Dataset: Healthy Lifestyle Cities Report 2021
Healthy lifestyle metrics of top 44 cities.

1. Sunshine hours(City)
2. Cost of a bottle of water(City)
3. Obesity levels(Country)
4. Life expectancy(years) (Country)
5. Pollution(Index score) (City)
6. Annual avg. hours worked
7. Happiness levels(Country)
8. Outdoor activities(City)
9. Number of take out places(City)
10. Cost of a monthly gym membership(City)

A. Understanding Data

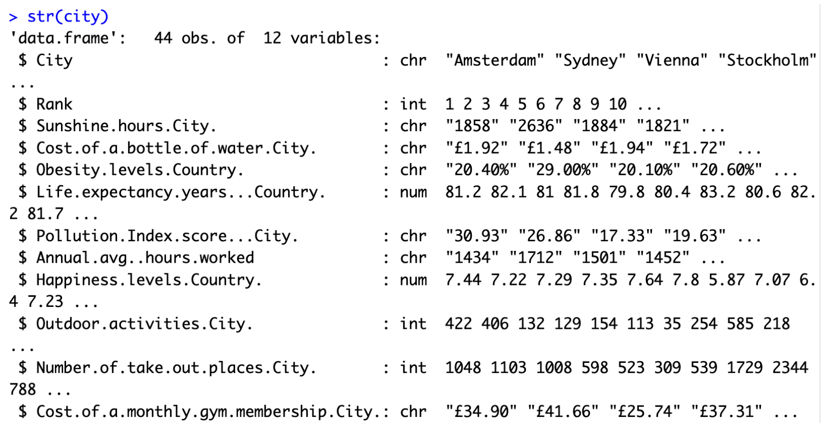
With the dataset containing information about data representing a healthy lifestyle of 44 cities, there are also 3 data about the countries of those cities: Obesity Level, Happiness level and Life Expectancy. It is possible that the data in the dataset are variables that affect the happiness level and life expectancy of those countries.

Since the cities in the list are all large cities such as the capital or economic center of the country, the Happiness Level of the country here can be considered the Happiness Level of that city. Cities such as New York, Berlin, or Milan often account for the majority of the population and economy of the country, so the Happiness Level of the country can reflect the happiness level of these cities well.

In that data, it is divided into 2 categories:

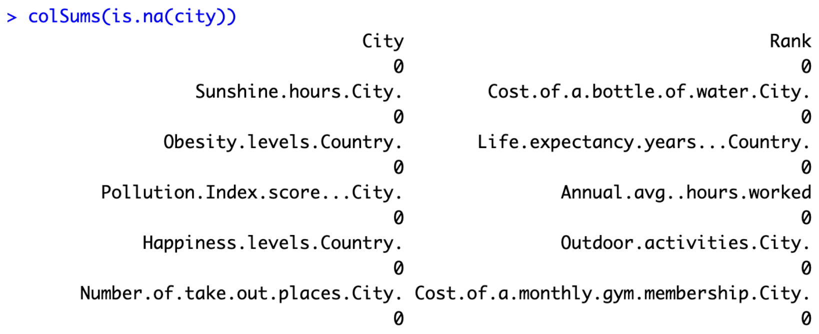
- Variables affecting mental health: Annual avg. hours worked, Number of take out places, Sunshine hours
- Variables affecting physical health: Cost of Bottle water, Pollution, Outdoor activities, Gym membership

View data types:

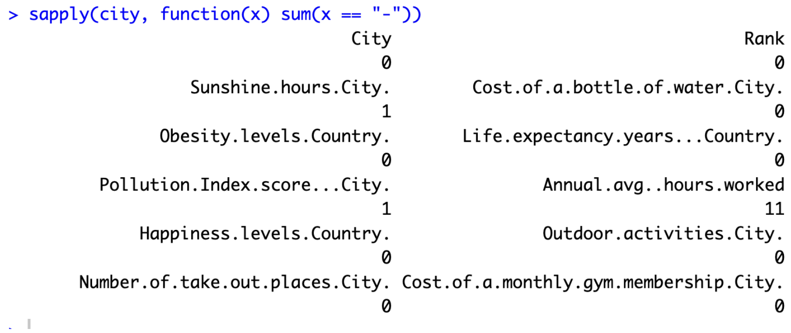


```
> str(city)
'data.frame':	44 obs. of  12 variables:
 $ City                                 : chr  "Amsterdam" "Sydney" "Vienna" "Stockholm"
...
 $ Rank                                 : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Sunshine.hours.City.                 : chr  "1858" "2636" "1884" "1821" ...
 $ Cost.of.a.bottle.of.water.City.      : chr  "£1.92" "£1.48" "£1.94" "£1.72" ...
 $ Obesity.levels.Country.              : chr  "20.40%" "29.00%" "20.10%" "20.60%" ...
 $ Life.expectancy.years...Country.     : num  81.2 82.1 81 81.8 79.8 80.4 83.2 80.6 82.
2 81.7 ...
 $ Pollution.Index.score...City.        : chr  "30.93" "26.86" "17.33" "19.63" ...
 $ Annual.avg..hours.worked             : chr  "1434" "1712" "1501" "1452" ...
 $ Happiness.levels.Country.            : num  7.44 7.22 7.29 7.35 7.64 7.8 5.87 7.07 6.
4 7.23 ...
 $ Outdoor.activities.City.             : int  422 406 132 129 154 113 35 254 585 218
...
 $ Number.of.take.out.places.City.      : int  1048 1103 1008 598 523 309 539 1729 2344
788 ...
 $ Cost.of.a.monthly.gym.membership.City.: chr  "£34.90" "£41.66" "£25.74" "£37.31" ...
```

