UNIVERSITY OF WESTMINSTER

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

5DATA006C.1 Data Visualisation and Communication Portfolio

Ruchintha Dias

Blackboard Name : - Ruchintha Dias

UOW ID: W20844558

IIT No: 20232724

Research Question and Data Sourcing

Research Question

- What factors affect the happiness level of a country/city.

Why is this relevant?

- The relevance of this research question lies in its focus on happiness. This is because it can be identified as one of the key indicators of social well-being. By understanding the involvement in the factors like work life balance, health and pollution can guide policymakers and infrastructure developers in prioritizing the work that needs to be done to promote better living conditions. Insights from this study can help prioritize initiatives that improve mental health, environmental quality and lifestyle satisfaction. This will ultimately enhance the happiness of people living in cities and countries. Moreover, this can serve as a basis for creating future urban policies aiming to promote prosperous and sustainable communities.

Data Sourcing

To continue the studies on this research question, a dataset named 'Healthy Lifestyle Cities Report 2021' was taken from Kaggle website. This dataset contains data on 44 cities worldwide, covering metrics like Sunshine hours, Cost of a bottle of water, Obesity levels, Life expectancy years, Pollution Index score, Annual average hours worked, Happiness levels, Outdoor activities, Number of take-out places and Cost of a monthly gym membership.

Reference;

https://www.kaggle.com/datasets/prasertk/healthy-lifestyle-cities-report-2021

Data Preparation

Evaluation of the dataset:

- The dataset is not tidy. It has missing values and unnecessary symbols included in the dataset.

Variables,

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

Observations:

Eash row of the dataset shows the data related to a city.

Transformations:

- Replacing missing value sign ‘-’ with ‘NaN’

Before

	City	Rank	Sunshine hours(City)	Cost of a bottle of water(City)	Obesity levels(Country)	Life expectancy(years) (Country)	Pollution(Index score) (City)	Annual avg. hours worked
0	Amsterdam	1	1858	£1.92	20.40%	81.2	30.93	1434
1	Sydney	2	2636	£1.48	29.00%	82.1	26.86	1712
2	Vienna	3	1884	£1.94	20.10%	81.0	17.33	1501
3	Stockholm	4	1821	£1.72	20.60%	81.8	19.63	1452
4	Copenhagen	5	1630	£2.19	19.70%	79.8	21.24	1380
5	Helsinki	6	1662	£1.60	22.20%	80.4	13.08	1540
6	Fukuoka	7	2769	£0.78	4.30%	83.2	-	1644
7	Berlin	8	1626	£1.55	22.30%	80.6	39.41	1386
8	Barcelona	9	2591	£1.19	23.80%	82.2	65.19	1686
9	Vancouver	10	1938	£1.08	29.40%	81.7	24.26	1670

Code & Result

```
#replace '-' to NaN
import numpy as np
df.replace("-", np.nan, inplace = True)
df.head(10)
```

	City	Rank	Sunshine hours(City)	Cost of a bottle of water(City)	Obesity levels(Country)	Life expectancy(years) (Country)	Pollution(Index score) (City)	Annual avg. hours worked
0	Amsterdam	1	1858	£1.92	20.40%	81.2	30.93	1434
1	Sydney	2	2636	£1.48	29.00%	82.1	26.86	1712
2	Vienna	3	1884	£1.94	20.10%	81.0	17.33	1501
3	Stockholm	4	1821	£1.72	20.60%	81.8	19.63	1452
4	Copenhagen	5	1630	£2.19	19.70%	79.8	21.24	1380
5	Helsinki	6	1662	£1.60	22.20%	80.4	13.08	1540
6	Fukuoka	7	2769	£0.78	4.30%	83.2	NaN	1644

The code, ‘df.replace("-",np.nan,inplace=True)’ replace ‘-‘ symboles in ‘df’ dataframe with the help of ‘.replace()’ method of python to ‘np.nan’, a null value recognize by the python server. ‘inplace=True’ set the changes to be made directly to the ‘df’ dataframe.

- Removing spaces from column names

Code

```
#Removing spaces from columns
df.columns = df.columns.str.replace(' ', '_')
```

Here, by using the '.str' to convert string operations on the column names and using '.replace(' ', '_')' replace all columns with spaces (' ') in the column names to underscores ('_').

- Identifying columns and count of missing data

Code & Result

```
[4]: #Identifying missing data
     df.isnull().sum()

[4]: City                                    0
     Rank                                    0
     Sunshine_hours(City)                    1
     Cost_of_a_bottle_of_water(City)         0
     Obesity_levels(Country)                 0
     Life_expectancy(years)_(Country)        0
     Pollution(Index_score)_(City)           1
     Annual_avg._hours_worked               11
     Happiness_levels(Country)               0
     Outdoor_activities(City)                0
     Number_of_take_out_places(City)         0
     Cost_of_a_monthly_gym_membership(City)  0
     dtype: int64
```

Here, by using the '.isnull' function to get Boolean expression of true=1 and false=0 and using '.sum' function counts all the null values within the columns of 'df' dataframe.

- Filling missing data with median value

Code & Result

```
#Filling up missing values

# Calculate the round up median value for "Sunshine_hours(City)"
Sunshine_hrs_median = round(df["Sunshine_hours(City)"].astype("float").median())
print("Sunshine_hours(City):", Sunshine_hrs_median)
# Replace NaN with the integer mean value in the "Sunshine_hours(City)" column
df["Sunshine_hours(City)"].replace(np.nan, Sunshine_hrs_median, inplace=True)
df["Sunshine_hours(City)"] = df["Sunshine_hours(City)"].astype("int")

# Calculate the round up median value for "Pollution(Index_score)_(City)"
```

The code ‘Sunshine_hrs_median=round(df["Sunshine_hours(City)"].astype("float").median())’, use ‘.astype’ function to convert the type of data into ‘float’ and use ‘.median()’ fuction to find median of the ‘Sunshine_hours(City)’ column and store it to variable called ‘Sunshine_hrs_median’.

