Group Project

Group 11

19/12/2019

# Importing the dataset———————————————————————————————————————————————–

d1 <- read.csv("C:/Users/souvi/Documents/R/IM507/apy.csv")  
  
View(d1)  
head(d1)

## State\_Name District\_Name Crop\_Year Season  
## 1 Andaman and Nicobar Islands NICOBARS 2000 Kharif   
## 2 Andaman and Nicobar Islands NICOBARS 2000 Kharif   
## 3 Andaman and Nicobar Islands NICOBARS 2000 Kharif   
## 4 Andaman and Nicobar Islands NICOBARS 2000 Whole Year   
## 5 Andaman and Nicobar Islands NICOBARS 2000 Whole Year   
## 6 Andaman and Nicobar Islands NICOBARS 2000 Whole Year   
## Crop Area Production  
## 1 Arecanut 1254 2000  
## 2 Other Kharif pulses 2 1  
## 3 Rice 102 321  
## 4 Banana 176 641  
## 5 Cashewnut 720 165  
## 6 Coconut 18168 65100000

dim(d1)

## [1] 246091 7

names(d1)

## [1] "State\_Name" "District\_Name" "Crop\_Year" "Season"   
## [5] "Crop" "Area" "Production"

summary(d1)

## State\_Name District\_Name Crop\_Year   
## Uttar Pradesh : 33306 BIJAPUR : 945 Min. :1997   
## Madhya Pradesh: 22943 TUMKUR : 936 1st Qu.:2002   
## Karnataka : 21122 BELGAUM : 925 Median :2006   
## Bihar : 18885 HASSAN : 895 Mean :2006   
## Assam : 14628 BELLARY : 887 3rd Qu.:2010   
## Odisha : 13575 DAVANGERE: 886 Max. :2015   
## (Other) :121632 (Other) :240617   
## Season Crop Area   
## Autumn : 4949 Rice : 15104 Min. : 0   
## Kharif :95951 Maize : 13947 1st Qu.: 80   
## Rabi :66987 Moong(Green Gram): 10318 Median : 582   
## Summer :14841 Urad : 9850 Mean : 12003   
## Whole Year :57305 Sesamum : 9046 3rd Qu.: 4392   
## Winter : 6058 Groundnut : 8834 Max. :8580100   
## (Other) :178992   
## Production   
## 1 : 4028   
## = : 3727   
## 0 : 3526   
## 100 : 3521   
## 2 : 2964   
## 3 : 2311   
## (Other):226014

str(d1)

## 'data.frame': 246091 obs. of 7 variables:  
## $ State\_Name : Factor w/ 33 levels "Andaman and Nicobar Islands",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ District\_Name: Factor w/ 646 levels "24 PARAGANAS NORTH",..: 428 428 428 428 428 428 428 428 428 428 ...  
## $ Crop\_Year : int 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...  
## $ Season : Factor w/ 6 levels "Autumn ",..: 2 2 2 5 5 5 5 5 5 5 ...  
## $ Crop : Factor w/ 124 levels "Apple","Arcanut (Processed)",..: 3 76 99 8 23 29 39 110 112 113 ...  
## $ Area : num 1254 2 102 176 720 ...  
## $ Production : Factor w/ 51628 levels "=","0","0.01",..: 15675 28 26753 42028 11396 42385 56 15648 9254 11919 ...

class(d1$Production)

## [1] "factor"

# Class of Production has been taken as factor which needs to be converted to double ---  
  
d1$Production <- as.double(d1$Production)  
str(d1)

## 'data.frame': 246091 obs. of 7 variables:  
## $ State\_Name : Factor w/ 33 levels "Andaman and Nicobar Islands",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ District\_Name: Factor w/ 646 levels "24 PARAGANAS NORTH",..: 428 428 428 428 428 428 428 428 428 428 ...  
## $ Crop\_Year : int 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...  
## $ Season : Factor w/ 6 levels "Autumn ",..: 2 2 2 5 5 5 5 5 5 5 ...  
## $ Crop : Factor w/ 124 levels "Apple","Arcanut (Processed)",..: 3 76 99 8 23 29 39 110 112 113 ...  
## $ Area : num 1254 2 102 176 720 ...  
## $ Production : num 15675 28 26753 42028 11396 ...

class(d1$Production)

