SOLUTION: Sample 1

library(tidyverse)

library(dplyr)

library(tidyr)

library(ggplot2)

library(openxlsx)

library(ggrepel)

library(scales)

#loading all the libraries needed for this project

install.packages('nycflights13')

library(nycflights13)

#installing and loading package for nyc flights

flights = nycflights13::flights #saving as 'flights'

Tb = tidyr::who

#saving the who data under the name TB

######################### PART 1 ################################

TB = Tb %>%

pivot\_longer(

cols = new\_sp\_m014:newrel\_f65,

names\_to = "key",

values\_to = "cases",

values\_drop\_na = TRUE

) %>%

mutate(

key = stringr::str\_replace(key, "newrel", "new\_rel")

) %>%

separate(key, c("new", "var", "sexage")) %>%

select(-new, -iso2, -iso3) %>%

separate(sexage, c("sex", "age"), sep = 1)

#tidying data by creating less and more descriptive columns.

write.xlsx(TB, 'TB.xlsx')

#exporting to an excel file.

######################### PART 2 ################################

########################## Question 1 #################

fct\_count(flights$origin)

#tallying the flights out of the three NYC airports which shows EWR is busiest.

########################## Question 2 #################

n\_distinct(flights$dest)

#Total dest airports

fct\_count(flights$dest) %>%

arrange(-n) %>%

head()

#using pipe operator to show the top 6 most popular dest.

#ORD is busiest dest.

########################## Question 3 #################

flights %>%

filter(month == 11 & day == 28 & origin == 'LGA') %>%

count()

#totaling flights out of LGA on Thanksgiving

########################## Question 4 #################

flights %>%

filter(origin == 'JFK') %>%

count(month, day) %>%

slice(which.max(n))

#busiest day was 7/11 with 332 flights out of JFK

########################## Question 5 #################

flights %>%

filter(origin == 'LGA') %>%

count(month) %>%

slice(which.max(n))

#busiest month was October with 9642 flights out of LGA

########################## Question 6 #################

flights %>%

filter(month == 6) %>%

slice(which.max(air\_time))

#longest flight in June is JFK to HNL with airtime of 650.

########################## Question 7 #################

flights %>%

filter(month == 5) %>%

slice(which.min(air\_time))

#shortest flight in may is EWR to BDL with airtime of 21.

########################## Question 8 #################

fct\_count(flights$carrier) %>%

arrange(n)

# OO had the least amount of flights with 32

######################### PART 3 ################################

########################## Plot 1 #################

flights %>%

group\_by(month, origin) %>%

summarize(count = n()) %>%

ggplot(mapping = aes(x = month, y = count, col = origin)) +

geom\_line() +

geom\_point() +

scale\_x\_continuous(breaks = 1:12)

#plotting the departures of all 3 airports per month

########################## Plot 2 #################

JFK = flights[flights$origin == "JFK",]

#refining data to only include JFK airport

p2 = JFK %>%

ggplot(mapping = aes(x=month, y=arr\_delay)) +

geom\_line(stat="summary", fun.y="mean") +

scale\_x\_continuous(breaks = 1:12)

p2 + ylab('avg. delay') #changing label on y axis to avg. delay

#plotting the avg. delay per month

########################## Plot 3 #################

flights %>%

filter (origin == 'JFK')

group\_by(carrier) %>%

summarize(count = n()) %>%

ggplot(mapping = aes(x=carrier, y=count, fill=carrier)) +

geom\_col(stat="identity", width=1, color="white") +

geom\_text(aes(label = count), vjust = 0)

#chart of all the carriers flights out of NYC

########################## Plot 4 #################

top5 <- flights %>%

filter(carrier == "UA" | carrier == "B6" | carrier == "EV" |

carrier == "DL" | carrier == "AA" ) %>%

ggplot(mapping = aes(x=carrier, y=dep\_delay, color=carrier)) +

geom\_boxplot()

top5

########################## Plot 5 #################

flights %>%

filter(origin == "LGA") %>%

group\_by(month) %>%

summarize(count = n()) %>%

ggplot(mapping = aes(x = month, y = count)) +

geom\_line() +

geom\_point() +

scale\_x\_continuous(breaks = 1:12)

#plotting the departures from LGA per month

######################### PART 4 ################################

########################## Question 1 #################

TB12 = TB %>%

filter(year == 2012) %>%

group\_by(country) %>%

summarise(sum(cases))

#refining the data to country and cases in 2012

colnames(TB12)[2] = "cases"

