DATA ANALYSIS AND VISUALIZATION HW4

Name: Shubham Sharma

DATASET – Gapminder_after1952

```
In [1]: import pandas as pd
        import plotly.express as px

In [6]: df = pd.read_csv("C:/Users/PC/Downloads/gapminder_after1952.csv")

In [7]: df.head()
```

Out[7]:

	country	continent	year	lifeExp	pop	gdpPercap	iso_alpha	iso_num
0	Afghanistan	Asia	1952	28.801	8425333	779.445314	AFG	4
1	Afghanistan	Asia	1957	30.332	9240934	820.853030	AFG	4
2	Afghanistan	Asia	1962	31.997	10267083	853.100710	AFG	4
3	Afghanistan	Asia	1967	34.020	11537966	836.197138	AFG	4
4	Afghanistan	Asia	1972	36.088	13079460	739.981106	AFG	4

```
In [8]: df.shape
Out[8]: (1692, 8)

In [9]: df.country.unique().shape
Out[9]: (141,)

In [10]: df.year.unique()
Out[10]: array([1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 1997, 2002,
                2007], dtype=int64)

In [11]: df.year.unique().shape
Out[11]: (12,)
```

The dataset contains 8 columns and 1692 rows, each representing information about various countries. The following are the column details:
Country – Countries name
Continent – Provides information on the continent to which each country belongs.
Year – Years from 1952 till 2007
LifeExp – Life Expectancy of the country
pop – Population of the country
gdpPercap – GDP per capita of the country
iso_alpha – Country shortform (acronym)
iso_num – ISO country codes

The data ranges from the year 1952 to 2007, and provides valuable insights into the health and economic conditions of different countries. With this dataset, researchers and analysts can explore the relationships between different variables and draw meaningful conclusions about the patterns and trends observed over time.

After running the codes to get a proper MDS, TSNE, Kmeans and clustering projection:

```
In [67]: projections.head()
```

Out[67]:

	MDS_x	MDS_y	country	iso_alpha	continent	tsne_x	tsne_y	scluster6	kmeans6
0	-1.067966	7.216552	Afghanistan	AFG	Asia	-8.855268	10.517956	5	3
12	-2.393537	-2.059857	Albania	ALB	Europe	2.377563	-0.302766	1	5
24	-0.966520	0.449756	Algeria	DZA	Africa	-1.030125	1.896615	1	1
36	1.122454	6.674040	Angola	AGO	Africa	-8.622373	11.268964	5	3
48	1.046018	-2.964628	Argentina	ARG	Americas	2.662710	-3.897974	4	5

```
In [70]: projections.to_csv("C:/Users/PC/Downloads/gapminder_after1952_projections.csv")
```