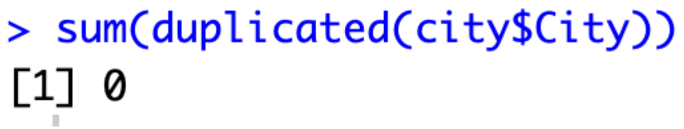
Checking for missing data



```
> colSums(is.na(city))
                                  City                                   Rank 
                                     0                                      0 
                  Sunshine.hours.City.        Cost.of.a.bottle.of.water.City. 
                                     0                                      0 
               Obesity.levels.Country.       Life.expectancy.years...Country. 
                                     0                                      0 
         Pollution.Index.score...City.               Annual.avg..hours.worked 
                                     0                                      0 
             Happiness.levels.Country.               Outdoor.activities.City. 
                                     0                                      0 
       Number.of.take.out.places.City. Cost.of.a.monthly.gym.membership.City. 
                                     0                                      0 
```



```
> sapply(city, function(x) sum(x == "-"))
                                  City                                   Rank 
                                     0                                      0 
                  Sunshine.hours.City.        Cost.of.a.bottle.of.water.City. 
                                     1                                      0 
               Obesity.levels.Country.       Life.expectancy.years...Country. 
                                     0                                      0 
         Pollution.Index.score...City.               Annual.avg..hours.worked 
                                     1                                     11 
             Happiness.levels.Country.               Outdoor.activities.City. 
                                     0                                      0 
       Number.of.take.out.places.City. Cost.of.a.monthly.gym.membership.City. 
                                     0                                      0 
```

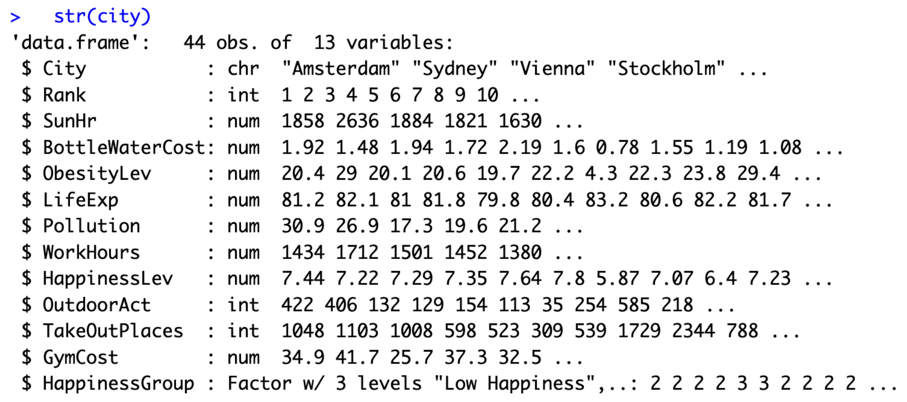


```
> sum(duplicated(city$City))
[1] 0
```

Missing data is filled as ‘-’, then I fill in missing values using the mean method.

Remove unnecessary symbols in columns.

Rename columns.



```
>   str(city)
'data.frame':   44 obs. of  13 variables:
 $ City          : chr  "Amsterdam" "Sydney" "Vienna" "Stockholm" ...
 $ Rank          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ SunHr         : num  1858 2636 1884 1821 1630 ...
 $ BottleWaterCost: num  1.92 1.48 1.94 1.72 2.19 1.6 0.78 1.55 1.19 1.08 ...
 $ ObesityLev    : num  20.4 29 20.1 20.6 19.7 22.2 4.3 22.3 23.8 29.4 ...
 $ LifeExp       : num  81.2 82.1 81 81.8 79.8 80.4 83.2 80.6 82.2 81.7 ...
 $ Pollution     : num  30.9 26.9 17.3 19.6 21.2 ...
 $ WorkHours     : num  1434 1712 1501 1452 1380 ...
 $ HappinessLev  : num  7.44 7.22 7.29 7.35 7.64 7.8 5.87 7.07 6.4 7.23 ...
 $ OutdoorAct    : int  422 406 132 129 154 113 35 254 585 218 ...
 $ TakeOutPlaces : int  1048 1103 1008 598 523 309 539 1729 2344 788 ...
 $ GymCost       : num  34.9 41.7 25.7 37.3 32.5 ...
 $ HappinessGroup : Factor w/ 3 levels "Low Happiness",..: 2 2 2 2 3 3 2 2 2 2 ...
```

B. Objective

I am interested in the Happiness Level variable here, whether it is influenced more by variables that affect mental or physical health, so the target question is which variables will have an influence on the Happiness Level of a city.

My assumption:

- Outdoor activities and SunHr are the two variables that have the greatest influence on Happiness Level
- Gymcost and Waterbottlecost are 2 variables that have a positive correlation with Obesity Level.