The code ‘df["Sunshine_hours(City)"].replace(np.nan,Sunshine_hrs_median, inplace=True)’ select ‘Sunshine_hours(City)’ column in ‘df’ dataframe and find ‘np.Nan’ values or null values to replace with above assigned variable named ‘Sunshine_hrs_median’. ‘inplace=True’ set the changes to be made directly to the ‘df’ dataframe.

Then the code ‘df["Sunshine_hours(City)"]=df["Sunshine_hours(City)"].astype("int")’ change the data type of the ‘Sunshine_hours(City)’ column to integer.

- Adding and removing symbols

Code & Result

```
#Adding symbols to cloumns
df.rename(columns={'Cost_of_a_bottle_of_water(City)': 'Cost_of_a_bottle_of_water(City)(£)'}, inplace=True)
df.rename(columns={'Cost_of_a_monthly_gym_membership(City)': 'Cost_of_a_monthly_gym_membership(City)(£)'}, inplace=True)
df.rename(columns={'Obesity_levels(Country)': 'Obesity_levels(Country)(%)'}, inplace=True)

#Converting object to str
df["Cost_of_a_bottle_of_water(City)(£)"] = df["Cost_of_a_bottle_of_water(City)(£)"].astype("str")
df["Cost_of_a_monthly_gym_membership(City)(£)"] = df["Cost_of_a_monthly_gym_membership(City)(£)"].astype("str")
df["Obesity_levels(Country)(%)"] = df["Obesity_levels(Country)(%)"].astype("str")

#Removing symbols from data
df['Cost_of_a_bottle_of_water(City)(£)'] = df['Cost_of_a_bottle_of_water(City)(£)'].str.replace("£", "").astype(float)
df['Cost_of_a_monthly_gym_membership(City)(£)'] = df['Cost_of_a_monthly_gym_membership(City)(£)'].str.replace("£", "").astype(float)
```

The code ‘df.rename(columns={'Cost_of_a_bottle_of_water(City)' : 'Cost_of_a_bottle_of_water(City)(£)' }, inplace=True) use ‘df.rename()’ method to rename 'Cost_of_a_bottle_of_water(City)' to 'Cost_of_a_bottle_of_water(City)(£)' and ultimately adding ‘(£)' symbol to the column.

Code ‘df["Cost_of_a_bottle_of_water(City)(£)"]=df["Cost_of_a_bottle_of_water(City)(£)"].astype("str")’ convert ‘Cost_of_a_bottle_of_water(City)(£)’ column into string data type.

Code ‘df['Cost_of_a_bottle_of_water(City)(£)']=df['Cost_of_a_bottle_of_water(City)(£)'].str.replace("£", "").astype(float)’ use ‘str.replace’ string fuction to remove ‘£’.

Normalizing happiness levels from 0 to 1

Code

```
#Normalizing Happiness_levels(Country) figuers
df['Happiness_levels(Country)'] = round((df['Happiness_levels(Country)']-df['Happiness_levels(Country)'].min())
                                        /(df['Happiness_levels(Country)'].max()-df['Happiness_levels(Country)'].min()),2)
```

This code finds minimum and maximum values by using ‘.min()’ and ‘.max()’ function and use them to scale the columns data into ‘0’ to ‘1’. The ‘round()’ function round the answer with 2 decimal places.

Final result

```
df
```

	City	Rank	Sunshine_hours(City)	Cost_of_a_bottle_of_water(City)(£)	Obesity_levels(Country)(%)	Life_expectancy(years)_(Country)	Pollution(Index_score)_(City)	Ann
0	Amsterdam	1	1858	1.92	20.4	81.2	30.93	
1	Sydney	2	2636	1.48	29.0	82.1	26.86	
2	Vienna	3	1884	1.94	20.1	81.0	17.33	
3	Stockholm	4	1821	1.72	20.6	81.8	19.63	
4	Copenhagen	5	1630	2.19	19.7	79.8	21.24	
5	Helsinki	6	1662	1.60	22.2	80.4	13.08	
6	Fukuoka	7	2769	0.78	4.3	83.2	51.12	
7	Berlin	8	1626	1.55	22.3	80.6	39.41	
8	Barcelona	9	2591	1.19	23.8	82.2	65.19	
9	Vancouver	10	1938	1.08	29.4	81.7	24.26	
10	Melbourne	11	2363	1.57	29.0	82.1	25.90	
11	Beijing	12	2671	0.26	6.2	75.4	85.43	
12	Bangkok	13	2624	0.22	10.0	74.1	76.64	
13	Buenos Aires	14	2525	0.57	28.3	75.9	52.64	
14	Toronto	15	2066	1.09	29.4	81.7	37.83	