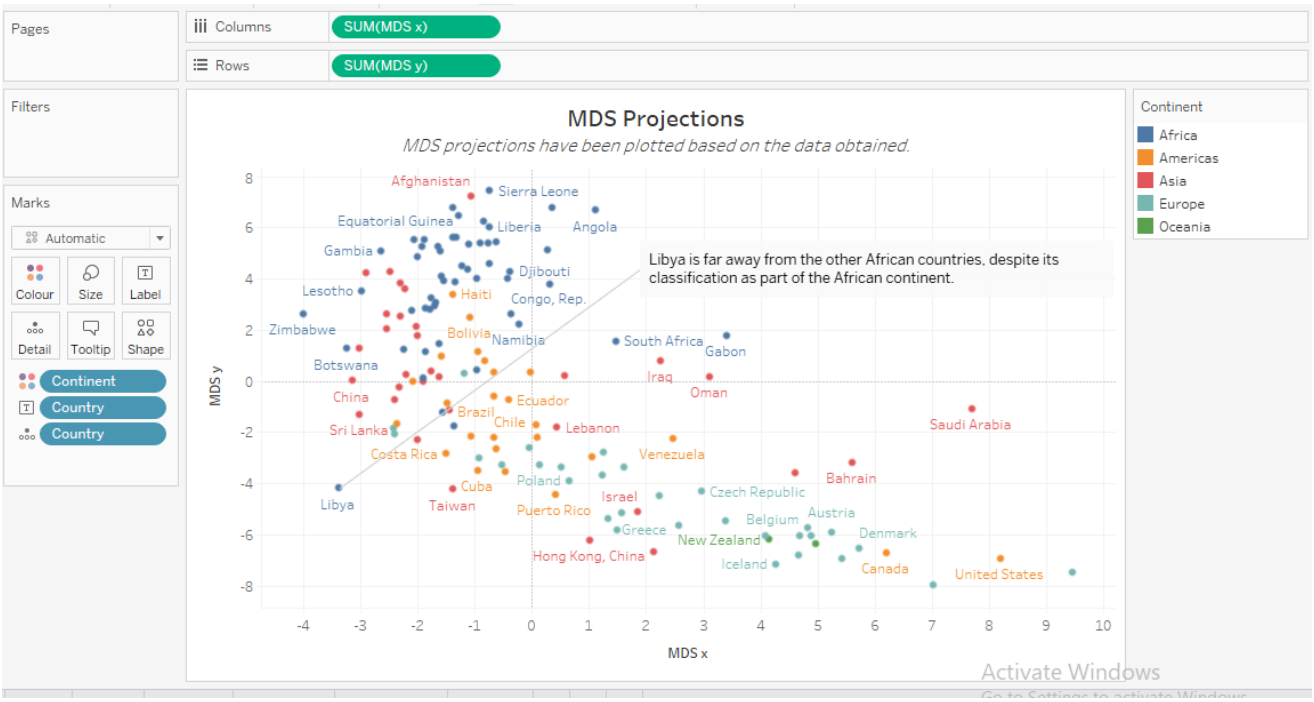
The above image shows the new dataset which includes MDS, TSNE, Kmeans and Scluster values according to the countries.

MDS -

The multidimensional scaling (MDS) analysis performed on a dataset containing information about various countries. MDS is a statistical technique used to analyze similarities or dissimilarities between data points in a high-dimensional space and represent them in a lower-dimensional space.

In this MDS chart, each data point represents a country and its position is determined by the similarity or dissimilarity of its attributes with other countries. Countries that are similar in terms of their attributes are placed closer together in the chart, while countries with dissimilar attributes are placed farther apart.

The MDS projection in this chart provides a visual representation of the relationships between different countries based on their attributes, such as life expectancy, population, and GDP per capita. It can be a useful tool for identifying patterns and trends in the data, and for comparing the attributes of different countries.



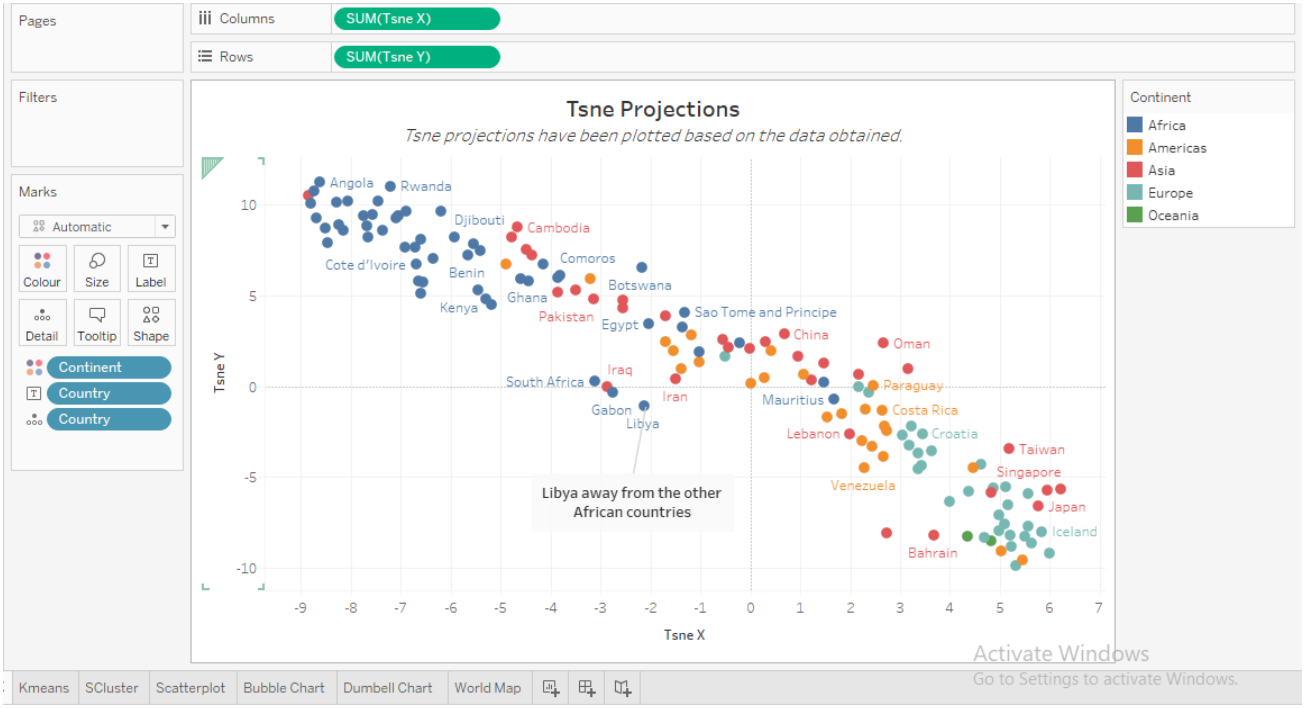
The MDS chart is a projection of a multidimensional dataset that uses similarities or dissimilarities between countries based on their attributes. The countries are colour coded based on the continents as given in the image. It's possible that **Libya**'s attributes are significantly different from other African countries, resulting in its position being farther away from the African cluster in the chart. Alternatively, there could be other factors at play that are not immediately apparent, such as historical or geopolitical factors, which have influenced Libya's position in the chart.

