Brian Lambert

STA 404

Module 6 Homework

1. **View UN tourism data, 1995-2014**
   1. **Head**

> head(hw6\_analysis)

country year inbound\_tourists expenditure

33 ALBANIA 2007 1062 1479

34 ALBANIA 2008 1247 1848

35 ALBANIA 2009 1711 2014

36 ALBANIA 2010 2191 1780

37 ALBANIA 2011 2469 1833

38 ALBANIA 2012 3156 1623

* 1. **Tail**

> tail(hw6\_analysis)s

country year inbound\_tourists expenditure

4274 VENEZUELA, BOLIVARIAN REPUBLIC OF 2008 699 1097

4275 VENEZUELA, BOLIVARIAN REPUBLIC OF 2009 562 1055

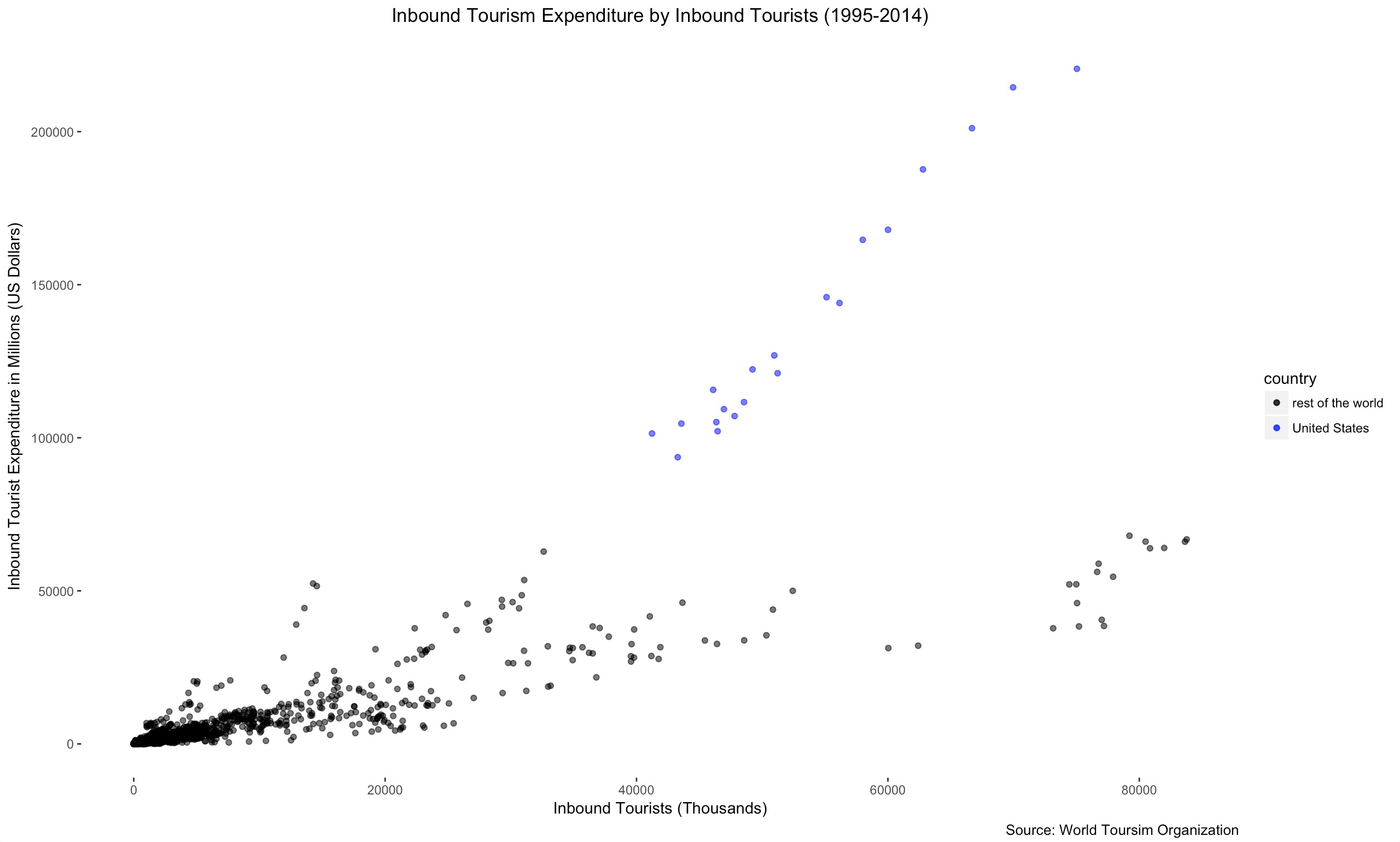
4276 VENEZUELA, BOLIVARIAN REPUBLIC OF 2010 526 794

