ANOVA and Correlation

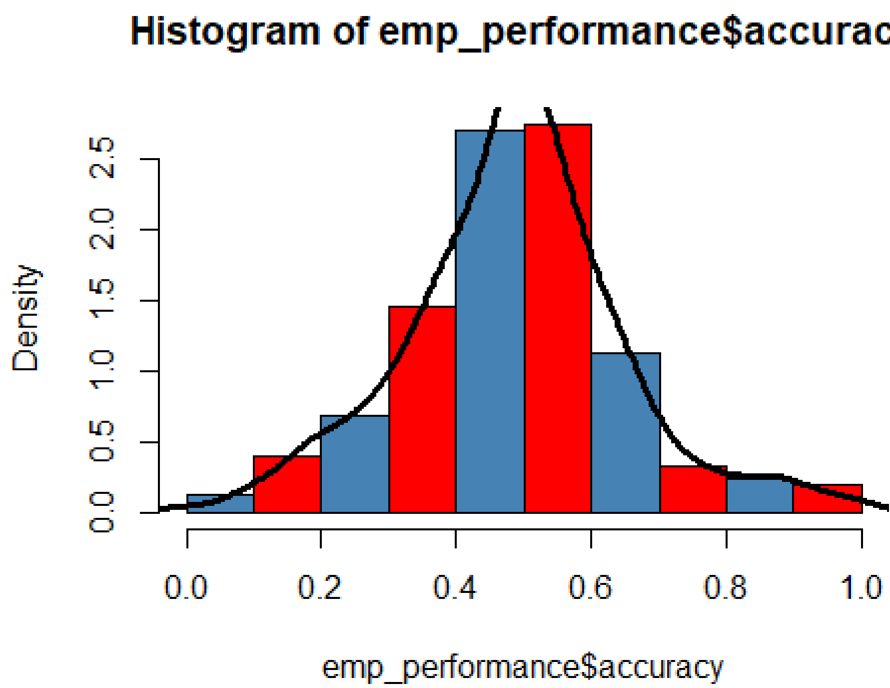
Suhail Shaikh

12/22/2019

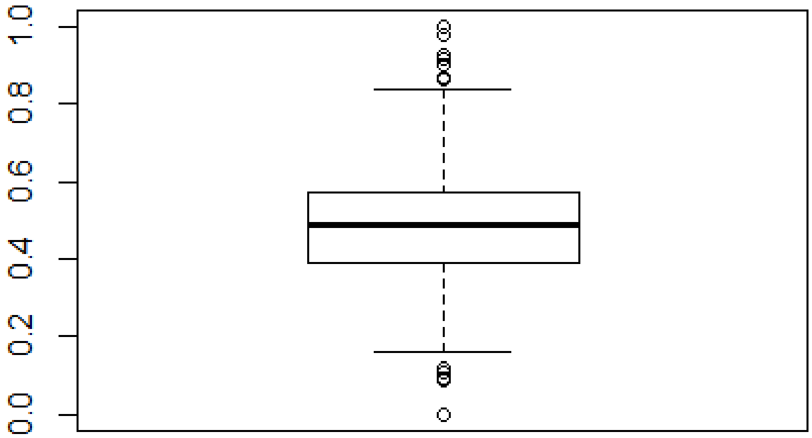
Q.1) Perform the Univariate Analysis of all the variables in the dataset.

```
setwd("C:\\Users\\shaik\\Desktop\\projects data mining\\General\\Correlation
and ANOVA")
load("Correlation and ANOVA.RData")

head(emp_performance)

##      accuracy actual_prod target_prod season   type work_area
## 107 0.5473852        8.57         7.5 Winter  Local   English
## 280 0.6628743       22.14        16.0 Winter  Local   Spanish
## 64  0.4758483        7.45         7.5 Winter Remote   English
## 43  0.3772455        6.30         8.0 Summer  Local     Other
## 193 0.4260051       12.45        14.0 Winter Remote     Other
## 253 0.4970060       16.60        16.0 Winter  Local   Spanish

str(emp_performance)

## 'data.frame':    248 obs. of  6 variables:
##  $ accuracy   : num  0.547 0.663 0.476 0.377 0.426 ...
##  $ actual_prod: num  8.57 22.14 7.45 6.3 12.45 ...
##  $ target_prod: num  7.5 16 7.5 8 14 16 7.5 7.5 8 7.5 ...
##  $ season     : Factor w/ 4 levels "Summer","Winter",..: 2 2 2 1 2 2 2 4 3
4 ...
##  $ type       : Factor w/ 2 levels "Local","Remote": 1 1 2 1 2 1 1 1 1 2
...
##  $ work_area  : Factor w/ 3 levels "Spanish","English",..: 2 1 2 3 3 1 2 2
3 2 ...
##  - attr(*, "na.action")= 'omit' Named int  5 10
##   ..- attr(*, "names")= chr  "168" "97"

#1.Variable name-accuracy
library(psych)
summary(emp_performance$accuracy)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3929  0.4886  0.4872  0.5741  1.0000

describe(emp_performance$accuracy)

##    vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 248 0.49 0.16   0.49    0.49 0.14   0   1     1 0.16      0.8 0.01
```

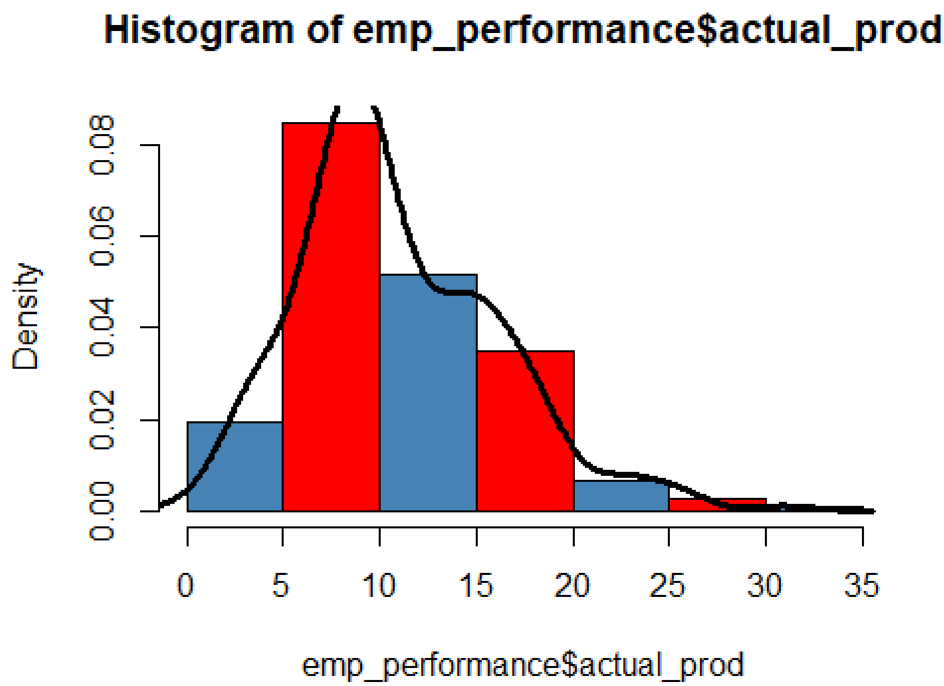
```r
#Here in accuracy mean is equal to median and sd is very small compared to the mean Hence Accuracy seems to have normally distributed. Lets check with the plots
hist(emp_performance$accuracy,probability=T,col=c("steelblue", "red"))
lines(density(emp_performance$accuracy),col="black",lwd=3)
```