TSNE –

The t-SNE (t-Distributed Stochastic Neighbor Embedding) chart is a 2D projection of a high-dimensional dataset containing information about various countries. t-SNE is a dimensionality reduction technique that is commonly used to visualize high-dimensional data in a lower-dimensional space, while preserving the similarities between data points. The countries are colour coded based on the continents as given in the image.

In this t-SNE chart, each data point represents a country, and its position in the chart is determined by its similarity to other countries based on the attributes included in the dataset. Countries that are similar in terms of their attributes are placed closer together in the chart, while countries with dissimilar attributes are placed farther apart.

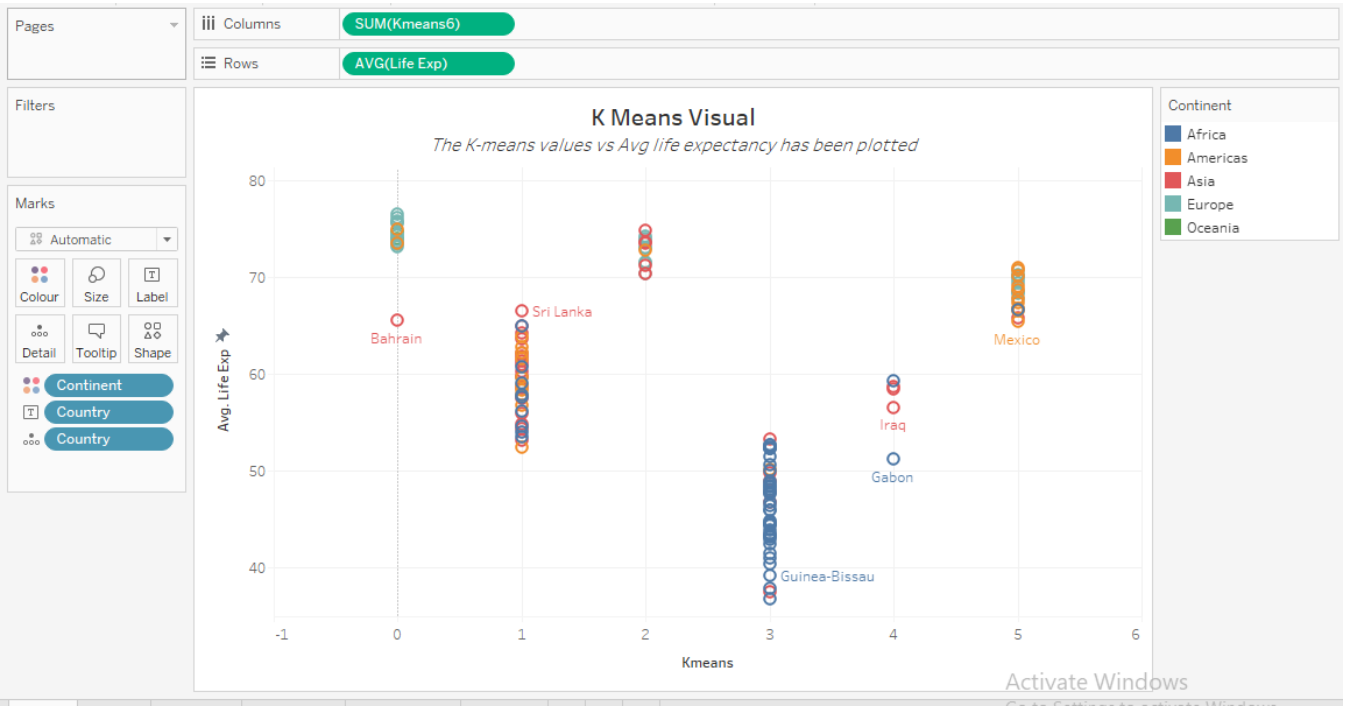
The t-SNE chart provides a visual representation of the relationships between different countries based on their attributes, such as life expectancy, population, and GDP per capita. It can be a useful tool for identifying clusters of countries with similar attributes and for exploring patterns and trends in the data.