## [1] "numeric"

typeof(d1$Production)

## [1] "double"

# Check for missing values ---  
  
sum(is.na(d1))

## [1] 0

# No missing values in the raw data. So, initially no need for treatment.

# Analysis of data —————————————————————————————————————————————————

## Finding out a year having record for all states———————————

library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

library(tidyr)  
  
# Rearrenging data state wise and year wise ---  
  
d\_year <- d1 %>% group\_by(Crop\_Year,State\_Name)%>%  
 summarize(Total\_Production = sum(Production))  
  
yrr <- spread(d\_year, key = "State\_Name", value = "Total\_Production" )  
  
View(yrr)  
  
# yrr shows the missing records, also 2004 and 2005 has all records. So, let us subset all data of 2005 (say) to analyse crop-wise.

# Finding out highest produced crop for that year ———————————

## Subsetting 2005 —————————————————————————————————————————————————

y\_2005 <- d1 %>% filter(Crop\_Year==2005)  
  
# Finding out highest produced crop for 2005 ---  
  
y\_2005\_crop <- y\_2005 %>% group\_by(Crop = Crop) %>%   
 summarize(Total = sum(Production)) %>% arrange(desc(Total))  
  
head(y\_2005\_crop)

## # A tibble: 6 x 2  
## Crop Total  
## <fct> <dbl>  
## 1 Rice 22986231  
## 2 Maize 16597967  
## 3 Moong(Green Gram) 13164070  
## 4 Sesamum 12824630  
## 5 Urad 12366603  
## 6 Wheat 12019284

# Rice is the highest produced crop for 2005. So, let us analyse the data for Rice for all states and crop years.

# Find out state wise and year wise production of the highest produced crop ——-

## Subsetting data for Rice——————————————————– ———————————————————————————–

rice <- d1 %>% filter(Crop=="Rice")  
  
# Rearrenging data year and state wise ---  
  
rice\_state <- rice %>% group\_by(Crop\_Year,State\_Name)%>%  
 summarize(Total\_Production = sum(Production))  
  
View(rice\_state)  
  
yrr\_rice <- spread(rice\_state, key = "State\_Name", value = "Total\_Production" )  
  
dim(yrr\_rice)

## [1] 19 34

# Check for missing values ——————————————————– ———————————————————————————–

sum(is.na(yrr\_rice))

## [1] 114

# Year wise missing values - in the data we can see so many missing values in 2015, so we should remove record for 2015 which is the last row  
nrow(yrr\_rice)

## [1] 19

yrr\_rice <- yrr\_rice[-19,]  
  
# State wise missing values ---  
sapply(yrr\_rice, function(x)sum(is.na(x)))

## Crop\_Year Andaman and Nicobar Islands   
## 0 10   
## Andhra Pradesh Arunachal Pradesh   
## 0 0   
## Assam Bihar   
## 0 0   
## Chandigarh Chhattisgarh   
## 5 3   
## Dadra and Nagar Haveli Goa   
## 1 6   
## Gujarat Haryana   
## 2 2   
## Himachal Pradesh Jammu and Kashmir   
## 6 8   
## Jharkhand Karnataka   
## 12 0   
## Kerala Madhya Pradesh   
## 1 2   
## Maharashtra Manipur   
## 0 7   
## Meghalaya Mizoram   
## 0 5   
## Nagaland Odisha   
## 0 1   
## Puducherry Punjab   
## 1 0   
## Rajasthan Sikkim   
## 4 1   
## Tamil Nadu Telangana   
## 1 1   
## Tripura Uttar Pradesh   
## 1 0   
## Uttarakhand West Bengal   
## 3 0

# There are records for 18 years where Andaman Nikobar and Jharkhand has more than 50% missing values

# Treating missing values ——————————————————— ———————————————————————————–

# Removing columns for Andaman Nikobar and Jharkhand  
  
rice\_new <- yrr\_rice %>% select(-`Andaman and Nicobar Islands`,-Jharkhand)  
  
sum(is.na(rice\_new))

## [1] 61

sapply(rice\_new, function(x)sum(is.na(x)))