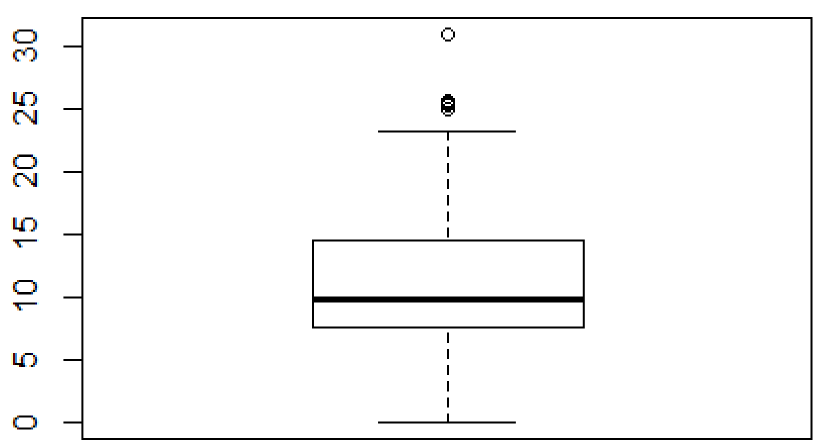

```r
#The histogram shows that the accuracy is normally distributed

boxplot(emp_performance$accuracy)
```




```
#Boxplot shows that there are some outliers in the variable accuracy.

#Hence the accuracy is normaly distributed with some outliers in the data.

#2.Variable name-actual_prod
library(psych)
summary(emp_performance$actual_prod)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.60    9.74   10.94   14.40   30.95

describe(emp_performance$actual_prod)

##    vars   n  mean   sd median trimmed  mad min   max range skew kurtosis
## X1    1 248 10.94 5.18   9.74   10.62 4.41   0 30.95 30.95 0.74     0.61
##      se
## X1 0.33

#Here in actual_prod mean is roughly equal to median and sd is very small
compared to the mean Hence actual_prod seems to have  normally distributed.
Lets check with the plots

hist(emp_performance$actual_prod,probability=T,col=c("steelblue", "red"))
lines(density(emp_performance$actual_prod),col="black",lwd=3)
```




```r
#The histogram shows that the it is right skew

boxplot(emp_performance$actual_prod)
```



```r
#Boxplot shows that there are some outliers in the variable accuracy.

#Hence the actual_prod is normaly distributed with some outliers in the data.

#68 Test
mean(emp_performance$actual_prod)+sd(emp_performance$actual_prod)

## [1] 16.12524

mean(emp_performance$actual_prod)-sd(emp_performance$actual_prod)

## [1] 5.761088

nrow(emp_performance[which(emp_performance$actual_prod <
mean(emp_performance$actual_prod)+sd(emp_performance$actual_prod) &
emp_performance$actual_prod > mean(emp_performance$actual_prod)-
sd(emp_performance$actual_prod)),]) /nrow(emp_performance)

## [1] 0.6975806

#Above test satisfy as approximately 68 percent of data lies within 1 sd from
the mean.

#95 test
nrow(emp_performance[which(emp_performance$actual_prod <
mean(emp_performance$actual_prod)+2*sd(emp_performance$actual_prod)  &
emp_performance$actual_prod >  mean(emp_performance$actual_prod)-
2*sd(emp_performance$actual_prod) ),]) /nrow(emp_performance)

## [1] 0.9516129

#Above test satisfy as approximately 95 percent of data lies within 2 sd from
the mean.

#99.7 test
nrow(emp_performance[which(emp_performance$actual_prod <
mean(emp_performance$actual_prod)+3*sd(emp_performance$actual_prod) &
emp_performance$actual_prod > mean(emp_performance$actual_prod)-
3*sd(emp_performance$actual_prod)),]) /nrow(emp_performance)

## [1] 0.9959677

#Above test satisfy as approximately 99.7 percent of data lies within 3 sd
from the mean.

#Since all aove test roughly satisfy actual_prod  is roughly normally
distributed with some outliers.


#3.Variable name-target_prod
```

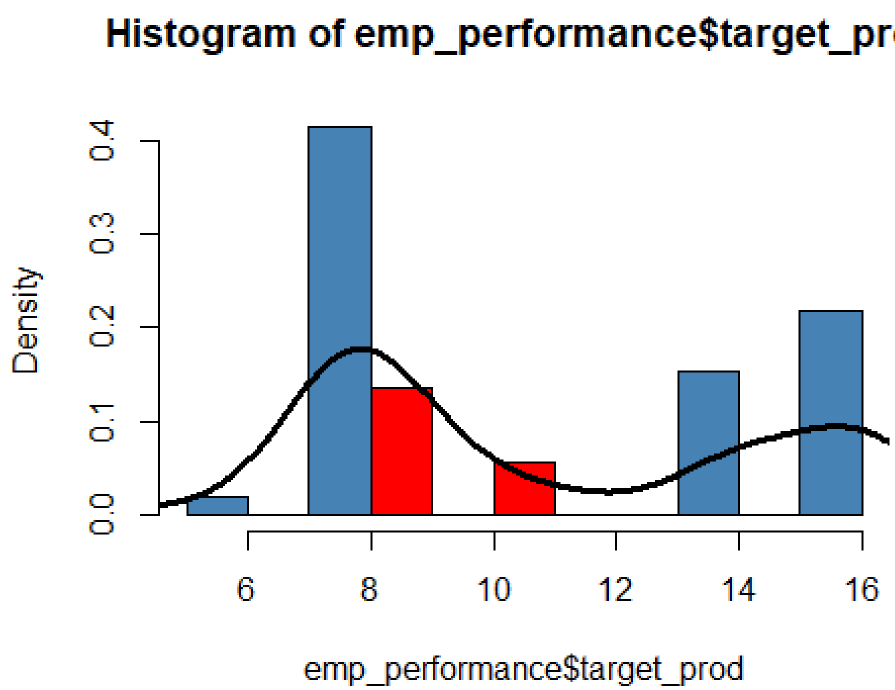
```r
library(psych)
summary(emp_performance$target_prod)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    5.00    7.50    9.00   10.75   14.00   16.00
```

