**CAP 2751 – Tools for Data Science Midterm Exam Fall 2023**

**Part A** (60pts) Name\_\_\_\_\_\_\_\_\_\_SOLUTION\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Identify **all** the following statements about R that are **not** true.

A: R-Console is an integrated development environment (IDE) for R.

B: R does not require datasets to be of rectangular dimension.

C: R is fast and free but not very user-friendly.

D: R provides excellent commercial support to its users.

E: R relies on packages available from CRAN

1. Which operation can best be used to determine if the integer A is divisible by another integer B?

  A: A%in%B          B: A%%B             C: A/B              D: A%/%B             E: None of these

1. How many of the R statements below will generate and store a vector of all the even numbers between 1 and 20 in a variable called **even\_num**.

A: even\_num <- c(1:10) \* 2

B: even\_num <- seq(2,20,2)

# C:  x = 2:20

even\_num <- x[which(x%%2==0)]

D: All are correct

E: Only two are correct

1. Which of the following is a legitimate use of exploratory data analysis?

A: Exploratory data analysis is used to determine if there are outliers in the dataset.

B: Exploratory data analysis is used to explore the relationship between multiple datasets.

C: Exploratory data analysis is used to determine the underlying structure of the data.

D: Two of the above are true

E: All the above are true.

1. Which of the following R functions is the best used to determine the number of values in a data frame?

A: na.rm( )    B: n\_distinct( )     C:  sum(!is.na( ))    D: sum(is.na( ))    E: which(is.na( ))

1. Which R-statement will correctly print the content of the variables X and Y?

A: print(X,Y) B: print(c(X,Y)) C: print(c(X,Y)) D: input(X,Y) E: Two of these are correct

1. A student writes the R statement below to find the average of the two columns **cty** and **hwy** of the dataset**mpg** but the code generates the error message shown when executed.

Write the correct statement(s) to give the average of the two columns.

**> mean(mpg$cty,mpg$hwy)**

          Error in mean.default(mpg$cty, mpg$hwy) : ‘trim”  must be numeric of length one

Three possible solutions:

    > mean(c(mpg$cty,mpg$hwy))

> colMeans(mpg[,c(cty,hwy)], na.rm = TRUE)

mpg%>% group\_by(cty,hwy)%>% summarise(mean\_mpg = mean())

1. Write the R statements to determine and print the number of rows and columns in the R dataset **mpg**.

Possible solutions:

ncol(mpg)

nrow(mpg)

dim(mpg)

1. Consider the R-package **nycflights13** that contains the data set **flights**, which lists all the flights out of the three airports in New York City in 2013. Briefly explain what the following R statements will do when executed.

**Number\_fight <-flights %>%**

**group\_by(month ==1) %>%**

**filter(dep\_delay > 120) %>%**

**summarize(count=n())**

The code counts and stores the number of flights in January that had a departure delay of more than 2 hours in a variable called Number\_flight.

1. The R code below should generate a **horizontal bar** plot of countries and the number of NBA players from the countries in the 2017-2018 NBA season. However, there are as many as six errors in the code. Identify five of the errors.

##download the data file and create a horizontal bar graph.

#salaries = read.csv("NBA\_salary.csv") # read.csv should be read\_csv

my.data <= salaries

ggplot2(my.data, aes(x = NBA\_Country, y = player\_count)) +  
geom\_bar(stat = "identity") +  
geom\_flip() + #geom\_flip should be coord\_flip or geom\_col()  
labs(title = "Num of NBA Players from Each Country",  
x = "Country",  
y = "Number of Players")

**Part B** (50points)

**Problem 1 (20points)**

The R dataset **mpg** in R contains fuel economy data from 1999 to 2008 for 38 popular models of cars. We are interested in selecting the five best automobiles in two vehicle classes (SUV and Pickup) based on a new column called **ave\_mpg** that takes the average of the city mileage per gallon (**cty**) and the highway mileage per gallon (**hwy**) for each vehicle. Write a code that reads the **mpg** dataset into the R workspace and generate a five-column data frame called **new\_mpg** that contains the manufacturer, model name, engine displacement, year of manufacture, and the average (take the average of the city and highway mileage) per gallon for each vehicle. Use your R code to answer the following questions.

**Question 1** (10points): Revise your data frame **new\_mpg** to include only SUVs that have the highest average mileage per gallon and store it in a variable called**best\_SUV** that contains the manufacturer, model, average mileage. Print **best\_SUV.**

**Question 2** (10points): Revise your data frame **new\_mpg** to include only pickup trucks that have the best average mileage per gallon and store it in a variable called **best\_Pickup** that contains the manufacturer, model, average mileage. Print **best\_Pickup.**

**Question 3** (5 points): (OPTIONAL) Use your data frame **new\_mpg** to generate a scatter plot of the average fuel efficiency (**ave\_mpg**) versus engine size (**displ**) with vehicles labeled as class, in an aesthetically pleasant and meaningful way.

**Problem 2 (20 points)**

The R dataset **nycflights13** contains all flights that departed from three airports in New York in 2013. Write a code in R that reads the data into a new variable called **flights\_nyc\_2013**. Revise the file to contain only flights going to Washington DC. Note that, like New York City, there are three major airports in the Washington area, namely, Ronal Reagan Washington Airport (DCA),

Washington Dulles International Airport (IAD), and Baltimore/Washington International Airport (BWI). We are only interested in flights that took place on Valentine’s Day (February 14, 2013) and on Thanksgiving Day (November 28, 2013). Call your revised dataset **flights\_DC\_holiday** and answer the following questions.

