TidyVerse_Assignment1

Vishal Arora & Samriti Malhotra

May 1, 2019

The Ask:- from this task is to show the usage of one of more tidy verse packages using the data set from wither kaggle | 538 site

We got the the data from Kaggle World Happiness Report | Spotify Top 100 of 2018

The world happiness has 3 csv files and spotify has one csv file, we are using the list.files function of base package to get all the csv matching the pattern and load them using the map_df function of purrr , passing map_df the read_csv function of readr library to read csv files and load them into respective Data Frames.

```
files <- list.files(".", pattern = "[^201?]{+}.csv", full.names = TRUE)

kable(files) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
  row_spec(0, background ="gray")
```

x
./2015.csv
./2016.csv
./2017.csv
./top2018.csv

```
happinessDF <- map_df(files[-4],read_csv)
spotifyDF <- map_df(files[4], read_csv)



kable(head(happinessDF)) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
  row_spec(0, background ="gray")
```

Country	Region	Happiness Rank	Happiness Score	Standard Error	Economy (GDP
Switzerland	Western Europe	1	7.587	0.03411	
Iceland	Western Europe	2	7.561	0.04884	
Denmark	Western Europe	3	7.527	0.03328	
Norway	Western Europe	4	7.522	0.03880	
Canada	North America	5	7.427	0.03553	
Finland	Western Europe	6	7.406	0.03140	

```
kable(head(spotifyDF)) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
  row_spec(0, background ="gray")
```

id	name	artists	genre	danc
6DCZcSspjsKoFjzjrWoCd	God's Plan	Drake	Hip-Hop/Rap	
3ee8Jmje8o58CHK66QrVC	SAD!	XXXTENTACION	Hip-Hop/Rap	
0e7ipj03S05BNilyu5bRz	rockstar (feat. 21 Savage)	Post Malone	Hip-Hop/Rap	
3swc6WTsr7rl9DqQKQA55	Psycho (feat. Ty Dolla $ign)	Post Malone	Hip-Hop/Rap	
2G7V7zsVDxg1yRsu7Ew9R	In My Feelings	Drake	Hip-Hop/Rap	
7dt6x5M1jzdTEt8oCbisT	Better Now	Post Malone	Hip-Hop/Rap	

Taking spotifyDF data frame, we use split function to split the data based on column value(mode),then there are three calls to purrr functions in the below code . The first map(~ lm()) call creates a list of "lm" objects; the second map(summary) call creates a list of "summary.lm" objects; the third map_dbl() creates a vector of double-precision values.

We can clearly the r-square values based for "MODE" based on danceability & Energy factor for the song.

```
spotifyDF %>%
  split(.$mode) %>%
  map(~ lm(danceability ~ energy, data = .)) %>%
  map(summary) %>%
  map_dbl("r.squared")
```

```
##          0          1
## 0.04018521 0.09347519
```

Taking happinessDF data frame , and filtering out the rows based on !is.na(Region) i.e. any row which has na for column Region should be left out of the data frame, and then using the select method to subset the data frame and the using the split function of base package to split my new subset of original dataframe based on Region. Then using three calls to purrr functions in the below code , The first map(~ lm()) call creates a list of "lm" objects; the second map(summary) call creates a list of "summary.lm" objects; the third map_dbl() creates a vector of double-precision values.

We can clearly see the R2 (r-square) for various Regions based on Family and Economy(GDP per Capita).

