Lab 2

Maya Casper

2024-02-19

## Section 1

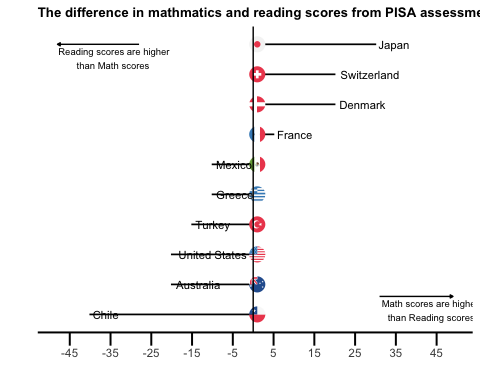
In Section 1, we will focus on analyzing PISA data we used in class. This dataset has math and reading scores from PISA assessment for 10 countries.

To begin working with this dataset, you are required to import and preprocess the data using the following code snippet provided below:

require(countrycode)  
library(here)  
library(ggplot2)  
library(ggtext)  
  
pisa <- read.csv(here('Lab 2/pisa.csv'))  
pisa$iso <- tolower(codelist[codelist$country.name.en %in% pisa$Country,]$iso2c)  
pisa$diff <- pisa$Math - pisa$Reading

#install.packages("ggflags", repos = c(  
 # "https://jimjam-slam.r-universe.dev",  
 #"https://cloud.r-project.org"))  
  
library(ggflags)  
  
pisa <- read.csv(here('Lab 2/pisa.csv'))  
pisa$iso<- tolower(codelist[codelist$country.name.en %in% pisa$Country,]$iso2c)  
pisa$diff <- pisa$Math - pisa$Reading  
countlabel<-pisa [1:10,1]  
  
  
ggplot(pisa, aes(x = reorder(Country, diff, decreasing = FALSE), y = diff)) +  
 geom\_bar(stat = "identity", width = 0.01, color="black")+  
 geom\_flag(aes(country = iso), size = 5, y = 1, color = "black") + #### Need to determine the local of flag   
 geom\_text(aes(label = Country), hjust = -0.1, vjust = 0.5, size = 3, color = "black") + ## need to move them   
 coord\_flip() +  
 scale\_y\_continuous(breaks = seq(-45, 45, by = 10),  
 limits = c(-48, 49)) +  
 geom\_hline(yintercept = 0, color = "black", linetype = "solid")+  
 labs(title = '\*\*The difference in mathmatics and reading scores from PISA assessment\*\*',  
 x = '',  
 y = '') +  
 theme(plot.title= element\_markdown(margin = margin(b=5),hjust=0, size = 10),  
 axis.text.y = element\_blank(),  
 panel.background = element\_rect(fill='white',colour='white'),  
 axis.ticks= element\_blank(),  
 plot.caption= element\_text(hjust = 0.01,   
 size = 9,  
 margin=margin(t=0)),  
 axis.line.x = element\_line(color = "black", linewidth = 0.7),  
 axis.ticks.x = element\_line(color='black',linewidth = .7),   
 axis.ticks.length=unit(.35, "cm"))+  
annotate('text',  
 x = 9.2,  
 y = -45,  
 hjust = 0.1,  
 vjust = 0,  
 label = 'Reading scores are higher   
 than Math scores',  
 size = 2.5,  
 color = 'black')+   
 annotate('text',  
 x = .8,  
 y = 34,  
 hjust = 0.1,  
 vjust = 0,  
 label = 'Math scores are higher   
 than Reading scores',  
 size = 2.5,  
 color = 'black')+   
 annotate("segment", x = 10, y = -28, xend = 10, yend=-48,  
 arrow = arrow(type = "closed", length = unit(0.01, "npc")))+  
 annotate("segment", x = 1.6, y = 31, xend = 1.6, yend=49,  
 arrow = arrow(type = "closed", length = unit(0.01, "npc")))

## Warning in geom\_flag(aes(country = iso), size = 5, y = 1, color = "black"):  
## Ignoring unknown parameters: `colour`



## Section 2

In Section 2, our analysis will focus on world population. This dataset has the population for 266 countries from 1960 to 2022.

