PSY6422 Data Management and Visualisation

220225650

2023-05-11

#Why are we interested in travel?

The aim of this project is to better understand whether there were any changes in the destinations people were travelling from the United Kingdom (UK) to over time. This was chosen because over the past 10 years there have been a number of events in the UK that are likely to have influenced destination choice. Examples include: austerity measures, the UK exiting the European Union and the Coronavirus pandemic.

##Key research question

Is travel from the UK in 2021 similar to travel in 2009?

#What is the data source for this project?

The data was provided by the UK's Office for National Statistics and accessed in May 2023 at their Travel Trends Dataset Pages. The specific files used were: * 2017 to 2021 edition of the data set and * 2009 to 2019 edition of the data set.

I have used the "Number of visits to specified countries: by main country visited and nationality 2017 to 2021" as this allowed a more nuanced view of travel which a region level may not pick up.

##How was the data collected?

This data provides estimated visits based on the International Passenger Survey (IPS). This survey uses interviews at all major airports, sea routes and Eurostar/ Eurotunnel terminals with approximately 250,000 people (although up to 800,000 interviews take place). However in 2021 the numbers takling part in the IPS were much lower. There is no further information on the sampling methods however at the scale and coverage these are trustworthy estimates. Additionally no other data sources exist at this level of detail.

A new method for calculating the 2021 estimates were introduced which superceeds previous years. The estimates that overlap are the same in the 2009 to 2019 datasets so I have continued with the merge but this should be noted.

##Is there anything that may impact the visualisation?

The data for 2021 is more limited than previous years of the IPS. The smaller number of travellers and issues collecting the data at key terminals led to an overall smaller sample of passengers in the IPS. This means that there is more error in the estimation than other years (excluding 2020 where no country level data is available). Countries are consistent every year but if this changes then some of the data importing and colours will need changed.

##Codebook

A full codebook describing the variables from the origional dataset and any calculated values are available on the github page.

#How was the data prepared for visualisation?

##Libraries and RENV

The following packages were used in this project:

```
renv::restore()
#Libraries
library(here)#Here for loading in the data
library(tidyverse)#Tidyverse for data management
library(ggplot2) #chart making
library(readxl)#Reading in Excel sheets as that is how the data is stored
library(plotly)#Making scroll over charts
library(htmlwidgets)#To save interactive chart
```

To support future replication the this R Project used RENV (version:0.17.3) with package versions available in the following file link

##Importing the data

###Cleaning variables

As an initial step I created a set of values that supported the importation and data cleaning processes. In theory these will allow for a more recent dataset to also be used by changing the input file and changing the years that we are keeping.

```
# 2. Load the data in

#Variables required for data cleaning of the ONS files

ONSList <- 68 #The length of the ONS data table, these are consistent

yearsdf1721 <- c("2017","2018","2019",
                 "2020","2021") # These are the years that we will want to take from this data frame

yearsdf0919 <- c("2009","2010","2011",
                 "2012","2013","2014",
                 "2015","2016") #Years wanted, always choose most recent

summaries <- list('Total World',
                  'Other Countries',"Europe",
                  '- of which EU',
                  '- of which EU Oth',
                  '- of which EU15',
                  "North America") #summary values used in the dataset

# Raw data locations
data1721 = (here("raw_data","section3ukresidentsvisitsabroad2017to2021.xlsx")) #data for 2017 to 2021
data0919 = (here ("raw_data","section3ukresidentsvisitsabroad2009to2019.xlsx")) #data for 2009 to 2017
```

###Importing the data

The following code imports the data into R. It then renames the columns to match those in the original ONS data file and removes the rows that are blank.

```
# Load in latest dataset 2017-2021
df1721<- read_excel(data1721,
                    sheet="3.06",
                    skip=10,
                    n_max=ONSList)
```

```
##rename columns 2017-2021
df1721_n <- tibble(x = 1:9, y = c("country", "coltoremove", "2017","2018",
                                  "2019","2020","2021","blankroremove",
                                  "avgrowth1519"))#list of column names

names(df1721) <- df1721_n %>% select(y) %>% pull()#function to change the names

##Remove NAs in rows and select years only 2017 to 2021
df1721 <- df1721 %>%
          select("country",all_of(yearsdf1721)) %>% #Select correct years
          drop_na("country")# remove blank rows

## Read the excel file for 2009 to 2019
df0919 <- read_excel(data0919,
                     sheet="3.10",
                     skip=10,
                     n_max=ONSList)

###rename columns 2009 - 2019
df0919_n <- tibble(x=1:19, y= c("country",
                                "coltoremove",
                                "2009","2010","2011",
                                "2012","2013","2014",
                                "2015","2016","2017",
                                "2018","2019",
                                "blank1",
                                "change1819",
                                "blank2",
                                "growth1819",
                                "blank3",
                                "Avgrowth1519"))#list of column names

names(df0919) <- df0919_n %>% select(y) %>% pull() #function to change the names

###Remove NAs in rows and select years only to 2017
df0919 <- df0919 %>%
          select("country",all_of(yearsdf0919)) %>% #Select correct years
          drop_na("country")# remove blank rows
```

###Merge the files Once both data sets were in the same shape so that I could then merge into a data set that has all available years. I then removed the summary variables.

```
#Merge the datasets
df0921 <- left_join(df0919, df1721,by="country")

##Remove summaries in the dataset
df0921 <-df0921 %>%
         filter((!country %in% summaries))
```

###Use of ranks As 58 countries is tricky to visualise I decided to choose a suitable cut off based on ranks for the visualisation. These were added here using the following code.I decided on 2021 to be the key year for this cut off because it is the most recent and presents a “now” vs the “past” view that is more intuitive.

Add ranking variable for 2021 (this will be used to decide on the cases kept)

```
## Add ranking variable for 2021 (this will be used to decide on the cases kept)
df0921 <- df0921 %>%
          mutate(rank_cut = rank(-`2021`, ties.method = "average"))

#save combined data set
comData_n = paste(here("created_data"), "/travel0921.csv", sep = "")
write.csv(df0921, comData_n)   #Write data to CSV
```

```
#check data
head(df0921)
```

```
## # A tibble: 6 x 15
##   country  `2009` `2010` `2011` `2012` `2013` `2014` `2015` `2016` `2017` `2018`
##   <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Canada     517.   498.   541.   471.   463.   552.   528.   658.   684.   779.
## 2 USA       3538.  3922.  4071.  3744.  3831.  4197.  4398.  4382.  4439.  4943.
## 3 Austria    760.   704.   624.   620.   714.   630.   666.   909.   700.   695.
## 4 Belgium   1541.  1660.  1798.  2074.  2013.  2189.  2047.  1897.  2063.  2087.
## 5 Bulgaria   268.   254.   325.   337.   325.   378.   343.   477.   596.   691.
## 6 Czech R~   469.   423.   453.   414.   422.   378.   512.   600.   687.   625.
## # i 4 more variables: `2019` <dbl>, `2020` <chr>, `2021` <dbl>, rank_cut <dbl>
```

#Visualisation 1 The idea with visualisation 1 was to answer the question using the number

#What data are we using for this project? ##About the IPS ##Raw data source ##Aditional data created #How was the data prepared? Explain that due to the repeating nature hoped to make a set of code that could add in the ##Packages and RENV Include a table of versions ##Removing additional information from the Excel sheets ##Merge the data ##Creating a ranking variable

#What would you have done differently?