#renaming the second column for syntax

TB12 %>%

filter(cases == min(cases))

TB12 %>%

filter(cases == max(cases))

#returning the max and min values

########################## Question 2 #################

Aus = TB %>%

filter(country == 'Australia') %>%

group\_by(sex, year) %>%

summarize(sum(cases))

#refining the data to be for Australia

colnames(Aus)[3] = "cases"

#renaming the second column for syntax

ggplot(Aus, aes(x=year, y=cases, fill = sex)) +

geom\_bar(position="dodge", stat="identity") +

scale\_x\_continuous(breaks = 1997:2013) +

geom\_text(aes(label = cases), vjust = 0)

#plotting the data by year and gender

########################## Question 3 #################

Afg = TB %>%

filter(country == 'Afghanistan') %>%

filter(year %in% (2000:2013)) %>%

group\_by(sex, year) %>%

summarize(sum(cases))

#refining the data to be for Afghanistan

colnames(Afg)[3] = "cases"

#renaming the second column for syntax

ggplot(Afg, aes(x=year, y=cases, fill = sex)) +

geom\_bar(position="dodge", stat="identity") +

scale\_x\_continuous(breaks = 2000:2013) +

geom\_text(aes(label = cases), vjust = 0)

#plotting the data by year and gender

########################## Question 4 #################

NorthA = TB %>%

filter(country == "United States of America" | country == "Canada" | country == "Mexico") %>%

filter (year %in% (1996:2013)) %>%

group\_by(age) %>%

summarise(sum(cases))

#refining dataset

colnames(NorthA)[2] = "cases"

#renaming the second column for syntax

NAcases = NorthA %>%

mutate(percent = cases / sum(NorthA[,2]))

#adding a percent column that shows the cases per age group

NAcases$percent = percent(NAcases$percent, accuracy = NULL)

#formatting the percent column as percentages.

NAcases

#outputting results of percentages

NAcases1 = NAcases %>%

mutate(csum = rev(cumsum(rev(cases))),

pos = cases/2 + lead(csum, 1),

pos = if\_else(is.na(pos), cases/2, pos))

ggplot(NAcases1, aes(x="", y=cases, fill=age)) +

geom\_bar(stat="identity", width=1) +

coord\_polar("y", start=0) +

theme\_void() +

geom\_label\_repel(data = NAcases1,

aes(y = pos, label = percent),

size = 4.5, nudge\_x = 1, show.legend = FALSE)

#plotting the percentages in a pie chart.

########################## Question 5 #################

# 3 countries with largest population determined by Google

SouthA = TB %>%

filter(country == "Brazil" | country == "Colombia" | country == "Argentina") %>%

filter (year %in% (1999:2013)) %>%

group\_by(age) %>%

summarise(sum(cases))

#refining dataset

colnames(SouthA)[2] = "cases"

#renaming the second column for syntax

SAcases = SouthA %>%

mutate(percent = cases / sum(SouthA[,2]))

#adding a percent column that shows the cases per age group

SAcases$percent = percent(SAcases$percent, accuracy = NULL)

#formatting the percent column as percentages.

SAcases

#outputting results of percentages

SAcases1 = SAcases %>%

mutate(csum = rev(cumsum(rev(cases))),

pos = cases/2 + lead(csum, 1),

pos = if\_else(is.na(pos), cases/2, pos))

ggplot(SAcases1, aes(x="", y=cases, fill=age)) +

geom\_bar(stat="identity", width=1) +

coord\_polar("y", start=0) +

theme\_void() +

geom\_label\_repel(data = SAcases1,

aes(y = pos, label = percent),

size = 4.5, nudge\_x = 1, show.legend = FALSE)

#plotting the percentages in a pie chart.

########################## Question 6 #################

# 3 countries with largest population determined by Google

Europe = TB %>%

filter(country == "Russian Federation" | country == "Germany" | country ==

"United Kingdom of Great Britain and Northern Ireland") %>%

filter (year %in% (1997:2013)) %>%

group\_by(age) %>%

summarise(sum(cases))

#refining dataset

colnames(Europe)[2] = "cases"

#renaming the second column for syntax

EUcases = Europe %>%

mutate(percent = cases / sum(Europe[,2]))

#adding a percent column that shows the cases per age group

EUcases$percent = percent(EUcases$percent, accuracy = NULL)

#formatting the percent column as percentages.

