proj_447

zihan zhou

include libraries for data cleaning, data visuilazation

```
library(data.table)
library(curl)
```

```
## Using libcurl 7.68.0 with GnuTLS/3.6.13
```

```
library(tidyverse)  # package for data manipulation
```

```
## ── Attaching packages ───────────────────────────────────────────────────
─────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.5     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.4     ✓ stringr 1.4.0
## ✓ readr   2.0.2     ✓ forcats 0.5.1
```

```
## ── Conflicts ──────────────────────────────────────────────────────────────
────── tidyverse_conflicts() ──
## x dplyr::between()    masks data.table::between()
## x dplyr::filter()     masks stats::filter()
## x dplyr::first()      masks data.table::first()
## x dplyr::lag()        masks stats::lag()
## x dplyr::last()       masks data.table::last()
## x readr::parse_date() masks curl::parse_date()
## x purrr::transpose()  masks data.table::transpose()
```

```
library(ggrepel)   # packages for plotting
library(sf)# package for working with spatial data; sf has functions compatible with ggplot
```

```
## Linking to GEOS 3.8.0, GDAL 3.0.4, PROJ 6.3.1
```

```
library(rnaturalearth)  # package for loading world map
library(rnaturalearthdata) #  same as above
library(countrycode)
library(dplyr)
library(ggplot2)
```

Implementing data cleaning and data wrangling

```
happiness <- fread("World Happiness Report (2021).csv")
```

Using Package(countrycode) to assign each country to according region, and add the column to data.table(happiness)

```
happiness_ <- as_tibble(data.table::fread("World Happiness Report (2021).csv"))
source<- pull(happiness_,Entity)
country_code <- countrycode(source, origin = "country.name",destination = "region")
happiness$countrycode_dt <- as.data.table(country_code)
names(happiness)[names(happiness) == colnames(happiness)[3] ] <- "life_satisfaction"
names(happiness)[names(happiness) == colnames(happiness)[4] ] <- "region"
```