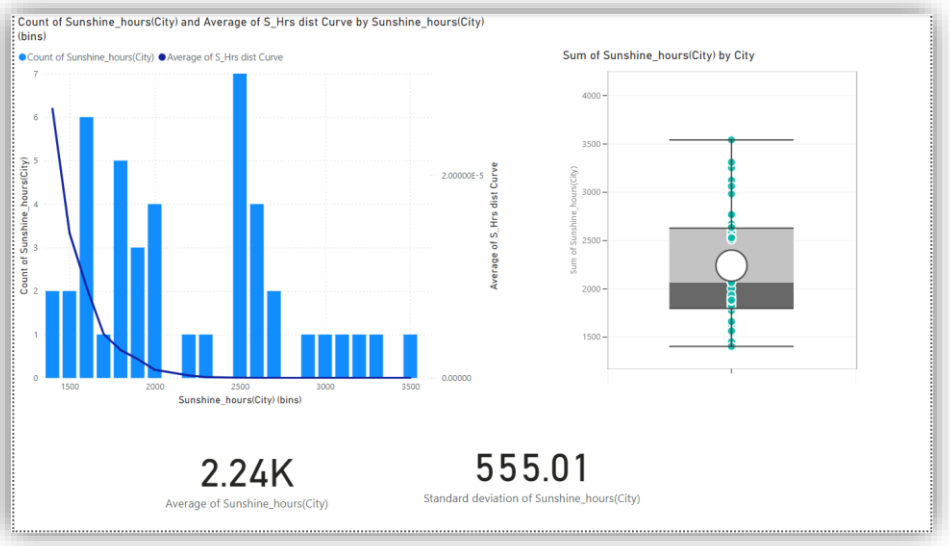
Before data cleaning

```
df
```

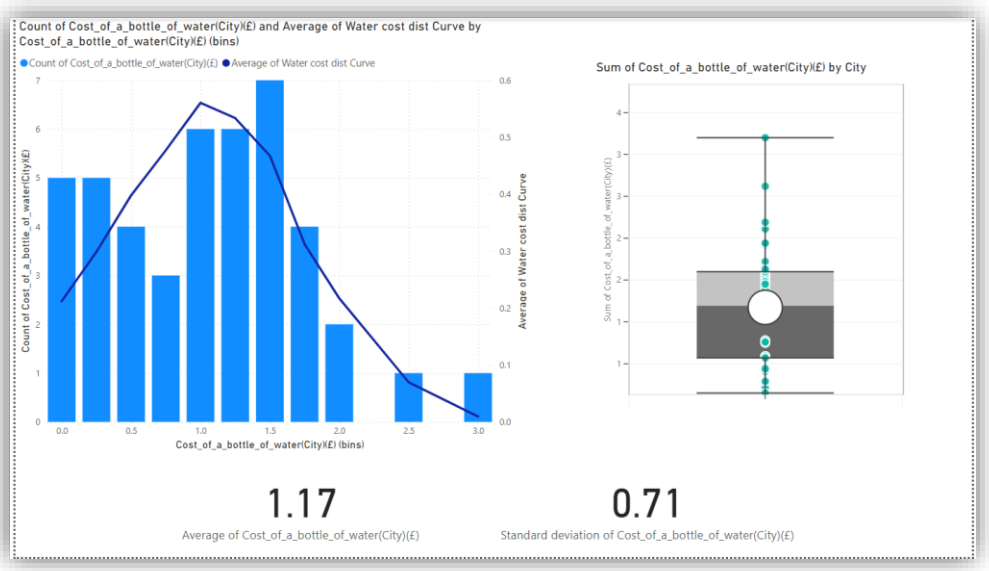
	City	Rank	Sunshine hours(City)	Cost of a bottle of water(City)	Obesity levels(Country)	Life expectancy(years) (Country)	Pollution(Index score) (City)	Annual avg. hours worked	Happiness levels(Country)	Outdoor activities(City)	Number of take out places(City)	Cost of a mont g membership(Ci
0	Amsterdam	1	1858	£1.92	20.40%	81.2	30.93	1434	7.44	422	1048	£34
1	Sydney	2	2636	£1.48	29.00%	82.1	26.86	1712	7.22	406	1103	£41
2	Vienna	3	1884	£1.94	20.10%	81.0	17.33	1501	7.29	132	1008	£25
3	Stockholm	4	1821	£1.72	20.60%	81.8	19.63	1452	7.35	129	598	£37
4	Copenhagen	5	1630	£2.19	19.70%	79.8	21.24	1380	7.64	154	523	£32
5	Helsinki	6	1662	£1.60	22.20%	80.4	13.08	1540	7.80	113	309	£35
6	Fukuoka	7	2769	£0.78	4.30%	83.2	-	1644	5.87	35	539	£55
7	Berlin	8	1626	£1.55	22.30%	80.6	39.41	1386	7.07	254	1729	£26
8	Barcelona	9	2591	£1.19	23.80%	82.2	65.19	1686	6.40	585	2344	£37
9	Vancouver	10	1938	£1.08	29.40%	81.7	24.26	1670	7.23	218	788	£31
10	Melbourne	11	2363	£1.57	29.00%	82.1	25.9	1712	7.22	243	813	£36
11	Beijing	12	2671	£0.26	6.20%	75.4	85.43	-	5.12	223	261	£38
12	Bangkok	13	2624	£0.22	10.00%	74.1	76.64	-	5.99	377	1796	£50
13	Buenos Aires	14	2525	£0.57	28.30%	75.9	52.64	-	5.97	246	1435	£22
14	Toronto	15	2066	£1.09	29.40%	81.7	37.83	1670	7.23	174	1656	£32