EUcases

#outputting results of percentages

EUcases1 = EUcases %>%

mutate(csum = rev(cumsum(rev(cases))),

pos = cases/2 + lead(csum, 1),

pos = if\_else(is.na(pos), cases/2, pos))

ggplot(EUcases1, aes(x="", y=cases, fill=age)) +

geom\_bar(stat="identity", width=1) +

coord\_polar("y", start=0) +

theme\_void() +

geom\_label\_repel(data = EUcases1,

aes(y = pos, label = percent),

size = 4.5, nudge\_x = 1, show.legend = FALSE)

#plotting the percentages in a pie chart.

########################## Question 7 #################

# 3 countries with largest population determined by Google

Africa = TB %>%

filter(country == "Nigeria" | country == "Ethiopia" | country == "Egypt") %>%

filter (year %in% (1995:2013)) %>%

group\_by(age) %>%

summarise(sum(cases))

#refining dataset

colnames(Africa)[2] = "cases"

#renaming the second column for syntax

AFcases = Africa %>%

mutate(percent = cases / sum(Africa[,2]))

#adding a percent column that shows the cases per age group

AFcases$percent = percent(AFcases$percent, accuracy = NULL)

#formatting the percent column as percentages.

AFcases

#outputting results of percentages

AFcases1 = AFcases %>%

mutate(csum = rev(cumsum(rev(cases))),

pos = cases/2 + lead(csum, 1),

pos = if\_else(is.na(pos), cases/2, pos))

ggplot(AFcases1, aes(x="", y=cases, fill=age)) +

geom\_bar(stat="identity", width=1) +

coord\_polar("y", start=0) +

theme\_void() +

geom\_label\_repel(data = AFcases1,

aes(y = pos, label = percent),

size = 4.5, nudge\_x = 1, show.legend = FALSE)

#plotting the percentages in a pie chart.

########################## Question 8 #################

# 3 countries with largest population determined by Google

Asia = TB %>%

filter(country == "China" | country == "India" | country == "Indonesia") %>%

filter (year %in% (1995:2013)) %>%

group\_by(age) %>%

summarise(sum(cases))

#refining dataset

colnames(Asia)[2] = "cases"

#renaming the second column for syntax

AScases = Asia %>%

mutate(percent = cases / sum(Asia[,2]))

#adding a percent column that shows the cases per age group

AScases$percent = percent(AScases$percent, accuracy = NULL)

#formatting the percent column as percentages.

AScases

#outputting results of percentages

AScases1 = AScases %>%

mutate(csum = rev(cumsum(rev(cases))),

pos = cases/2 + lead(csum, 1),

pos = if\_else(is.na(pos), cases/2, pos))

ggplot(AScases1, aes(x="", y=cases, fill=age)) +

geom\_bar(stat="identity", width=1) +

coord\_polar("y", start=0) +

theme\_void() +

geom\_label\_repel(data = AScases1,

aes(y = pos, label = percent),

size = 4.5, nudge\_x = 1, show.legend = FALSE)

#plotting the percentages in a pie chart.

SOLUTION: Sample 2

library(nycflights13)

library(tidyverse)

library(ggplot2)

library(dplyr)

library("xlsx")

tuberculosis = tidyr::who

tuberculosis =

tuberculosis %>%

pivot\_longer(cols = new\_sp\_m014:newrel\_f65, names\_to = "key", values\_to = "cases", values\_drop\_na = TRUE) %>%

mutate(key = stringr::str\_replace(key, "newrel", "new\_rel")) %>%

separate(key, c("new", "type", "sexage"), sep = "\_") %>%

select(-new, -iso2, -iso3) %>%

separate(sexage, c("sex", "age"), sep = 1)

write.xlsx(tuberculosis, file = "assignment5\_test.xlsx", sheetName = "myData", append = FALSE)

###Part 2###

flights = nycflights13::flights

flights = na.omit(flights)

#question 1

NYCCount =

flights %>%

count(origin)

busiest = which.max(NYCCount$n) #grabs row location of busiest airport

summarize(NYCCount, "Busiest Airport" = NYCCount[busiest,1], "Number of flights" = NYCCount[busiest,2])

#question 2

Destinations =

flights %>%

count(dest)

popular = which.max(Destinations$n) #grabs row location of most popular destination

summarize(Destinations, "Number of Destinations" = n\_distinct(Destinations,"dest"), "Most Popular Destination" = Destinations[popular,1])