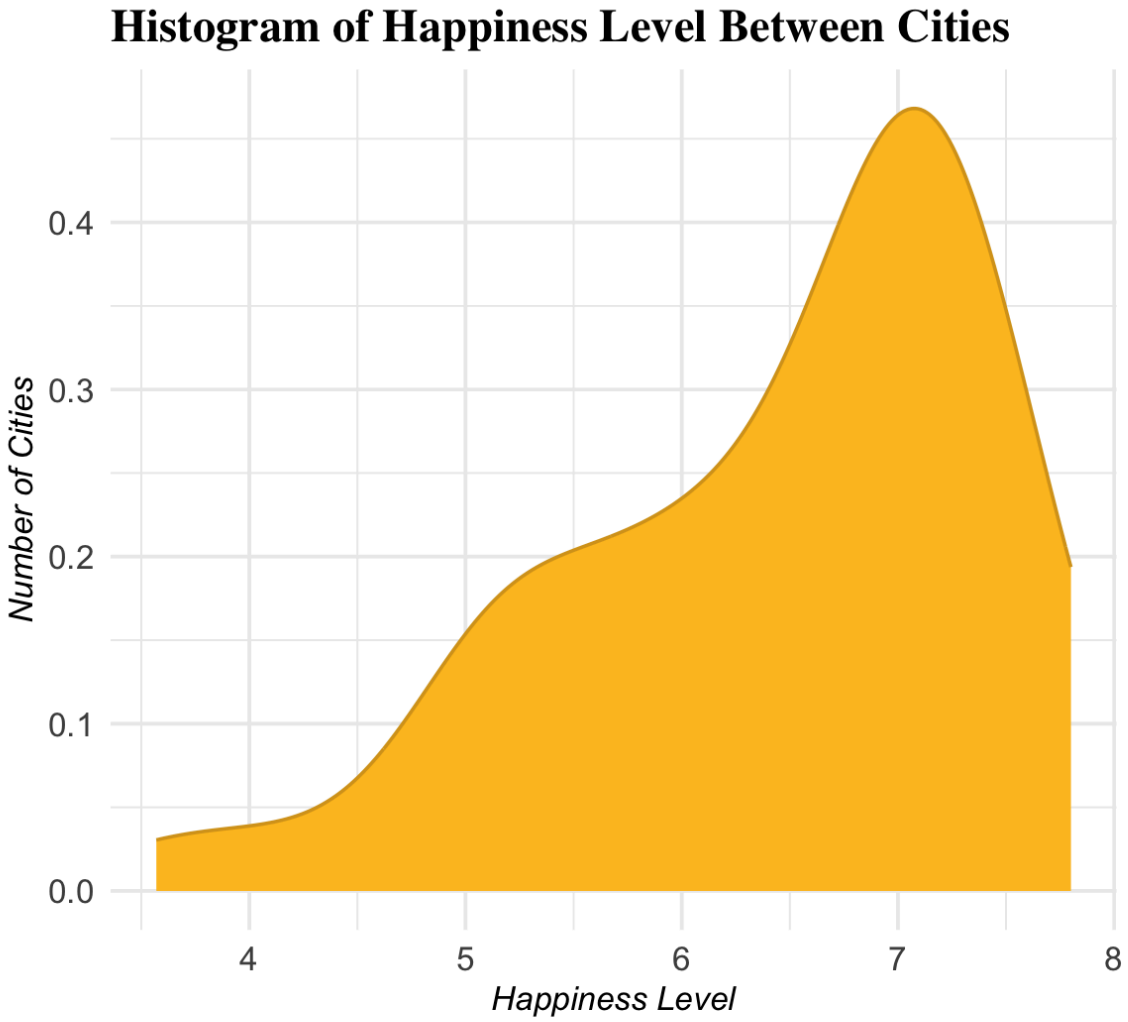
C. Data Visualization

1. Assumption for a healthy lifestyle within a city

- Great amount of SunHr and Outdoor Act
- Low Pollution Index
- High Life Expectancy
- Maybe low cost on gym and water bottle

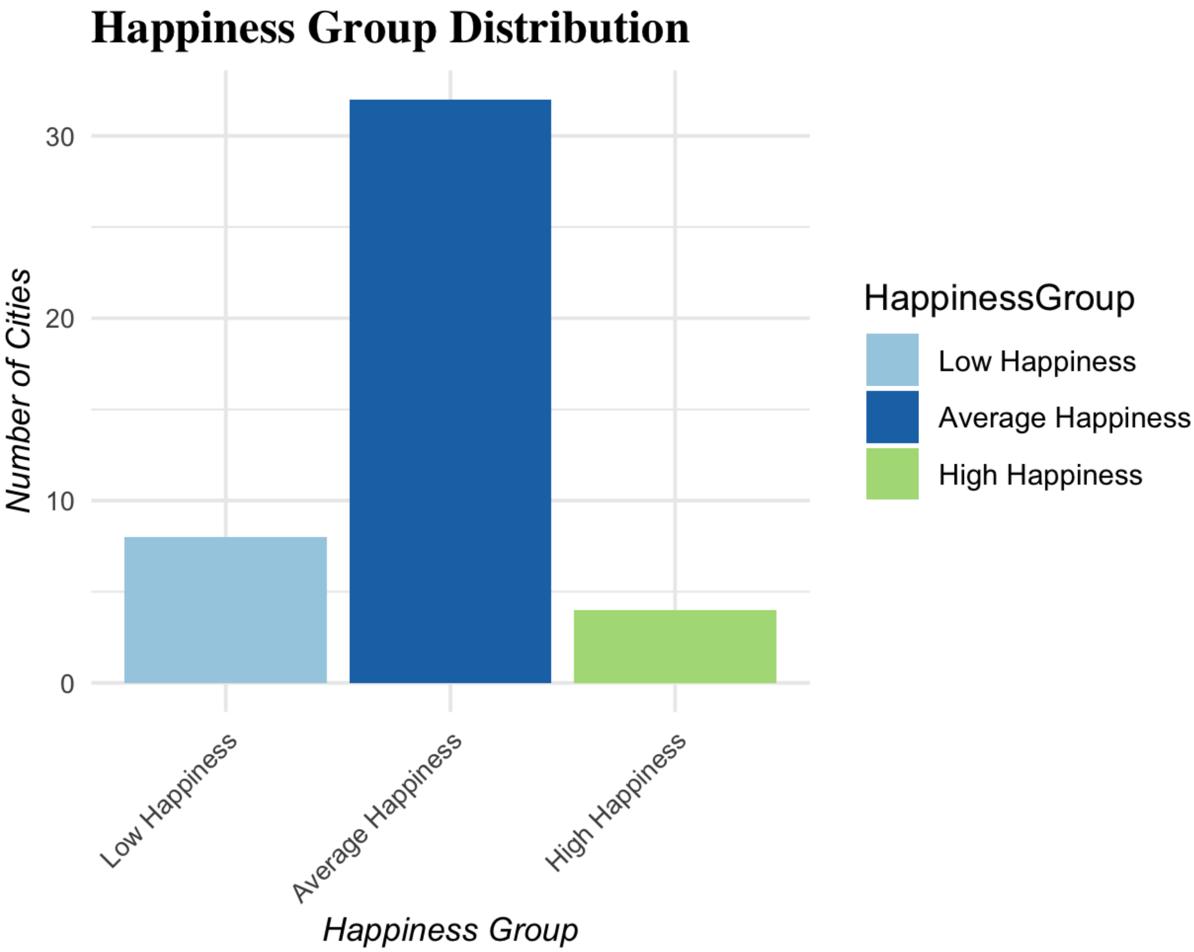
2. Grouping according to Happiness Level

Happiness Level world wide was 5.54 points based on 134 countries in 2022



Graph 1.

