# TDP

###################################################

# AT&T Technology Development Program

# Data Analysis Interview: Assignment

# Anna Connolly, anna.connolly@sbcglobal.net

###################################################

# Import Datasets

# BayerAG\_BAYZF data

bayer <- read.csv("~/Desktop/Datasets/BayerAG\_BAYZF.csv")

View(bayer)

# Bee.Colony.Loss data

bee <- read.csv("~/Desktop/Datasets/Bee Colony Loss.csv")

View(bee)

# Dupont\_DD data

dupont <- read.csv("~/Desktop/Datasets/Dupont\_DD.csv")

View(dupont)

# US\_Almonds data

almonds <- read.csv("~/Desktop/Datasets/US\_Almonds.csv")

View(almonds)

# US\_States\_Blueberries data

blue <- read.csv("~/Desktop/Datasets/US\_States\_Blueberries.csv")

View(blue)

###################################################

# Download packages and load libraries

# ggplot2

library(ggplot2)

# dplyr

library(dplyr)

# tidyverse

library(tidyverse)

# lubridate

library(lubridate)

###################################################

# Part A:

###################################################

# Question 1-Calculate correlation between adjusted closing prices for the Dupont and BayerAG stocks.

###################################################

# Look at dupont data

head(dupont)

# Adj.Close variable

View(dupont$Adj.Close)

# Make dupont a data frame

dupont\_df <- data.frame(dupont)

# Check data frame

View(dupont\_df)

# Look at bayer data

head(bayer)

# Adj.Close variable

View(bayer$Adj.Close)

# Make bayer a data frame

bayer\_df <- data.frame(bayer)

# Check data frame

View(bayer\_df)

# Combine dupont\_df and bayer\_df

question1 <- merge(dupont\_df, bayer\_df,

by = 'Date',

suffixes = c(".dupont", ".bayer"))

# Check combination

View(question1)

head(question1)

question1

# Visualize correlation between adjusted closing prices for the Dupont and BayerAG stocks.

plot(question1$Adj.Close.dupont, question1$Adj.Close.bayer)

# Correlationb looks positive, linear, moderately strong

# Calculate correlation between adjusted closing prices for the Dupont and BayerAG stocks.

cor(question1$Adj.Close.dupont, question1$Adj.Close.bayer)

# Correlation = 0.737419

### The correlation between adjusted closing prices for the Dupont and BayerAG stocks is 0.737419.

###################################################

# Question 2-In the blueberry dataset, which state had the highest increase of yield per acre for a given year?

###################################################

# Look at blue data set

View(blue)

head(blue)

# Visualize increase of yield per acre for a given year for each state

ggplot(blue, aes(x=Year, y=Yield.per.acre..Pounds.)) +

geom\_point() +

facet\_wrap(~State)

# Select variables of interest

question2 <- select(blue, Year, Yield.per.acre..Pounds., State)

View(question2)

# Arrange yield per acre

arrange(question2, desc(Yield.per.acre..Pounds., na.rm = TRUE))

# Oregon 2016 9,760

# Check

check2 <- filter(question2, "Yield.per.acre..Pounds" > 9,760)

View(check2)

### Oregon had the highest increase of yield per acre for the year 2016.

###################################################

# Question 3-Which year had the highest standard deviation of colonies?

###################################################

# Look at bee data set

View(bee)

head(bee)

# Year variable

# X.Colonies variable

# Look at colonies variable

View(bee$X.Colonies)

mean(bee$X.Colonies)

# Mean = 37175.53

# Select year, volonies, and state

# group by state

# arrange by colonies

question3 <-bee %>%

select(Year, X.Colonies, X.State) %>%

group\_by(X.State) %>%

mutate(popmean = 37175.53) %>%

mutate(popsd = sqrt(((X.Colonies-popmean)\*\*2)/(365))) %>%

arrange(desc(popsd))

# View new object

View(question3)

# n value

length(bee$X.Colonies)

str(question3)

# 365 observations

### The year 2012-2013 had the highest standard deviation of colonies.

###################################################

# Question 4-For the year 2014, visualize the histograms of Blueberry Yield Per Acre for states that had a Total Colony Loss less than 35% and states that had a Total Colony Loss more than 35%.

###################################################

# Look at data

View(bee)

View(blue)

View(blue$Yield.per.acre..Pounds.)

# Percent to numeric for X.Total.Annual.Loss

bee$totalloss <- as.numeric(sub("%","",bee$X.Total.Annual.Loss))/100

View(bee$totalloss)

as.numeric(sub("/", "", bee$Year))

# States that had Total Colony Loss < 35%

less35 <- bee %>%

select(X.State, X.Total.Annual.Loss, Year) %>%

mutate(loss = as.numeric(sub("%","",bee$X.Total.Annual.Loss))/100) %>%

filter(loss < 0.35) %>%

filter(Year=="2013/14" | Year=="2014/15")

# Check

View(less35)

# States that had Total Colony Loss > 35%

greater35 <- bee %>%

select(X.State, X.Total.Annual.Loss, Year) %>%

mutate(loss = as.numeric(sub("%","",bee$X.Total.Annual.Loss))/100) %>%

filter(loss > 0.35) %>%

filter(Year=="2013/14" | Year=="2014/15")

# Check

View(greater35)

# States that had Total Colony Loss < 35%

View(less35$X.State)

# Hawaii, Oregon South, Carolina, Nevada, Idaho,Michigan, Wyoming, Utah, Vermont, Missouri, Georgia, Montana, Massachusetts, Nebraska, Colorado, North Carolina, Arkansas, Minnesota, Louisiana, South Dakota, California, New Mexico, Wisconsin

# States that had Total Colony Loss > 35%

View(greater35$X.State)

# South Dakota, Tennessee, Nebraska, North Dakota, Colorado, California, Mississippi, Washington, New Hampshire, Montana, North Carolina, Kansas, Louisiana, Virginia, Massachusetts, New Mexico, Texas, Rhode Island, Kentucky, Alabama, New Jersey, Arkansas, Indiana, Minnesota, New York, Florida, Connecticut, Arizona, Ohio, Wisconsin, Maine, West Virginia, Maryland, Delaware, Pennsylvania, Iowa, Illinois, Oklahoma, Nevada, North Dakota, New Jersey, Missouri, Michigan, Georgia, New York, Vermont, District of Columbia

# Question 4-For the year 2014, visualize the histograms of Blueberry Yield Per Acre for states that had a Total Colony Loss less than 35% and states that had a Total Colony Loss more than 35%.