To begin, you should import the dataset using the following code snippet. This code will first filter the countries and include only European Union countries with at least 10 million people. It will also compute a % change in population from 1960 to 2022.

library(pacman)  
p\_load(ggtext)  
library(tidyr)  
library(ggplot2)  
library(ggtext)  
pop <- read.csv(here('Lab 2/population.csv'))  
  
EU <- c("Austria", "Belgium", "Bulgaria", "Croatia", "Cyprus",   
 "Czechia", "Denmark", "Estonia", "Finland", "France",   
 "Germany", "Greece", "Hungary", "Ireland", "Italy", "Latvia",   
 "Lithuania", "Luxembourg", "Malta", "Netherlands", "Poland",  
 "Portugal", "Romania", "Slovak Republic", "Slovenia", "Spain","Sweden")  
  
pop <- pop[pop$Country.Name %in% EU,]  
pop <- pop[which(pop$X2022>10000000),]  
pop$per\_change <- ((pop$X2022 - pop$X1960)/pop$X2022)\*100  
  
  
colnames(pop) <- gsub('X', '', colnames(pop))  
  
pop\_full <- pop %>%   
 pivot\_longer(cols = starts\_with(c('1','2')),  
 names\_to = 'Year',  
 values\_to = 'Population')   
  
pop\_full$Year <- as.numeric(pop\_full$Year)  
  
  
my\_color <- rep('gray',length(unique(pop\_full$Country.Name))) #creates everything grey   
 names(my\_color) <- unique(pop\_full$Country.Name)  
 my\_color["Spain"] <- 'blue'  
 my\_color["Netherlands"] <- 'orange'

plot<- ggplot(pop\_full,  
 aes(x = Year, y = Population, colour = Country.Name)) +  
 geom\_line(stat = 'identity')+  
 scale\_color\_manual(values=my\_color)+  
 guides(color='none', size = 'none') + # Hide the legends for color and size  
 labs(title="Spain and Netherlands are the two countries with largest population growth in European Union (1960 -2022)",  
 subtitle = '(Among countries with at least 10 million people)',  
 x = '',  
 y = ' Population (in millions)',  
 caption = "Source: OECD Stats \nhttps://stats.oecd.org/") +  
 theme(plot.title= element\_markdown(margin = margin(b=5),hjust=0, size = 15, face ='bold'),  
 panel.background = element\_rect(fill='white',colour='white'),  
 axis.ticks= element\_blank(),  
 panel.grid.major.y = element\_line(color = "grey80",   
 linewidth = 0.5,   
 linetype = "dashed"),  
 plot.caption= element\_text(hjust = 0.01,   
 size = 9,  
 margin=margin(t=0)),  
 axis.line.x = element\_line(color = "black", linewidth = 0.7),  
 axis.ticks.x = element\_line(color='black',linewidth = .7),   
 axis.ticks.length=unit(.35, "cm"))  
print(plot)

## A graph of different countries/regions with blue and orange lines Description automatically generatedSection 3

In Section 3, our analysis will focus on the relationship between online hotel revenue and the number of travel agents over time using a connected scatterplot. To begin, you should import the dataset using the following code snippet.

hotel <- read.csv('hotel.csv',fileEncoding="UTF-8-BOM")  
hotel$travel\_agents <- hotel$travel\_agents/1000  
  
 # number of travel agents are in thousands  
 # hotel revenue is in billion dollars  
  
hotel

year travel\_agents hotel\_revenue  
1 2000 123.385 12.95  
2 2001 110.583 19.95  
3 2002 104.046 28.02  
4 2003 103.501 40.12  
5 2004 90.428 51.16  
6 2005 88.521 64.10  
7 2006 87.431 79.81  
8 2007 85.252 89.79  
9 2008 86.070 94.46  
10 2009 76.809 90.00  
11 2010 70.272 99.76  
12 2011 67.276 116.11  
13 2012 64.552 124.60  
14 2013 64.280 143.49  
15 2014 63.975 155.38

library(ggplot2)  
library(ggtext)  
first <- hotel[1:5,1:3]  
second<- hotel [5:9, 1:3]  
third<-hotel[9:10, 1:3]  
forth<-hotel[10:15, 1:3]  
label1<-hotel[1:15,1]  
   