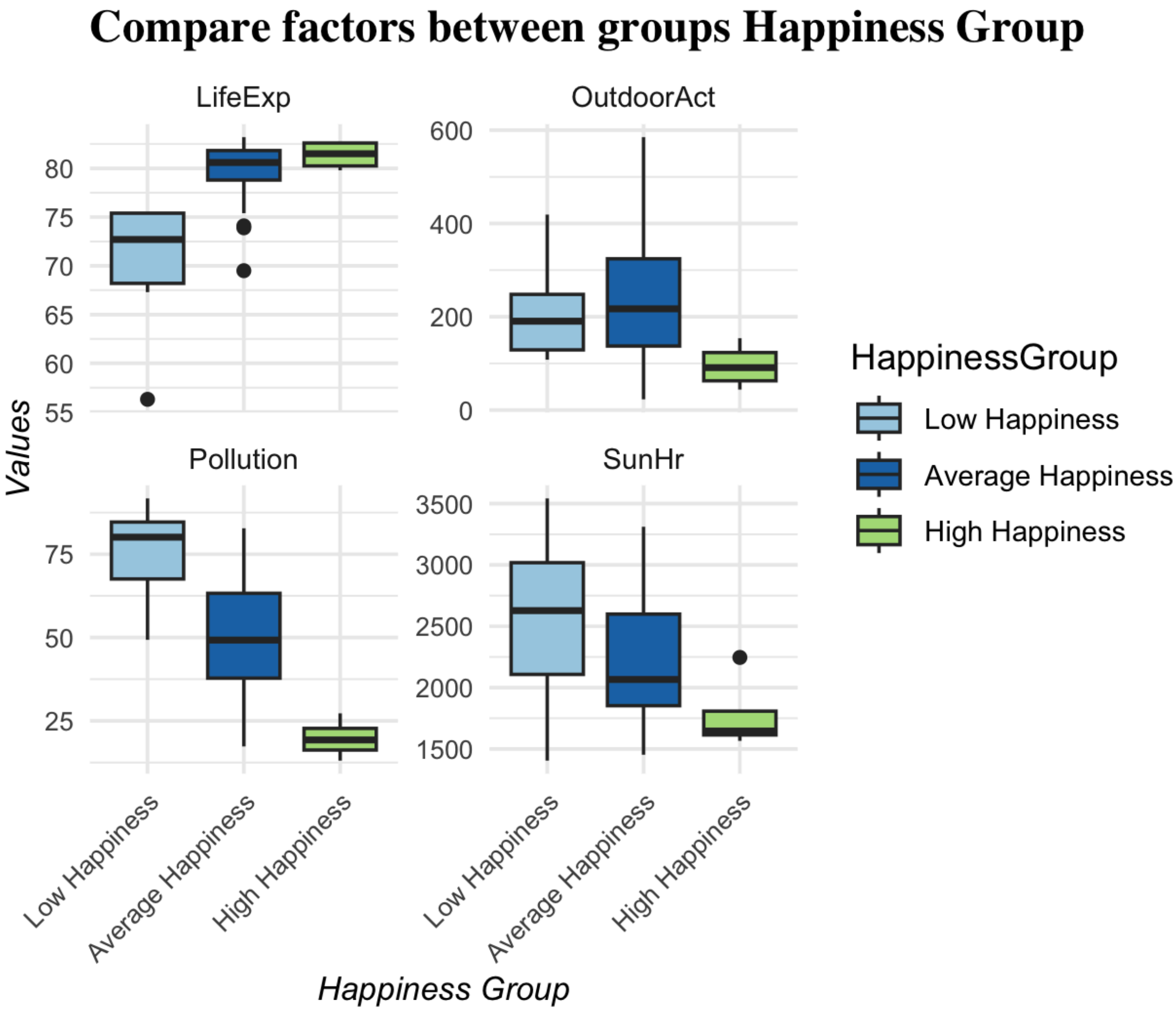
- Data concentrated in bins with happiness level = 7, uneven distribution, skewed to the right
- Grouping based on breaks (0, 5.5, 7.5, Inf)



Graph 2.

The average_happiness group has the highest number of cities (more than 30), followed by the low_happiness group. So the number of cities with high happiness level is very low.

3. Analyzing Groups of Happiness Level



Graph 3.

a. Life Expectancy

- The first thing we notice is that the Low Happiness Group has a much lower Median compared the two others. While the Average and High Happiness group have Life Expectancy over 80 years old, the Low Happiness just get over 70 years. Moreover the Lowest Group also have IQR three times bigger than other groups, and Outliers way down to 55 years old, which shows that data distributed dispersed and some cities have very low life expectancy.

-> There is an observation that the lower the Happiness Level, the lower the Life Expectancy, assuming Happiness Level has a latent effect on Life Expectancy.

b. Pollution and SunHr

There are some similarities between Pollution and SunHr features.
The higher the Happiness Level, the lower their Pollution Index and Sun Hours. The Low Happiness Group is at the top for Pollution Index and SunHr, while the High Happiness Group is at the bottom.The cities in the low happiness group have Pollution Index Median almost 4 times higher than the high happiness cities, and almost twice as high for the Average group.

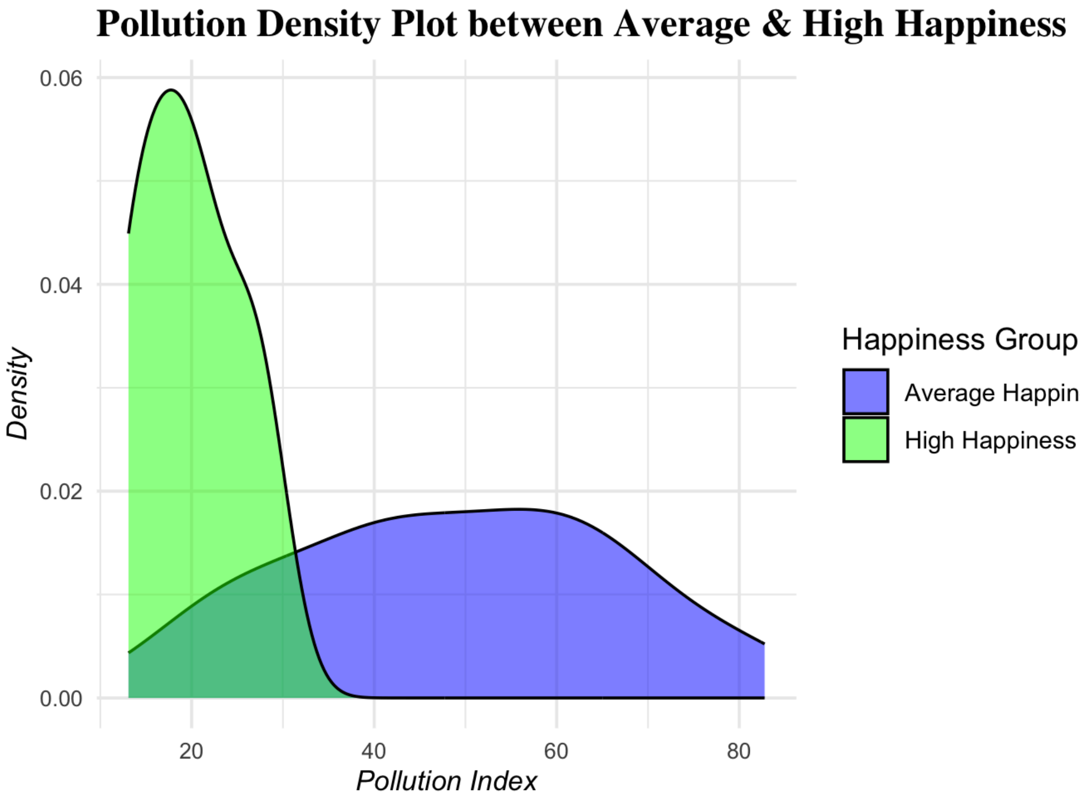
All 3 groups have no outliers for the Pollution value, which suggests that the data is normally distributed and balanced, with the average and high media groups both lying exactly in the middle which is an interesting point, *we will look at the density plot for further analysis.*
For the SunHr value, the surprise here is that the high High Happiness Level group has the lowest SunHr, as well as the smallest IQR, suggesting that the cities in this group have similar Sun Hours in the range of 1500-2000 hours.