4277 VENEZUELA, BOLIVARIAN REPUBLIC OF 2011 595 805

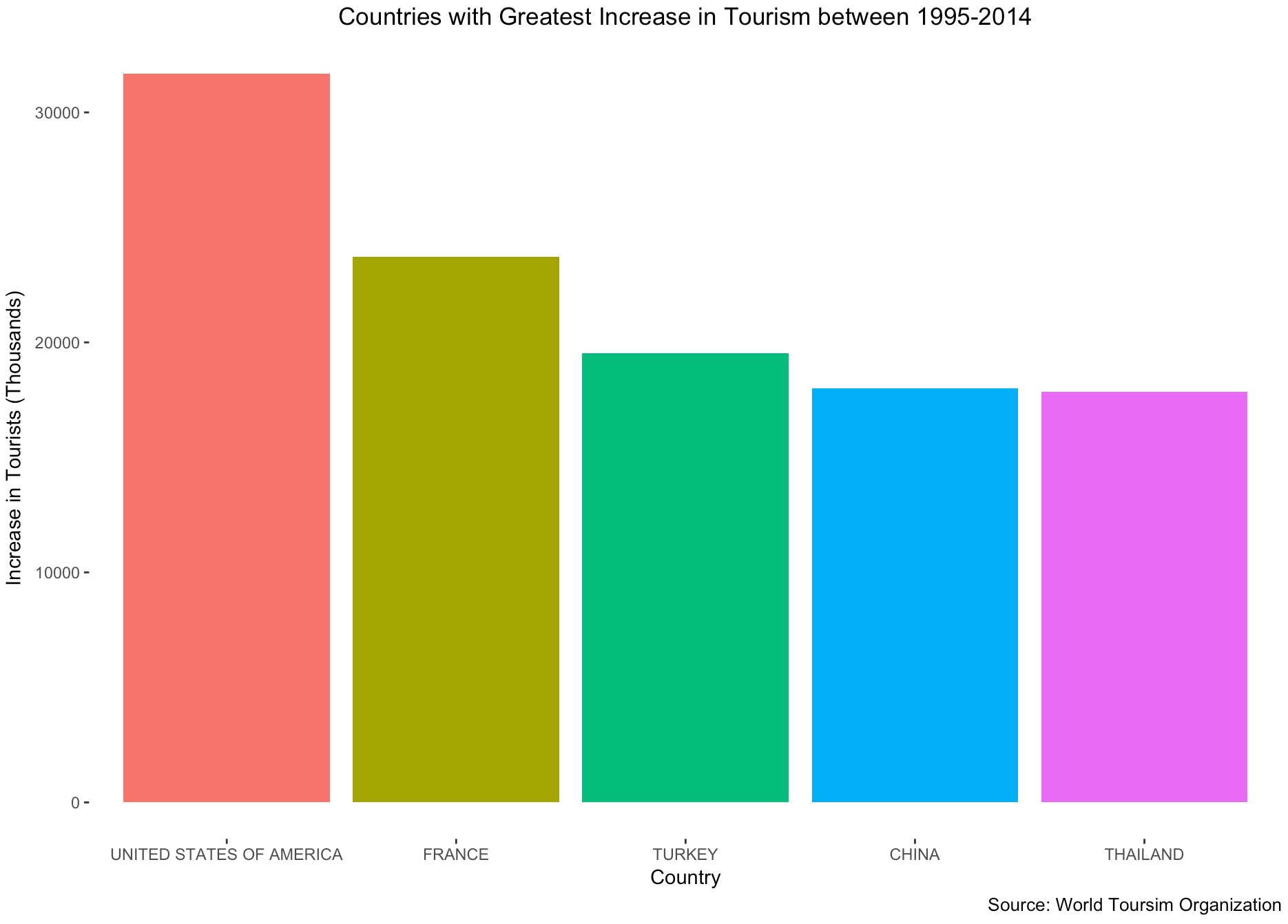
4278 VENEZUELA, BOLIVARIAN REPUBLIC OF 2012 988 904

