Statistical Programming 600 Final

Objective Statement:

Previously, we were given data on seeding rate and yield and asked to calculate required replicates for future testing in order to have an adequately powered test. Now, we are given data centered around seeding and harvest focusing on variables of time, date, moisture, and yield. Other variables are longitude and lattitude as well as distance.

First, we want to see what the range of planting dates were for the fields. Second, we want to see if there are any large time gaps in harvest time. This second objective looks at if it took more than a day to a harvest a given field and if there were any hour or longer gaps in harvest data acquisition.

Next, we want to see if the effect size of planting dates is large enough to consider it as a confounding variable. This will let us know if future experiments need to have harvesting constrained to tighter dates. We also want a fist approximation of the number of fields necessary to observe an effect of date planted on yield.

We also want to know how moisture changes over the course of a day and if this affects the recorded yield.

Data Manipulation:

Reading in filepaths as a list will allow us to scale up the amount of files we can work on at a time if we decide to add more files later.

#This will read in the files as dataframes in a list  
filenamesSeed <- list.files(path = "C://data/StatProgFinal/", pattern = "Finals", full.names = TRUE)  
filenamesHarvest <- list.files(path = "C://data/StatProgFinal/", pattern = "Finalh", full.names = TRUE)  
filenamesSeed

## [1] "C://data/StatProgFinal/FinalseedA.csv"  
## [2] "C://data/StatProgFinal/FinalseedB.csv"  
## [3] "C://data/StatProgFinal/FinalseedC.csv"  
## [4] "C://data/StatProgFinal/FinalseedD.csv"

filenamesHarvest

## [1] "C://data/StatProgFinal/FinalharvestA.csv"  
## [2] "C://data/StatProgFinal/FinalharvestB.csv"  
## [3] "C://data/StatProgFinal/FinalharvestC.csv"  
## [4] "C://data/StatProgFinal/FinalharvestD.csv"

Now we use read.csv to read in the list of paths. This will result in a list of dataframes. Here we have one list for seeding data and one list for harvest data.

Seed.Data <- lapply(filenamesSeed, read.csv)  
Harvest.Data <- lapply(filenamesHarvest, read.csv)

Using lapply allows us to run a function over a list of dataframes. Here we manipulate data to separate Date and Time from the Timestamp.

# This function will split the timestamp into Date & Time  
Harvest.Data <- lapply(Harvest.Data, function(x){x$Date = sub("T.\*", "", x$Timestamp)  
x$Date = as.Date(x$Date)  
DateNum = as.numeric(x$Date)  
#DateInt = as.factor(x$Date)  
#DateInt = as.integer(DateInt)  
x$Time = sub(".\*T", "", x$Timestamp)  
x$Time = sub(":.\*", "", x$Time)  
TimeNum = gsub(":", "", x$Time)  
TimeNum = as.numeric(TimeNum)  
x <- data.frame(cbind(x, x$Date, DateNum, TimeNum), stringsAsFactors = F)  
} )   
  
Seed.Data <- lapply(Seed.Data, function(x){x$Date = sub("T.\*", "", x$Timestamp)  
x$Date = as.Date(x$Date)  
DateNum = as.numeric(x$Date)  
x$Time = sub(".\*T", "", x$Timestamp)  
x$Time = sub("\\..\*", "", x$Time)  
TimeNum = gsub(":", "", x$Time)  
TimeNum = as.numeric(TimeNum)  
ControlRate = as.integer(x$ControlRate)  
x <- data.frame(cbind(x, x$Date, DateNum, TimeNum), stringsAsFactors = F)  
} )

Here I will specifically keep Date and Time together from Timestamp, but transform in into a more workable format for graphing, again using lapply to make this more reproducible and scalable.