## Crop\_Year Andhra Pradesh Arunachal Pradesh   
## 0 0 0   
## Assam Bihar Chandigarh   
## 0 0 5   
## Chhattisgarh Dadra and Nagar Haveli Goa   
## 3 1 6   
## Gujarat Haryana Himachal Pradesh   
## 2 2 6   
## Jammu and Kashmir Karnataka Kerala   
## 8 0 1   
## Madhya Pradesh Maharashtra Manipur   
## 2 0 7   
## Meghalaya Mizoram Nagaland   
## 0 5 0   
## Odisha Puducherry Punjab   
## 1 1 0   
## Rajasthan Sikkim Tamil Nadu   
## 4 1 1   
## Telangana Tripura Uttar Pradesh   
## 1 1 0   
## Uttarakhand West Bengal   
## 3 0

# Filling missing values with mean   
  
library(imputeTS)  
rice\_new <- na\_mean(rice\_new)  
  
sum(is.na(rice\_new))

## [1] 0

# Now, missing values have been treated

# From the treated dataset, finding out top 5 rice producing states and graphically represent the variation of production with respect to years ———————– ———————————————————————————–

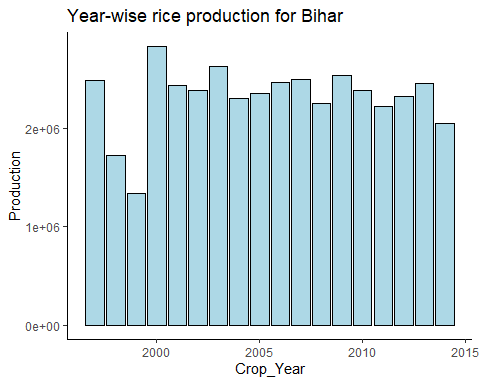
# Restructuring the treated data (rice\_new) for further analysis ---  
  
rice\_str <- gather(rice\_new, key = "State\_Name", value = "Production",c(-Crop\_Year))  
  
View(rice\_str)  
  
# Plotting rice\_new (Treated dataset) for further analysis ---  
  
# Total production of top 5 states in rice production throughout years ---  
  
rice\_str %>% group\_by(State = State\_Name) %>%  
 summarize(Total = sum(Production)) %>% arrange(desc(Total))

## # A tibble: 31 x 2  
## State Total  
## <chr> <dbl>  
## 1 Bihar 41620760   
## 2 Uttar Pradesh 39361116   
## 3 Assam 31963519   
## 4 Odisha 31725782.  
## 5 Karnataka 30311260   
## 6 West Bengal 26622834   
## 7 Madhya Pradesh 22361273.  
## 8 Kerala 19171274.  
## 9 Maharashtra 16599793   
## 10 Tamil Nadu 14874474.  
## # ... with 21 more rows

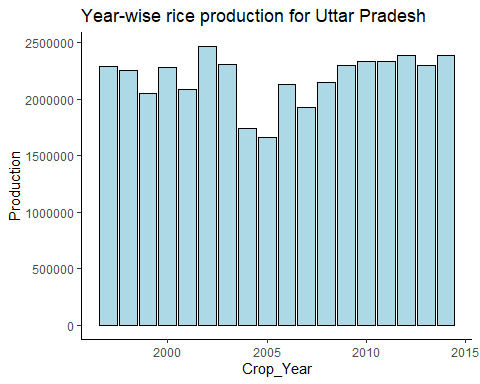
# Bihar, Uttar Pradesh, Assam, Odisha, Karnataka are the top 5 rice producers.

# Plotting data using ggplot2 —————————————————– ———————————————————————————–

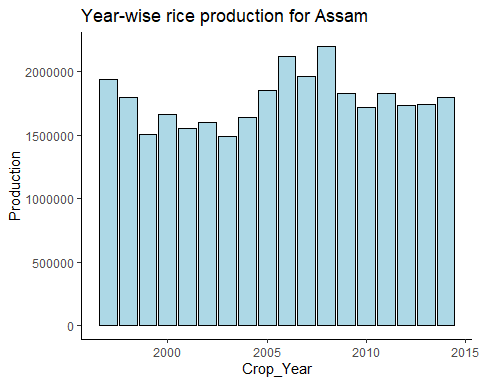
library(ggplot2)  
  
# Year-wise rice production for Bihar -   
rice\_str %>% filter(State\_Name=="Bihar") %>%  
 ggplot(aes(Crop\_Year,Production))+  
 geom\_bar(stat="identity", color="black", fill = "light blue") +  
 ggtitle("Year-wise rice production for Bihar") +  
 theme\_classic()