It is possible that Libya's position in the TSNE chart, which shows it as being distant from other African countries, could be due to its higher GDP rate or life expectancy compared to other African

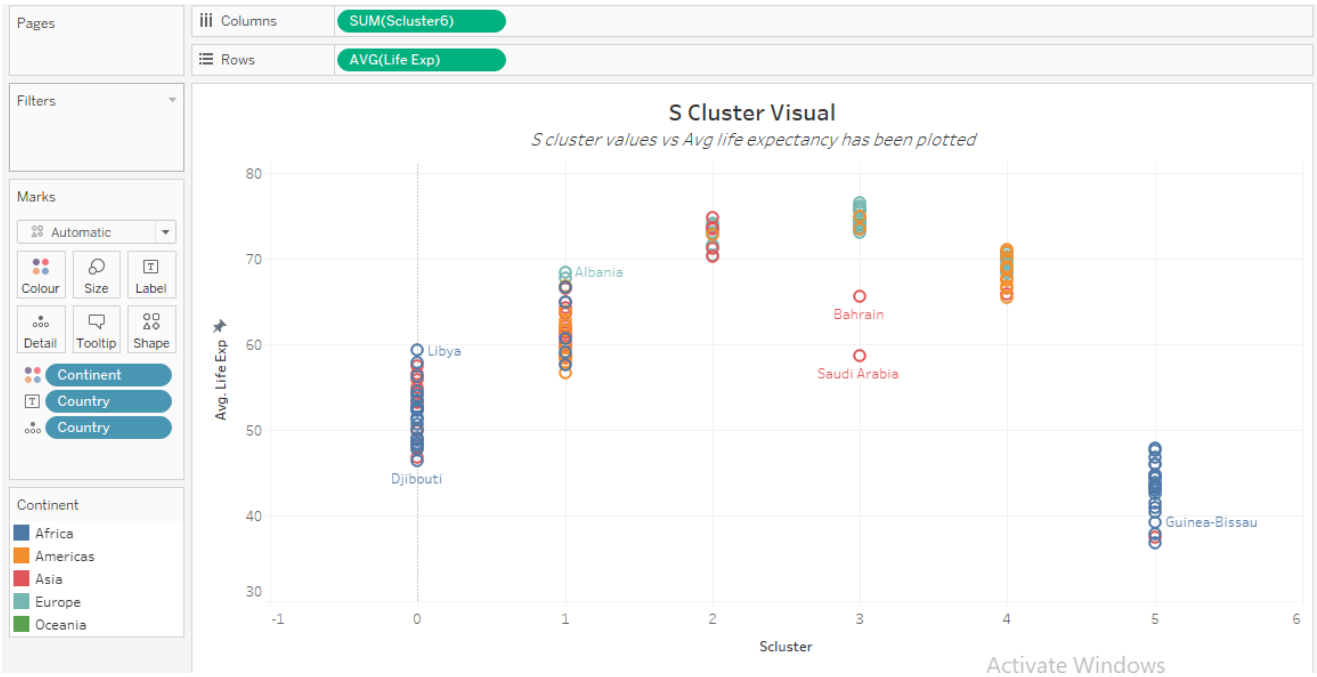
countries. However, it is important to note that the TSNE chart considers all the attributes included in the dataset, and other factors could also be contributing to Libya's position in the chart. It's also possible that there are some similarities or dissimilarities between Libya and other non-African countries that influenced its position in the chart.