Harvest <- lapply(Harvest.Data, function(x){x$Date = sub("T", " ", x$Timestamp)  
x$Date = sub("Z", " ", x$Date)  
x$Date = sub("\\..\*", "", x$Date)  
x$Date = as.POSIXct(x$Date)  
x$Time = sub(".\*T", "", x$Timestamp)  
x$Time = sub("\\..\*", "", x$Time)  
TimeNum = gsub(":", "", x$Time)  
TimeNum = as.numeric(TimeNum)  
x <- data.frame(cbind(x, x$Date, TimeNum), stringsAsFactors = F)  
} )

Data Analysis:

First, we want to see what the range of planting dates were for the fields. Below, we see that field B and field C were not planted in one day and we also see that there was a 5 day period over which the fields were planted. We will look more closely at the effect of seeding date on yield later.

lapply(Seed.Data, function(x) {levels(as.factor(x$Date))})

## [[1]]  
## [1] "2018-05-17"  
##   
## [[2]]  
## [1] "2018-05-20" "2018-05-21"  
##   
## [[3]]  
## [1] "2018-05-18" "2018-05-19"  
##   
## [[4]]  
## [1] "2018-05-20"

Moving forward we are going to look at yield. I will subset the data keeping only points with desired increment as on the midterm, but now with more scalable code using lapply.

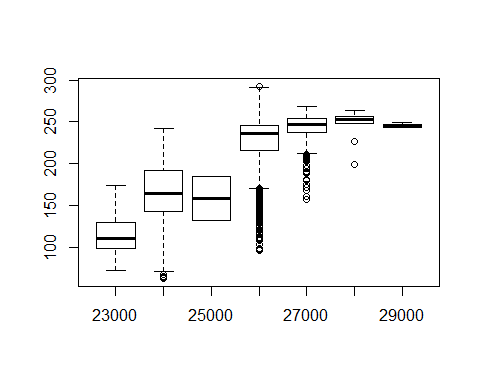
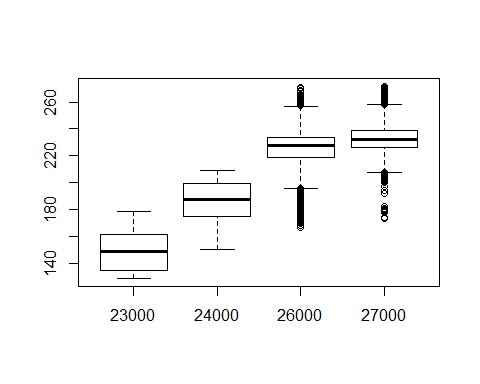
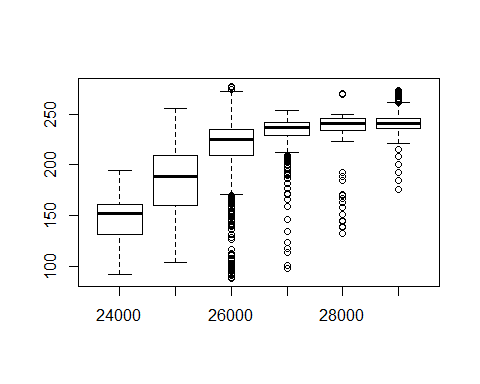
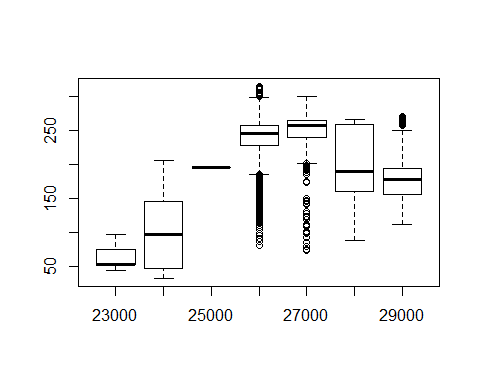
new <- lapply(Seed.Data, function(x){  
merged <- subset(x, ControlRate ==23000 | ControlRate == 24000 | ControlRate == 25000 | ControlRate == 26000 | ControlRate == 27000 | ControlRate == 28000 | ControlRate == 29000)})

Combine the desired data from the list into a new dataframe.

new <- lapply(new, function(x) {y <- data.frame(x$Yield, x$DateNum, x$ControlRate)})

First, lets check the effect of seeding rate on yield regardless of planting date. Similar to my midterm I will select 2700 as the best control rate and use this to see the effect of seeding date on yield.

lapply((new), function(y) { boxplot(y$x.Yield ~ y$x.ControlRate)})

