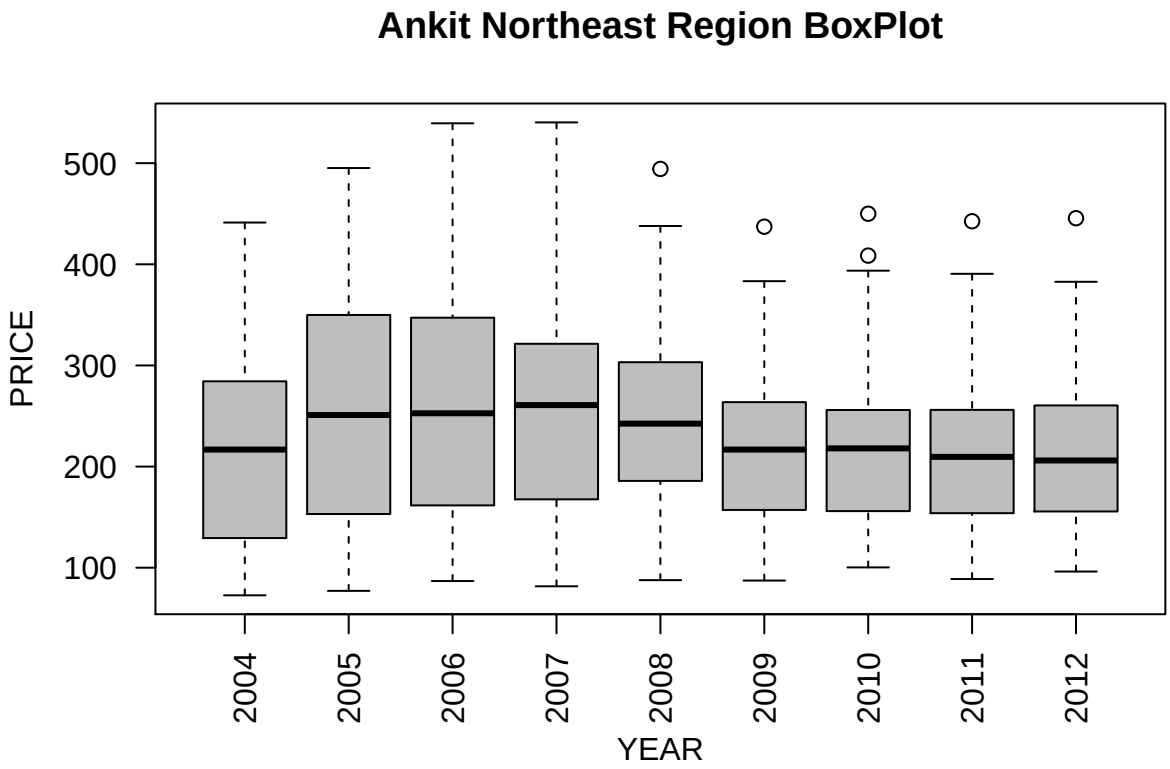
```
northeast<-read.csv('Northeastzone.csv')
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec =
## dec, : embedded nul(s) found in input
```

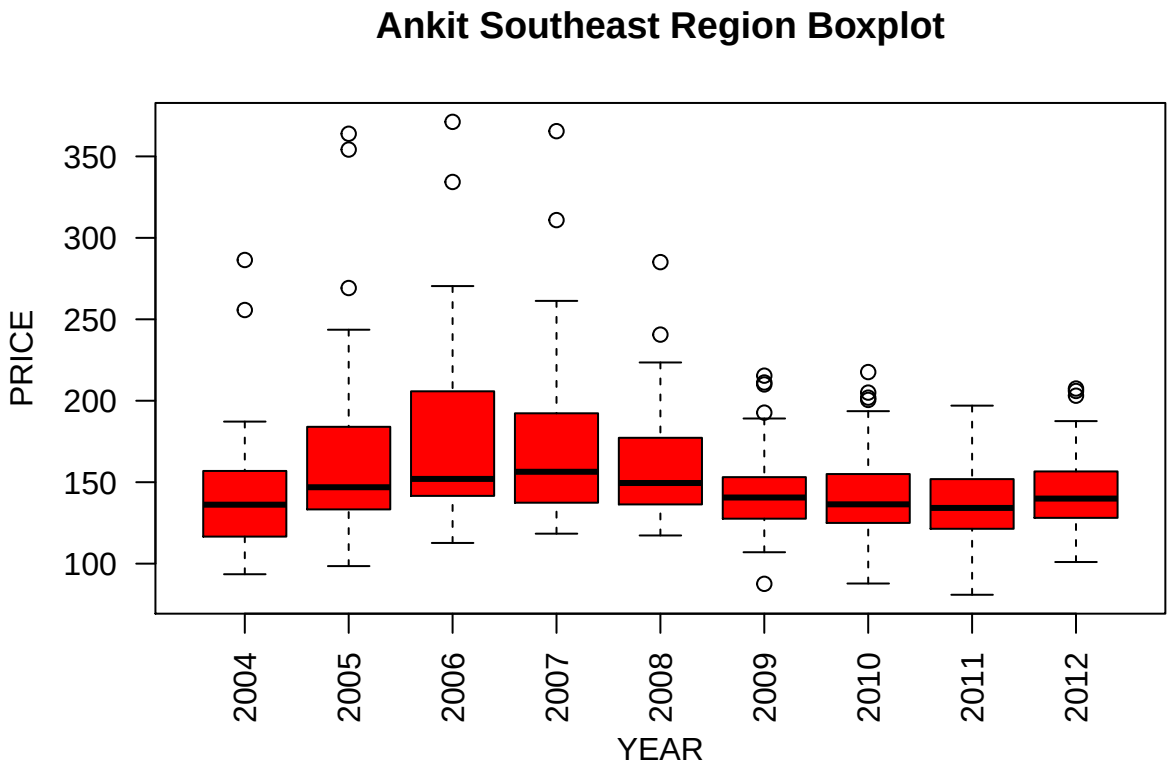
```
southeast<-read.csv('Southeastzone.csv')
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec =
## dec, : embedded nul(s) found in input
```