Exploratory data analysis

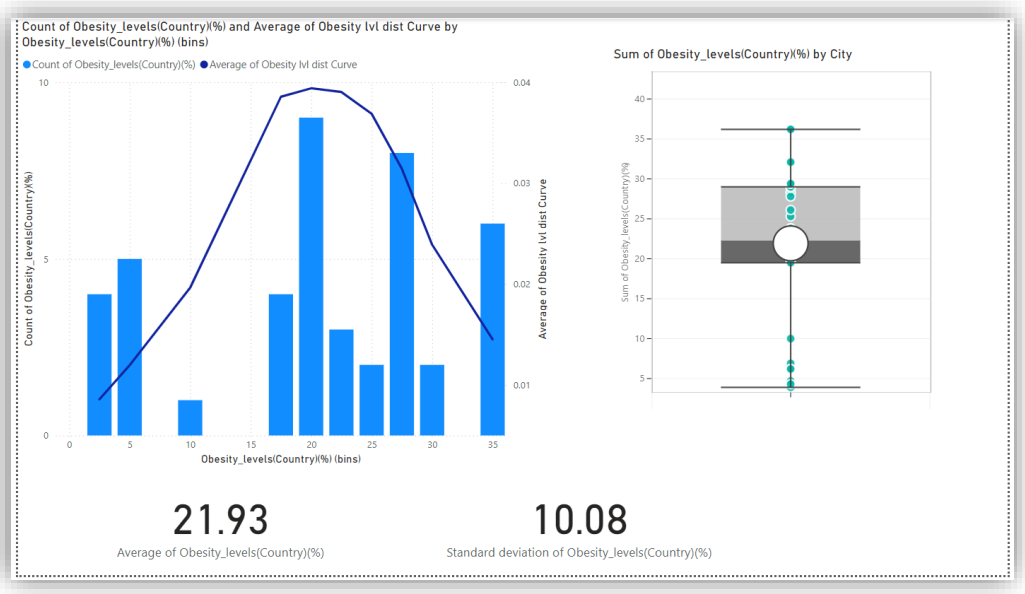
Univariate analysis



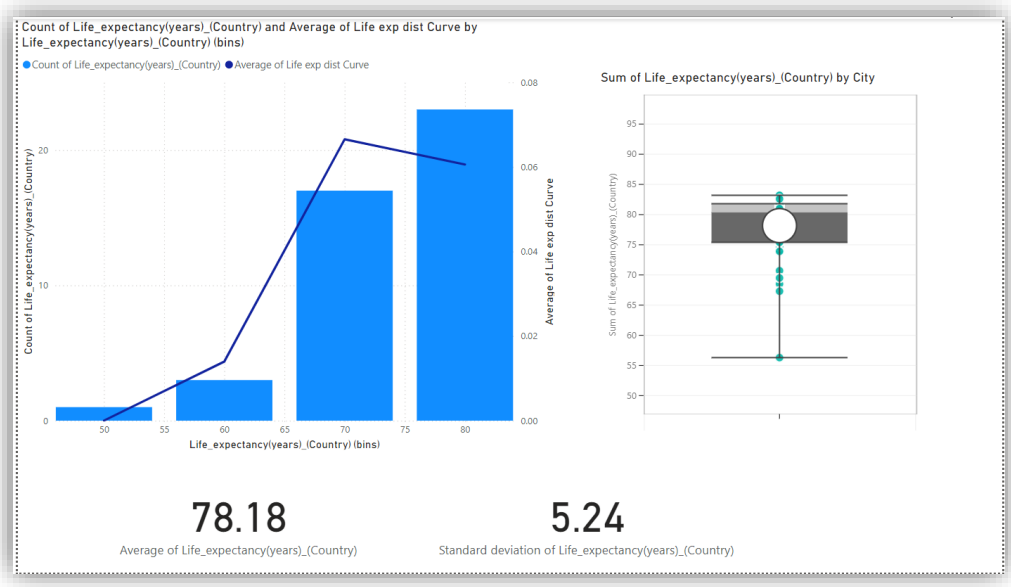
The data shows an exponential distribution with an average of 2.24 sunshine hours and there's no outliers evident in the boxplot.



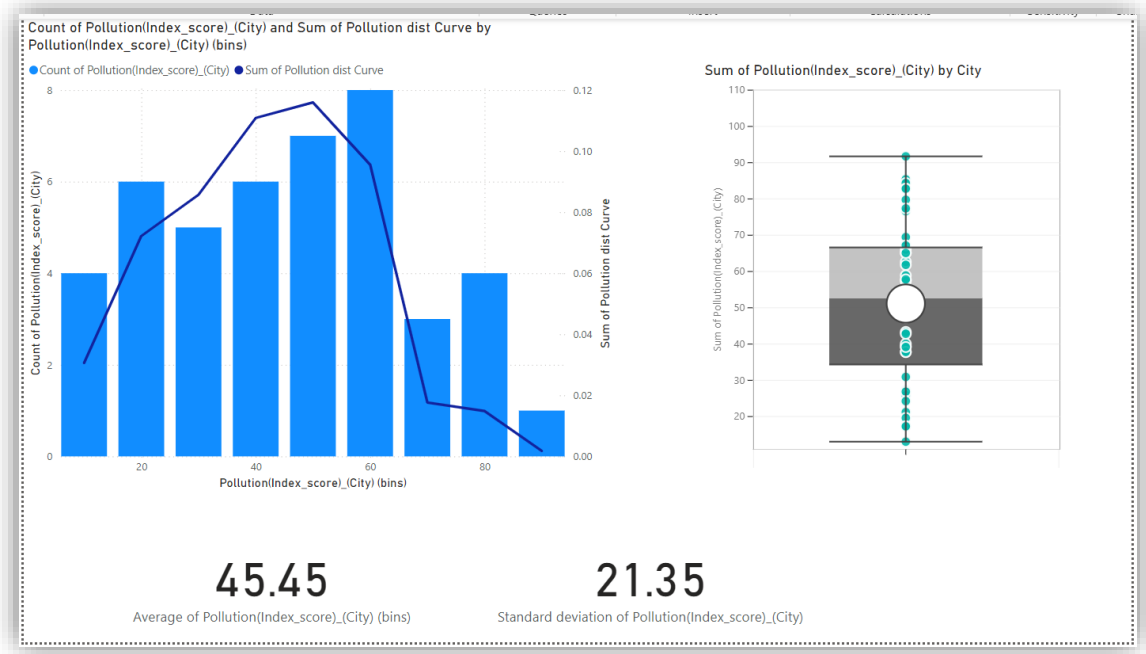
The data shows a slightly normal distribution with an average of 1.17 water bottle cost and there's no outliers present in the boxplot.



