Google Data Analytics Capstone – Case Study

Title - Exploring the analysis of global well-being

Author - Pyae Phyo Maung

Date - January 2, 2024

1. Introduction

This is Google Data Analytics Capstone – Case Study 3 (Follow Your Own Case Study Path**)** which can be found Google Data Analytics Capstone: Complete a Case Study course. In this section, I had a chance to choose my own case study that is of particular interest to me. The case study involves world happiness report of 2023 and further more I will also join with other datasets such as continents and global country information dataset 2023. The data has been made available by Sustainable Development Solutions Network under this license.

2. Scenario

I am a junior data analyst working for a business intelligence consultant. I have been at my job for six months, and my boss feels I am ready for more responsibility. He has asked me to lead a project for a brand-new client — this will involve the world happiness report, a landmark survey of the state of global happiness, which describe the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness. I will choose the topic, ask the right questions, identify a fresh dataset and ensure its integrity, conduct analysis, create compelling data visualizations, and prepare a presentation.

3. Analyzing Data

In this analysis, the 6 phases of the Data Analysis process: Ask, Prepare, Process, Analyze, Share and Act are as follow.

3.1 Ask

The client is non-profit organization and they want to know about the global well-being based on the findings of the World Happiness Report.

The problem I am trying to solve is - How can nations collaborate globally to enhance well-being based on the findings of the World Happiness Report?

The main objective is to share experiences and best practices in promoting happiness and well-being. It fosters international collaboration and the exchange of ideas for creating policies that enhance the overall life satisfaction of citizens.

In this project, the key stakeholders are

- CEO
- Project manager
- Data analytic team

The clear statement of the business task is to better understand how relationship between happiness, money, health, and many other metrics.

3.2 Prepare

A total 3 CSV files of datasets is chosen as follow.

1. World happiness Report 2023 (link)
2. Continent dataset (link)
3. Global Country Information Dataset 2023 (link)

Data cleaning in spreadsheets will be time-consuming and slow compared to R. I am choosing R simply because I could do both data wrangling and analysis/ visualizations in the same platform.

The data is located in R studio because I will use R studio for data analysis. It is not bias and cleaned. But when merging with other datasets, I only use about three or columns which can be supplied into main dataset. It's ROCCC because it's reliable, original, comprehensive, current and cited.

The company has their own license over the datasets. Besides that, the dataset doesn't have any personal information. All the files have consistent columns and each column has the correct type of data. It may have some key insights about the countries and their happiness. There is no problem with data.

3.3 Process

This step will prepare the data for analysis. All the csv files will be merged into one file to improve workflow.

3.3.1 Install packages and load libraries

```
#installing packages
install.packages("skimr")
install.packages("janitor")
install.packages("tidyverse")
install.packages("dplyr")
install.packages("lubridate")
install.packages("hms")
install.packages("tidyr")
install.packages("ggplot2")
install.packages("plotly")
install.packages("htmltools")

#Usage library
library(dplyr)
library(skimr)
library(janitor)
library(tidyverse)
library(lubridate)
library(hms)
library(tidyr)
library(ggplot2)
library(plotly)
```

3.3.2 Read all csv files and combine into one data frame

Read all csv files from exact location and extracting specific columns from continents and world dataset and merging into main dataset.

```
#Read 3 csv files
wh2023 <- read.csv("World happiness 2023/world-happiness-2023.csv")
wd2023 <- read.csv("World happiness 2023/world-data-2023.csv")
continents2023 <- read.csv("World happiness 2023/continents-data-2023.csv")

#extract specific columns from continent dataset
continent2023 <- continents2023 %>%
  select(name,Country_code,Region,Sub_region) %>%
  rename(Country = name)
wh2023 <- wh2023 %>%
  rename(Country = Country_name)
#extract specific columns from world-data dataset
wd2023 <- wd2023 %>%
  select(Country,Abbreviation,Unemployment.rate,Urban_population,Population,Official.language) %>%
rename(Unemployment_rate =Unemployment.rate,Official_language = Official.language)


#merge continents,world data and world happiness dataset
wh2023 <- merge(wh2023,continent2023,by = "Country",all.x = TRUE)
wh2023 <- wh2023 %>%
  relocate(Country_code,Region,Sub_region,.after = Country)
wh2023 <- merge(wh2023,wd2023,by = "Country",all.x = TRUE)%>%
  relocate(Abbreviation,.after = Country)
View(wh2023)
```