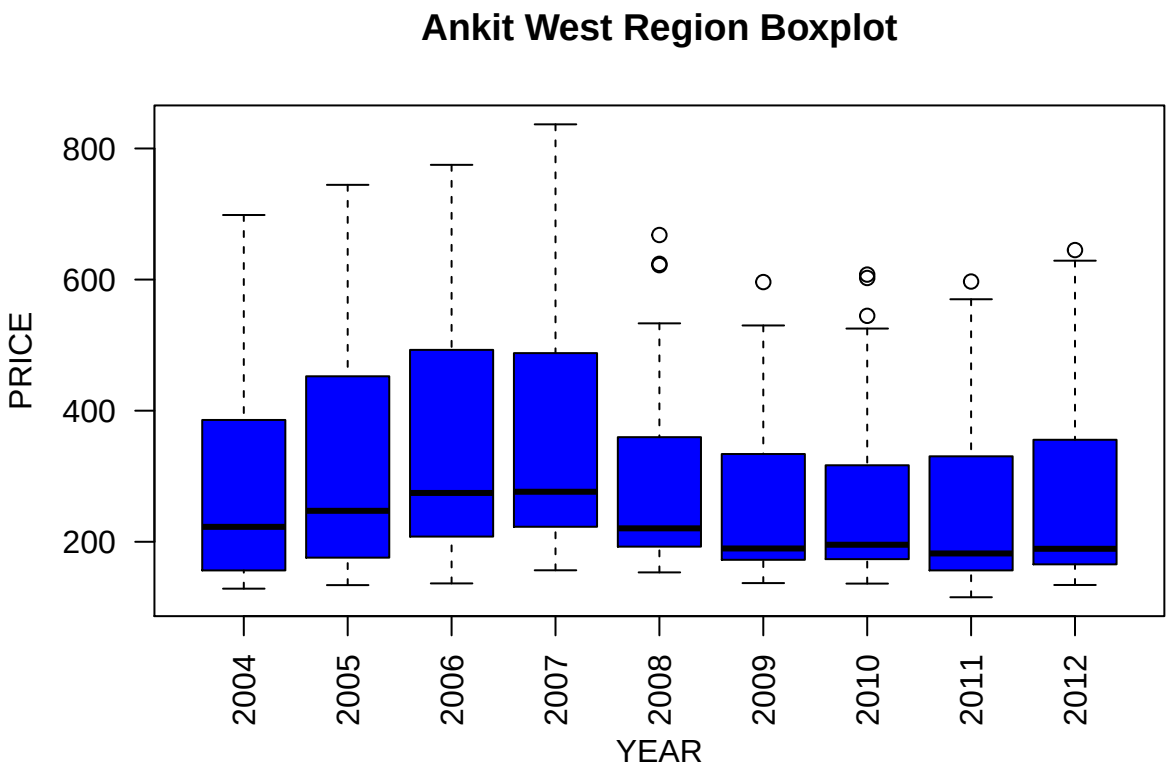
```
#Plotting the boxplot by each region boxplot
boxplot(northeast,las=2,names=c("2004","2005","2006","2007","2008","2009","2010","2011","2012"),col="gr
```