4279 VENEZUELA, BOLIVARIAN REPUBLIC OF 2013 986 926

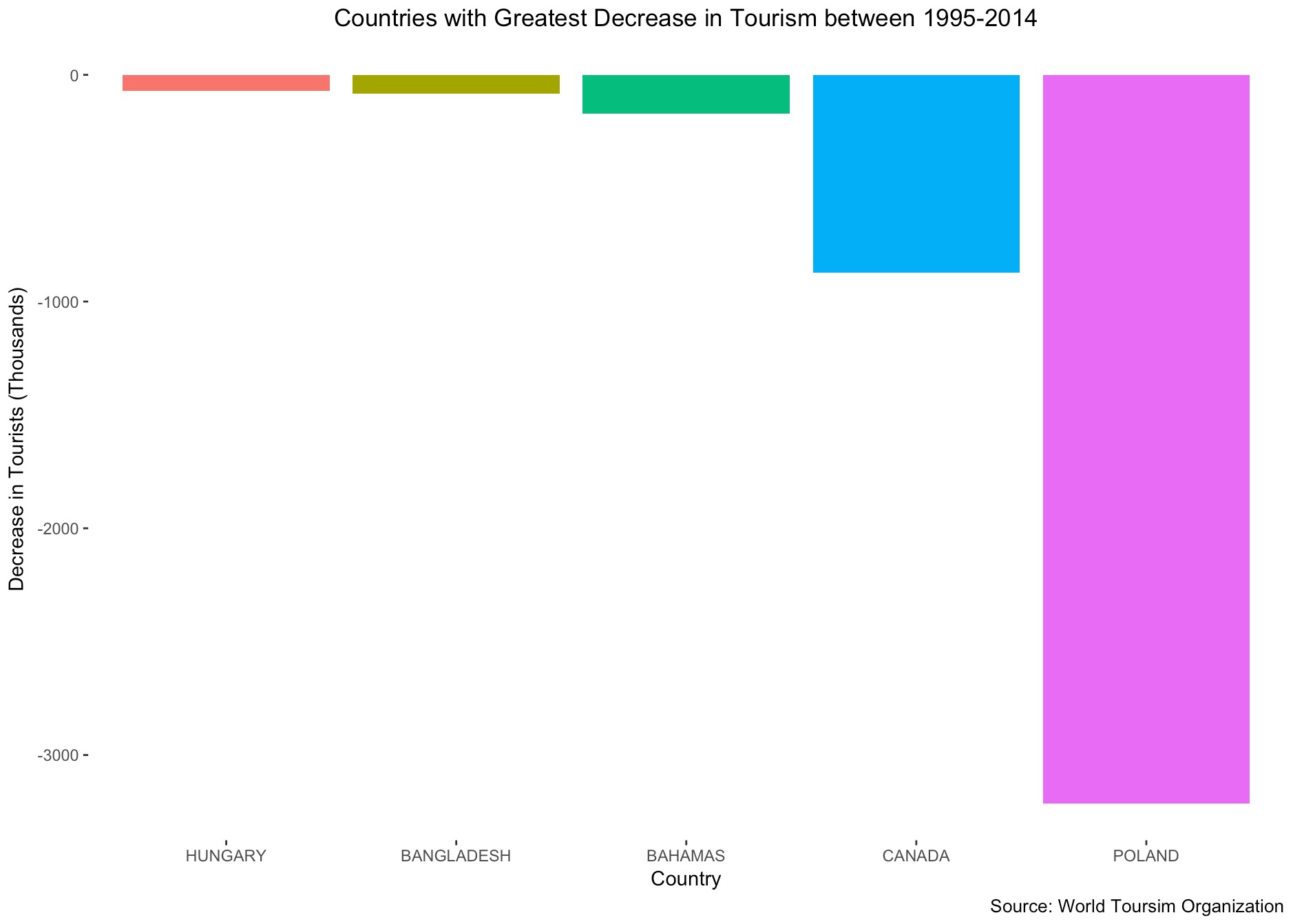
1. **Expenditure (inbound tourism US$million) vs. Tourists (inbound tourists thousands)**

****

1. **Other Plots**
   1. **Countries with greatest growth in tourism, 1995-2014**