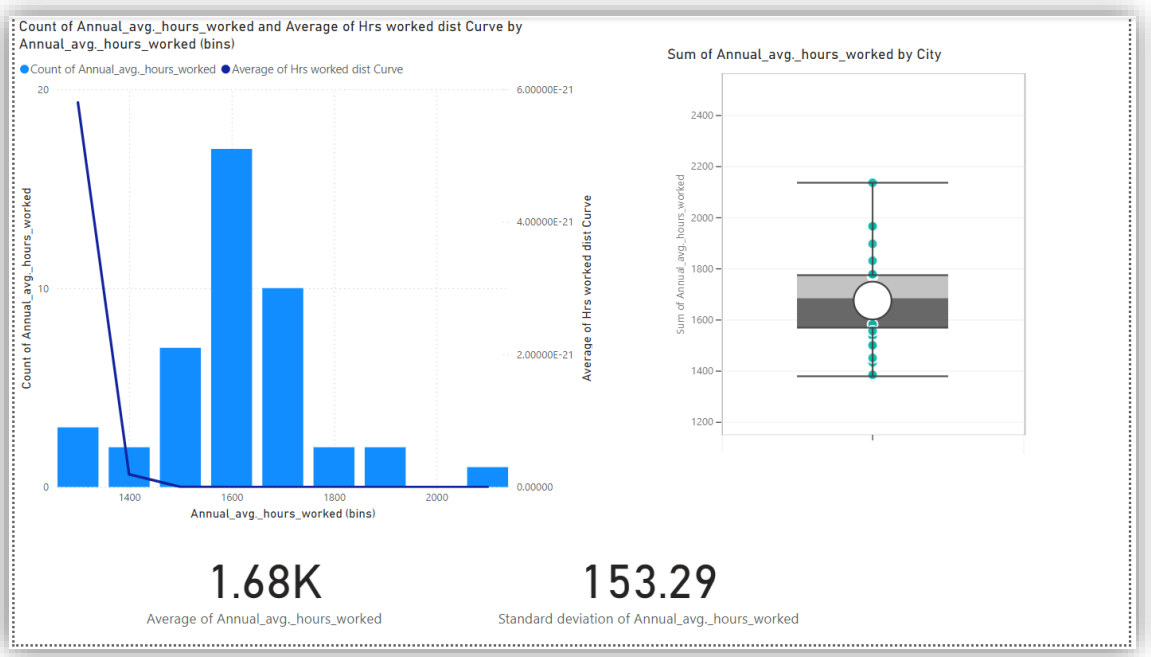
The data shows a slightly normal distribution with an average of 21.93 average obesity level and there's no outliers present in the boxplot.



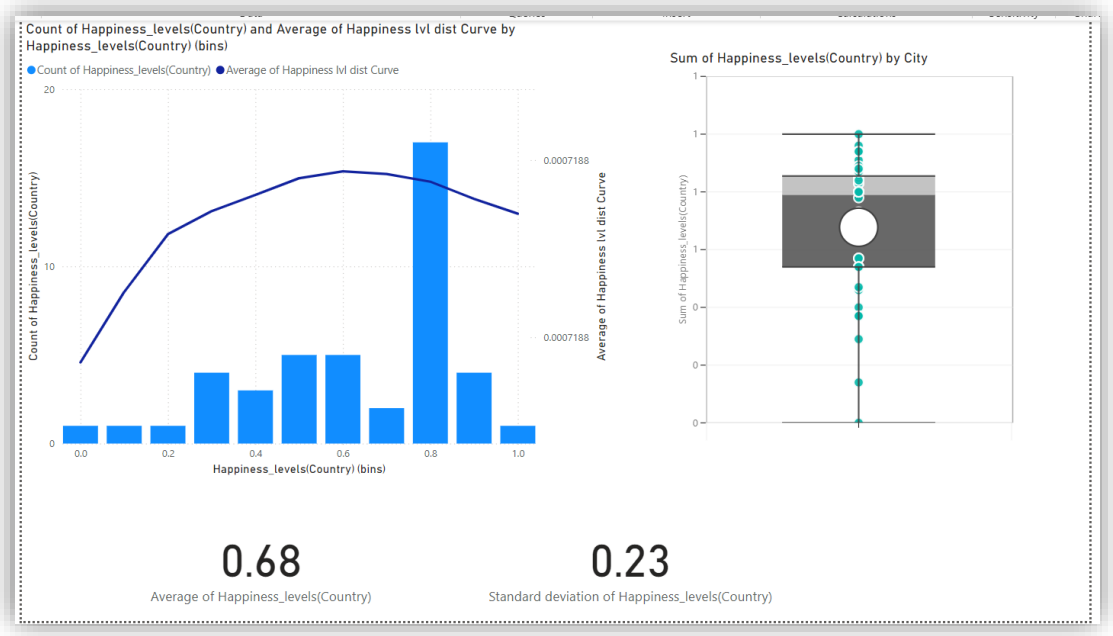
The data shows a negatively skewed distribution with an average of 78.18 average life expectancy and there's no outliers present in the boxplot.



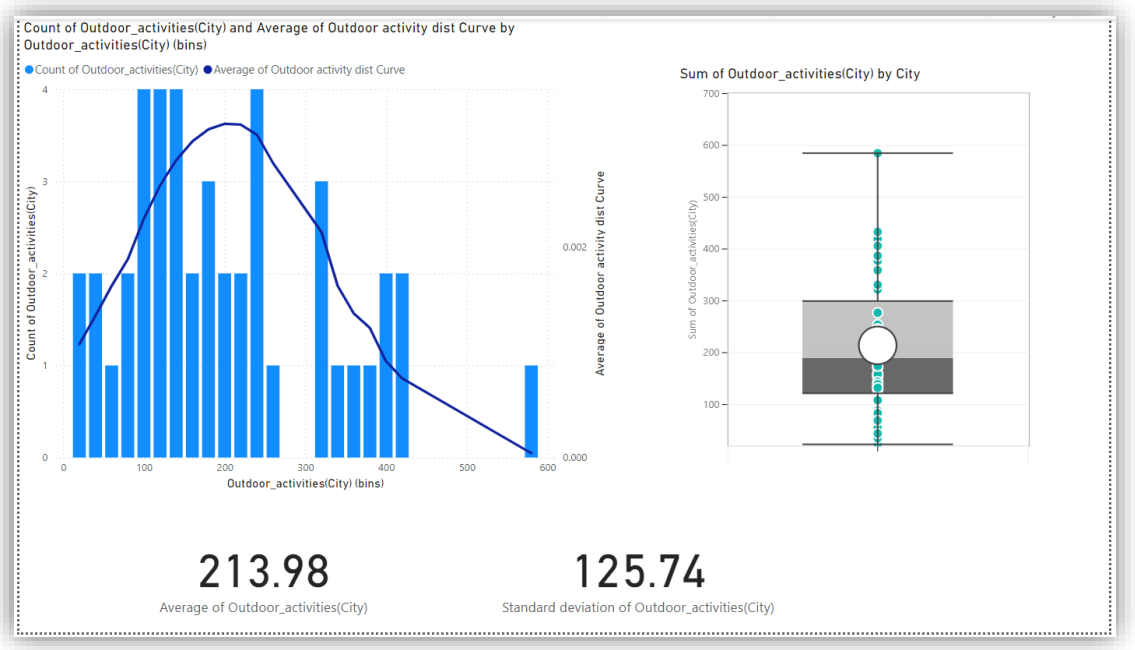
The data shows a slightly normal distribution with an average of 45.45 average pollution index score and there's no outliers present in the boxplot.