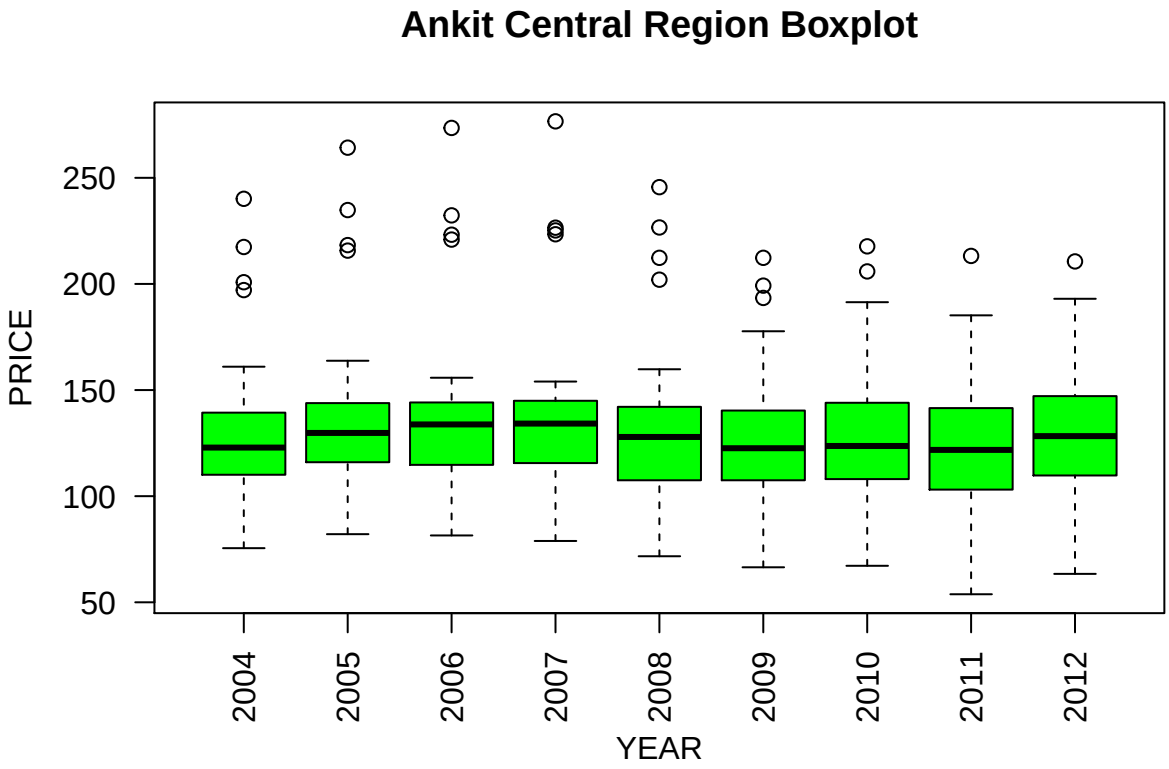
```
boxplot(southeast,las=2,names=c("2004","2005","2006","2007","2008","2009","2010","2011","2012"),col="red
```




```
boxplot(west,las=2,names=c("2004","2005","2006","2007","2008","2009","2010","2011","2012"),col="blue",xl
```




```
boxplot(central,las=2,names=c("2004","2005","2006","2007","2008","2009","2010","2011","2012"),col="gree
```




```
# Importing the Dataset
Home<-read.csv('Homeprice.csv')
attach(Home)
#Summarize the dataset by region and Descriptive Statistics
summary(Home)
```

```
##     Central        Northeast       Southeast          West
##  Min.   : 13.0   Min.   : 72.7   Min.   : 80.9   Min.   :115.4
##  1st Qu.:108.5   1st Qu.:155.4   1st Qu.:128.7   1st Qu.:173.6
##  Median :127.0   Median :216.9   Median :142.8   Median :228.9
##  Mean   :128.8   Mean   :234.3   Mean   :151.7   Mean   :300.7
##  3rd Qu.:142.6   3rd Qu.:289.9   3rd Qu.:164.2   3rd Qu.:374.4
##  Max.   :276.6   Max.   :540.3   Max.   :371.2   Max.   :836.8
##  NA's   :93      NA's   :185     NA's   :75      NA's   :439
```

```
# ggplot2 and plyr for trend analysis for the price variation between the states
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(plyr)
```

```
## Loading required package: plyr
```

```
# Importing Dataset for Price Trend Analysis
centralplot<-read.csv('T_centralplot.csv')
head(centralplot)
```