3.3.3 Data cleaning process

- Check null values and duplicated rows from table.

```
#check null rows
wh2023[wh2023==""]<-NA
null_values <- colSums(is.na(wh2023))
print(null_values)

#check duplicate
duplicate_index <- anyDuplicated(wh2023)
# Print the results
if (duplicate_index > 0) {
  print(paste("Duplicate rows found at index:",duplicate_index, "\n"))
} else {
  cat("No duplicate rows found.\n")
}
```

- Cleaning null values

```
#Clear null values
wh2023 <- wh2023[complete.cases(wh2023), ]
view(wh2023)
```

3.3.4 Saving the result as txt file

```
#Save as txt file
write.table(wh2023,"World happiness 2023/world_happiness_report.txt",fileEncoding = "UTF-8", quote = FALSE)
```

3.4 Analyze

The data frame is now ready for descriptive analysis that will help us to build a profile for global well-being by regions and their stats.

3.4.1 Summary of dataset

To quick start, let's generate a summary of the dataset.

```
#Analyze
head(wh2023)
#summary of dataset
summary(wh2023)
```

3.4.2 Some Calculations for data analysis

Run a few calculations in one file to get a better sense of the data layout.

- Calculate the highest and lowest 10 data points of **ladder_score**.
- Calculate the **ladder_score** by continents.
- Calculate the **ladder_score** by **Sub_regions**.

```
#top 10 highest ladder score
highest_ladder <- wh2023 %>%
  select(Country,Ladder_score) %>%
  arrange(desc(Ladder_score)) %>%
  top_n(10,wt=Ladder_score)
view(highest_ladder)

#top 10 lowest ladder score
lowest_ladder <- wh2023 %>%
  select(Country,Ladder_score) %>%
      arrange(desc(Ladder_score)) %>%
      tail(10)
view(lowest_ladder)
```

Result dataset.

	Country	Ladder_score
1	Finland	7.804
2	Denmark	7.586
3	Iceland	7.530
4	Israel	7.473
5	Netherlands	7.403
6	Sweden	7.395
7	Norway	7.315
8	Switzerland	7.240
9	Luxembourg	7.228
10	New Zealand	7.123

	Country	Ladder_score
114	Madagascar	4.019
115	Zambia	3.982
116	Tanzania	3.694
117	Comoros	3.545
118	Malawi	3.495
119	Botswana	3.435
120	Zimbabwe	3.204
121	Sierra Leone	3.138
122	Lebanon	2.392
123	Afghanistan	1.859

```r
#average,max,min ladder score by Continents
regions <- wh2023 %>%
  group_by(Region) %>%
  summarise(avg_ladder = mean(Ladder_score),
            sum_ladder = sum(Ladder_score),
            country_count = length(Country),
            max_ladder = max(Ladder_score),
            min_ladder = min(Ladder_score))
view(regions)

#average,max,min ladder score by sub regions
sub_region <- wh2023 %>%
  group_by(Sub_region) %>%
  summarise(avg_ladder = mean(Ladder_score),
            sum_ladder = sum(Ladder_score),
            country_count = length(Country),
            max_ladder = max(Ladder_score),
            min_ladder = min(Ladder_score))
view(sub_region)
```

Result dataset.

	Region	avg_ladder	sum_ladder	country_count	max_ladder	min_ladder
1	Africa	4.408875	141.084	32	5.902	3.138
2	Americas	6.057619	127.210	21	6.961	5.211
3	Asia	5.295765	180.056	34	7.473	1.859
4	Europe	6.516794	221.571	34	7.804	5.071
5	Oceania	7.109000	14.218	2	7.123	7.095

	Sub_region	avg_ladder	sum_ladder	country_count	max_ladder	min_ladder
1	Australia and New Zealand	7.109000	14.218	2	7.123	7.095
2	Central Asia	5.828250	23.313	4	6.144	5.330
3	Eastern Asia	5.934500	23.738	4	6.129	5.818
4	Eastern Europe	5.922000	47.376	8	6.589	5.071
5	Latin America and the Caribbean	5.966053	113.355	19	6.609	5.211
6	Northern Africa	4.724750	18.899	4	5.329	4.170
7	Northern America	6.927500	13.855	2	6.961	6.894
8	Northern Europe	7.095222	63.857	9	7.804	6.213
9	South-eastern Asia	5.431222	48.881	9	6.587	4.372
10	Southern Asia	4.201429	29.410	7	5.360	1.859
11	Southern Europe	6.095800	60.958	10	6.650	5.277
12	Sub-Saharan Africa	4.363750	122.185	28	5.902	3.138
13	Western Asia	5.471400	54.714	10	7.473	2.392
14	Western Europe	7.054286	49.380	7	7.403	6.661