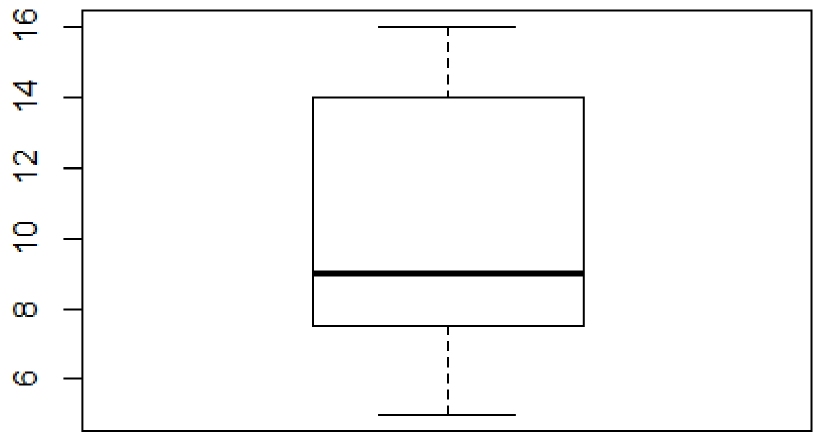
```r
describe(emp_performance$target_prod)
```

```
##    vars   n  mean   sd median trimmed  mad min max range skew kurtosis
## X1    1 248 10.75 3.59      9   10.57 2.22   5  16    11 0.42    -1.51
##      se
## X1 0.23
```

```r
#Here in target_prod mean is greater than the median. Hence target_prod seems 
to not have  normally distributed. Lets check with the plots
hist(emp_performance$target_prod,probability=T,col=c("steelblue", "red"))
lines(density(emp_performance$target_prod),col="black",lwd=3)
```




```r
#The histogram shows that the it is not normally distributed

boxplot(emp_performance$target_prod)
```




```
#Boxplot shows that there is left skew in the data

#4.Variable name-season
# Examine the data
stab<- table(emp_performance$season)
addmargins(stab)

##
## Summer Winter Spring   Fall    Sum
##     31    102     45     70    248

ptab<-prop.table(stab)
round(ptab,2)

##
## Summer Winter Spring   Fall
##   0.12   0.41   0.18   0.28

addmargins(round(ptab,2))

##
## Summer Winter Spring   Fall    Sum
##   0.12   0.41   0.18   0.28   0.99

#It shows that out of entire data we have maximum data for the season Fall
and leaset data for summer season
```

```
# Plot the data
dev.off()

## null device
##           1

barplot(ptab, col=c("orange", "gray","blue","red"))
legend("topright", c("summer", "winter","spring","Fall"), lty=1,lwd=4,
col=c("orange", "gray","blue","red"), cex=0.7)


#4.Variable name-type
# Examine the data
stab<- table(emp_performance$type)
addmargins(stab)

##
##  Local Remote    Sum
##    163     85    248

ptab<-prop.table(stab)
round(ptab,2)

##
##  Local Remote
##   0.66   0.34

addmargins(round(ptab,2))

##
##  Local Remote    Sum
##   0.66   0.34   1.00

#It shows that out of entire data we have 66% data for type Local and 34%
data for type Remote

# Plot the data
dev.off()

## null device
##           1

barplot(ptab, col=c("orange", "gray"))
legend("topright", c("Local", "Remote"), lty=1,lwd=4, col=c("orange",
"gray"), cex=0.7)


#4.Variable name-work_area
# Examine the data
stab<- table(emp_performance$work_area)
addmargins(stab)
```

```
##
## Spanish English   Other     Sum
##      54      79     115     248
```

```r
ptab<-prop.table(stab)
round(ptab,2)
```

```
##
## Spanish English   Other
##    0.22    0.32    0.46
```

```r
addmargins(round(ptab,2))
```

```
##
## Spanish English   Other     Sum
##    0.22    0.32    0.46    1.00
```

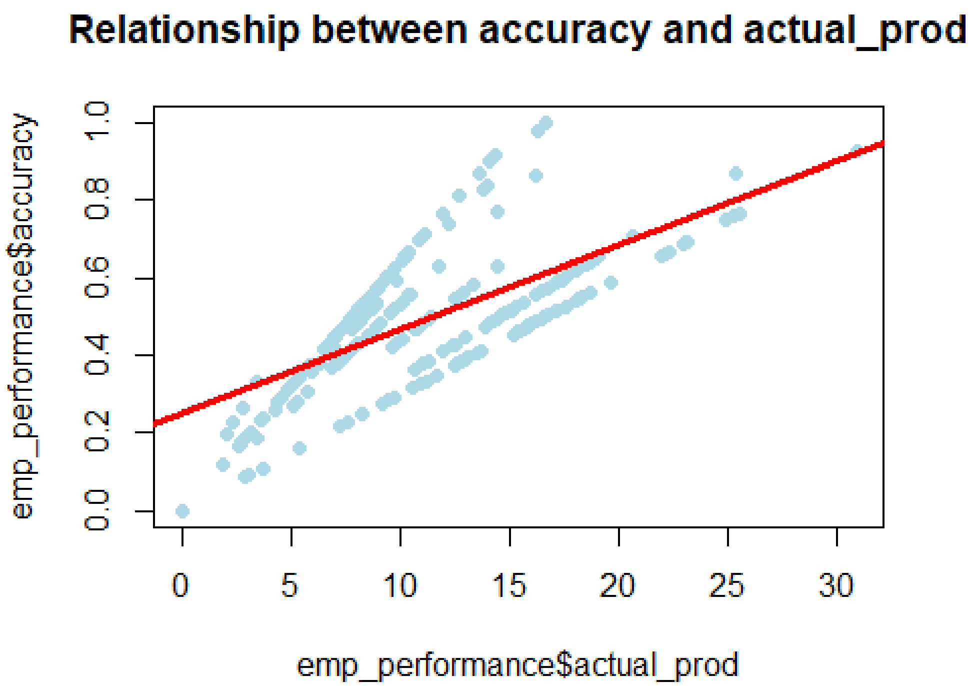
```r
#It shows that out of entire data we have 22% data for workarea using Spanish
and 32% data for workarea using  English and 46% data have workarea using
Other language

# Plot the data
dev.off()
```

```
## null device
##           1
```