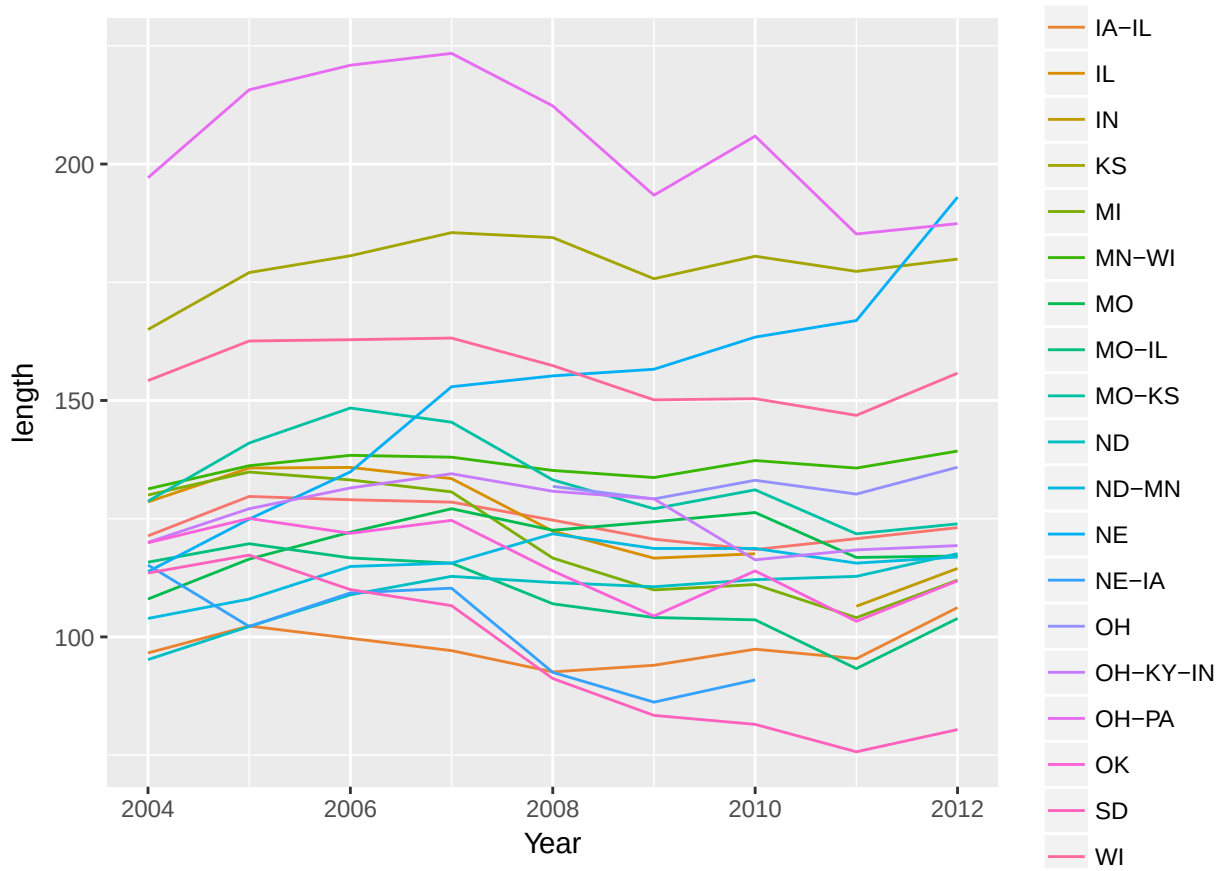
```
##   State Year Central
```

```
## 1    IA 2004   127.2
## 2    IA 2004    93.6
## 3    IA 2004   143.3
## 4    IA 2005   137.7
## 5    IA 2005    96.6
## 6    IA 2005   154.8
```

```
#Removing Missing values(NA)
a<-na.omit(centralplot)
```

```
# Plotting the graph
central<-ddply(centralplot,c("State","Year"),summarise, length=mean(Central))
ggplot(data=central,mapping=aes(x=Year,y=length,colour=State, main="Ankit Trend")) + geom_line()
```

```
## Warning: Removed 15 rows containing missing values (geom_path).
```