#question 4

dates =

flights %>%

filter(origin == "JFK") %>%

group\_by(year,month,day) %>%

count(year,month,day)

busiestDay = which.max(dates$n) #finds the row value of the busiest day of the year

answer4 = data.frame(day = c(dates[busiestDay,3]), month = c(dates[busiestDay,2]))

summarize(answer4, "Busiest day of the year for JFK" = answer4[,1], "The corresponding month" = answer4[,2])

#question 5

flightMonths =

flights %>%

filter(origin == "LGA") %>%

group\_by(year,month) %>%

count(month)

busiestMonth = which.max(flightMonths$n)

answer5 = data.frame(month = c(flightMonths[busiestMonth,2]))

summarize(answer5, "Busiest month of the year for LGA" = answer5[1,1])

#question 8

carrierList = count(flights,carrier) #creates a dataframe with the amount of flights per carrier

leastBusyCarrier = which.min(carrierList$n)

summarize(carrierList, "Least Busy Carrier" = carrierList[leastBusyCarrier,1])

###Part 3###

#Plot 1

plot1Data =

flights %>%

group\_by(year,month,origin) %>%

count()

ggplot(plot1Data) + geom\_col(mapping = aes(x = month, y = n, fill = origin), position = position\_dodge()) +

labs(title = "Departures per month per origin", y = "Flights")

#Plot 3

plot3Data =

flights %>%

filter(origin == "JFK") %>%

group\_by(carrier) %>%

count(carrier)

ggplot(plot3Data) + geom\_col(mapping = aes(x = carrier, y = n)) + labs(title = "Flights departing JFK per carrier", y = "Flights")

#plot 5

plot5Data =

flights %>%

filter(origin == "LGA") %>%

group\_by(month) %>%

count(month)

ggplot(plot5Data) + geom\_col(mapping = aes(x = month, y = n)) + labs(title = "Flights departing from LGA per month", y = "Flights")

###part 4###

#question 1

TBQ1 =

tuberculosis %>%

filter(year == 2012) %>%

group\_by(country) %>%

summarize("Cases\_2012" = sum(cases))

TBQ1Min =

TBQ1 %>%

filter(Cases\_2012 == 0)

TBQ1Min

TBQ1Max = which.max(TBQ1$Cases\_2012)

paste("The country,", TBQ1[TBQ1Max,1], "has the most amount of cases for the year 2012. Total number of cases:",TBQ1[TBQ1Max,2])

#question 2 (plot)

TBQ2PlotData =

tuberculosis %>%

filter(country == "Australia") %>%

group\_by(sex, year) %>%

summarize("Cases" = sum(cases))

ggplot(TBQ2PlotData) + geom\_line(mapping = aes(x = year, y = Cases, color = sex)) + labs(title = "Tuberculosis cases in Australia per gender from 1980-2013")

#question 3 (plot)

TBQ3PlotData =

tuberculosis %>%

filter(country == "Afghanistan", year >= 2000 & year <= 2013) %>%

group\_by(sex, year) %>%

summarize("Cases" = sum(cases))

ggplot(TBQ3PlotData) + geom\_line(mapping = aes(x = year, y = Cases, color = sex)) + labs(title = "Tuberculosis cases in Afghanistan per gender from 2000-2013")

#question 9 (plot)

TBQ9PlotData =

tuberculosis %>%

group\_by(year,age) %>%

summarize("Cases" = sum(cases))

ggplot(TBQ9PlotData) + geom\_line(mapping = aes(x = year, y = Cases, color = age)) + labs(title = "Tuberculosis cases worldwide per age group")

#question 10

TBQ10Data =

tuberculosis %>%

filter(year == 1997 | year == 2010) %>%

group\_by(country) %>%

summarize("Cases\_1997" = sum(cases[year==1997]), "Cases\_2010" = sum(cases[year==2010]), "percentChange" =

((sum(cases[year==2010]) - sum(cases[year==1997])) / sum(cases[year==1997])) \* 100)

TBQ10ReductionData = filter(TBQ10Data,percentChange == -100)

TBQ10Sharpest = which.max(TBQ10ReductionData$Cases\_1997)

paste("The country with the sharpest reduction is:", TBQ10ReductionData[TBQ10Sharpest,1], "The percent change is:",

TBQ10ReductionData[TBQ10Sharpest,4], "The number of cases in 1997 were:", TBQ10ReductionData[TBQ10Sharpest,2],

"and the number of cases in 2010 was:", TBQ10ReductionData[TBQ10Sharpest,3])