The Low Happiness group has the highest number of sunshine hours, ranging from 1500-3500 hours, with the Median above around 2500 hours. This may be because cities in this group often have to work outdoors such as agriculture, construction, **...** or High sunshine hours can cause inconvenience or **stress**, such as: Higher temperatures (heat stress) when working in the sun or lack of rest time or harsh working environment.
This shows that High sunshine hours may not mean high happiness. Although high sunshine hours, other factors such as pollution, work pressure, or lack of rest time can negatively affect happiness. Moreover, Pollution Index can be an important feature that affects the Happiness Level of cities

c. OutdoorAct

The OutdoorAct feature has the most different distribution of data compared to the other three variables, with the Average group having the highest number of outdoor activities, ranging from around 50 to over 3000.
The High Happiness group is at the bottom in terms of number of outdoor activities

d. Significant observation of Pollution variables



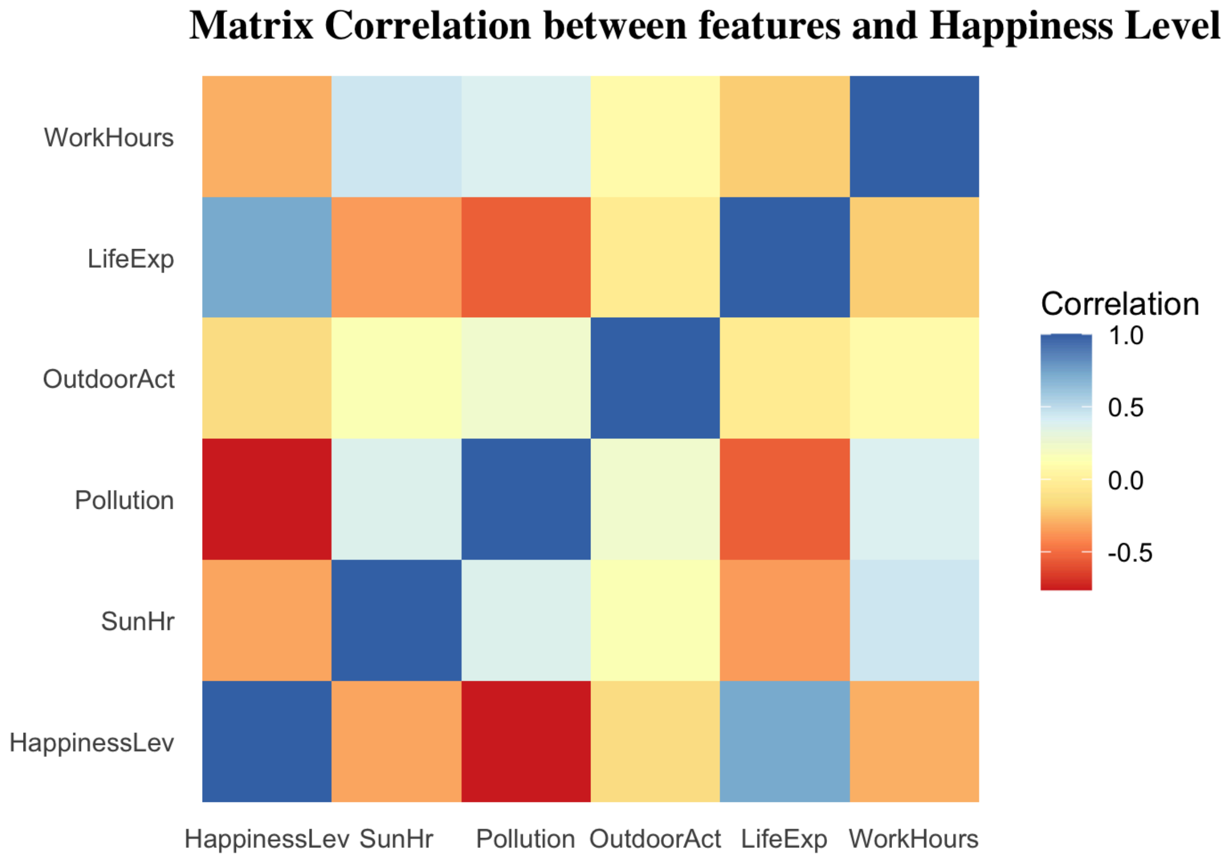
Graph 4.

The Pollution Index of the Average and High Happiness groups is very evenly distributed or what you might call a normal distribution, so I decided to take a closer look at it.

The Mode of Average is high and narrow, indicating that the data is less volatile, cities with **high levels of happiness often have similar Pollution index** and are **around 20**, as well as the left-skewed curve shows that the Pollution index of these cities has a lower value than the common level.

Unlike the density plot of the High Happiness group, Average Happiness has a symmetrical curve and the mode is 3 times smaller, the spread is also wider, indicating a more scattered data distribution. **The Mode of is around the pollution index value of 50**, the spread also shows that the data of this group is more diverse, reflecting **the cities with very different environmental conditions**. However, this group has 32 cities, 4 times more than the High Happiness group, so it is also part of the reason why it has a wider spread.

4. Correlation Between Features



Graph 5.