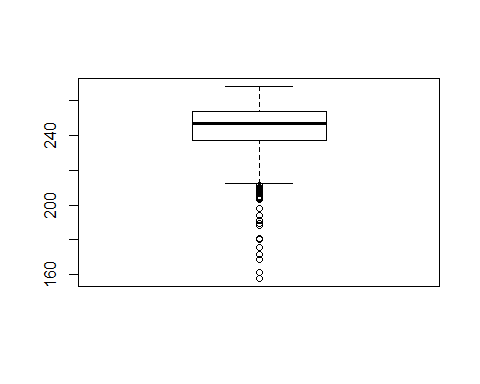
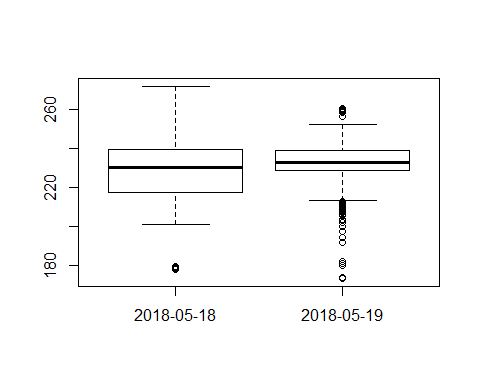
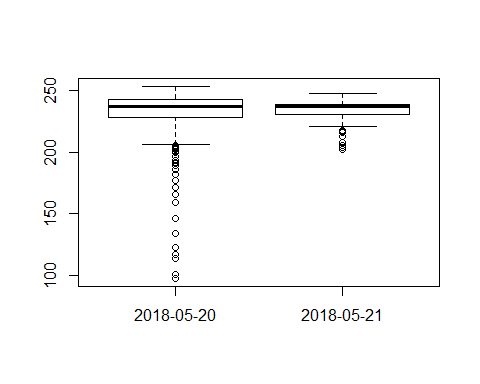
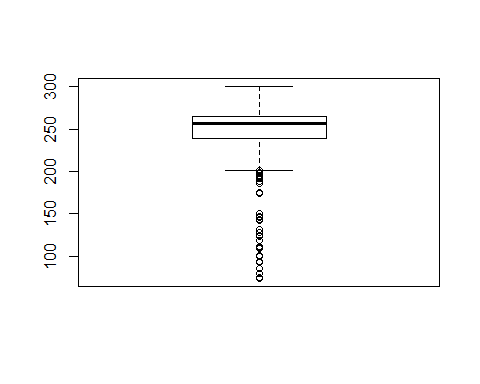


Picking just the seeding data with 27000 control rate to look at effects of seeding date.

new2 <- lapply(Seed.Data, function(x){  
merged <- subset(x, ControlRate == 27000)})

What is the effect of seeding date on yield? Here we see that there is more variation between fields than there is within fields due to seeding date, showing that seeding date has a relatively small effect.

lapply((new2), function(y) { boxplot(y$Yield ~ y$Date)})



Moving on to Harvest…

Let’s check on the range of harvest dates. It appears there is much larger range of harvest dates, up to about half a month in field A and over a month in field D.

lapply(Harvest.Data, function(x) {levels(as.factor(x$Date))})

## [[1]]  
## [1] "2018-10-27" "2018-10-28" "2018-11-15"  
##   
## [[2]]  
## [1] "2018-11-02"  
##   
## [[3]]  
## [1] "2018-11-11" "2018-11-12"  
##   
## [[4]]  
## [1] "2018-10-01" "2018-11-12" "2018-11-13"