Kmeans –



The K-means clustering visualization has partitioned similar countries into six distinct groups. It is apparent that countries that are closely positioned to each other in both the MDS and TSNE visualizations are predominantly clustered together in the same group within the K-means plot. In most cases, countries located in the same continent are grouped in the same cluster. In particular, **group 3 is primarily composed of African countries** with the exception of a handful.

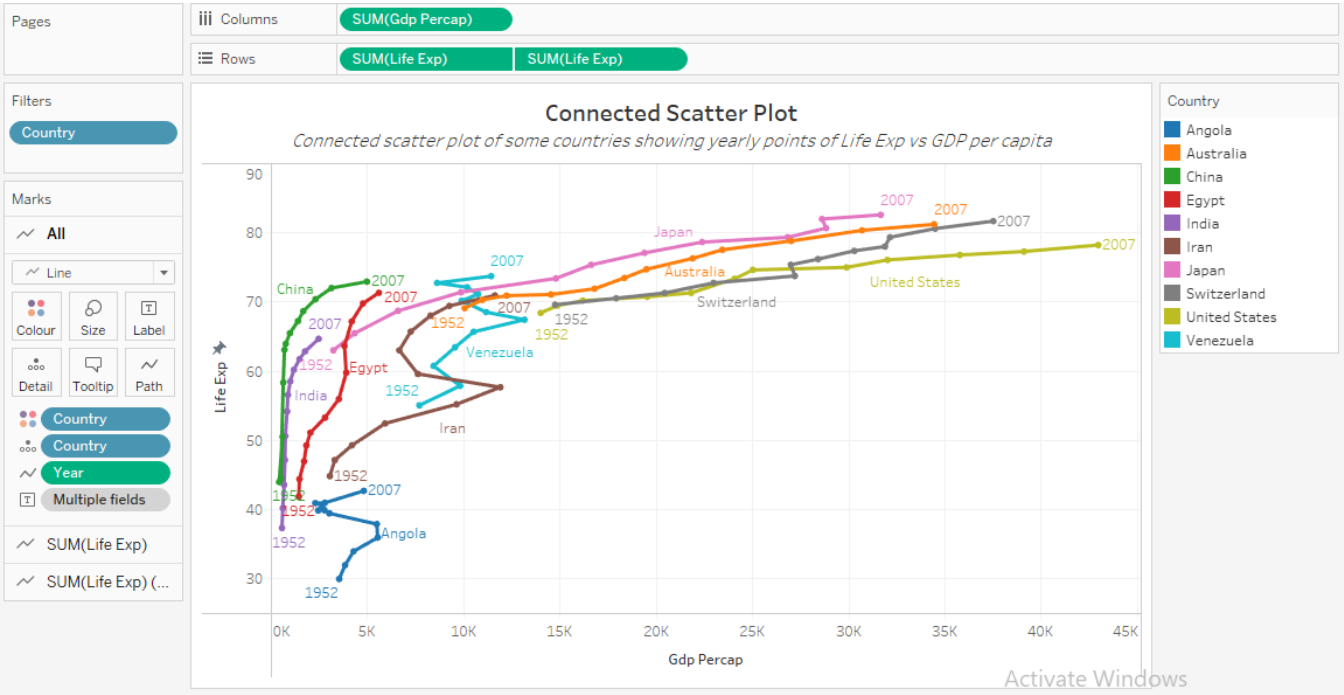
S Cluster –



The S Cluster visualization has organized similar countries into six distinct groups. Notably, countries located closely to one another in both the MDS and TSNE visualizations are generally assigned to the same cluster in the S Cluster plot. Furthermore, countries situated within the same continent are mostly grouped together within the same cluster. A noteworthy observation is that **group 5 primarily consists of African countries** with only a few exceptions.