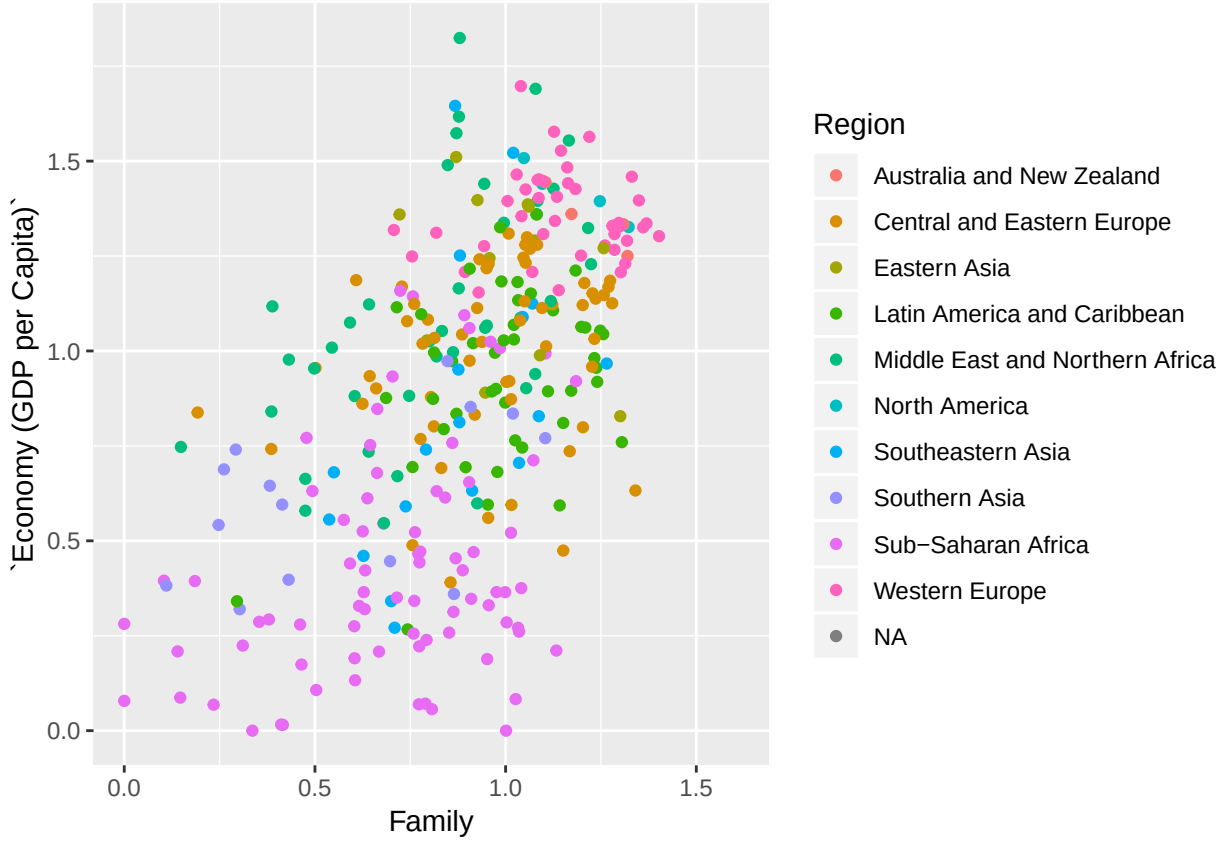
```
regionRSquare <- happinessDF %>%
  filter(!is.na(Region)) %>%
  select(Region,Family,`Economy (GDP per Capita)`) %>%
  split(.$Region) %>%
  map(~ lm(Family ~ `Economy (GDP per Capita)`, data=.)) %>%
  map(summary) %>%
  map_dbl("r.squared")

  kable(regionRSquare) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
  row_spec(0, background ="gray")
```