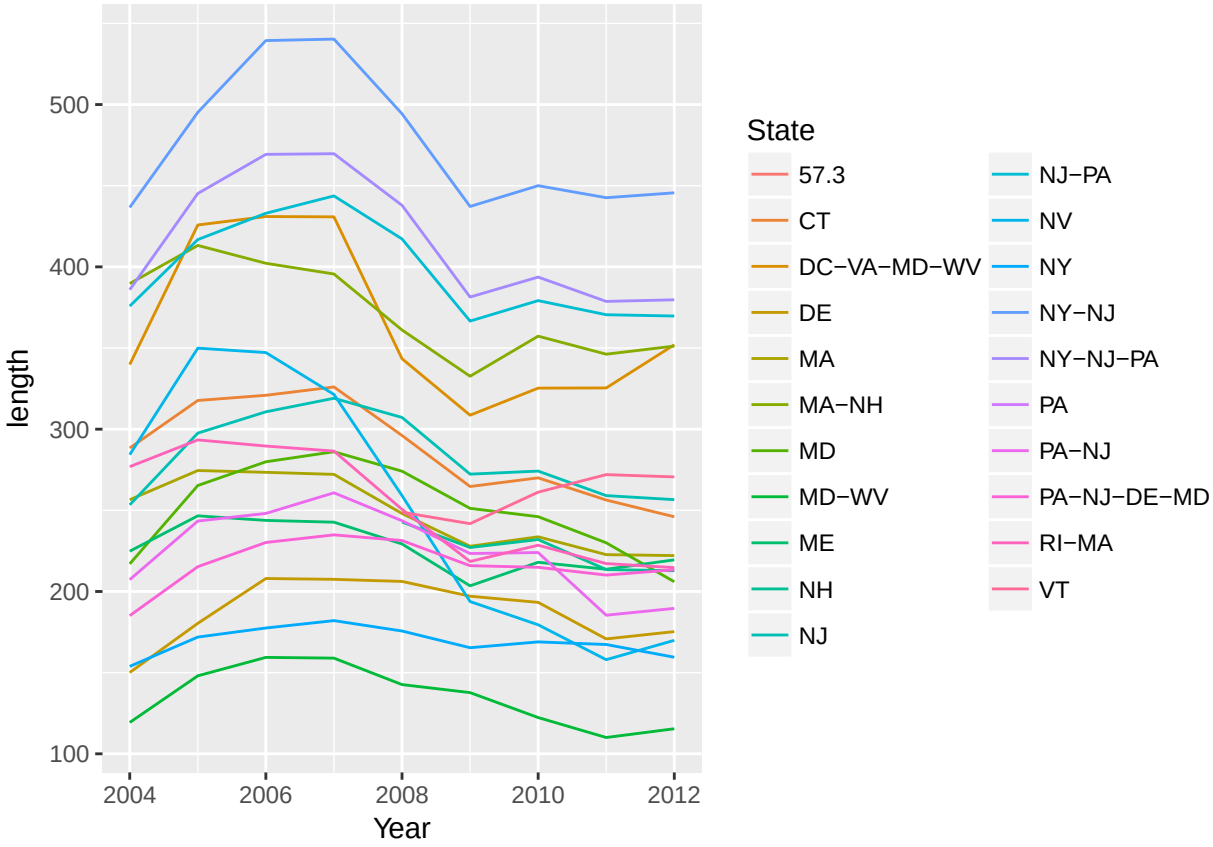

```
# Importing the dataset
northplot<-read.csv('T_northplot.csv')
head(northplot)
```

```
##   State Year Northeast
## 1    CT 2004     441.3
## 2    CT 2004     231.6
## 3    CT 2004     249.2
## 4    CT 2004     231.5
## 5    CT 2005     482.4
## 6    CT 2005     253.3
```

```
# Removing missing values(NA)
b<- na.omit(northplot)
```

```
# Plotting the graph
north<-ddply(northplot,c("State","Year"),summarise, length=mean(Northeast))
ggplot(data=north,mapping=aes(x=Year,y=length,colour=State, main="Ankit Trend")) + geom_line()
```

```
## Warning: Removed 8 rows containing missing values (geom_path).
```



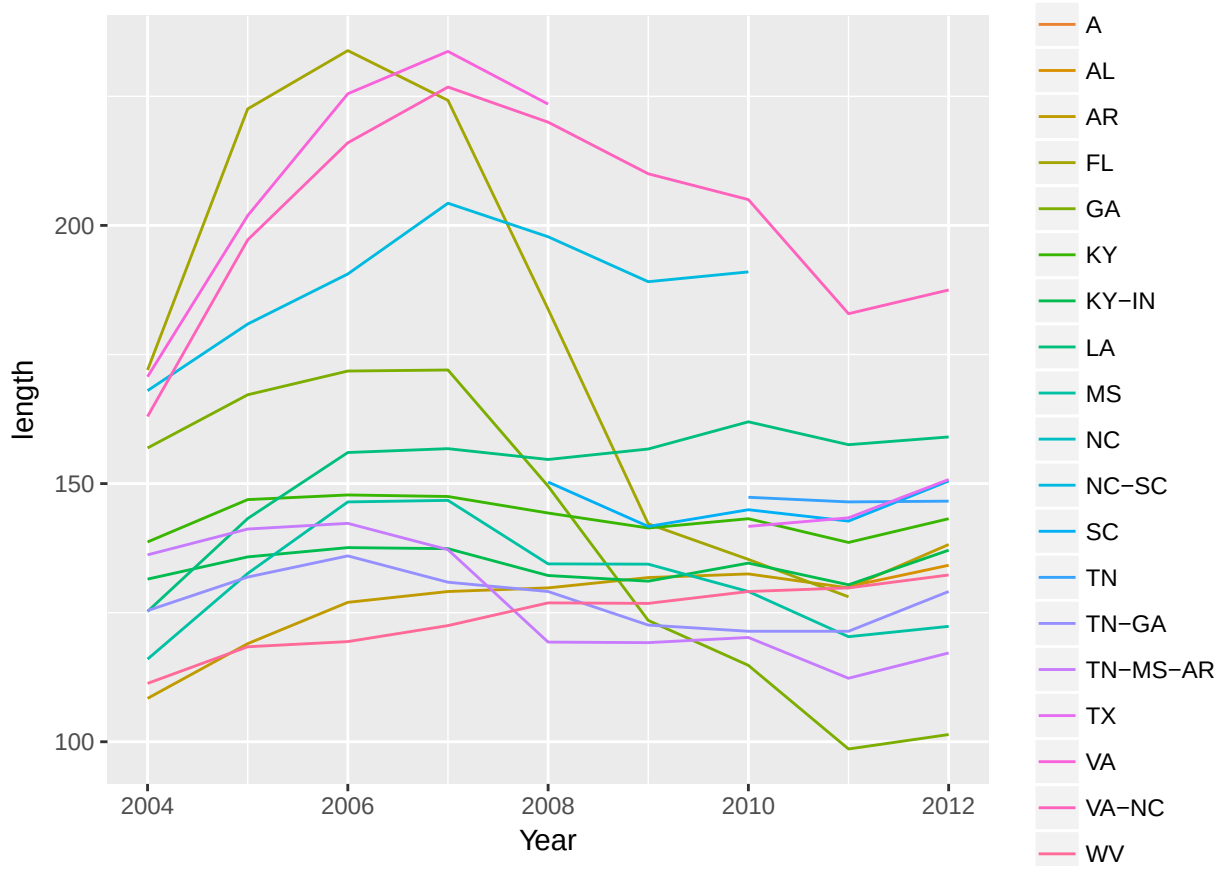

```
# Importing the dataset
southplot<-read.csv('T_southplot.csv')
head(southplot)
```

```
##   State Year Southeast
## 1    AL 2004     146.6
## 2    AL 2004        NA
## 3    AL 2004        NA
## 4    AL 2004     115.2
## 5    AL 2004     116.6
## 6    AL 2005     157.0
```

```
# Removing missing values(NA)
c<- na.omit(southplot)
```

```
# Plotting the graph
south<-ddply(southplot,c("State","Year"),summarise, length=mean(Southeast))
ggplot(data=south,mapping=aes(x=Year,y=length,colour=State, main="Ankit Trend")) + geom_line()
```

```
## Warning: Removed 38 rows containing missing values (geom_path).
```




```
# Importing the dataset
west.plot<-read.csv('T_westplot.csv')
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec =
## dec, : embedded nul(s) found in input
```