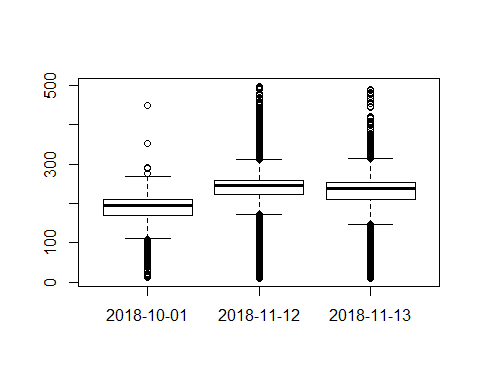
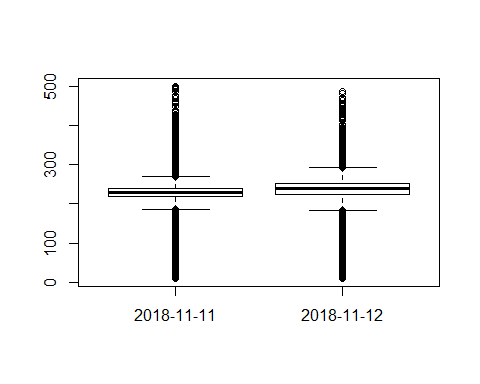
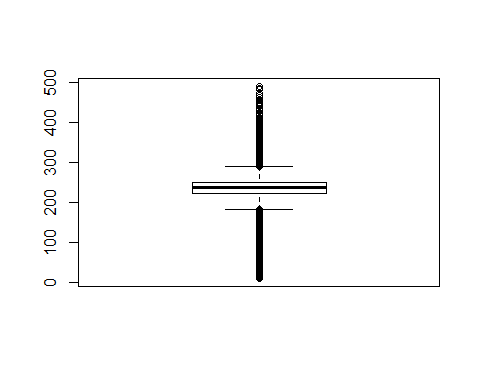
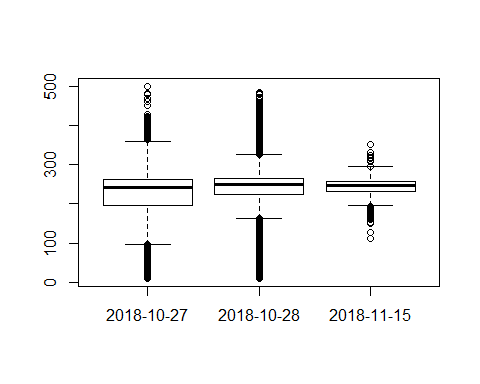
How about gaps in time? When looking at the hour level it appears that there aren’t any gaps in time longer than 2 hours.

lapply(Harvest.Data, function(x) {levels(as.factor(x$Time))})

## [[1]]  
## [1] "16" "17" "18" "19" "20" "21" "22" "23"  
##   
## [[2]]  
## [1] "15" "16" "17" "18" "19" "20" "21" "22" "23"  
##   
## [[3]]  
## [1] "00" "01" "15" "16" "17" "18" "19" "20" "21" "22" "23"  
##   
## [[4]]  
## [1] "15" "16" "17" "18" "19" "20" "21" "22" "23"

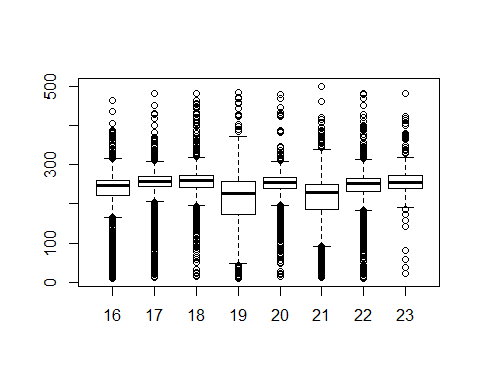
What is the effect of Harvest date on yield? Here we see that most fields were showed little variation in yield when harvested up to a month apart, however field D did show some visually significant variation in yield between harvest times over a month apart.

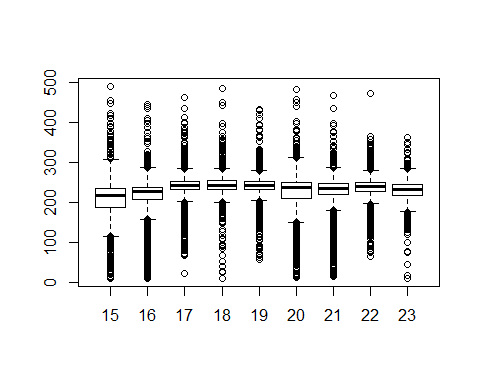
lapply((Harvest.Data), function(y) { boxplot(y$VRYIELDVOL ~ y$Date)})