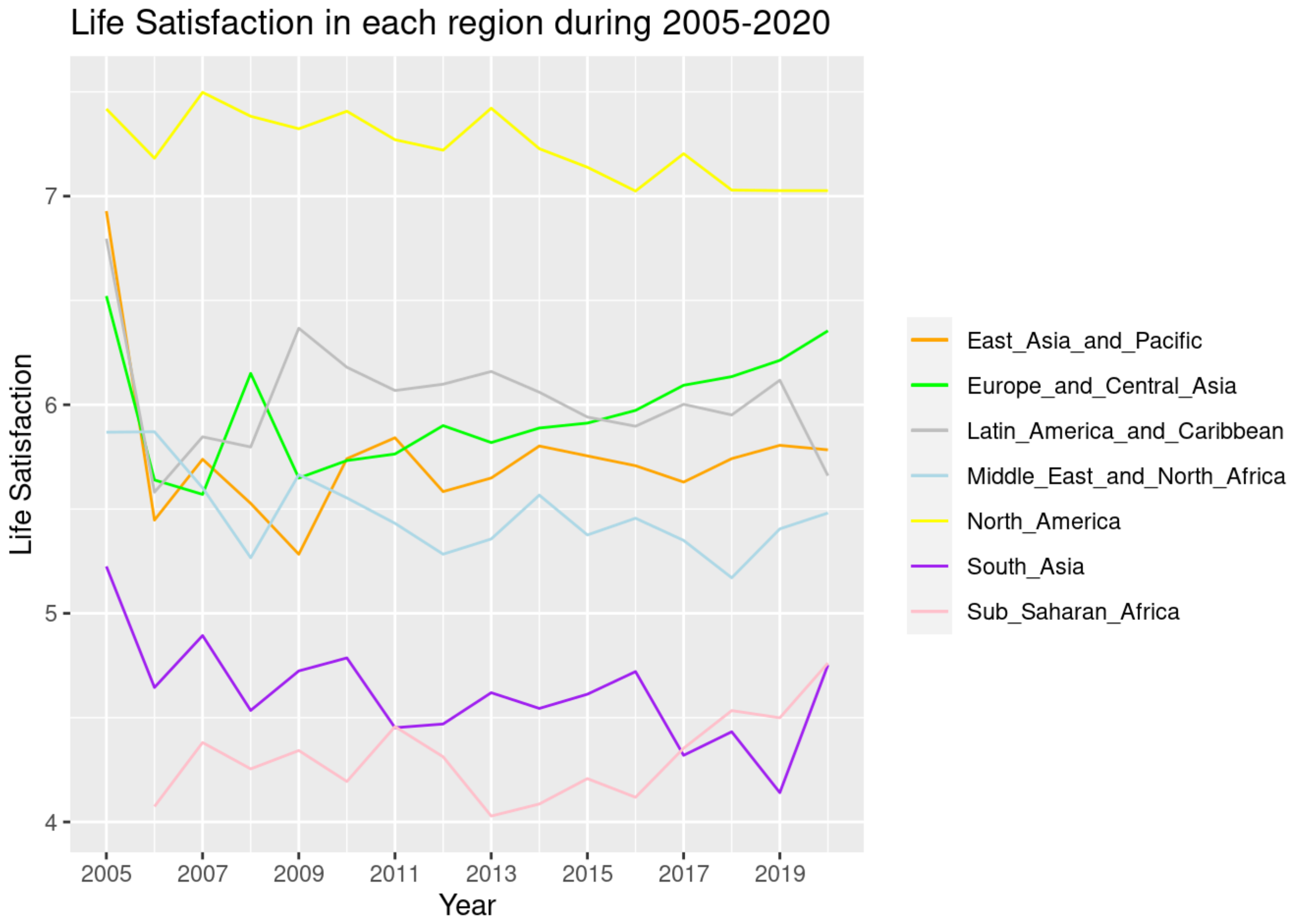
Implement ‘dcast’ method to create desired data.table, and change the columnnames to avoid “&” and blank space.

```
happiness_1<- dcast(happiness, Year~region, value.var = "life_satisfaction", fun= list(mean), d
rop= FALSE)
colnames(happiness_1)[2] = "East_Asia_and_Pacific"
colnames(happiness_1)[3] = "Europe_and_Central_Asia"
colnames(happiness_1)[4] = "Latin_America_and_Caribbean"
colnames(happiness_1)[5] = "Middle_East_and_North_Africa"
colnames(happiness_1)[6] = "North_America"
colnames(happiness_1)[7] = "South_Asia"
colnames(happiness_1)[8] = "Sub_Saharan_Africa"
```

Data visulization

plot “Life Satisfaction in each region during 2005-2020”