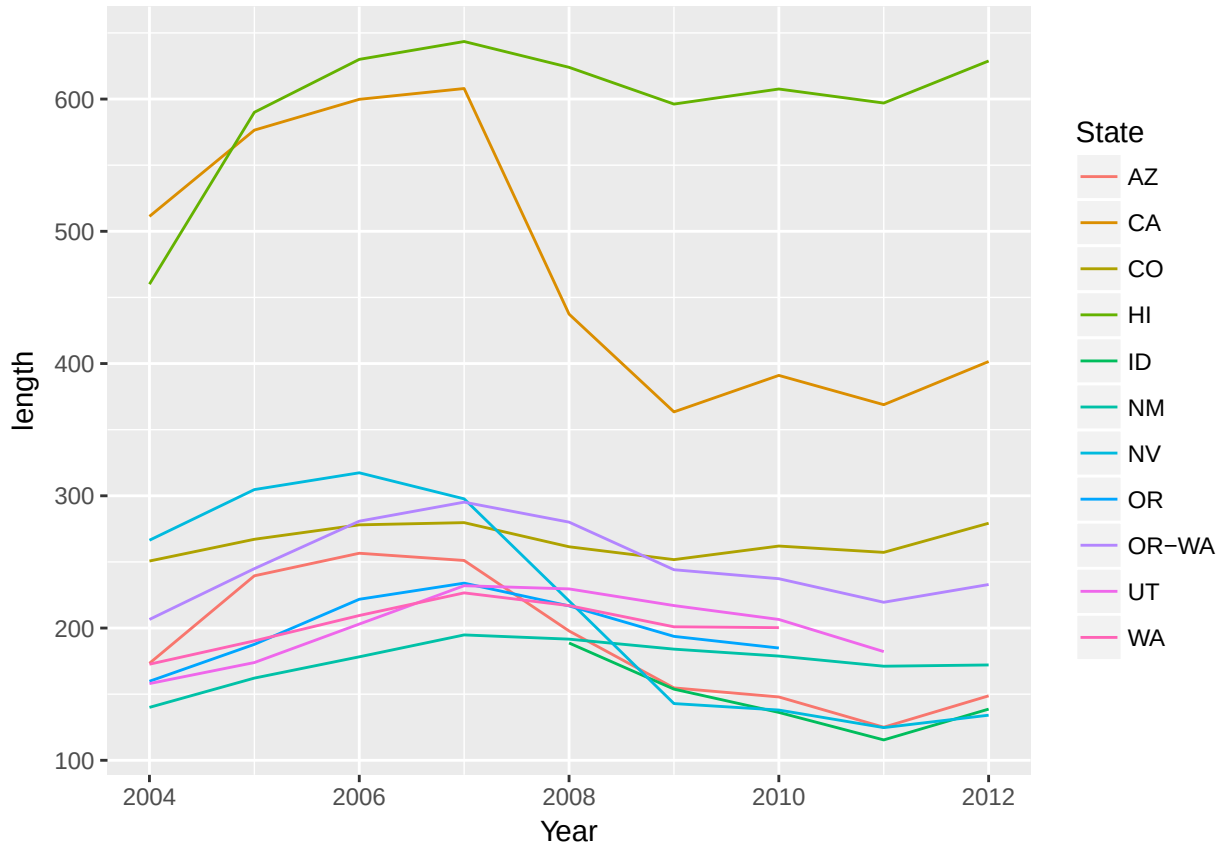
```
head(west.plot)
```

```
##   State Year  West
## 1    AZ 2004 169.4
## 2    AZ 2004 177.3
## 3    AZ 2005 247.4
## 4    AZ 2005 231.6
## 5    AZ 2006 268.2
## 6    AZ 2006 244.9
```

```
# Removing Missing values(NA)
d<-na.omit(west.plot)
```

```
#Ploting the graph
west<-ddply(west.plot,c("State","Year"),summarise, length=mean(West))
ggplot(data=west,mapping=aes(x=Year,y=length,colour=State, main="Ankit Trend")) + geom_line()
```

```
## Warning: Removed 7 rows containing missing values (geom_path).
```




```
#HYPOTHESIS TESTING:
#NULL HO- Assuming there is no difference between the two population mean of region
#ALTERNATVE HA- There is a difference beween the means of region
#Performing Two independent sample t-test on central and northeast region
cn<-t.test(Central,Northeast)
cn
```

```
##
##  Welch Two Sample t-test
##
## data:  Central and Northeast
## t = -21.625, df = 570.71, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -115.09160  -95.92525
## sample estimates:
## mean of x mean of y
##  128.8095  234.3180
```

```
#Performing two independent sample t-test on central and southeast region
cs<-t.test(Central,Southeast)
cs
```

```
##
##  Welch Two Sample t-test
##
## data:  Central and Southeast
## t = -10.805, df = 1155.4, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -27.08353 -18.75899
## sample estimates:
## mean of x mean of y
##  128.8095  151.7308
```

```
#Performing two independent sample t-test on central and west region
cw<-t.test(Central,West)
cw
```

```
##
##  Welch Two Sample t-test
##
## data:  Central and West
## t = -14.867, df = 235.92, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -194.6189 -149.0759
## sample estimates:
## mean of x mean of y
##  128.8095  300.6570
```

```
#Performing two independent sample t-test on west and southeast region
Ws<-t.test(West,Southeast)
Ws
```

```
##
##  Welch Two Sample t-test
##
## data:  West and Southeast
## t = 12.858, df = 237.82, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  126.1097 171.7426
## sample estimates:
## mean of x mean of y
##  300.6570  151.7308
```

```
#Performing two independent sample t-test on northeast and southeast region
ns<-t.test(Northeast,Southeast)
ns
```

```
##
##  Welch Two Sample t-test
##
## data:  Northeast and Southeast
## t = 16.739, df = 594.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  72.89705 92.27729
## sample estimates:
## mean of x mean of y
##  234.3180  151.7308
```

```
# Performing two independent sample t-test on northeast and west region
nw<-t.test(Northeast,West)
```

```
nw
```

```
##
##  Welch Two Sample t-test
##
## data:  Northeast and West
## t = -5.3554, df = 307.23, p-value = 1.676e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -90.71388 -41.96408
## sample estimates:
## mean of x mean of y
##   234.318   300.657
```

```
#Performing two independent sample t-test on southeast and west
sw<-t.test(Southeast,West)
sw
```

```
##
##  Welch Two Sample t-test
##
## data:  Southeast and West
## t = -12.858, df = 237.82, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -171.7426 -126.1097
## sample estimates:
## mean of x mean of y
##  151.7308  300.6570
```

```
# Performing one way annova for the regions because the dataset contains missing values
# Importing dataset
aov.central<-read.csv('acentral.csv')
head(aov.central)
```

```
##   Year Central
## 1 2004   129.5
## 2 2004   140.8
## 3 2004    95.2
## 4 2004   107.8
## 5 2004   147.8
## 6 2004   127.2
```

```
# Performing one - way ANOVA on central region
ANOVA.c<-aov(Central~Year ,aov.central)
summary(ANOVA.c)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Year          1   1039    1039   0.853  0.356
## Residuals   683 831734    1218
## 26 observations deleted due to missingness
```

```
# Importing dataset
aov.northeast<-read.csv('anortheast.csv')
head(aov.northeast)
```

```
##   Year Northeast
## 1 2004     161.3
```

```
## 2 2004     207.3
## 3 2004     197.9
## 4 2004     217.0
## 5 2004     377.2
## 6 2004      85.3
```

```
# Performing out one - way ANOVA on northeast region
ANOVA.n<-aov(Northeast~Year ,aov.northeast)
summary(ANOVA.n)
```

```
##              Df  Sum Sq Mean Sq F value  Pr(>F)
## Year          1   83606   83606   7.259 0.00723 **
## Residuals   671 7728318   11518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 38 observations deleted due to missingness
```

```
# Importing dataset
aov.southeast<-read.csv('southeast.csv')
head(aov.southeast)
```

```
##   Year Southeast
## 1 2004        NA
## 2 2004      97.1
## 3 2004     156.9
## 4 2004     154.7
## 5 2004     127.7
## 6 2004      93.5
```

```
# Performing out a one - way ANOVA on southeast region
ANOVA.s<-aov(Southeast~Year ,aov.southeast)
summary(ANOVA.s)
```