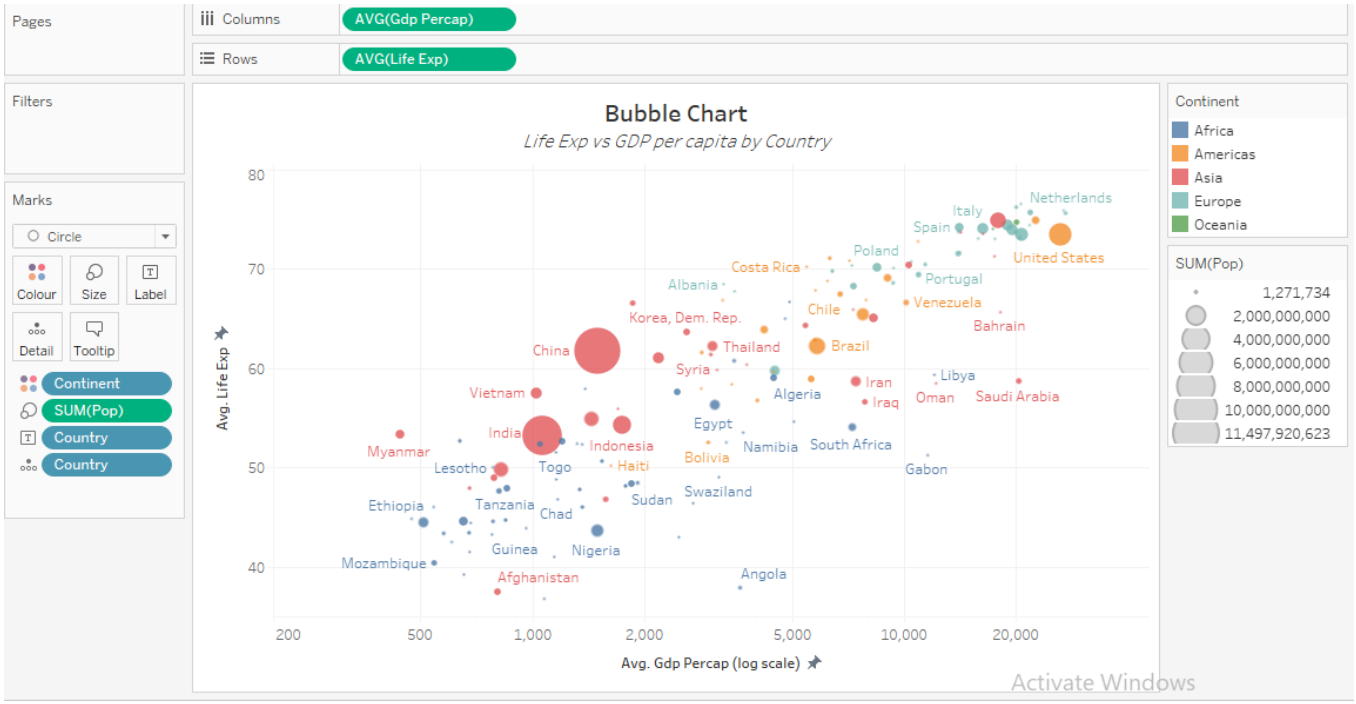
The positioning of certain African countries such as **Libya** is in different clusters (cluster 0) may be attributed to various factors, one of which is the average life expectancy of its population. Other factors such as GDP per capita, population, and geographical location may also have played a role in the clustering of countries in the visualizations.