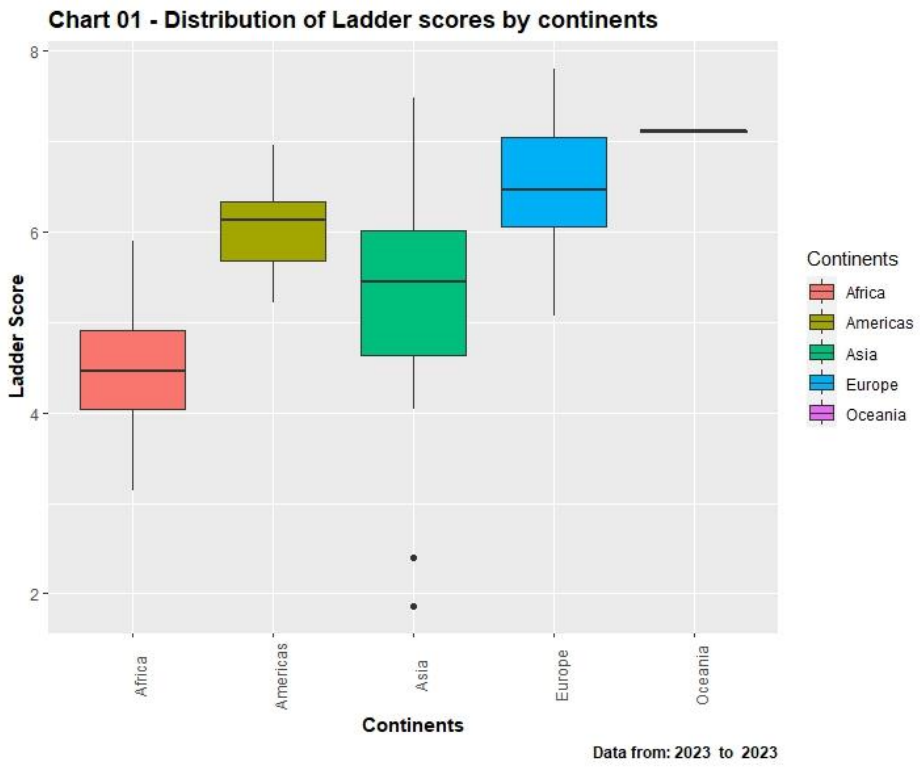
3.5 Share

The share phase of the data analysis process typically involves communicating findings, summarizing results using data visualizations, and creating a slideshow to present to stakeholders.