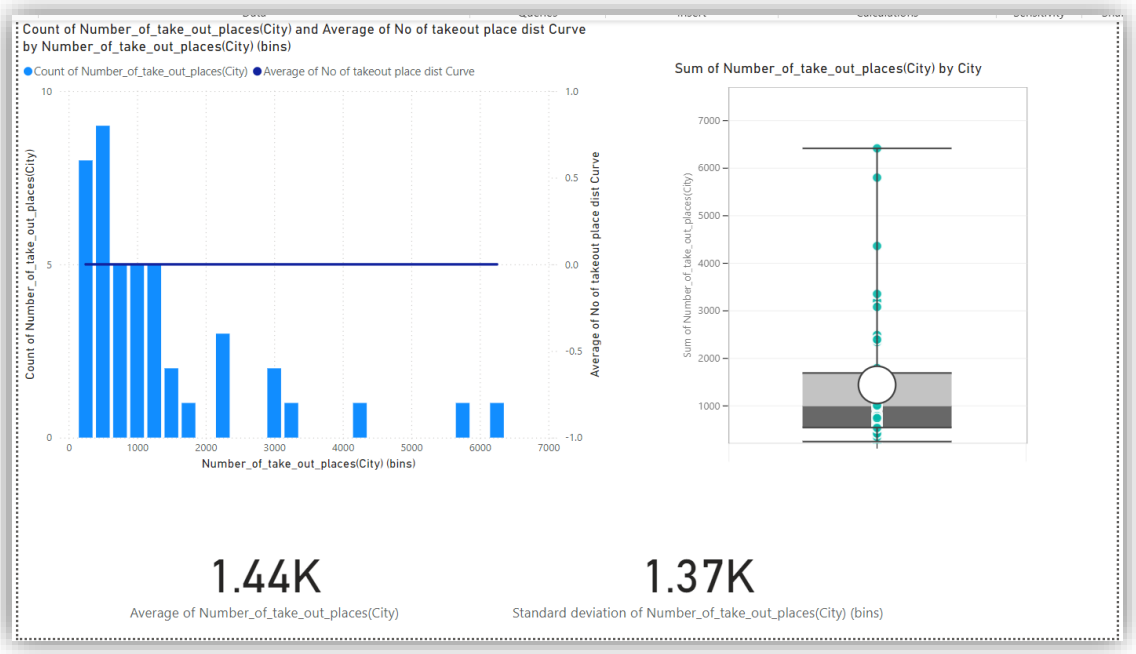
The data shows a exponential distribution with an average of 1.68 average annual average working hours and there's no outliers present in the boxplot.



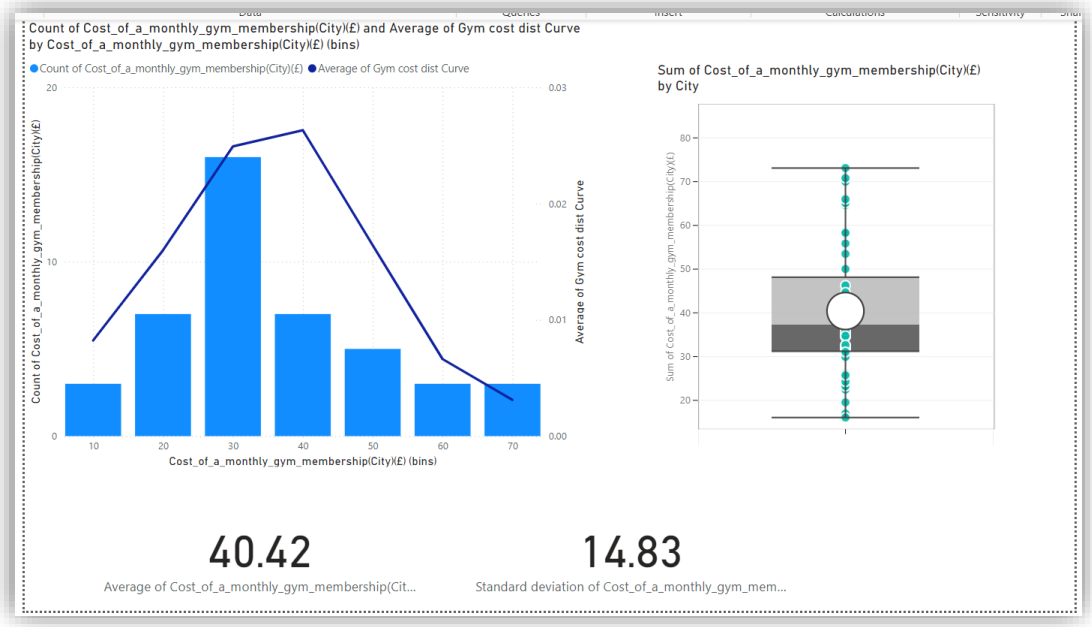
The data shows a negatively skewed distribution with an average of 0.68 happiness level and there's no outliers present in the boxplot.



The data shows a normal distribution with an average of 213.98 outdoor activities and there's no outliers present in the boxplot.