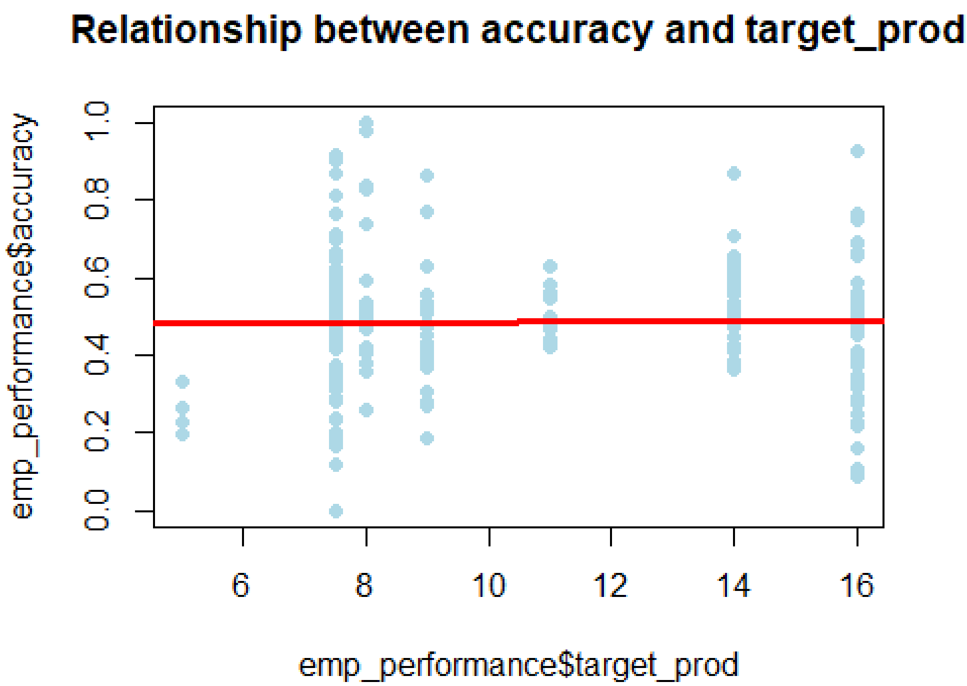
```r
barplot(ptab, col=c("orange", "gray","blue"))
legend("topright", c("Spanish", "English","Other"), lty=1,lwd=4,
col=c("orange", "gray","blue"), cex=0.7)
```

Q.2) Analyze the relatioship of all the numeric variables in the dataset with Accuracy

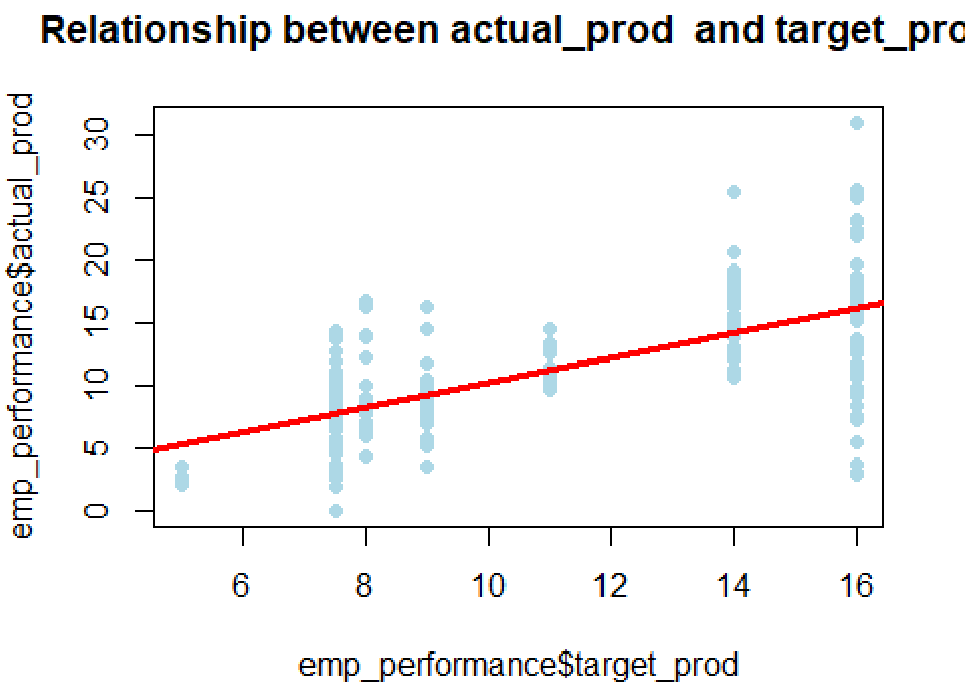
```r
#Examining Relation between accuracy and actual_prod
plot(emp_performance$accuracy~emp_performance$actual_prod, pch=16,
col="lightblue", main="Relationship between accuracy and actual_prod")
abline(lm(emp_performance$accuracy~emp_performance$actual_prod), lwd=3,
col="red")
```