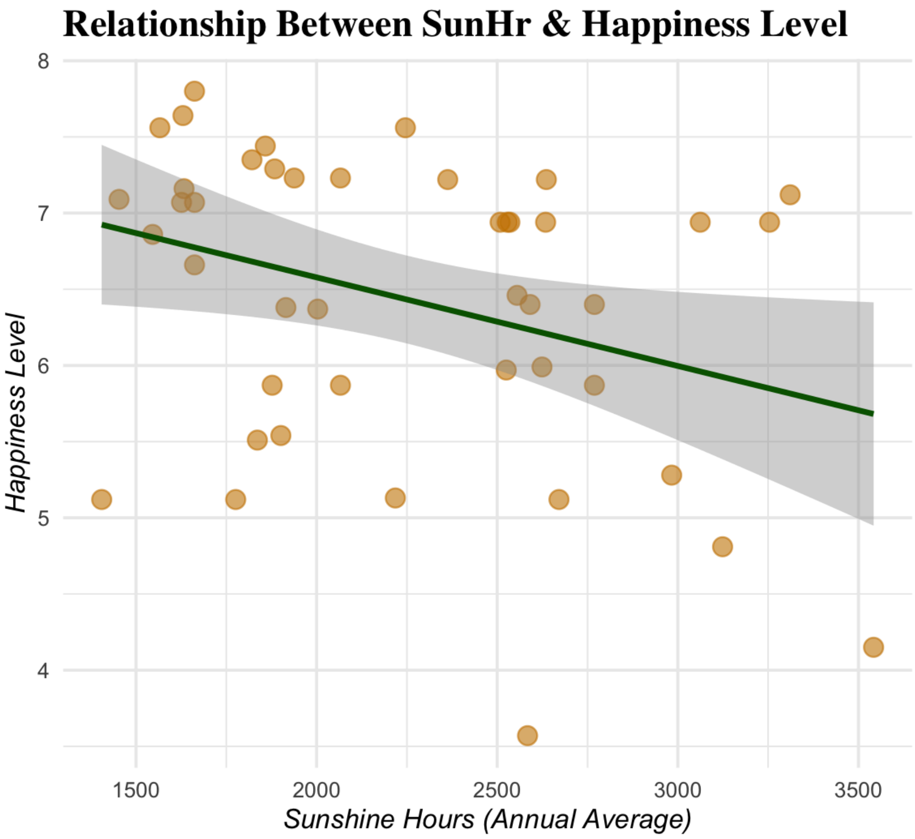
The value with a high positive correlation with Happiness Level is LifeExp, in contrast to Pollution.

SunHr and WorkHours have a slight negative correlation with HappinessLev, indicating a low influence on Happiness, but it has a negative influence.

- Positive Factors Affecting Happiness:
    - LifeExp: Higher life expectancy is often associated with higher happiness levels.
- Negative Factors Affecting Happiness:
    - Pollution: Cities with high pollution levels often have lower happiness levels.
    - SunHr

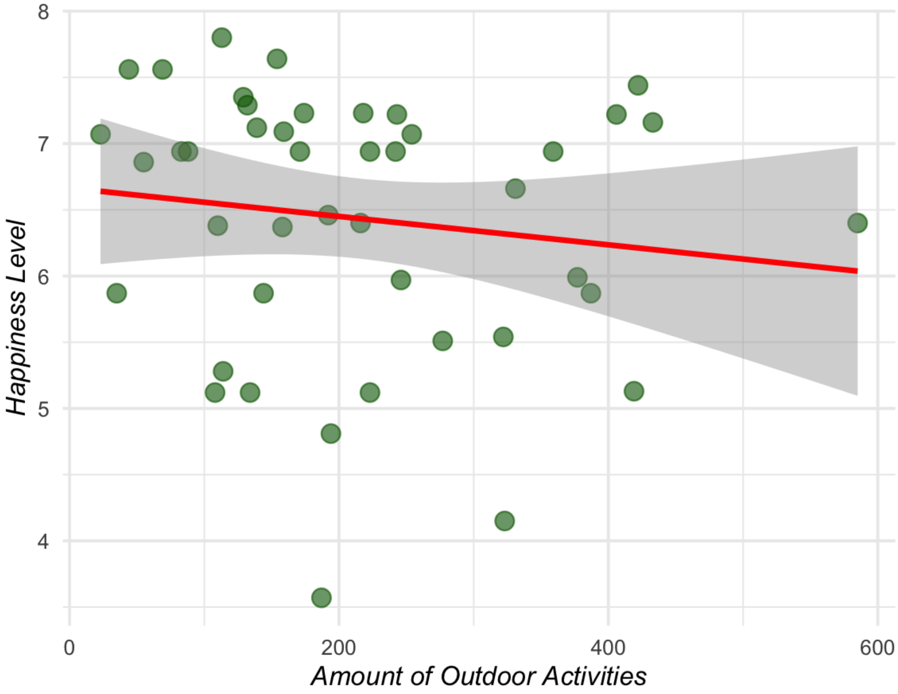
    - WorkHours: More working hours can reduce happiness.
- Low Impact Factors:
    - OutdoorAct: not important determinants of happiness.

5. Special Features with Happiness Level

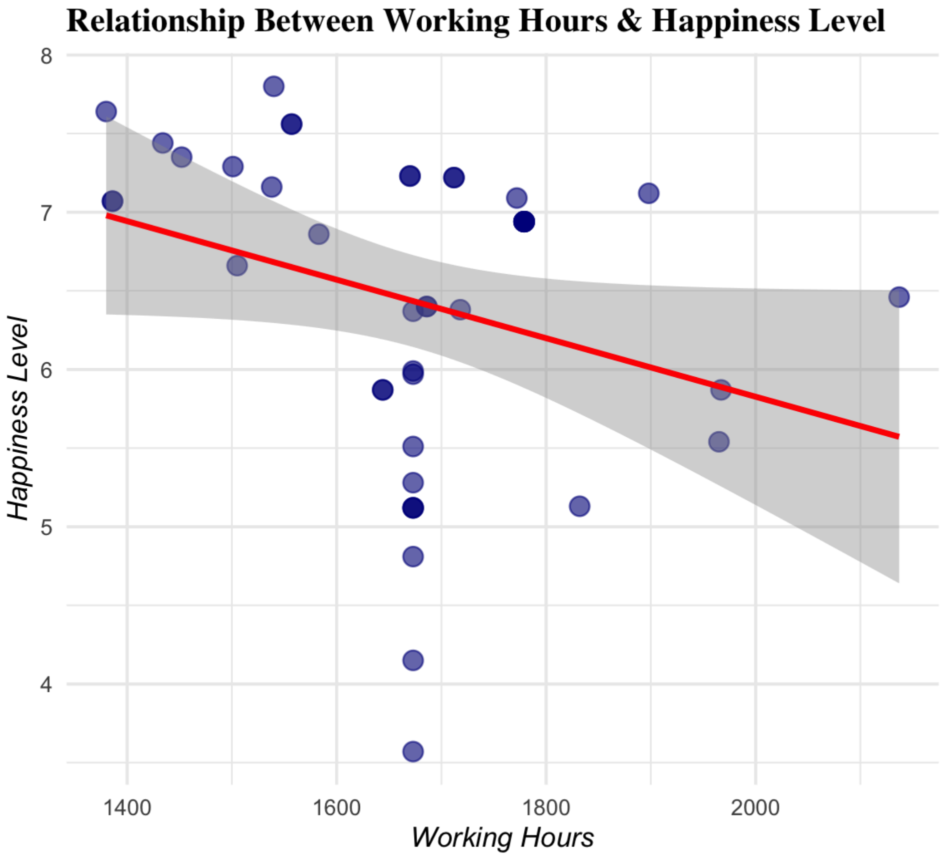


Graph 6.