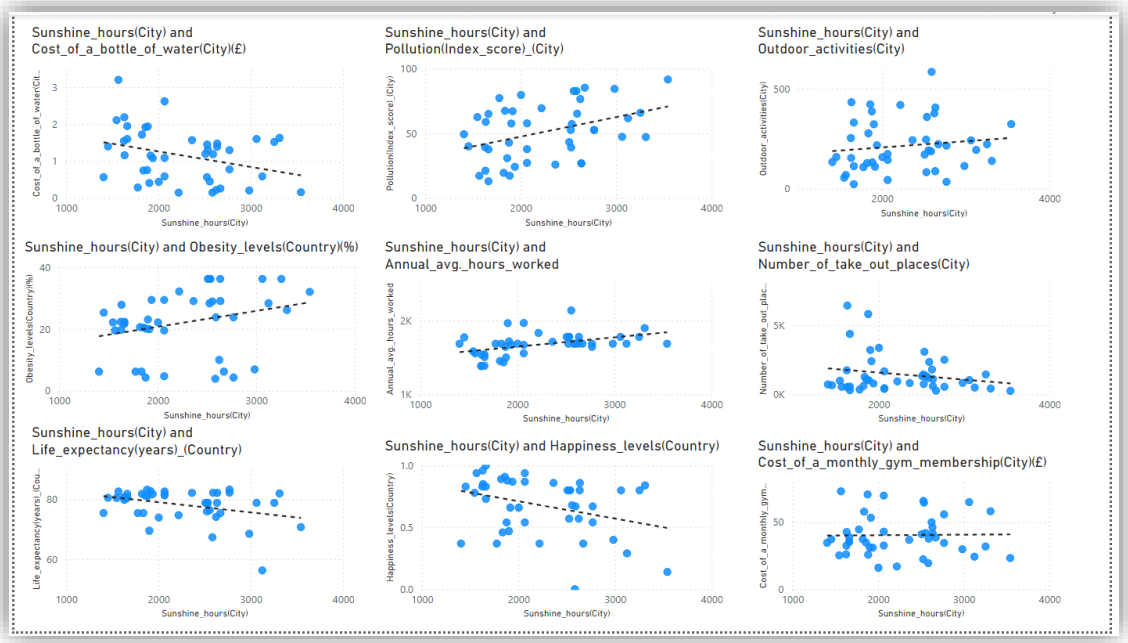


The data shows a uniform distribution with an average of 1.44 number of takeout places and there's no outliers present in the boxplot.

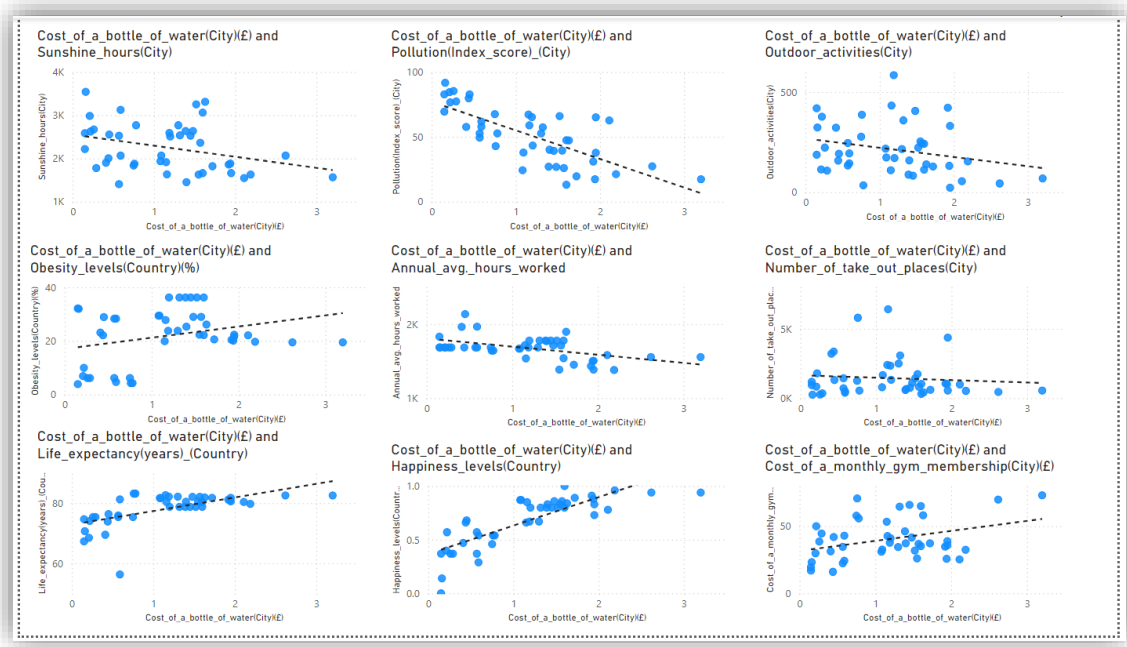


The data shows a normal distribution with an average of 40.42 cost of monthly gym membership and there's no outliers present in the boxplot.