3.5.1 Create data visualization with charts

Chart 01 - World happiness Report: Distribution of happiness scores by continents.

```
#chart 01 - World happiness Report: ladder scores by continents
ggplot(wh2023, aes(x = Region, y = Ladder_score  ,fill = Region)) +
  geom_boxplot() +
  labs(title="Chart 01 - Distribution of Ladder scores by continents",
       caption = paste0("Data from: ", 2023, "  to  ", 2023),
       fill="Continents",
       x = "Continents",
       y = "Ladder Score")+
  theme(axis.text.x = element_text(angle = 90),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```



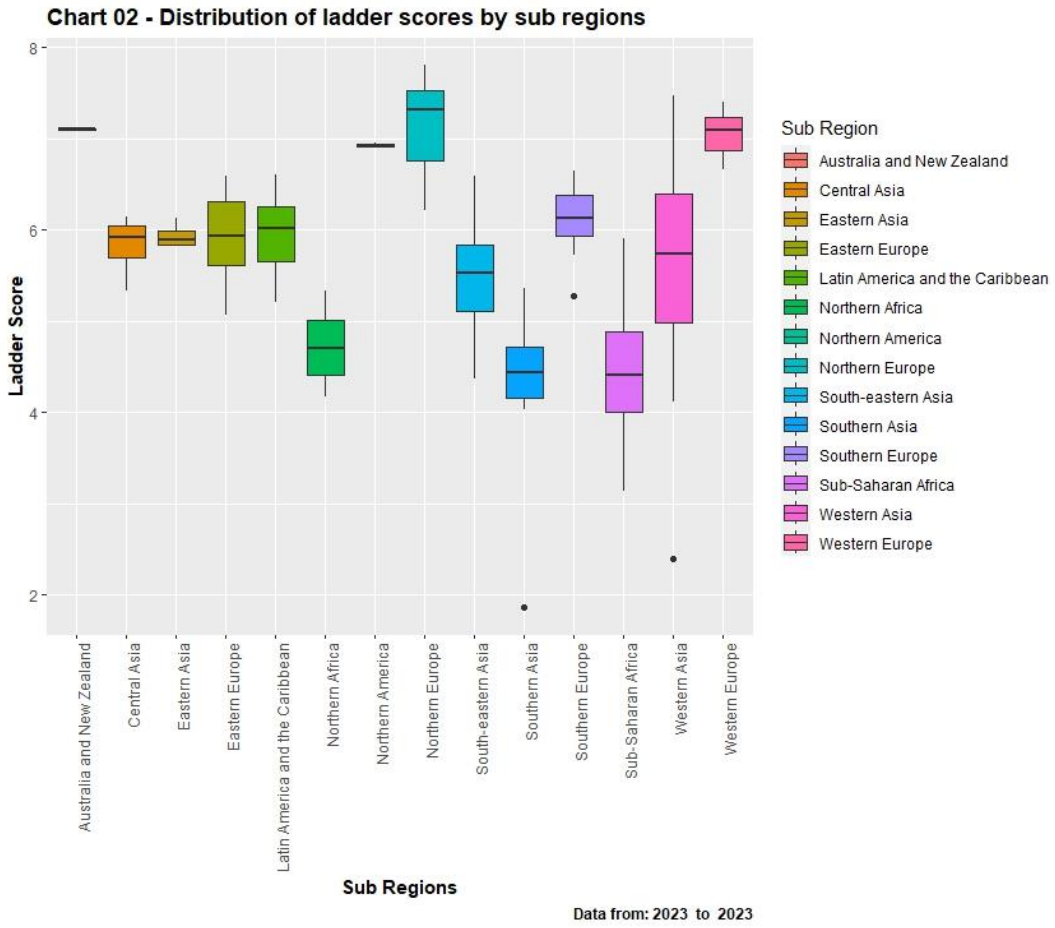
Some considerations can be taken by this chart:

- Oceania has only two countries with the highest average point **~7.1** of the dataset.

- Africa has the lowest average point with **~4.5** of the dataset.
- Europe has the highest point with **~7.8** of the dataset.
- Asia has the lowest point with **~1.8** of the dataset.

Chart 02 - World happiness Report: Distribution of happiness scores by sub regions.

```
#chart 02 - World happiness Report: ladder score by sub region
ggplot(wh2023, aes(x = Sub_region, y = Ladder_score  ,fill = Sub_region)) +
  geom_boxplot() +
  labs(title="Chart 02 - Distribution of ladder scores by sub regions",
       caption = paste0("Data from: ", 2023, "  to  ", 2023),
       fill= "Sub Region",
       x = "Sub Regions",
       y = "Ladder Score")+
  theme(axis.text.x = element_text(angle = 90,hjust = 1),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```