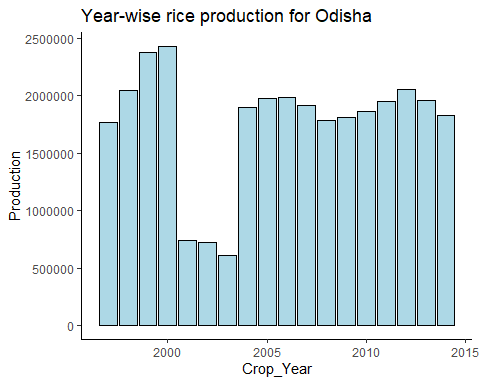
# Year-wise rice production for Uttar Pradesh -   
rice\_str %>% filter(State\_Name=="Uttar Pradesh") %>%  
 ggplot(aes(Crop\_Year,Production))+  
 geom\_bar(stat="identity", color="black", fill = "light blue") +  
 ggtitle("Year-wise rice production for Uttar Pradesh") +  
 theme\_classic()



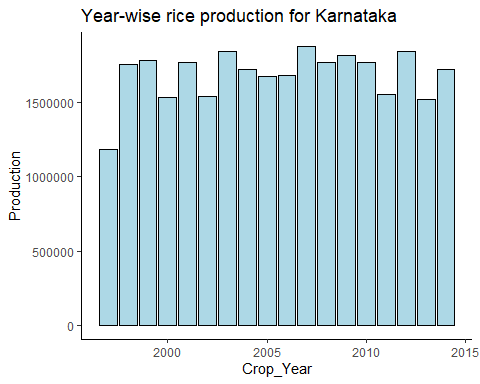
# Year-wise rice production for Assam -   
rice\_str %>% filter(State\_Name=="Assam") %>%  
 ggplot(aes(Crop\_Year,Production))+  
 geom\_bar(stat="identity", color="black", fill = "light blue") +  
 ggtitle("Year-wise rice production for Assam") +  
 theme\_classic()



# Year-wise rice production for Odisha -   
rice\_str %>% filter(State\_Name=="Odisha") %>%  
 ggplot(aes(Crop\_Year,Production))+  
 geom\_bar(stat="identity", color="black", fill = "light blue") +  
 ggtitle("Year-wise rice production for Odisha") +  
 theme\_classic()



# Year-wise rice production for Karnataka -   
rice\_str %>% filter(State\_Name=="Karnataka") %>%  
 ggplot(aes(Crop\_Year,Production))+  
 geom\_bar(stat="identity", color="black", fill = "light blue") +  
 ggtitle("Year-wise rice production for Karnataka") +  
 theme\_classic()



# Finding out top 5 highest produced crops for 2005 and graphically finding its relation to season —————————————————————-

# IN 2005 most produced crops are -   
head(y\_2005\_crop)

## # A tibble: 6 x 2  
## Crop Total  
## <fct> <dbl>  
## 1 Rice 22986231  
## 2 Maize 16597967  
## 3 Moong(Green Gram) 13164070  
## 4 Sesamum 12824630  
## 5 Urad 12366603  
## 6 Wheat 12019284

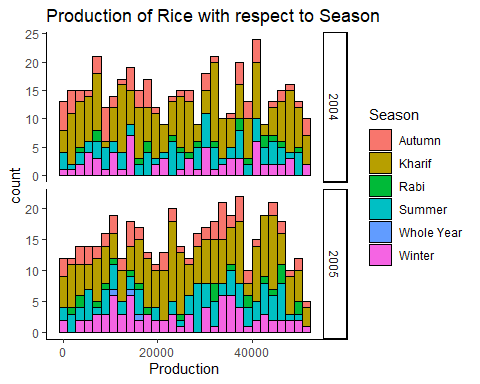
# Top 5 produced crops are - Rice, Maize, Moong(Green Gram), Sesamum, Urad  
  