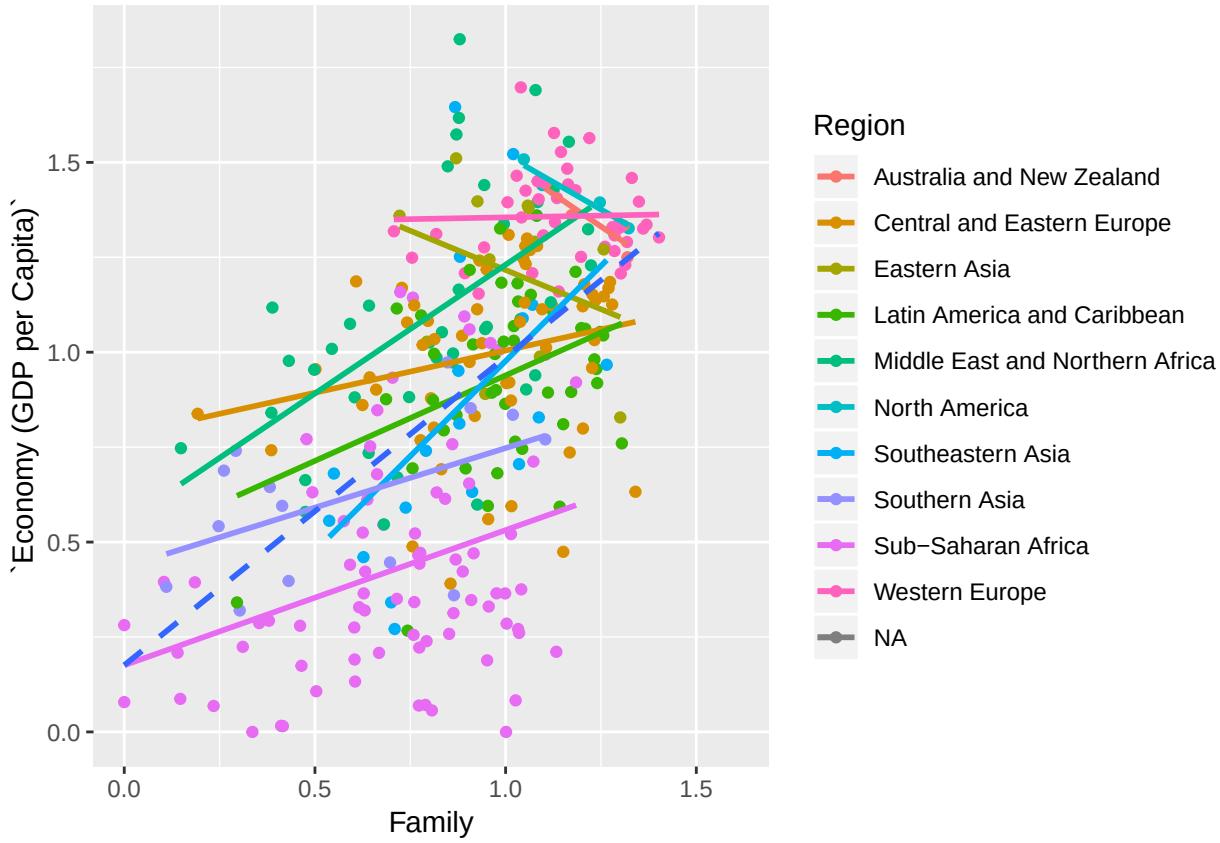
	x
Australia and New Zealand	0.8080644
Central and Eastern Europe	0.0532958
Eastern Asia	0.1007601
Latin America and Caribbean	0.1450172
Middle East and Northern Africa	0.3041670
North America	0.9323728
Southeastern Asia	0.2813623
Southern Asia	0.2457082
Sub-Saharan Africa	0.1071135
Western Europe	0.0007291

Using the ggplot2 library from tidyverse package to draw scatter plt for various Regions using Family vs Economy .
Then extending the same and plotting the linear model for each subgroup(i.e. Region) , setting the method=lm inside the geom_smooth function.

```
r  ggplot(happinessDF,aes(x=Family,y=`Economy (GDP per Capita)`,col=Region))+   geom_point()
```




```
r   ggplot(happinessDF,aes(x=Family,y=`Economy (GDP per Capita)`,col=Region))+   geom_point()+
geom_smooth(method="lm",se=FALSE) +   geom_smooth(aes(group=1),method="lm",se=FALSE,linetype=2)
```


`Economy (GDP per Capita)`
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
Family
Region
Australia and New Zealand
Central and Eastern Europe
Eastern Asia
Latin America and Caribbean
Middle East and Northern Africa
North America
Southeastern Asia
Southern Asia
Sub−Saharan Africa
Western Europe
NA