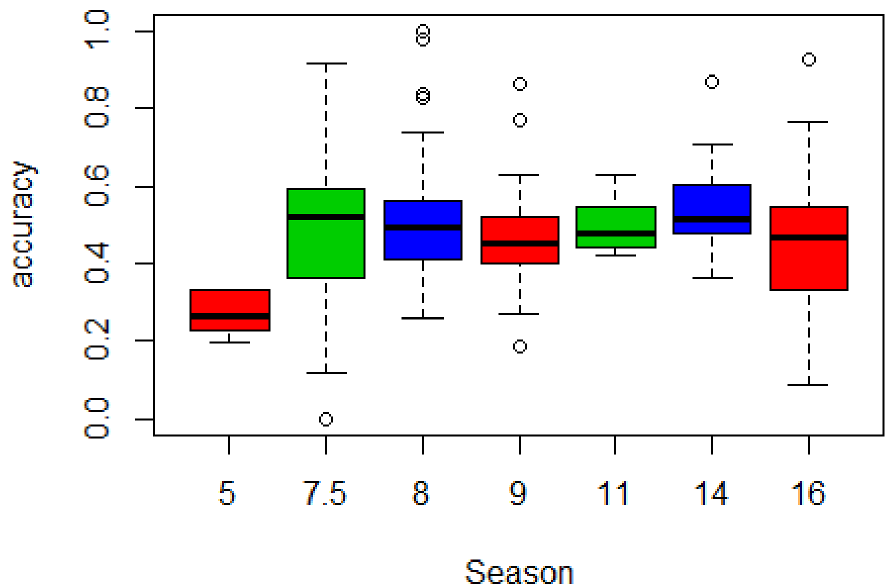

```
# Looks like there is a strong positive correlation between accuracy and
actual_prod

# Now, let's get a numeric value for the correlation
cor(emp_performance$accuracy,emp_performance$actual_prod, use="complete.obs")

## [1] 0.6883099

#It shows that correlation coefficient is 0.688. So there is positive
correlation between accuracy and actual_prod


#Examining Relation between accuracy and target_prod
plot(emp_performance$accuracy~emp_performance$target_prod, pch=16,
col="lightblue", main="Relationship between accuracy and target_prod")
abline(lm(emp_performance$accuracy~emp_performance$target_prod), lwd=3,
col="red")
```




```
# Looks like there is no correlation between accuracy and target_prod

# Now, let's get a numeric value for the correlation
cor(emp_performance$accuracy,emp_performance$target_prod, use="complete.obs")

## [1] 0.01022798

#It shows that correlation coefficient is 0.01022798 which is very less So
there is no correlation between accuracy and target_prod


#Examining Relation between actual_prod and target_prod
plot(emp_performance$actual_prod~emp_performance$target_prod, pch=16,
col="lightblue", main="Relationship between actual_prod  and target_prod")
abline(lm(emp_performance$actual_prod~emp_performance$target_prod), lwd=3,
col="red")
```

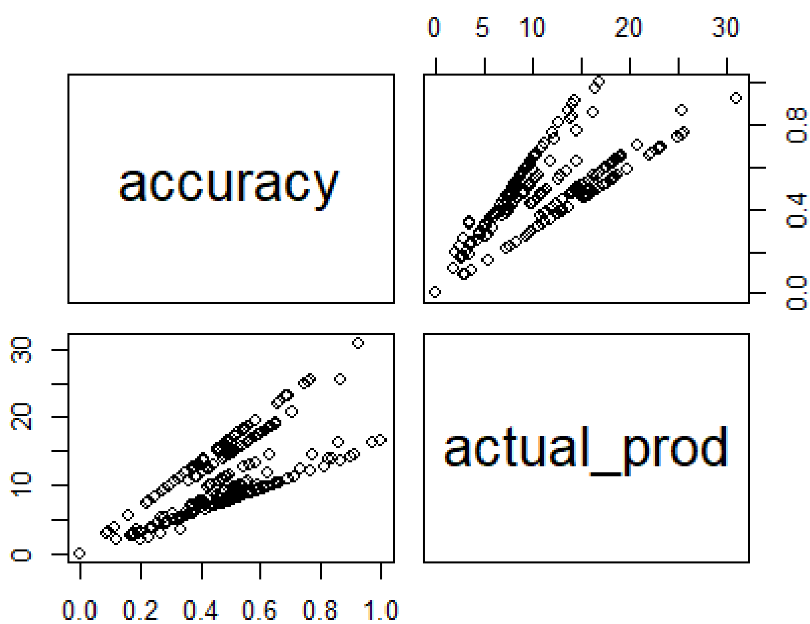



```
# Looks like there is no correlation between actual_prod and target_prod

#The plot shows that target_prod should be a factor variable with different
levels and not the numeric variable

str(emp_performance$target_prod)

##  num [1:248] 7.5 16 7.5 8 14 16 7.5 7.5 8 7.5 ...

emp_performance$target_prod <- as.factor(emp_performance$target_prod)
str(emp_performance$target_prod)

##  Factor w/ 7 levels "5","7.5","8",..: 2 7 2 3 6 7 2 2 3 2 ...

boxplot(accuracy~target_prod, data=emp_performance, col=2:4, xlab="Season")
```

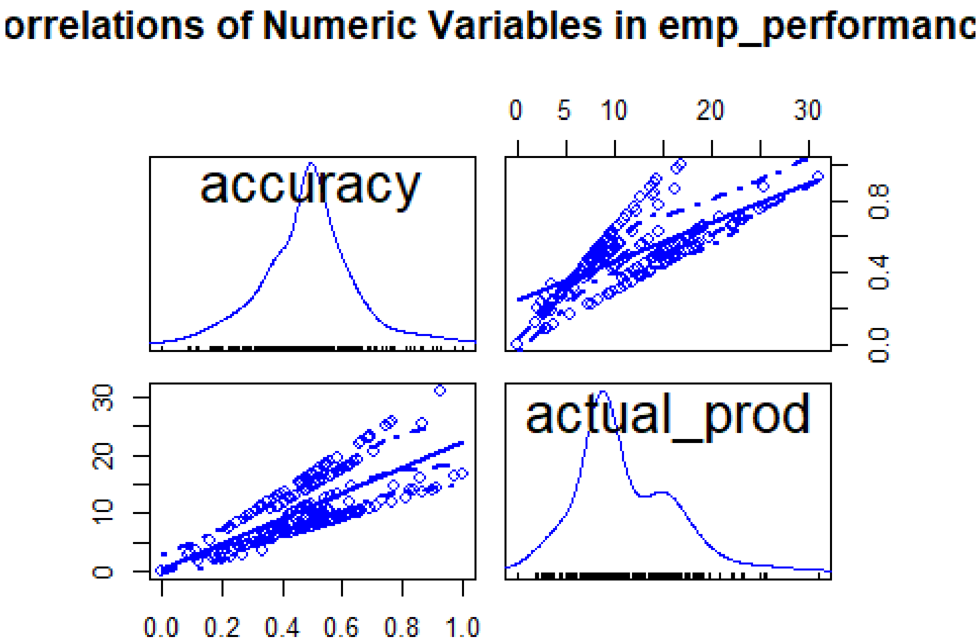



```
#Ho-there is no difference in levels of target_prod
#Ha-there is difference in levels of target_prod
emp_accuracy_target_prod.aov <- aov(accuracy~target_prod,
data=emp_performance)
emp_accuracy_target_prod.aov

## Call:
##    aov(formula = accuracy ~ target_prod, data = emp_performance)
##
## Terms:
##                 target_prod Residuals
## Sum of Squares     0.480008  6.113920
## Deg. of Freedom           6       241
##
## Residual standard error: 0.1592764
## Estimated effects may be unbalanced

summary(emp_accuracy_target_prod.aov)

##              Df Sum Sq Mean Sq F value  Pr(>F)
## target_prod   6  0.480 0.08000   3.154 0.00538 **
## Residuals   241  6.114 0.02537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Since p-value is 0.00538 which is less than 0.05 (at 95% confidence
interval) we reject the NULL hypothesis.
```

```r
#Hence there is difference in levels of target_prod i.e. There is statistical significance between accuracy and target_prod

#Correlation Matrix
emp_performance_1 <- emp_performance[,c("accuracy", "actual_prod")]
cormat <- cor(emp_performance_1)
round(cormat, 2)

##             accuracy actual_prod
## accuracy        1.00        0.69
## actual_prod     0.69        1.00

# scatterplots
pairs(emp_performance_1)
```




```r
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:psych':
##
##     logit

scatterplotMatrix(~accuracy+actual_prod, data=emp_performance_1,
main="Correlations of Numeric Variables in emp_performance")
```




```
#So only accuracy and actual_prod have significant positive correlation.
#And hence actual_prod is related to accuracy

#Important finding-target_prod should be a factor variable with different
levels and not the numeric variable
```