****

* 1. **Countries with decrease in tourism, 1995-2014**

****

**CODE:**

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

# #

# Author: Brian Lambert #

# Name: module\_6\_homework\_code.R #

# Description: Generating visualizations to analyze UN tourism #

# data from 1995-2014. #

# #

# setwd(“/Users/brianlambert/Desktop/STA404/Module\_6”) #

# #

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

#========================== Data ==========================

# Preprocessing of Excel file

# deleted 5 rows

# deleted last column ("notes")

# deleted first row (country, series, 1995--2014)

#

# Deleted rows at bottom of the spreadsheet

# The information is the same as that provided in the "Compendium of Tourism Statistics"

# Inbound tourism:

# TF Arrivals of non-resident tourists at national borders

# VF Arrivals of non-resident visitors at national borders

# THS Arrivals of non-resident tourists in hotels and similar establishments

# TCE Arrivals of non-resident tourists in all types of accommodation establishments

# Outbound tourism:

# TF Departures - trips abroad by resident tourists

# VF Departures - trips abroad by resident visitors

# IMF International Monetary Fund

# CB Central Bank

# .. Data not available

# Source: World Tourism Organization (UNWTO)

# read in data ================================================================

library(readr)

hw6\_subset\_csv <- read\_xlsx("/Users/brianlambert/Desktop/STA404/Module 6/tourism\_data.xlsx",

col\_names = FALSE)

hw6\_subset\_csv = tail(hw6\_subset\_csv, -6)

hw6\_subset\_csv = head(hw6\_subset\_csv, -14)

View(hw6\_subset\_csv)

# add country name to each row =================================================

hw6\_out <- data.frame(NULL)

for (ii in 1:nrow(hw6\_subset\_csv)) {

if (!is.na(hw6\_subset\_csv[ii,"X\_\_1"])) {

country <- hw6\_subset\_csv[ii,"X\_\_2"]

}

else {

hw6\_out <- rbind(hw6\_out,

cbind(country, hw6\_subset\_csv[ii,2:24]) )

} # end of 'else'

} # end of 'for'

# clean up names ================================================================

names(hw6\_out) <- c("country", "v2", "group", "series", paste0("X",1995:2014))

# issues: ======================================================================

# countries have different numbers of rows

# inbound and outbound tourism have different numbers of ros

hw6\_out %>%

group\_by(country) %>%

summarize(nvars = n())

# search for presence of patterns - grep

# http://www.statmethods.net/management/functions.html

# chaining if-else statements

# http://www.dummies.com/programming/r/how-to-chain-ifelse-statements-in-r/

hw6\_out2 <- data.frame(NULL)

for (ii in 1:nrow(hw6\_out)) {

if (length(grep("Inbound", hw6\_out[ii,"v2"]) )) {

prefix <- "Inbound"

} else if (length(grep("Outbound", hw6\_out[ii,"v2"]))) {

prefix <- "Outbound"

} else {

vars <- paste(prefix, hw6\_out[ii,"v2"], sep="\_")

hw6\_out2 <- rbind(hw6\_out2,

cbind(vars, hw6\_out[ii,c(1,3:24)]))

}

}

# examining a couple of example cases ===================================================================

as.data.frame(hw6\_out2 %>%

group\_by(country) %>%

summarize(nvars = n())

)

hw6\_out2 %>%

filter(country=="POLAND")

hw6\_out2 %>%

filter(country=="ZIMBABWE")

# clean up variable names ================================================================================

hw6\_out2 <- hw6\_out2 %>%

mutate(new\_vars = sub(" - ","\_",vars)) %>%

mutate(new\_vars = gsub(" ","\_",new\_vars)) %>%

mutate(new\_vars = paste(new\_vars,series,sep="\_"))

unique(hw6\_out2$new\_vars)

hw6\_wide <- hw6\_out2 %>%

select(country, new\_vars, X1995:X2014)

View(hw6\_wide)

# now can think about restructuring the data ============================================================

library(tidyr)

hw6\_long <- hw6\_wide %>%

group\_by(new\_vars) %>%

gather(key=cyear, value=cmeasure, X1995:X2014) %>%

mutate(year = as.numeric(substring(cyear,2)),

measure = as.numeric(sub(',','',cmeasure))) %>% # commas in cell entries!

select(-cyear, -cmeasure)

View(hw6\_long)

hw6\_analysis <- hw6\_long%>%

spread(key=new\_vars, value=measure)