# Year 2014 in blueberries data

blue2014 <- subset(blue\_df, Year==2014)

# select(State)

View(blue2014)

# States- Alabama, Arkansas, Florida, Georgia, Indiana, Maine, Michigan, Oregon, Washington

# < 35% States- Arkansas, Georgia, Oregon

lessblue2014 <- subset(blue2014, State == "Arkansas" |State == "Georgia" | State =="Oregon")

View(lessblue2014)

# > 35% States- Alabama, Arkansas, Florida, Georgia, Maine, Washington

moreblue2014 <- subset(blue2014, State == "Alabama" | State == "Arkansas" |State == "Florida" | State == "Georgia" | State =="Maine" | State=="Washington")

View(moreblue2014)

# Histograms of Blueberry Yield Per Acre for states that had a Total Colony Loss less than 35%

ggplot(lessblue2014, aes(Yield.per.acre..Pounds.)) +

geom\_bar(aes(fill=State)) +

coord\_flip() +

labs(x="Yield per acre (Pounds)", y="Count")

# Histograms of Blueberry Yield Per Acre for states that had a Total Colony Loss more than 35%

ggplot(moreblue2014, aes(Yield.per.acre..Pounds.)) +

geom\_bar(aes(fill=State)) +

coord\_flip()+

labs(x="Yield per acre (Pounds)", y="Count")

# I don't find these very helpful, so these are scatterplots that better visualize the data

# Blueberry Yield Per Acre for states that had a Total Colony Loss less than 35%

ggplot(lessblue2014, aes(State, Yield.per.acre..Pounds.)) +

geom\_point(aes(color=State)) +

labs(x="State", y="Yield per acre (Pounds)")

# Blueberry Yield Per Acre for states that had a Total Colony Loss more than 35%

ggplot(moreblue2014, aes(State, Yield.per.acre..Pounds.)) +

geom\_point(aes(color=State)) +

labs(x="State", y="Yield per acre (Pounds)")

###################################################

# Question 5-Calculate and visualize correlation matrix for variables below.

# Which two variables had the highest correlation?

###################################################

### Dupont Adjusted Closing Stock price

View(dupont\_df)

dupont\_df$Adj.Close

# Reformat dates

dupont\_df$Date <- format(as.Date(dupont\_df$Date, format="%Y-%m-%d"),"%Y")

View(dupont\_df$Date)

# Rename Date to Year

names(dupont\_df)[names(dupont\_df) == "Date"] <- "Year"

View(dupont\_df)

### BayerAG Adjusted Closing Stock price

View(bayer\_df)

bayer\_df$Adj.Close

# Reformat dates

bayer\_df$Date <- format(as.Date(bayer\_df$Date, format="%Y-%m-%d"),"%Y")

View(bayer\_df$Date)

# Rename Date to Year

names(bayer\_df)[names(bayer\_df) == "Date"] <- "Year"

View(bayer\_df)

### Average Colony Loss

# Make bee a data frame

bee\_df <- data.frame(bee)

# Check data frame

View(bee\_df)

# Reformat dates

bee\_df$Year <- format(as.Date(bee\_df$Year, format="%Y/%y"),"%Y")

View(bee\_df$Year)

bee\_df$totalloss

### Average Blueberry Yield per Acre

# Make blue a data frame

blue\_df <- data.frame(blue)

# Check data frame

View(blue\_df)

blue\_df$Yield.per.acre..Pounds.

### Almond Yield per Acre

# Make almonds a data frame

almonds\_df <- data.frame(almonds)

# Check data frame

View(almonds\_df)

almonds\_df$Yield.per.Acre..pounds.

# Possible combinations-6

### Dupont Adjusted Closing Stock price and BayerAG Adjusted Closing Stock price

question5a <- merge(dupont\_df, bayer\_df,

by = 'Year',

suffixes = c(".dupont", ".bayer"))

View(question5a)

head(question5a)

question5a

# Visualize correlation between Dupont Adjusted Closing Stock price and BayerAG Adjusted Closing Stock price

plot(question5a$Adj.Close.dupont, question5a$Adj.Close.bayer)

cor(question5a$Adj.Close.dupont, question5a$Adj.Close.bayer)

# Correlation = 0.737419

### Dupont Adjusted Closing Stock price and Average Blueberry Yield per Acre

### Dupont Adjusted Closing Stock price and Almond Yield per Acre

### BayerAG Adjusted Closing Stock price and Average Blueberry Yield per Acre

### BayerAG Adjusted Closing Stock price and Almond Yield per Acre

### Average Blueberry Yield per Acre and Almond Yield per Acre

# Which two variables had the highest correlation?

###################################################

###################################################

# Question 6-Using Dupont and BayerAG Adjusted Closing Stock price as input, forecast the number of Colonies for the states of California, North Dakota, and Texas if Dupont's stock prices closed at $90 and BayerAG at $80.

###################################################

###################################################

# Part B:

###################################################

# Datasets-

# Hypothesis-

# Conclusion-

###################################################