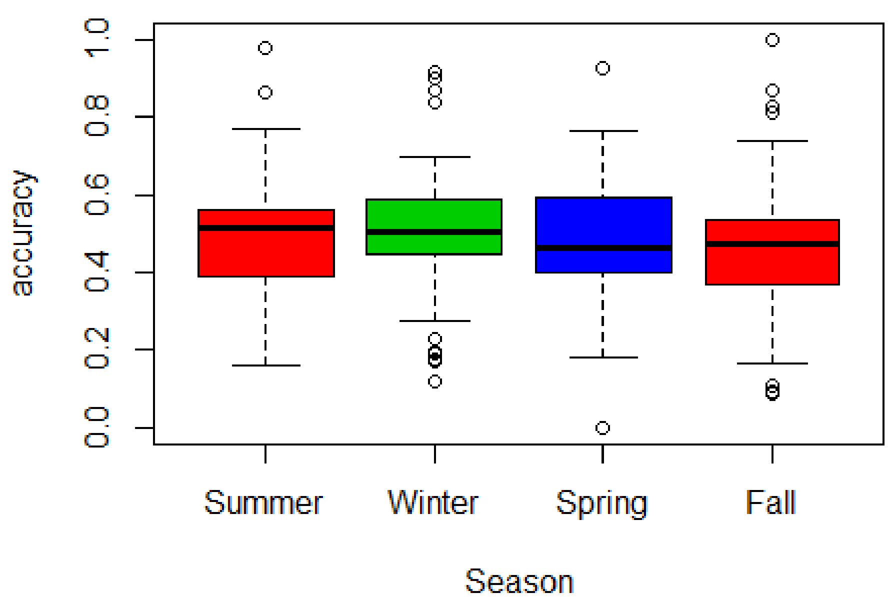
Q.3) Analyze the relationship of all the categorical variables in the dataset with the variable accuracy

```
#Relationship between accuracy and season

#Bivariate Analysis between accuracy and season
library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

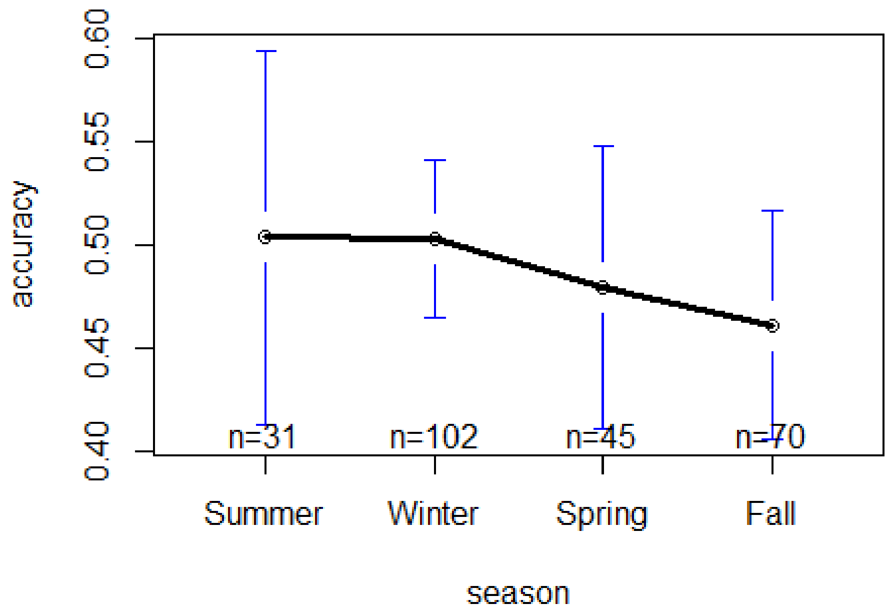
```r
emp_performance %>% group_by(season) %>% summarise(avg = mean(accuracy),
med=median(accuracy),std = sd(accuracy))
```

```
## # A tibble: 4 x 4
##   season   avg   med   std
##   <fct>  <dbl> <dbl> <dbl>
## 1 Summer 0.504 0.517 0.183
## 2 Winter 0.503 0.503 0.146
## 3 Spring 0.480 0.464 0.170
## 4 Fall   0.461 0.476 0.174
```

```r
# There is no considerable differences between average of summer and
winter.Average accuracy for summer is 0.504 and for winter is 0.503
#Also there is not much difference between average accuracy for Spring and
Fall.