Relationship Between Outdoor Activites & Happiness Level

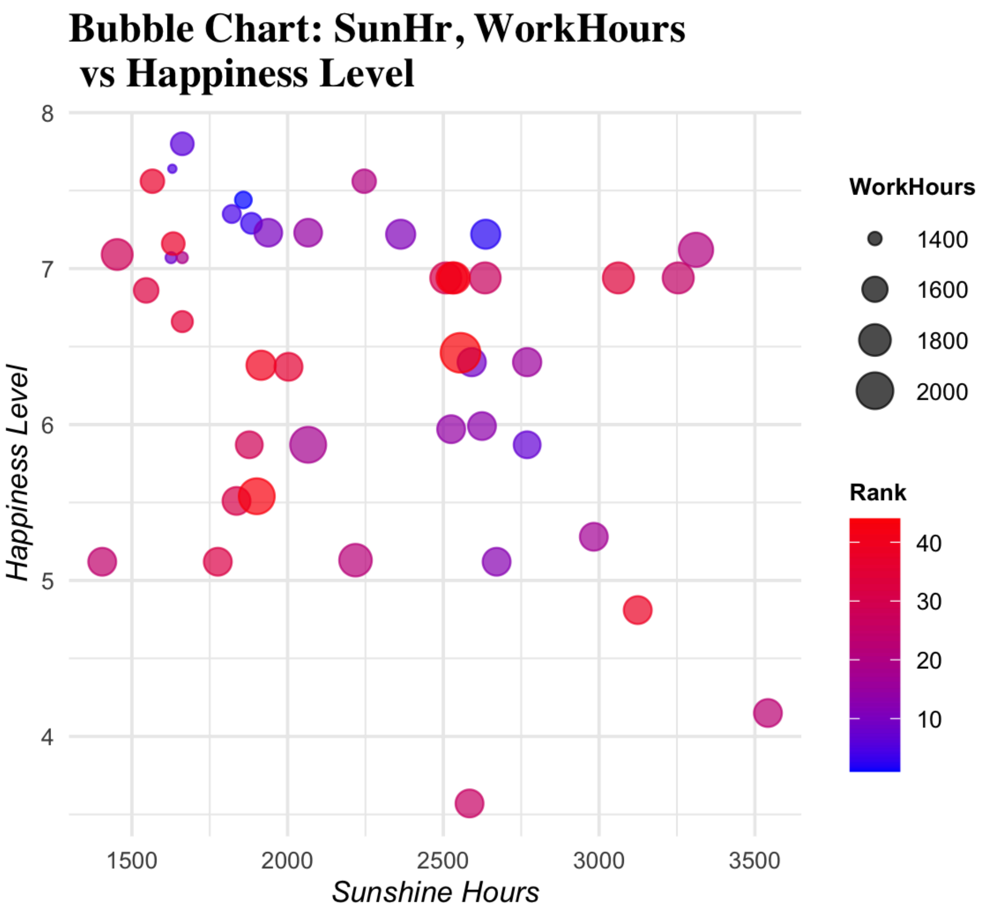


Graph 7.



Graph 8.

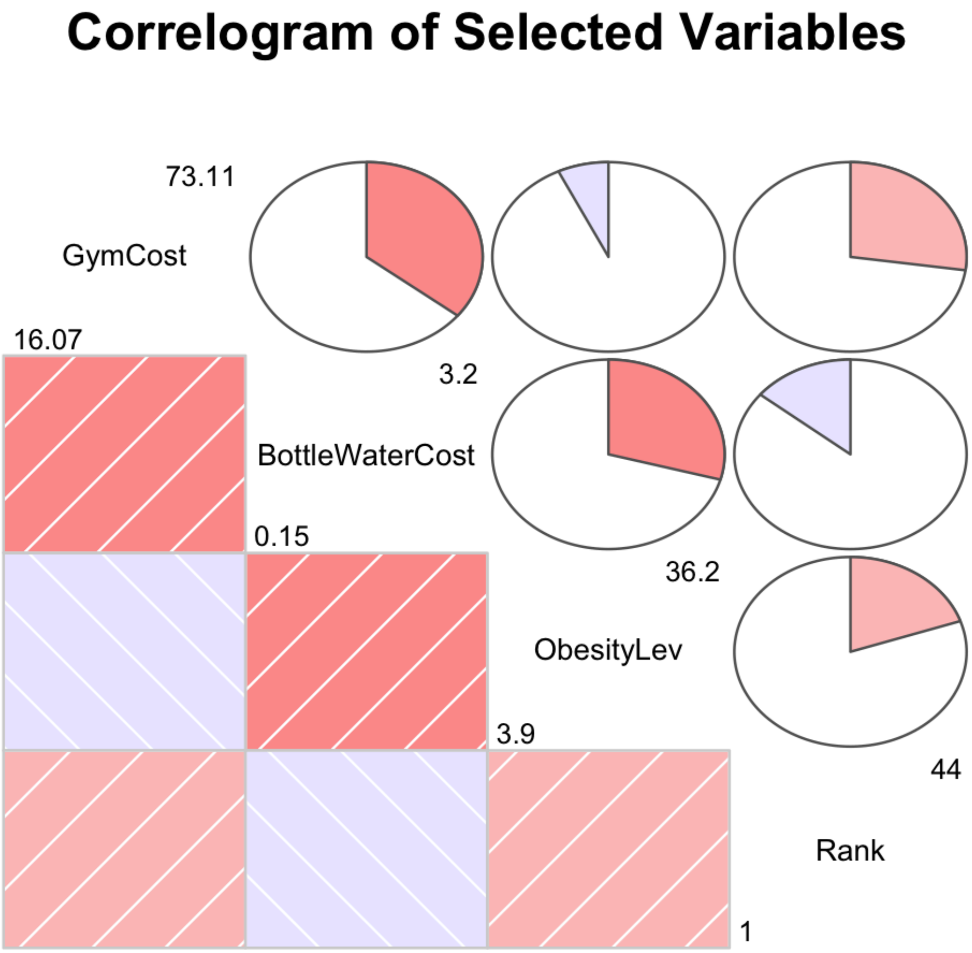
All three trend lines have negative slopes, indicating that both SunHr and Outdoor Activities as well as Working Hours have a negative impact on Happiness Level. SunHr and Working Hours have stronger slopes and scatter points closer to the trend line, so they have a stronger impact than OutdoorAct.