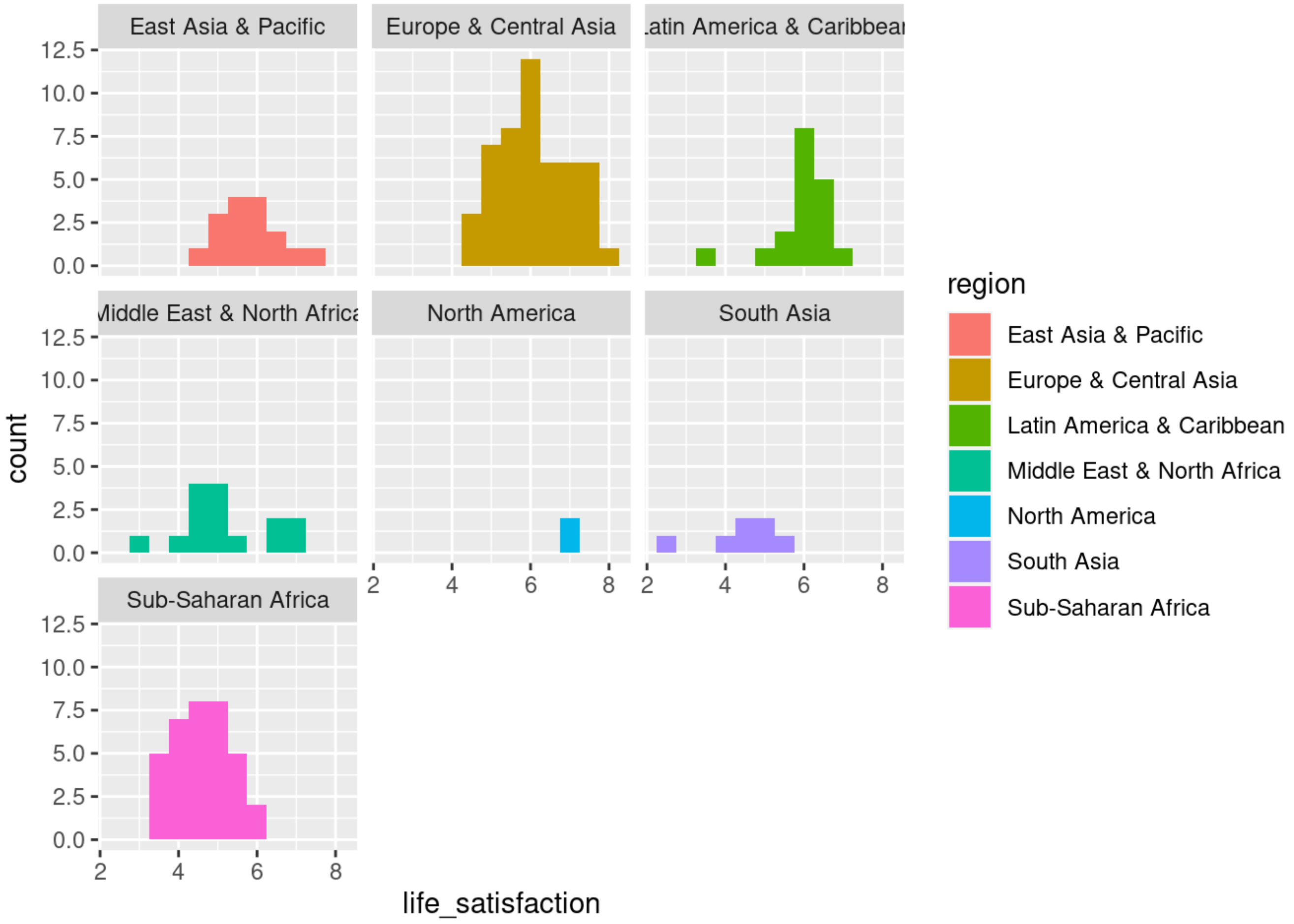
```
ggplot(happiness_1,aes(x=Year)) +
  geom_line(aes( y= East_Asia_and_Pacific,colour= "East_Asia_and_Pacific"))+
  geom_line(aes( y= Europe_and_Central_Asia,colour= "Europe_and_Central_Asia"))+
  geom_line(aes( y= Latin_America_and_Caribbean,colour= "Latin_America_and_Caribbean"))+
  geom_line(aes( y= Middle_East_and_North_Africa,colour= "Middle_East_and_North_Africa"))+
  geom_line(aes( y= North_America,colour= "North_America"))+
  geom_line(aes( y= South_Asia,colour= "South_Asia"))+
  geom_line(aes( y= Sub_Saharan_Africa,colour= "Sub_Saharan_Africa")) +
  scale_colour_manual("", values = c("East_Asia_and_Pacific" ="orange", "Europe_and_Central_Asi
a" ="green",
    "Latin_America_and_Caribbean"="grey",  "Middle_East_and_North_Africa"="lightblue",
    "North_America"="yellow", "South_Asia" ="purple", "Sub_Saharan_Africa" ="pink"))+
  scale_x_continuous("Year",breaks = seq(2005,2020,by=2))+ scale_y_continuous("Life Satisfactio
n",breaks = seq(4,8)) +
  labs(title = "Life Satisfaction in each region during 2005-2020")
```



From this graph, we can see that: Among all these 7 regions, North_America has the highest level of life-satisfaction, and sub_saharan_africa has the lowest level of life-satisfaction for almost each year except year 2017-2019. There is also an decrease for each region(except sub_saharan_africa) during 2005-2020.

plot “Life Satisfaction in 2018” for each country

```
ggplot(happiness[Year==2018], aes(x=life_satisfaction))+geom_histogram(aes(fill=region),binwidth = 0.5)+
  facet_wrap( ~region) + labs("Life Satisfaction in 2018")
```



From the histogram graph above, we can see that in year 2018, most countries in Europe&central asia, Lation America& Caribbran, and East Asia&Pacific has life_satisfaction over 5, while most of coutries in Sub-Saharan Africa and South Asia has life_satisfaction less than 5.There exits huge disparities worldwide.