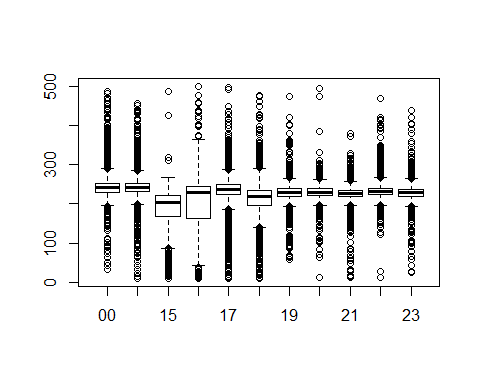


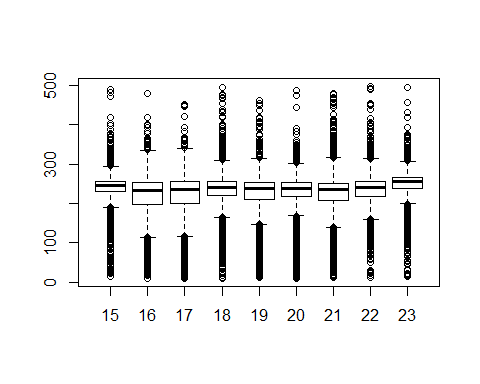
Does time of day have a large effect on harvest yield?

lapply(Harvest.Data, function(y) {x = boxplot(y$VRYIELDVOL ~ y$Time)  
print(x)  
})







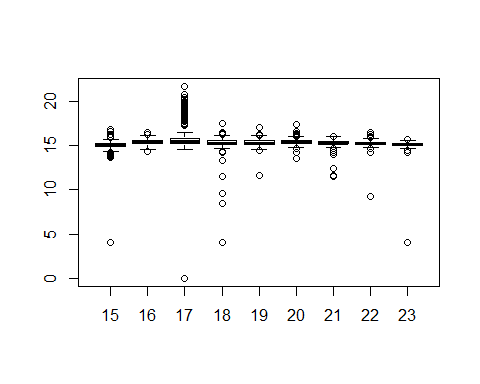
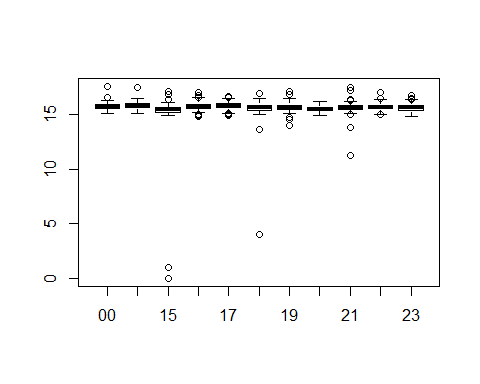
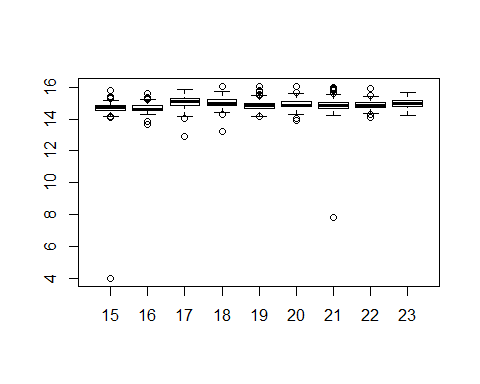
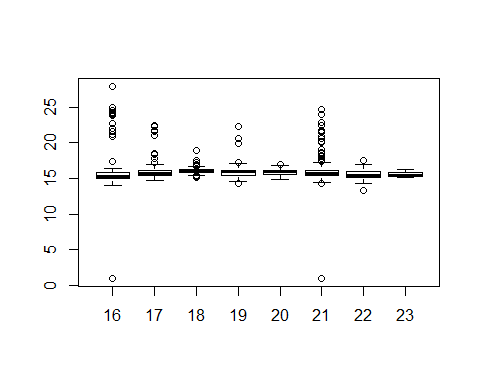


Next, we want to see if the effect size of planting dates is large enough to consider it as a confounding variable.