# Visualizing the data for relation between accuracy and season
boxplot(accuracy~season, data=emp_performance, col=2:4, xlab="Season")
```

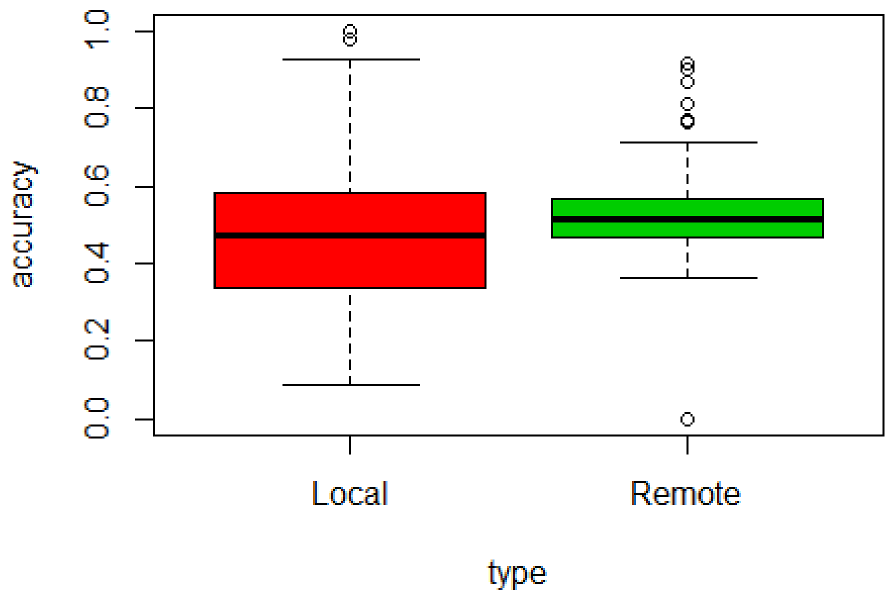



```r
#It shows that median accuracy for summer is highest followed by Winter, Fall
and least for spring season

# checking confidence interval with Plots means
gplots::plotmeans(emp_performance$accuracy~emp_performance$season,
xlab="season", ylab="accuracy", lwd=3, col="black", p=0.99)
```

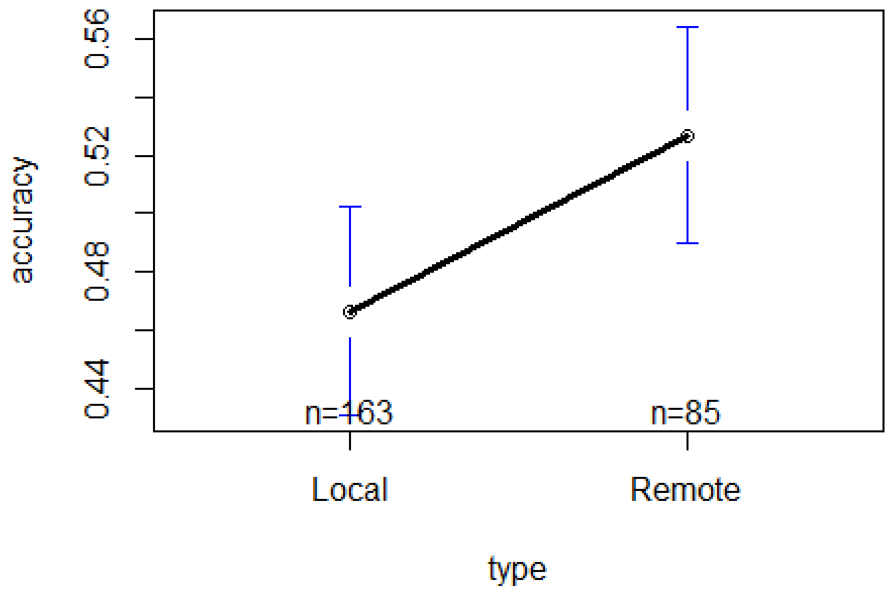



```
#plotmeans shows that there may not be stistical difference among all
seasons.

#Now since the independent variable is factor and Dependent variable is
numeric we can use ANOVA or linear regression.
#But advantage of using the ANOVA is Tukeyplot's post hoc test provides
information about the difference among the different levels of season.

#Ho-there is no difference in levels of season
#Ha-there is difference in levels of season
emp_accuracy_season.aov <- aov(accuracy~season, data=emp_performance)
emp_accuracy_season.aov

## Call:
##    aov(formula = accuracy ~ season, data = emp_performance)
##
## Terms:
##                      season Residuals
## Sum of Squares  0.082899  6.511029
## Deg. of Freedom        3       244
##
## Residual standard error: 0.1633541
## Estimated effects may be unbalanced

summary(emp_accuracy_season.aov)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## season         3  0.083 0.02763   1.036  0.377
## Residuals    244  6.511 0.02669

#Since p-value is 0.377 which is greater than 0.05 (at 95% confidence
interval) we cannot reject the NULL hypothesis.
#Hence there is no difference in levels of season i.e. There is no
statistical significance between accuracy and season



#Relationship between accuracy and type

#Bivariate Analysis between accuracy and type
library(dplyr)
emp_performance %>% group_by(type) %>% summarise(avg = mean(accuracy),
med=median(accuracy),std = sd(accuracy))

## # A tibble: 2 x 4
##   type     avg   med   std
##   <fct>  <dbl> <dbl> <dbl>
## 1 Local  0.466 0.472 0.175
## 2 Remote 0.527 0.516 0.131

# There is considerable differences between average of Local and Remote
employees. accuracy for Local employees is 0.466 and for Remote employees is
0.527

# Visualizing the data for relation between accuracy and type
boxplot(accuracy~type, data=emp_performance, col=2:4, xlab="type")
```




```
#It shows that median accuracy for remote employees is higher than Local
employees. Also IQR for Local emp is more than Remote emp.