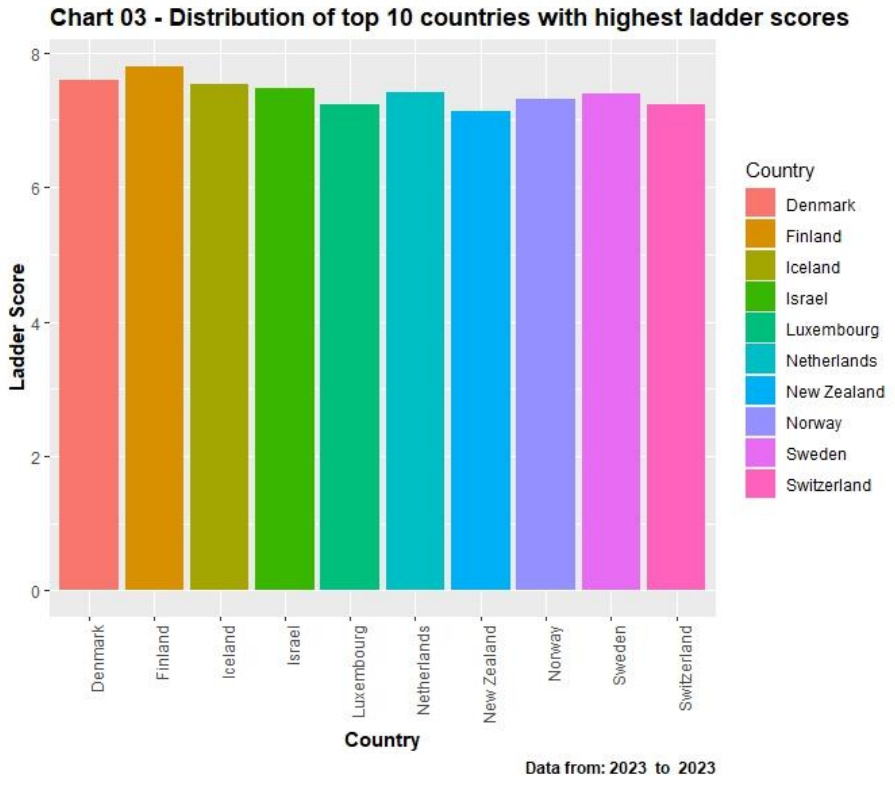
Some considerations can be taken by this chart:

- Australia and New Zealand has the highest average point with **~7.1** of the dataset.

- Southern Asia has the lowest average point with **~4.2** of the dataset.
- Northern Europe has the highest point with **~7.8** of the dataset.
- Southern Asia has the lowest point with **~1.8** of the dataset.

Chart 03 - World happiness Report: Countries with highest happiness scores.

```
#chart 03 - World happiness Report: highest ladder score by countries
ggplot(highest_ladder, aes(Country,y = Ladder_score,fill = Country)) +
  geom_bar(stat = "identity") +
  labs(title="Chart 03 - Distribution of top 10 countries with highest ladder scores",
       caption = paste0("Data from: ", 2023, "  to  ", 2023),
       fill= "Country",
       x = "Country",
       y = "Ladder Score")+
  theme(axis.text.x = element_text(angle = 90,hjust = 1),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```

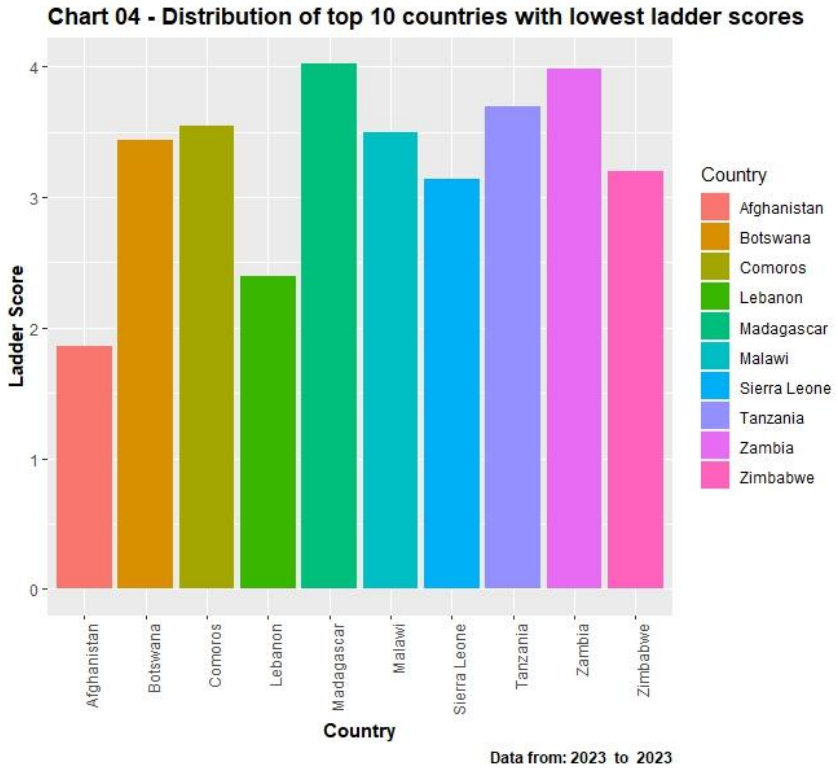


Some considerations can be taken by this chart:

- Finland has the highest score with **~7.8** of the dataset.
- Most of them are European countries except Israel and New Zealand.
- They have higher freedom life choices than average (**~0.79**).

Chart 04 - World happiness Report: Countries with lowest happiness scores.