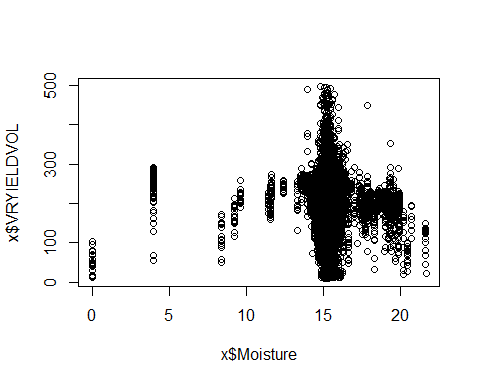
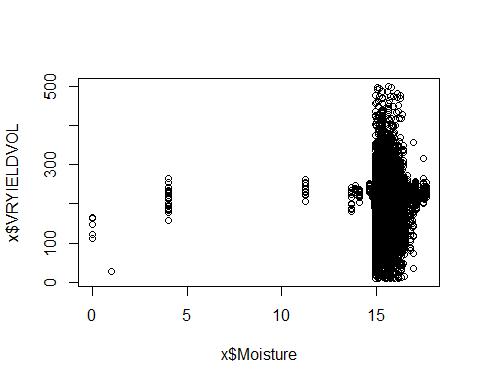
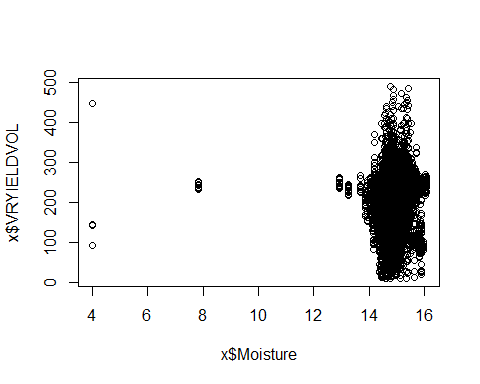
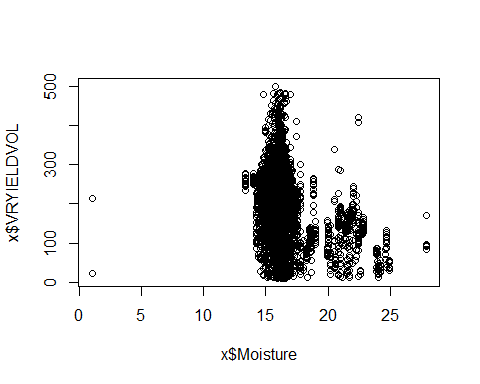
This will let us know if future experiments need to have harvesting constrained to tighter dates. We also want a fist approximation of the number of fields necessary to observe an effect of date planted on yield.

Finally we wanted to know how moisture changes over the course of a day during harvest and if this affects the recorded yield.

lapply(Harvest.Data, function(x) {boxplot(x$Moisture ~ x$Time)})



lapply(Harvest.Data, function(x) {plot(x$VRYIELDVOL ~ x$Moisture)})



And finally, what is the effect size of and replicates needed for seeding date and yield.

Effect size is found as it was previously described by the professor.

We know May 17 to may 21st was the planting period, 5 days total, which we will us to say %Diff is approximately 5% since it is 100 day corn. We will asssume SD is similar to that of the midterm and use this to make CV. We will then find effect size and required reps for planting date and yield.

We will find this using the equation previously provided by the professor.

Standard Deviation Pooled from the midterm.

sdpooled <- (sqrt(43\*\*2 + 35\*\*2 + 19\*\*2 + 8\*\*2 + 6\*\*2 + 24\*\*2 + 33\*\*2))  
sdpooled

## [1] 72.11103

and the average yield was 250 so

CV <- (sdpooled)/((250+250)/2)  
Requiredrep <- 2 \* ((CV/0.05)\*\*2) \* ((((1-(0.05/2))+(1-0.2)))\*\*2)  
Requiredrep

## [1] 209.7056

So a lot of reps (~200) would be required to find a significant difference between in yield when planting date with a range of with a range of 5 days is considered as the independent variable.