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## Year          1   52144   52144   30.81 4.16e-08 ***
## Residuals   643 1088189    1692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 88 observations deleted due to missingness
```

```
# Importing dataset
aov.w<-read.csv('awest.csv')
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec =
## dec, : embedded nul(s) found in input
```

```
head(aov.w)
```

```
##   Year  West
## 1 2004 145.4
## 2 2004 627.3
## 3 2004    NA
## 4 2004 325.3
## 5 2004 187.6
## 6 2004 239.1
```

```
# Performing out a oneway ANOVA on west region
ANOVA.west<-aov(West~Year ,aov.w)
summary(ANOVA.west)
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## Year          1  141040  141040   4.689 0.0315 *
## Residuals   215 6467056   30079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 8 observations deleted due to missingness
```

```
# Imporitng dataset of all regions
aov.regions<-read.csv('aregions.csv')
head(aov.regions)
```

```
##   Treatment Response
## 1   Central    129.5
## 2   Central    140.8
## 3   Central     95.2
## 4   Central    107.8
## 5   Central    147.8
## 6   Central    127.2
```

```
# Performing annova for all regions
ANOVA<-aov(Response~Treatment ,aov.regions)
summary(ANOVA)
```

```
##               Df   Sum Sq Mean Sq F value Pr(>F)
## Treatment      3  7122414 2374138   247.7 <2e-16 ***
## Residuals   1526 14624073    9583
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 107 observations deleted due to missingness
```

```
# install and require "agricolae" for Least Significant Difference(LSD) test
require(agricolae)
```

```
## Loading required package: agricolae
```

```
# comparing the means of the regions or treatment in order to the one causing the significance of the t
comp<-LSD.test(ANOVA,"Treatment")
comp
```

```
## $statistics
##    MSerror   Df     Mean       CV
##   9583.272 1526 195.9027 49.97081
##
## $parameters
##         test p.ajusted    name.t ntr alpha
##   Fisher-LSD      none Treatment   4  0.05
##
## $means
##           Response       std   r      LCL      UCL   Min   Max   Q25
## Central   130.4303  34.34571 390 120.7069 140.1536  53.8 276.6 109.3
## Northeast 238.8557 107.12569 341 228.4572 249.2543  72.7 540.3 155.7
## Southeast 151.3334  39.32290 506 142.7970 159.8698  80.9 371.2 128.2
## West      310.0304 180.20641 293 298.8124 321.2484 115.4 836.8 174.4
##              Q50    Q75
## Central   127.85 143.15
## Northeast 220.60 299.10
## Southeast 142.85 163.70
```

```
## West      232.40 376.20
##
## $comparison
## NULL
##
## $groups
##           Response groups
## West      310.0304      a
## Northeast 238.8557      b
## Southeast 151.3334      c
## Central   130.4303      d
##
## attr(,"class")
## [1] "group"
```

```
# Explananing Results

# Descriptive statistics results

# Price variation trend was plotted for each region which helps to show how prices vary among states in
# We can also know the Price variation at the states with the highest and least average price.

# Box Plot Results

# Central region box plot have the highest number of outliers
# North east have the least number of outliers.

# West region prices are higher with the average mean of 300.7
# Northeast region prices are at second place with the average mean of 234.3
# Southeast region prices are at third place with the average mean of 151.7
# Central region has lowest price at last with the average mean of 128.8

# Independent two sample t test results

# Hypothesis testing was done for two independent populations of regions with the mean differences
# Indepedent T test was performed between the two regions mean and it was found that there is no signif

# ANOVA Results

# one-way Anova was perfomed for all the regions and also between each regions
# As there were missing values, we perfomed one-way Anova
# In one-way annova we can have missing values
# we performed annova to check the means of each years price in the region and also within all the regi
# Southeast is significant at alpha 0.0001
# Norheast is significant at alpha 0.01
# West is significant at aplha 0.10
# Therefore the years in west region have high significant difference
# ANOVA between the regions have very high significant difference at alpha =0.001
# As a result, we need to perform Least Square Difference LSD test to know the region which causes the
# We found that west region caused the difference.
```