Graph 9.

Large bubbles represent high work hours, which are typically in the Happiness Level 5-7 range, while low work hours are typically blue and fall in the Happiness Level 7 and above range. The chart shows that cities with high healthy lifestyle rankings tend to have low work hours.

6. Gym Cost and Bottle Water Cost



Graph 10.

The positive correlation value shows that as gym cost (GymCost) increases, so does bottle water cost (BottleWaterCost). This may suggest that both variables reflect the high standard of living in cities where the general cost of living is high.

Relationship to ObesityLev

This observation suggests that **gym cost does not have a significant impact on obesity rates**. This may suggest that healthy lifestyles are not only dependent on gym cost, but also related to other factors (such as eating habits or opportunities for outdoor physical activity). However, there is a weak relationship between BottleWaterCost and ObesityLev. The slight positive correlation suggests that wealthier cities (high water cost) may have more unhealthy eating habits, leading to higher obesity rates. The low correlation between Rank and the cost variables suggests that it does not necessarily reflect the healthy lifestyles of these cities. Rank is more influenced by factors like lifestyle, health, or longevity than cost (GymCost, BottleWaterCost).

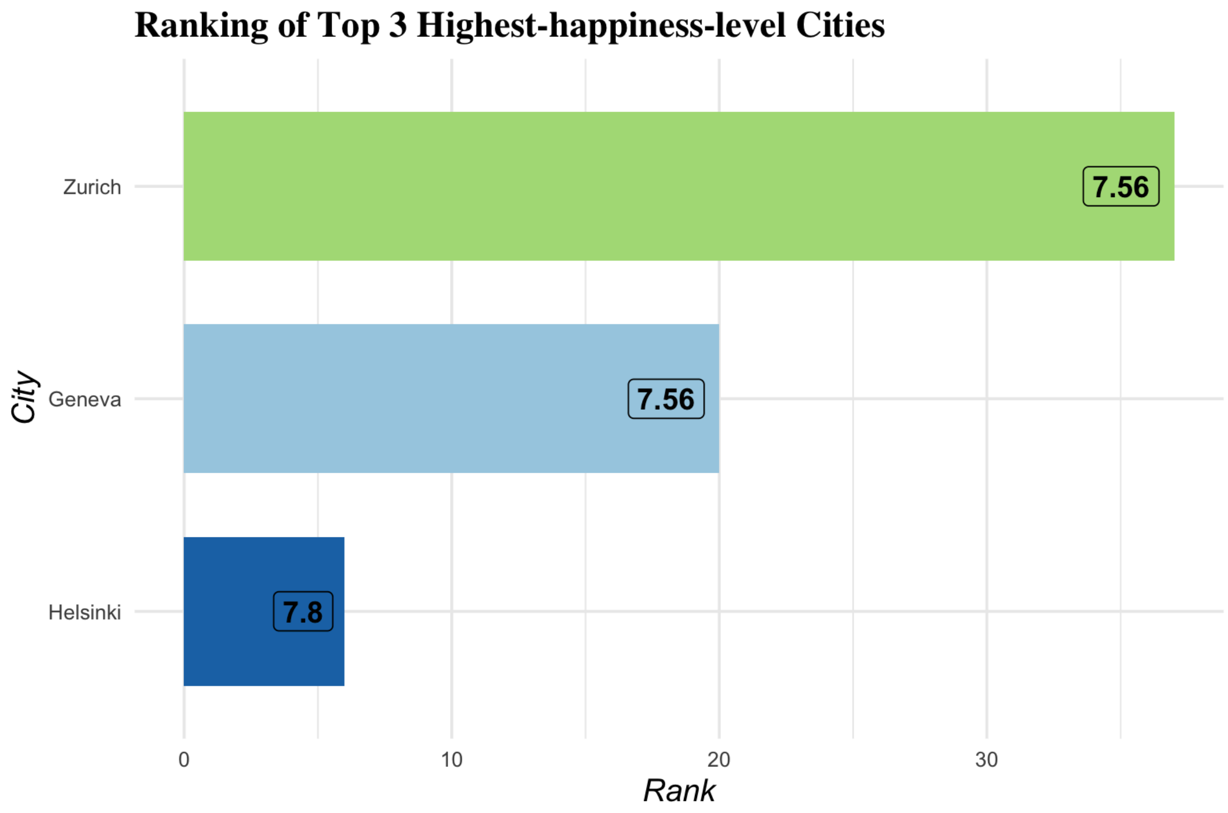
D. Conclusion:

These are the top 3 cities with high Happiness Levels as well as high indexes to fit the condition of cities with healthy lifestyles.

The analysis shows that **cities with high Happiness Levels often have the outstanding characteristics of healthy lifestyles: high average life expectancy, low pollution levels, short working hours, and sometimes moderate or lower sunshine hours.**

Among the three happiest cities, Helsinki, Geneva, and Zurich, there is a notable difference. Although Helsinki has the highest happiness index, it is not the top city for healthy living. Similarly, Geneva and Zurich, although having the same happiness index, have significantly different rankings for healthy living (Geneva is ranked 20th, while Zurich is ranked 30th). This suggests that happiness is not solely dependent on factors related to physical health.

This result emphasizes that, although a healthy lifestyle plays an important role in improving quality of life, it is not the only factor that determines happiness. Other factors such as community connection, personal freedom, culture, and mental health also contribute to satisfaction and happiness. This encourages a balanced approach, combining physical and mental aspects of health, to achieve overall happiness.



Graph 11.

Ref:

- https://www.theglobaleconomy.com/rankings/happiness/#:~:text=Happiness%20Index%2C%200%20(unhappy),countries%20where%20data%20are%20available.