Create world map for “life satisfaction” in each country in year 2018

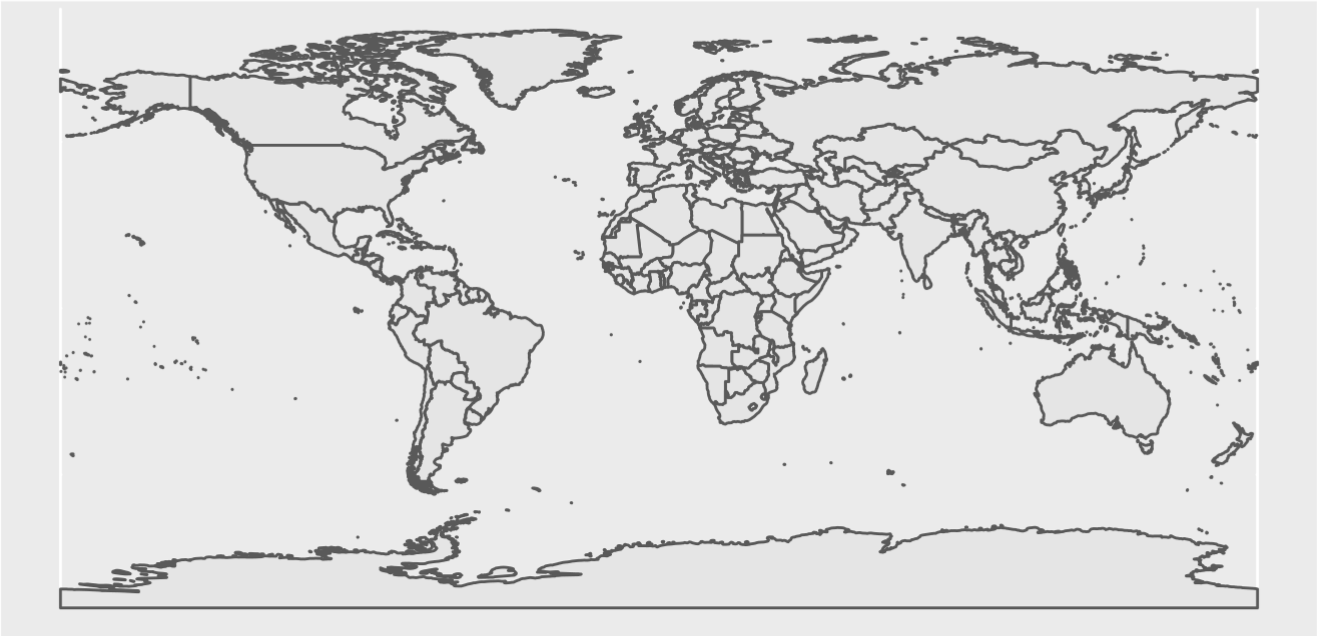
```
h1 <- happiness[Year==2018]
world <- ne_countries(scale = "medium", returnclass = "sf")
```

Merge dataset “world” and dataset “happiness”, and transform the output from data.frame to sf using method ‘st_as_sf’