View(hw6\_analysis)

as.data.frame(hw6\_analysis %>%

filter(country=="UNITED KINGDOM")

)

#========================== Head % Tail of UN Tourism Data ==========================

head(hw6\_analysis)

tail(hw6\_analysis)

#========================== Expenditure vs Tourists Plot ==========================

# more data preprocessing to extract relevant variables, remove rows with na values, and clean up variable names

hw6\_analysis = data.frame(hw6\_analysis$country, hw6\_analysis$year, hw6\_analysis$Inbound\_Arrivals\_Thousands\_TF, hw6\_analysis$`Inbound\_Tourism\_expenditure\_in\_the\_country\_US$\_Mn\_IMF`)

hw6\_analysis = na.omit(hw6\_analysis)

names(hw6\_analysis) = c("country", "year", "inbound\_tourists", "expenditure")

# comparing the inbound tourism expendidture with the number of tourists from 1995-2014

# with the United States data highlighted

ggplot(hw6\_analysis, aes(x = inbound\_tourists, y = expenditure)) +

geom\_jitter(data = hw6\_analysis[which(hw6\_analysis$country == "UNITED STATES OF AMERICA"),],

aes(color = "blue"),

alpha = 0.6) +

geom\_jitter(data = hw6\_analysis[which(hw6\_analysis$country != "UNITED STATES OF AMERICA"),],

aes(color = "black"),

alpha = 0.6) +

ggtitle("Inbound Tourism Expenditure by Inbound Tourists (1995-2014)") +

labs(x="Inbound Tourists (Thousands)", y="Inbound Tourist Expenditure in Millions (US Dollars)",

caption="Source: World Toursim Organization", size = 8) +

theme(plot.title = element\_text(hjust = 0.5),

panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(),

panel.border = element\_blank(),

panel.background = element\_blank()) +

# source: https://groups.google.com/forum/#!topic/ggplot2/GD2kZuOQS1Q

scale\_colour\_manual(name = 'country',

values =c('blue'='blue', 'black'='black'), labels = c('rest of the world','United States'))

#========================== Greatest Tourism Growth by Country Plot ==========================

# create data frame ordering the countries by their increase in tourists between their first year

# reflected and last year reflected in the data.

countries = unique(hw6\_analysis$country)

growth\_vector = c()

country\_vector = c()

for (c in countries) {

new\_df = hw6\_analysis[which(hw6\_analysis$country == c),]

min\_year = min(new\_df$year)

min\_year\_tourists = new\_df[which(new\_df$year == min\_year), 'inbound\_tourists']

max\_year = max(new\_df$year)

max\_year\_tourists = new\_df[which(new\_df$year == max\_year), 'inbound\_tourists']

growth\_vector = append(growth\_vector, max\_year\_tourists - min\_year\_tourists)

country\_vector = append(country\_vector, c)

}

country\_growth = data.frame(growth\_vector, country\_vector)

country\_growth = country\_growth[order(growth\_vector),]

names(country\_growth) = c("growth", "country")

country\_growth$country <- factor(country\_growth$country, levels = country\_growth$country[order(-country\_growth$growth)])

# plots showing the countries with the greatest growth in tourism from 1995-2014

ggplot(data = tail(country\_growth, n = 5), aes(x = country, y = growth, fill = country)) +

geom\_bar(stat = "identity") +

ggtitle("Countries with Greatest Increase in Tourism between 1995-2014") +

labs(x="Country", y="Increase in Tourists (Thousands)",

caption="Source: World Toursim Organization", size = 8) +

theme(plot.title = element\_text(hjust = 0.5),

panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(),

panel.border = element\_blank(),

panel.background = element\_blank(),

legend.position = "none")

#========================== Greatest Tourism Decline by Country Plot ==========================

# plots showing the countries with the greatest decline in tourism from 1995-2014

# plots showing the countries with the greatest growth in tourism from 1995-2014

ggplot(data = head(country\_growth, n = 5), aes(x = country, y = growth, fill = country)) +

geom\_bar(stat = "identity") +

ggtitle("Countries with Greatest Decrease in Tourism between 1995-2014") +

labs(x="Country", y="Decrease in Tourists (Thousands)",

caption="Source: World Toursim Organization", size = 8) +

theme(plot.title = element\_text(hjust = 0.5),

panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(),

panel.border = element\_blank(),

panel.background = element\_blank(),

legend.position = "none")