```
#chart 04 - World happiness Report: lowest ladder score by countries
ggplot(lowest_ladder, aes(Country,y = Ladder_score,fill = Country)) +
  geom_bar(stat = "identity") +
  labs(title="Chart 04 - Distribution of top 10 countries with lowest ladder scores",
       caption = paste0("Data from: ", 2023, "  to  ", 2023),
       fill= "Country",
       x = "Country",
       y = "Ladder Score")+
  theme(axis.text.x = element_text(angle = 90,hjust = 1),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```

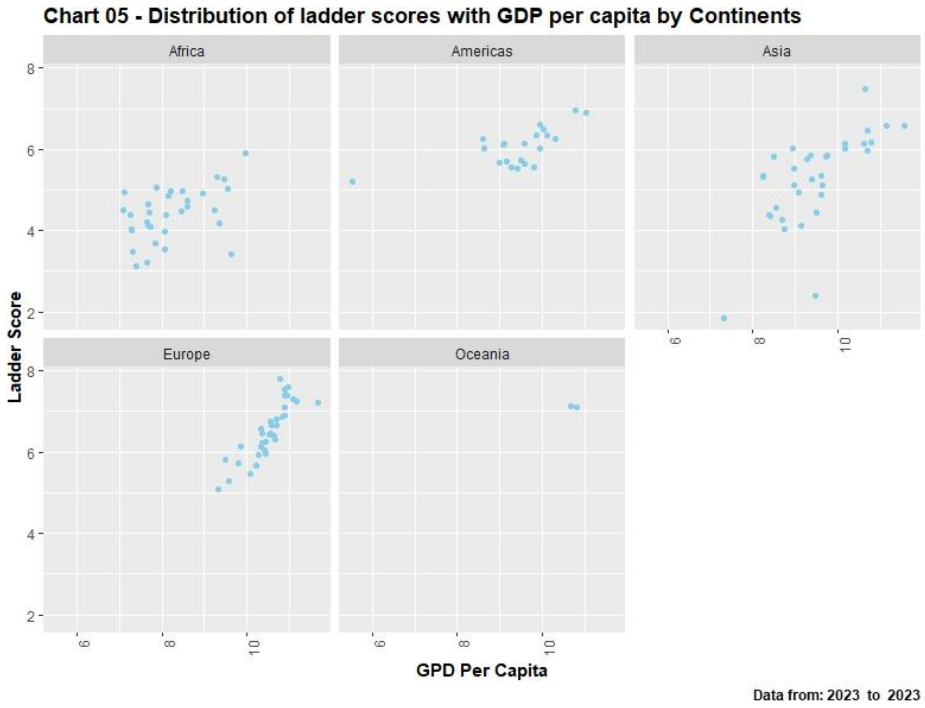


From this chart, we can see:

- Afghanistan has the lowest point with **~1.8** of the dataset.
- Most of them are African countries except lebanon and Afghanistan.
- Most of them have lower freedom life choices than average (**~0.79**) except Zambia and Tanzania.