Connected Scatter Plot –



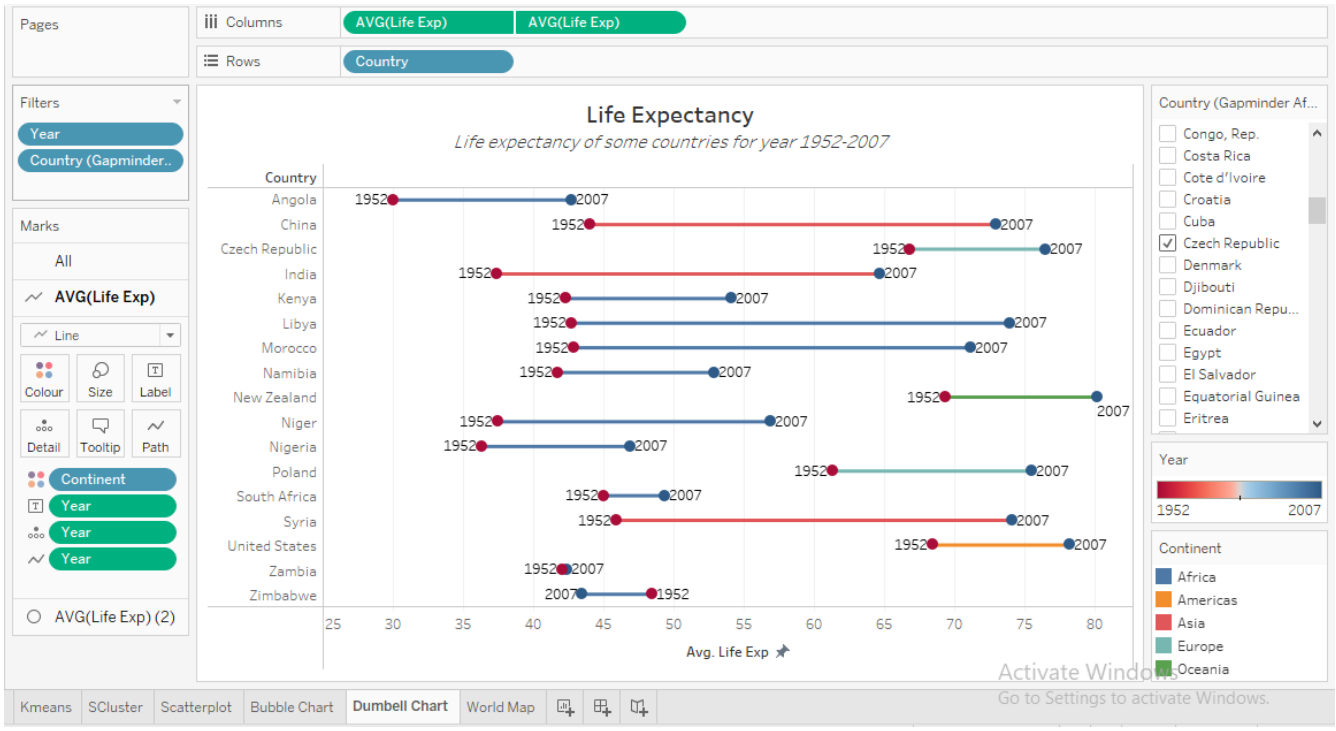
The scatter plot of GDP per capita versus life expectancy for different countries. The chart depicts the journey of different countries from 1952 to 2007. Each point in the plot represents a country, with the x-axis indicating its GDP per capita and the y-axis representing its life expectancy. The scatter plot is divided into several regions(10 countries), each of which is colored differently. It appears that countries which are **closer to each other in the MDS plot are also closer in the connected scatter plot**, suggesting some underlying similarity between these countries. From the plot, it is evident that **countries with higher GDP per capita tend to have a higher life expectancy**, suggesting a positive correlation between the two variables.