**Question 1** (10points): Rank the three Washington DC airports based on the number of flight arrivals. Which is the busiest holiday airport in Washington DC?

**Question 2** (10points): What are the five most popular carriers into Wahington DC? Which airline has the worst average holiday delay?

**Question 3** (5points): (OPTIONAL) Use data visualization (of your choice) to display the number of flight arrivals for each of the three Washington DC airports to compare the flight traffic for the two holidays.

SAMPLE SOLUTION

############# CAP 2751 R for Data Science Midterm Part B #############

#

# Loading necessary library

library(ggplot2)

library(dplyr)

################## Problem 1 #################

# Question 1

# Reading the mpg dataset

data("mpg")

# Calculating the average miles per gallon for each vehicle

mpg\_data <- mpg %>%

mutate(ave\_mpg = (cty + hwy) / 2)%>%

filter(class =="pickup" | class== "suv")

# Generate new\_mpg dataframe

new\_mpg <- mpg\_data%>%

select(manufacturer, model, displ, class, ave\_mpg)

# Filtering for only SUVs and getting the one(s) with the highest average mpg

best\_SUV <- new\_mpg %>%

filter(class == "suv") %>%

arrange(desc(ave\_mpg)) %>%

slice\_head(n = 5)

# Printing the variable -> best\_SUV

print(best\_SUV)

#Question 2

# Filtering for only PICKUP and getting the one(s) with the highest average mpg

best\_Pickup = new\_mpg %>%

filter(class == "pickup") %>%

arrange(desc(ave\_mpg)) %>%

slice\_head(n = 5)

# Printing the variable -> best\_Pickup

print(best\_Pickup)

#Question 3 (Extra Credit

# Createing a scatter plot

ggplot(new\_mpg, aes(x = displ, y = ave\_mpg, color = class)) +

geom\_point() +

labs(title = "Average MPG vs. Engine Displacement",

x = "Engine Displacement (displ",

y = "Average Fuel Efficiency (MPG)",

color = "Vehicle Class")

##################### Problem 2 ###############################

# Question 1

# Loading the data

library(nycflights13)

library(tidyverse)

data(flights)

flights\_nyc\_2013 = flights

# Filtering for flights going to Washington DC

flights\_nyc\_to\_DC = flights\_nyc\_2013 %>%

filter(dest %in% c("DCA", "IAD", "BWI"))

# Filtering for flights on Valentine's Day and Thanksgiving Day

flights\_DC\_holiday = flights\_nyc\_to\_DC %>%

filter(

(month == 2 & day == 14) | # February 14

(month == 11 & day == 28) ) # November 28

# Count the number of flights per airport and arrange in descending order

airport\_rankings = flights\_DC\_holiday %>%

group\_by(dest) %>%

summarise(flights\_count = n()) %>%

arrange(desc(flights\_count))

# View the airport flght arrival count

print(airport\_rankings)

# Determine the busiest airport based on the number of flight arrivals

busiest\_airport = airport\_rankings[1, ] # This selects the first row, which has the highest count

busiest\_airport\_name = busiest\_airport$dest

busiest\_airport\_count = busiest\_airport$flights\_count

cat("The busiest holiday airport in Washington DC is ", busiest\_airport\_name, " with ", busiest\_airport\_count, " flight arrivals.\n")

# Question 2

# Filter for flights going to Washington DC

flights\_nyc\_to\_DC = flights\_nyc\_2013 %>%

filter(dest %in% c("DCA", "IAD", "BWI"))

# Filter for flights on Valentine's Day and Thanksgiving Day

flights\_DC\_holiday = flights\_nyc\_to\_DC %>%

filter(

(month == 2 & day == 14) | (month == 11 & day == 28))

# Determine the five most popular carriers

top\_carriers = flights\_DC\_holiday %>%

group\_by(carrier) %>%

summarise(flights\_count = n()) %>%

arrange(desc(flights\_count)) %>%

top\_n(5) # selects the top 5 rows

print(top\_carriers)

# Question 2 part 2

# Determine the airline with the worst average delay

worst\_avg\_delay = flights\_DC\_holiday %>%

group\_by(carrier) %>%

summarise(average\_delay = mean(arr\_delay, na.rm = TRUE)) %>%

arrange(desc(average\_delay)) %>%

top\_n(1) # selects the top 1 row

print(worst\_avg\_delay)

#Extra Credit

# Load the data

data(flights)

flights\_nyc\_2013 = flights

# Filter for flights going to Washington DC

flights\_nyc\_to\_DC = flights\_nyc\_2013 %>%

filter(dest %in% c("DCA", "IAD", "BWI"))

# Filter for flights on Valentine's Day and Thanksgiving Day

flights\_DC\_holiday = flights\_nyc\_to\_DC %>%

filter(

(month == 2 & day == 14) |

(month == 11 & day == 28) )

# Create a new variable to distinguish between the holidays

flights\_DC\_holiday = flights\_DC\_holiday %>%

mutate(holiday = ifelse(month == 2 & day == 14, "Valentine's Day", "Thanksgiving"))

# Count the number of flights per airport and holiday

airport\_holiday\_counts = flights\_DC\_holiday %>%

group\_by(dest, holiday) %>%

summarise(flights\_count = n())

# Create the bar chart

ggplot(data = airport\_holiday\_counts, aes(x = dest, y = flights\_count, fill = holiday)) +

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

labs(title = "Number of flight arrivals at each Washington DC airport",

x = "Airport name",

y = "Number of Flights")