Chart 05 - World happiness Report: Happiness scores with GDP per capita by continents.

```
#chart 05 - World happiness Report: ladder score with GDP per capita
ggplot(wh2023, aes(x = Logged_GDP_per_capita, y = Ladder_score)) +
  geom_point(color = "skyblue")+
  facet_wrap(~Region)+
  labs(title="Chart 05 - Distribution of ladder scores with GDP per capita by Continents",
       caption = paste0("Data from: ", 2023, "  to  ", 2023),
       x = "GPD Per Capita",
       y = "Ladder Score")+
  theme(axis.text.x = element_text(angle = 90,hjust = 1),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```

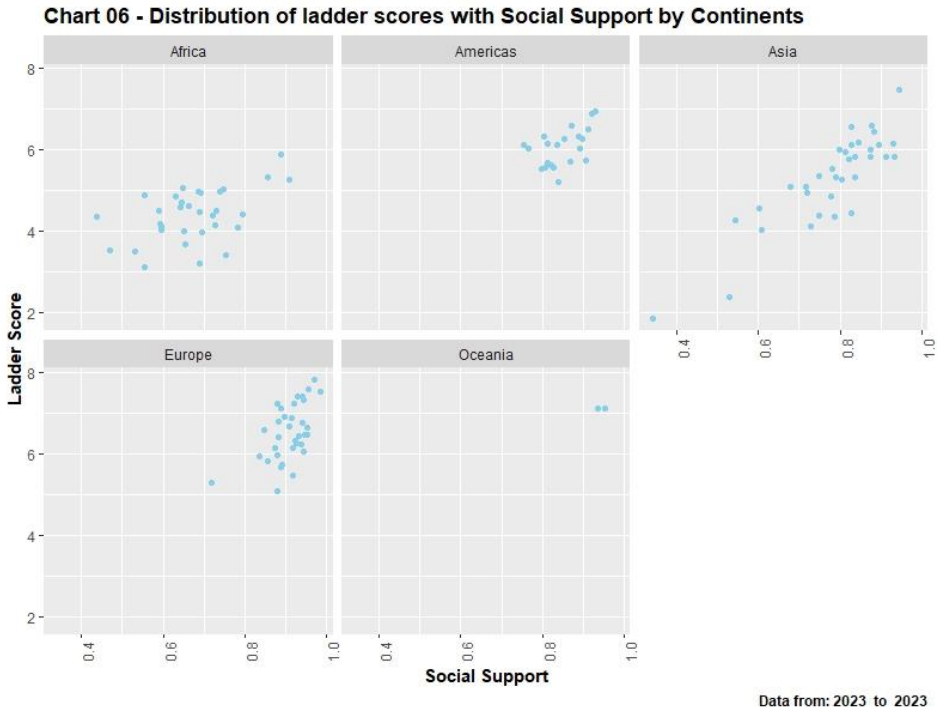


From this chart, we can see:

- The scatter plot in Africa and Asia is mostly around in the center.
- America, Europe and Oceania have higher happiness score with higher GDP per capita.

Chart 06 - World happiness Report: Happiness scores with social support by continents.

```
#chart 06 - World happiness Report: ladder score with Social support
ggplot(wh2023, aes(x = Social_support, y = Ladder_score)) +
  geom_point(color = "skyblue")+
  facet_wrap(~Region)+
  labs(title="Chart 06 - Distribution of ladder scores with Social Support by Continents",
       caption = paste0("Data from: ", 2023, "  to  ", 2023),
       x = "Social Support",
       y = "Ladder Score")+
  theme(axis.text.x = element_text(angle = 90,hjust = 1),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```