# Also, yrr\_rice shows that 2004, 2005 has data for all the states. So let us plot for 2004 and 2005

# Finding relation between produced crop with respect to season ——————- ———————————————————————————–

# Production of Rice -  
d1 %>% filter(Crop\_Year==c(2004,2005),Crop=="Rice") %>%  
 ggplot(aes(Production))+  
 geom\_histogram(color="black",aes(fill=Season)) + theme\_classic()+  
 facet\_grid(Crop\_Year~., scales = "free")+  
 ggtitle("Production of Rice with respect to Season")

## Warning in Crop\_Year == c(2004, 2005): longer object length is not a  
## multiple of shorter object length

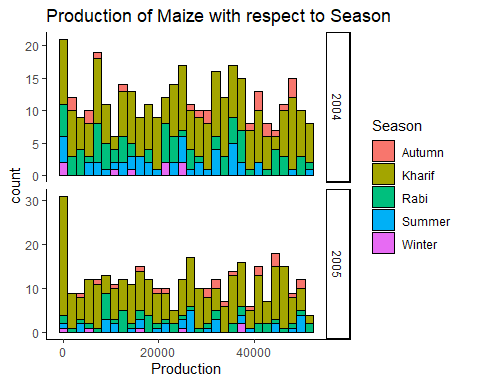
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Production of Maize -  
d1 %>% filter(Crop\_Year==c(2004,2005),Crop=="Maize") %>%  
 ggplot(aes(Production))+  
 geom\_histogram(color="black",aes(fill=Season)) + theme\_classic()+  
 facet\_grid(Crop\_Year~., scales = "free")+  
 ggtitle("Production of Maize with respect to Season")

## Warning in Crop\_Year == c(2004, 2005): longer object length is not a  
## multiple of shorter object length

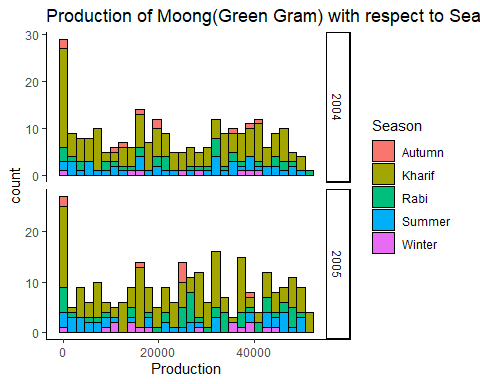
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Production of Moong(Green Gram) -  
d1 %>% filter(Crop\_Year==c(2004,2005),Crop=="Moong(Green Gram)") %>%  
 ggplot(aes(Production))+  
 geom\_histogram(color="black",aes(fill=Season)) + theme\_classic()+  
 facet\_grid(Crop\_Year~., scales = "free")+  
 ggtitle("Production of Moong(Green Gram) with respect to Season")

## Warning in Crop\_Year == c(2004, 2005): longer object length is not a  
## multiple of shorter object length

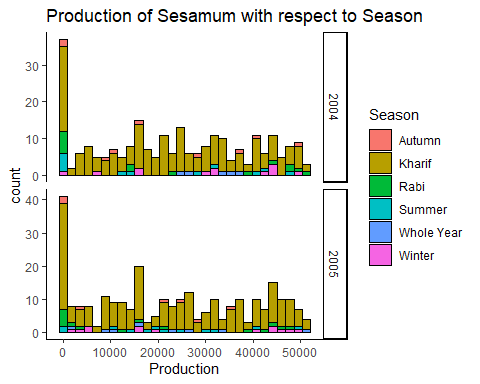
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Production of Sesamum -  
d1 %>% filter(Crop\_Year==c(2004,2005),Crop=="Sesamum") %>%  
 ggplot(aes(Production))+  
 geom\_histogram(color="black",aes(fill=Season)) + theme\_classic()+  
 facet\_grid(Crop\_Year~., scales = "free")+  
 ggtitle("Production of Sesamum with respect to Season")

## Warning in Crop\_Year == c(2004, 2005): longer object length is not a  
## multiple of shorter object length

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Production of Urad -  
d1 %>% filter(Crop\_Year==c(2004,2005),Crop=="Urad") %>%  
 ggplot(aes(Production))+  
 geom\_histogram(color="black",aes(fill=Season)) + theme\_classic()+  
 facet\_grid(Crop\_Year~., scales = "free")+  
 ggtitle("Production of Urad with respect to Season")

## Warning in Crop\_Year == c(2004, 2005): longer object length is not a  
## multiple of shorter object length

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