 ggplot() +   
 geom\_point(data=first,aes(x=hotel\_revenue,y=travel\_agents), size=2, shape=1) +  
 geom\_path(data=first,aes(x=hotel\_revenue,y=travel\_agents),color='#b75b37',linewidth=0.75) +   
 geom\_point(data=second,aes(x=hotel\_revenue,y=travel\_agents),size=2, shape=1) +  
 geom\_path(data=second,aes(x=hotel\_revenue,y=travel\_agents),color='#0e73b0',linewidth=0.75) +   
 geom\_point(data=third,aes(x=hotel\_revenue,y=travel\_agents),size=2, shape=1) +  
 geom\_path(data=third,aes(x=hotel\_revenue,y=travel\_agents),color='#57a6ac',linewidth=0.75) +   
 geom\_point(data=forth,aes(x=hotel\_revenue,y=travel\_agents),size=2, shape = 1) +  
 geom\_path(data=forth,aes(x=hotel\_revenue,y=travel\_agents),color='#b75b37',linewidth=0.75) +  
 geom\_text(data = hotel, aes(x = ifelse(label1 == "2008", hotel\_revenue + 2, hotel\_revenue), y = travel\_agents, label = label1),   
 hjust = 0, vjust = -1.5, size = 2, color = "black") +  
 scale\_x\_continuous(  
 breaks = seq(0,180,30),  
 limits = c(0,185),  
 expand = c(0,0),  
 labels = paste0(format(seq(0,180,30),digits = 1),'B')) +  
 scale\_y\_continuous(  
 breaks = seq(0,140,20),  
 limits = c(0,150),  
 expand = c(0,0),  
 labels = paste0(format(seq(0,150,20),digits = 1),'K'))+  
 labs(title = "\*\*Online Hotel Revenue vs. Number of Travel Agents\*\*",   
 subtitle = '',  
 x = 'Online Hotel Revenue (USD)',  
 y = 'Number of Travel Agents')+   
 theme(plot.title = element\_markdown(margin = margin(b=10),  
 hjust = -0.1),  
 axis.title.x = element\_markdown(size = 8,   
 vjust = 1.2,   
 hjust = 0),  
 axis.title.y = element\_markdown(size = 8,   
 vjust = 1.075,  
 margin = margin(r =0,b=0)),   
 panel.background = element\_blank(),  
 panel.grid.major.y = element\_line(color='gray',linetype='dashed',linewidth=0.25),  
 panel.grid.major.x = element\_line(color='gray',linetype='dashed',linewidth=0.25),  
 axis.ticks.x = element\_blank(),  
 axis.ticks.y = element\_blank(),   
 axis.line.y = element\_line(colour='lightgray'),   
 axis.text = element\_markdown(size= 7,   
 vjust = 1))+  
 annotate('text',  
 x = 30,  
 y = 120,  
 hjust = .1,  
 vjust = 0,  
 label = 'Between 2000 and 2004, online travel revenue increased   
 while the number of travel agents decreased',  
 size = 2.5,  
 color = '#b75b37') +  
 annotate('text',  
 x = 60,  
 y = 95,  
 hjust = 0.1,  
 vjust = 0,  
 label = 'From 2004 to 2008 , online hotel revenues continued to increase   
 while the number of travel agents decreased steadily',  
 size = 2.5,  
 color = '#0e73b0')+  
 annotate('text',  
 x = 35,  
 y = 58,  
 hjust = 0.1,  
 vjust = 0,  
 label = 'From 2008 to 2009 , online hotel revenue  
 ecreased a little but for the first time   
 ince 2000 along with a sudden decrease   
 in the travel of agents',  
 size = 2.5,  
 color = '#57a6ac')+  
 annotate('text',  
 x = 102,  
 y = 85,  
 hjust = 0.1,  
 vjust = 0,  
 label = 'Recession',  
 size = 2.5,  
 color = '#57a6ac')+  
 annotate('text',  
 x = 120,  
 y = 52,  
 hjust = 0.1,  
 label = 'Online hotel revenue recovered and continued   
 to increase while the number of travel agents   
 kept decreasing and stablized',  
 size = 2.5,  
 color = '#b75b37')+  
 annotate("segment", x = 94.5, y = 86.1, xend = 100, yend=86.1,  
 arrow = arrow(type = "closed", length = unit(0.01, "npc")))