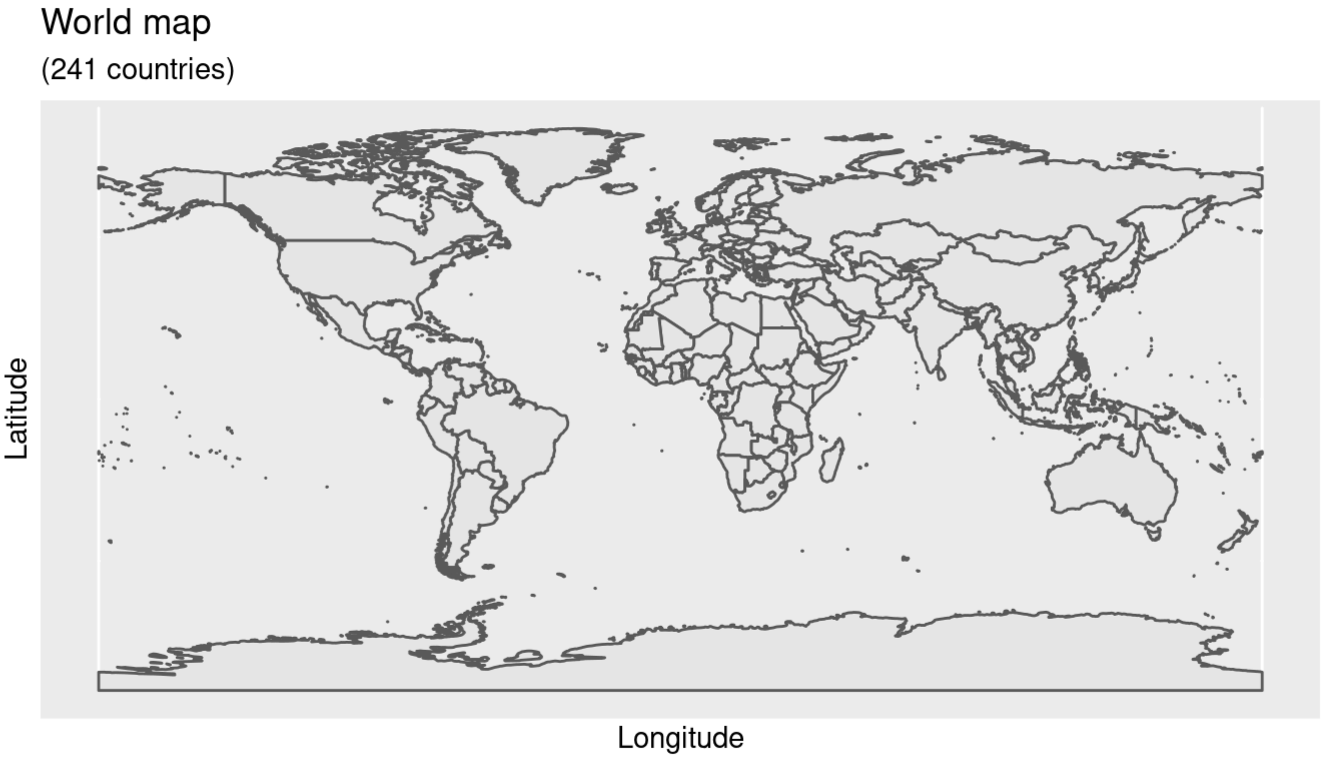
```
colnames(world)[9] = "Entity"
merge_data <- merge(h1, world, by="Entity",all.y=TRUE)
world_ <- st_as_sf(merge_data)
```

Plot the world map of happiness

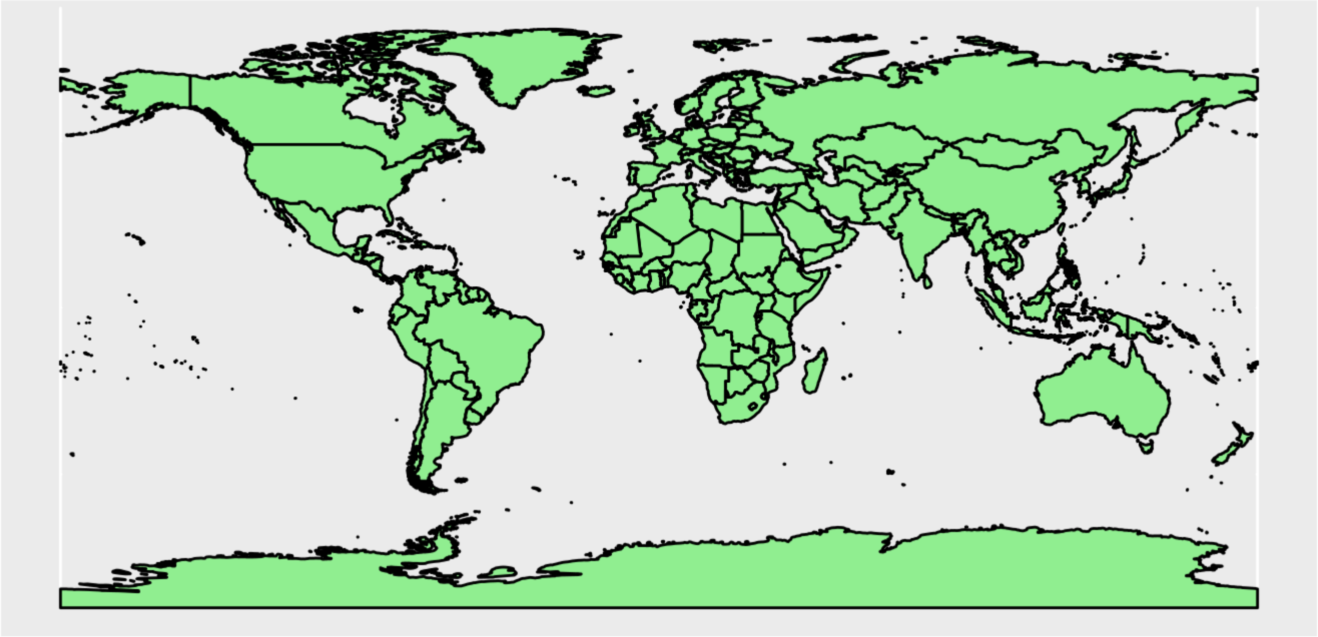
```
ggplot(data = world_) + geom_sf()
```