# checking confidence interval with Plots means
gplots::plotmeans(emp_performance$accuracy~emp_performance$type, xlab="type",
ylab="accuracy", lwd=3, col="black", p=0.99)
```



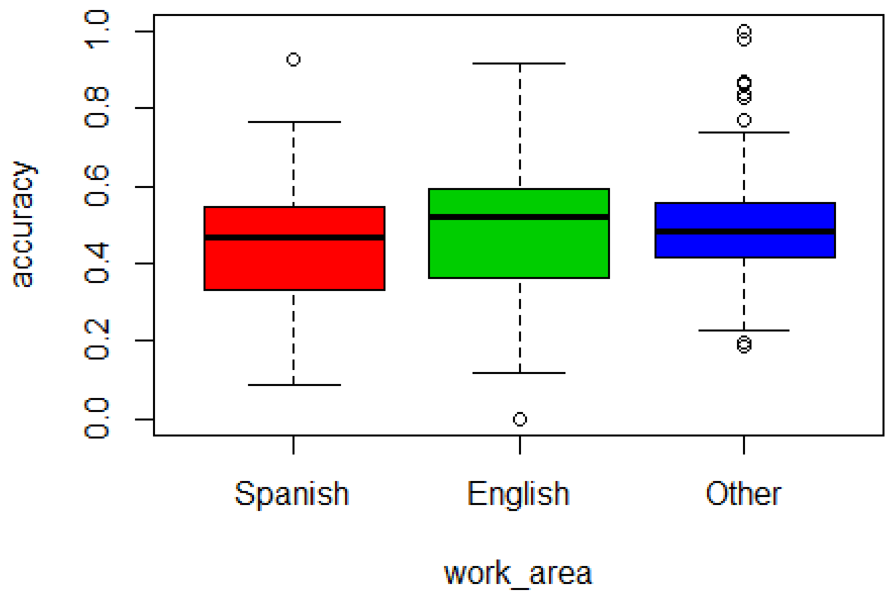

```
#plotmeans shows that there may be stistical difference among Local and
Remote employees.

#Now since the independent variable is factor and Dependent variable is
numeric we can use ANOVA or linear regression.
#But advantage of using the ANOVA is Tukeyplot's post hoc test provides
information about the difference among the different levels of season.

#Ho-there is no difference in levels of type (local employees and remote
employees)
#Ha-there is difference in levels of type((local and remote))
emp_accuracy_type.aov <- aov(accuracy~type, data=emp_performance)
emp_accuracy_type.aov

## Call:
##    aov(formula = accuracy ~ type, data = emp_performance)
##
## Terms:
##                     type Residuals
## Sum of Squares  0.205162  6.388766
## Deg. of Freedom        1       246
##
## Residual standard error: 0.1611539
## Estimated effects may be unbalanced

summary(emp_accuracy_type.aov)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)   
## type          1  0.205 0.20516     7.9 0.00534 **
## Residuals   246  6.389 0.02597                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

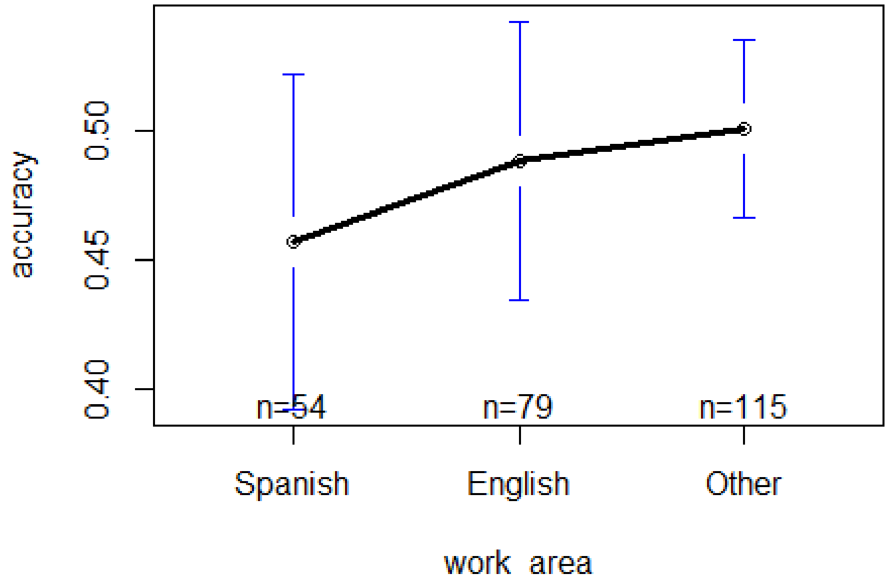
```r
#Since p-value is 0.00534 which is less than 0.05 (at 95% confidence interval) we reject the NULL hypothesis.
#Hence there is difference in levels of type (local employees and remote employees)i.e. There is statistical significance between accuracy and type


#Relationship between accuracy and work_area

#Bivariate Analysis between accuracy and work_area
library(dplyr)
emp_performance %>% group_by(work_area) %>% summarise(avg = mean(accuracy), med=median(accuracy),std = sd(accuracy))
```

```
## # A tibble: 3 x 4
##   work_area   avg   med   std
##   <fct>     <dbl> <dbl> <dbl>
## 1 Spanish   0.457 0.467 0.179
## 2 English   0.488 0.519 0.181
## 3 Other     0.500 0.484 0.141
```

```r
# There is considerable differences between average of levels of work_area. Average accuracy for wowork_area with Spanish is 0.457 , average accuracyfor English  is 0.488 and for Other is 0.5

# Visualizing the data for relation between accuracy and work_area
boxplot(accuracy~work_area, data=emp_performance, col=2:4, xlab="work_area")
```




```
#It shows that median accuracy for English is highest followed by Other and
then Spanish. Also IQR for other is least

# checking confidence interval with Plots means
gplots::plotmeans(emp_performance$accuracy~emp_performance$work_area,
xlab="work_area", ylab="accuracy", lwd=3, col="black", p=0.99)
```




```
#plotmeans shows that there may not be stistical difference among Local and
Remote employees.

#Now since the independent variable is factor and Dependent variable is
numeric we can use ANOVA or linear regression.
#But advantage of using the ANOVA is Tukeyplot's post hoc test provides
information about the difference among the different levels of season.

#Ho-there is no difference in levels of work_area
#Ha-there is difference in levels of work_area
emp_accuracy_workarea.aov <- aov(accuracy~work_area, data=emp_performance)
emp_accuracy_workarea.aov

## Call:
##    aov(formula = accuracy ~ work_area, data = emp_performance)
##
## Terms:
##                    work_area Residuals
## Sum of Squares   0.068796  6.525133
## Deg. of Freedom         2       245
##
## Residual standard error: 0.1631968
## Estimated effects may be unbalanced

summary(emp_accuracy_workarea.aov)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## work_area     2  0.069 0.03440   1.292  0.277
## Residuals   245  6.525 0.02663

#Since p-value is 0.277 which is greater than 0.05 (at 95% confidence interval) we cannot reject the NULL hypothesis.
#Hence there is no difference in levels of work_area (Spanish,English and Other)i.e. There is no statistical significance between accuracy and work_area
```