Bubble Chart –



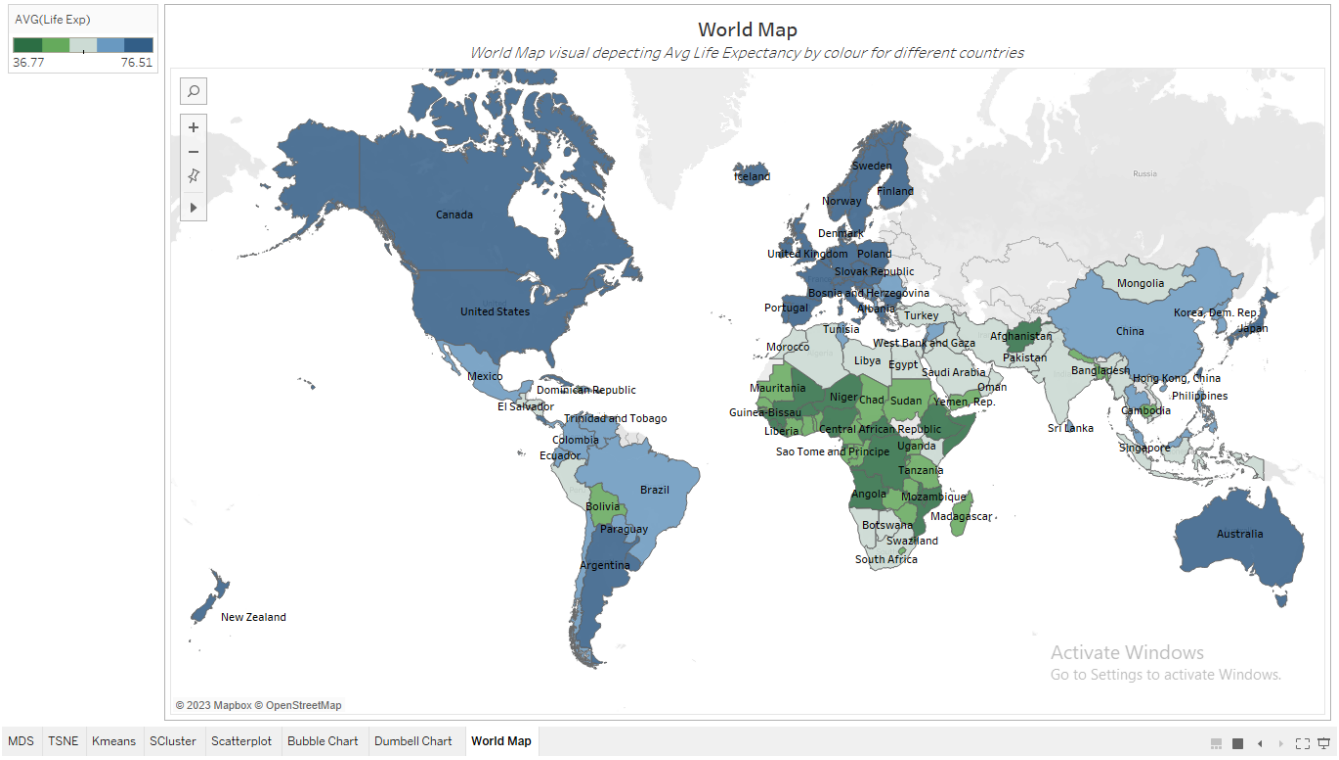
- The bubble chart depicts the relationship between GDP per capita, life expectancy, and population of each country.
- The countries considered in the previous plots have been marked in the bubble chart as well.
- The **size of the bubble** represents the **population** of the country.
- The x-axis represents the GDP per capita of each country, with countries **on the right having a higher GDP per capita.**
- The y-axis represents the life expectancy of each country, with countries **at the top having a higher life expectancy.**
- The bubbles are color-coded based on the continent in which the country is located.
- Most countries with higher GDP per capita also have higher life expectancy.
- The bubbles are distributed in a clustered pattern, with most of the bubbles concentrated in the bottom left corner of the chart.
- The **GDP per Capita is shown in a log scale for a clearer visualization**.
- Overall, the bubble chart provides a clear and comprehensive visualization of the relationship between GDP per capita, life expectancy, and population of each country.

Line Chart –



- The above chart displays the shift in life expectancy across a range of countries from 1952 to 2007, with the line graph color-coded according to the respective continents.
- Notably, the graph reveals that a significant number of African nations have a life expectancy below 50 years.
- Zimbabwe, in particular, stands out as the **sole country where life expectancy has decreased** over time.
- Additionally, countries that cluster together in both K means and S Cluster analyses display similar trends in life expectancy over time.
- On a positive note, **Libya and Morocco** are two African nations that exhibit comparatively **high average life expectancies**.

World Map –



- The world map chart above provides a visualization of the average life expectancy across different countries, with the map color-coded according to life expectancy.
- Notably, **African countries exhibit significantly lower life expectancies** compared to other regions of the world.
- This finding reinforces the grouping of African nations under a single cluster in both K means and S Cluster visuals.
- Apart from this, **Afghanistan** also has less life expectancy.
- In sum, the world map chart provides additional evidence to support the placement of countries in their respective groups(cluster) based on life expectancy.