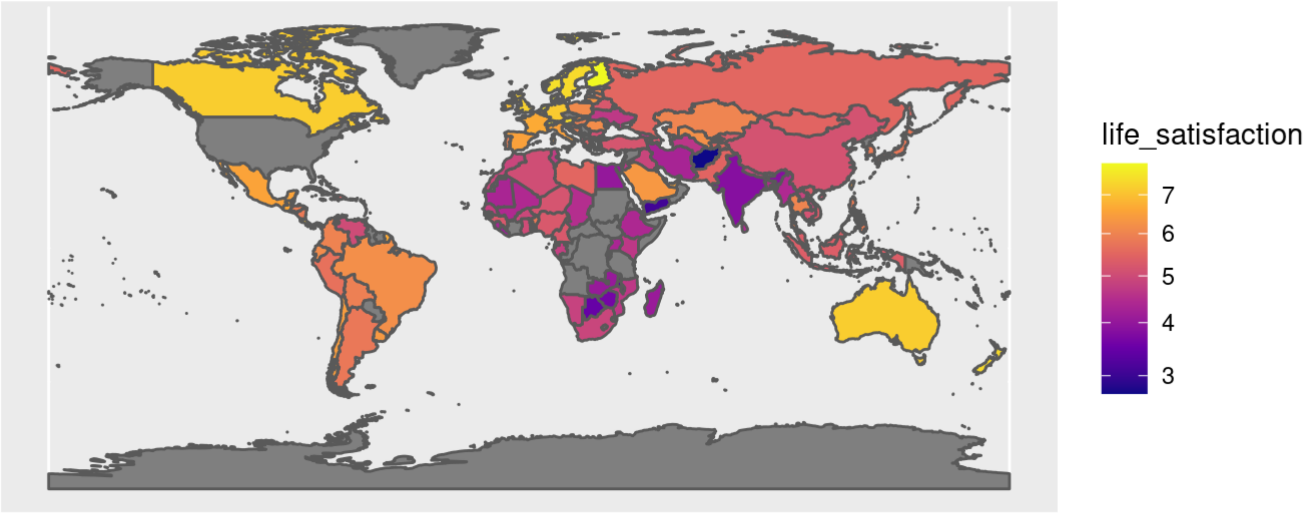
```
ggplot(data = world_) +geom_sf() + xlab(″Longitude″) + ylab(″Latitude″) +
ggtitle(″World map″, subtitle = paste0(″(″, length(unique(world$name)), ″ countries)″))
```



```
ggplot(data = world_) + geom_sf(color = "black", fill = "lightgreen")
```



```
ggplot(data = world_) + geom_sf(aes(fill = life_satisfaction)) +
scale_fill_viridis_c(option = "plasma", trans = "sqrt")
```



From the map above, we can see that: grey parts stand for countries without data recorded in year 2018; and for other parts in this world map,from colour yellow to blue, the darker the colour is for each country, the lower life_satisfaction score the coutry has. We can see that Canada and Australia and most parts of Europe has high level of life_satisfaction, while most parts of Africa has low level of life_satisfaction.