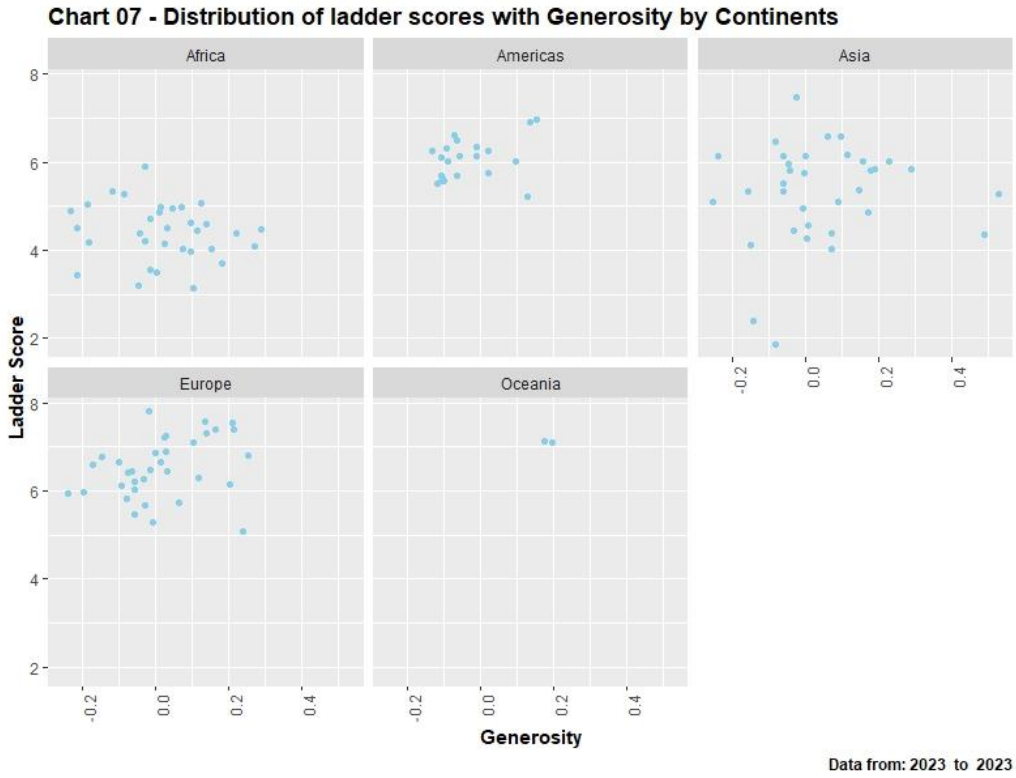


From this chart, we can see:

- The scatter plot in Africa and Asia is mostly around in the center.
- America, Europe and Oceania have higher happiness score with higher social support.

Chart 07 - World happiness Report: Happiness scores with Generosity by continents.

```
#chart 07 - World happiness Report: ladder score with Generosity
ggplot(wh2023, aes(x = Generosity, y = Ladder_score)) +
  geom_point(color = "skyblue")+
  facet_wrap(~Region)+
  labs(title="Chart 07 - Distribution of ladder scores with Generosity by Continents",
       caption = paste0("Data from: ", 2023, "  to  ", 2023),
       x = "Generosity",
       y = "Ladder Score")+
  theme(axis.text.x = element_text(angle = 90,hjust = 1),
        axis.title.x = element_text(face="bold"),
        axis.title.y = element_text(face="bold"),
        plot.caption = element_text(face="bold"),
        plot.title =element_text(face="bold"))
```



From this chart, we can see:

Mostly of the countries around the world have lower generosity even with higher happiness scores. But only two countries such as Myanmar and Indonesia have higher generosity with **~4.3** and **~5.2** of the dataset.

Chart 08 - World happiness Report: Happiness scores by countries.

```
# Create choropleth map
choropleth_map <- plot_ly(
  type = "choropleth",
  locationmode = "country names",
  locations = wh2023$Country,
  z = wh2023$Ladder_score,
  colorscale = "Viridis",
) %>%
  layout(
    title = list(
      text = "<b>World Happiness Report: Ladder score by country</b>"
    ),
    geo = list(showframe = FALSE, showcoastlines = FALSE),
    margin= list(b=-10,t=100)
  )%>%
  colorbar(
    title = list(
      text= "<b>Ladder Score</b>"
    )
  )
# Display the map
choropleth_map
```

Click image link for more details

The world map image links to the standalone web page which describes world map in which we can observe the ladder scores and country names.

3.5.2 Summary of Analysis

Trends or relationship I found in the data are as follow.

- Oceania has the highest average happiness scores with **~7.1** and Africa has the lowest average happiness scores with **~4.5** of the dataset.

- Finland has the highest happiness score with **~7.8** and Afghanistan has the lowest happiness score with **~1.8** of the dataset.
- For highest happiness scores, most of the European countries included and for lowest happiness scores, most of the African countries included.
- America, Europe and Oceania have higher happiness score with higher social support and higher GDP per capita.
- Mostly of the countries around the world have lower generosity even with higher happiness scores. But only two countries such as Myanmar and Indonesia have highest generosity with **~4.3** and **~5.2** of the dataset.

Conclusion.

According to the World Happiness Report, world happiness mostly depends on GDP per capita, social support, freedom to make life choices, and healthy life expectancy. European countries are happier than any other region as they have higher points in the above as described. And African countries have lower happiness scores compared to other regions. As for generosity, Myanmar and Indonesia have higher points even though they don't have high happiness scores.

3.6 Act

In the Act phase, we have to provide some recommendations for my client to improve their working strategies.

There are various factors affecting the happiness scores of a country. We have to consider every possibility of how we can improve the happiness scores of our country.

In this case, I would like to give the top three recommendations based on my analysis:

1. Investing in education can help to become a wealthier country and improve the happiness score in the future.
2. Enhance the healthcare system to ensure physical and mental well-being.
3. Promote strong social support to provide emotional support and help individuals during challenging times.

4. Reference

1. **lucyallan** from Kaggle.
2. **Sevvalsimsek** from Kaggle.