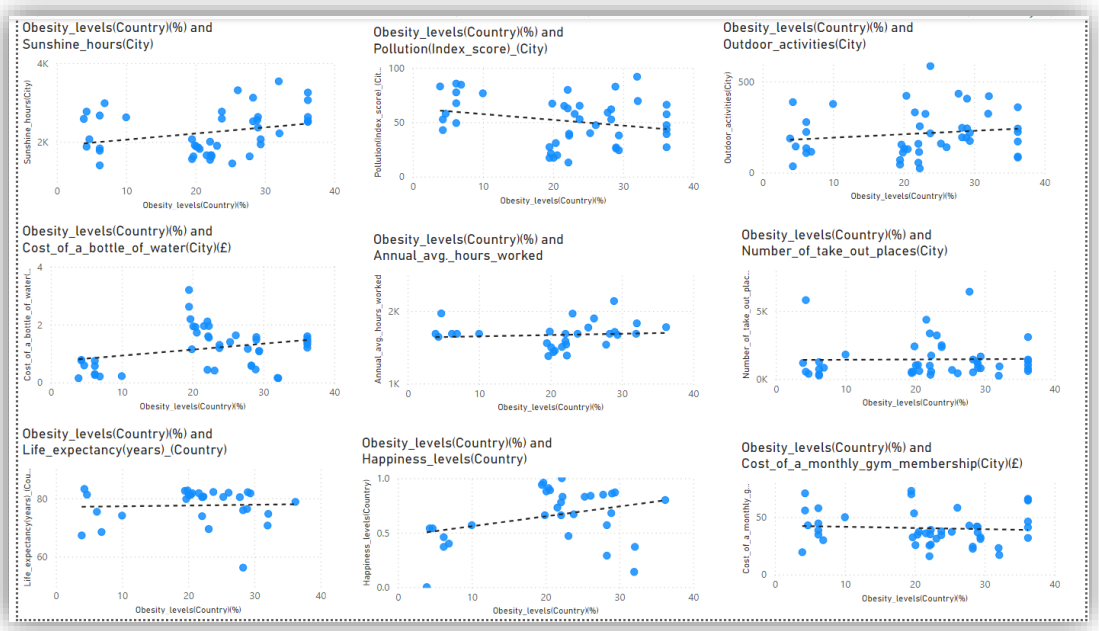
Multivariate analysis



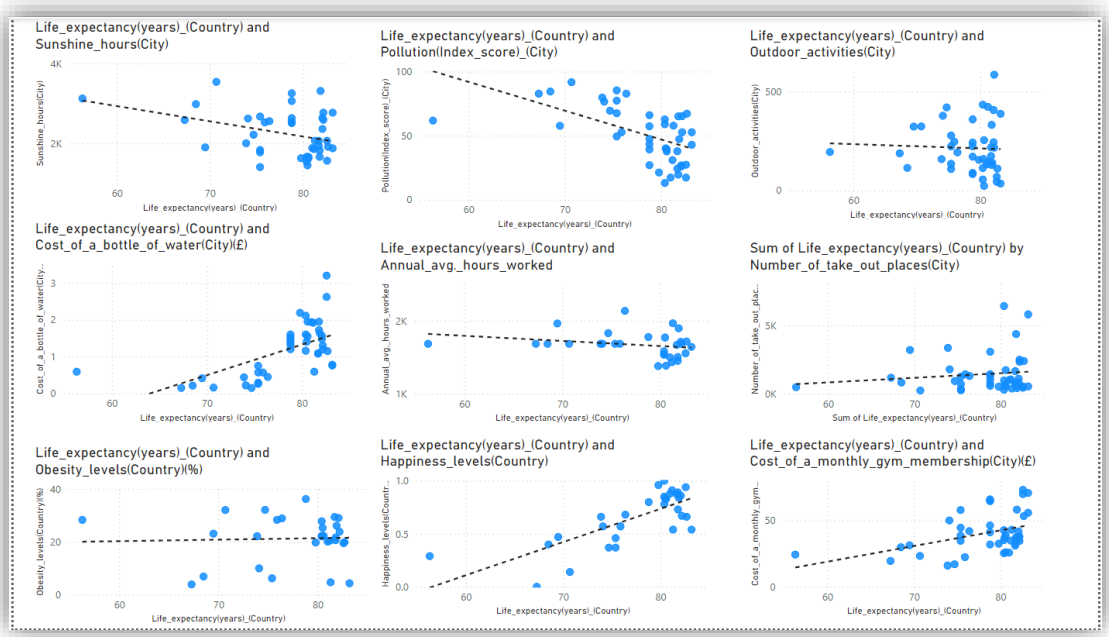
Sunshine hours positively correlate with life expectancy, pollution and outdoor activities, show weak trends with most variables and negatively relate to water bottle costs.



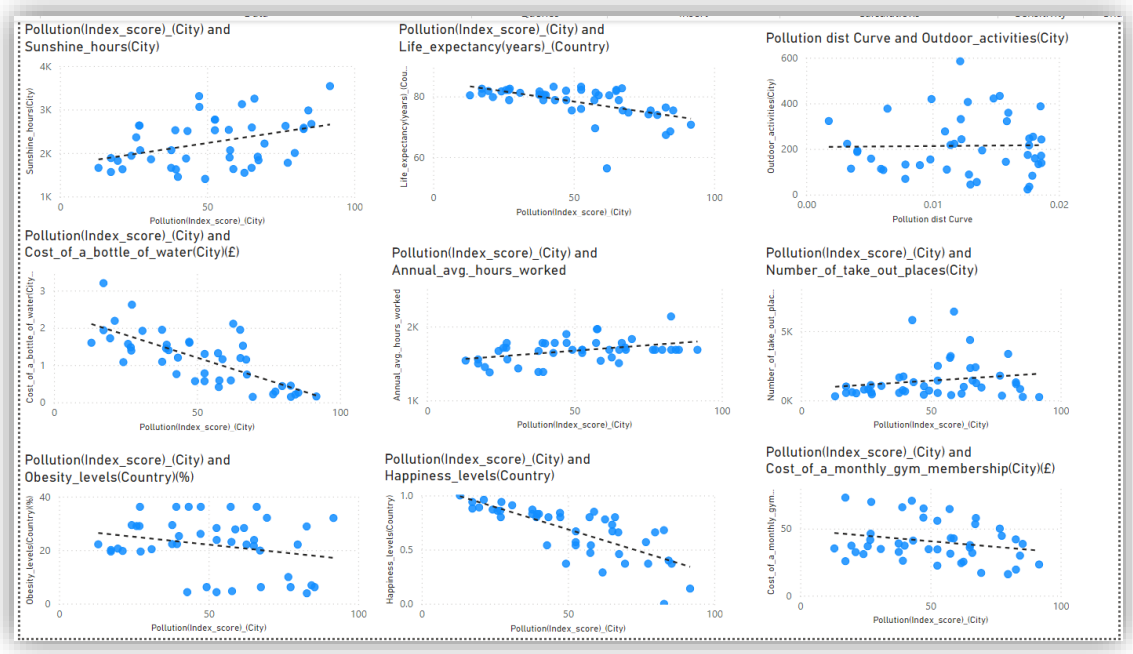
Water bottle cost positively correlate with happiness level, show weak trends with most variables and negatively relate to pollution index.



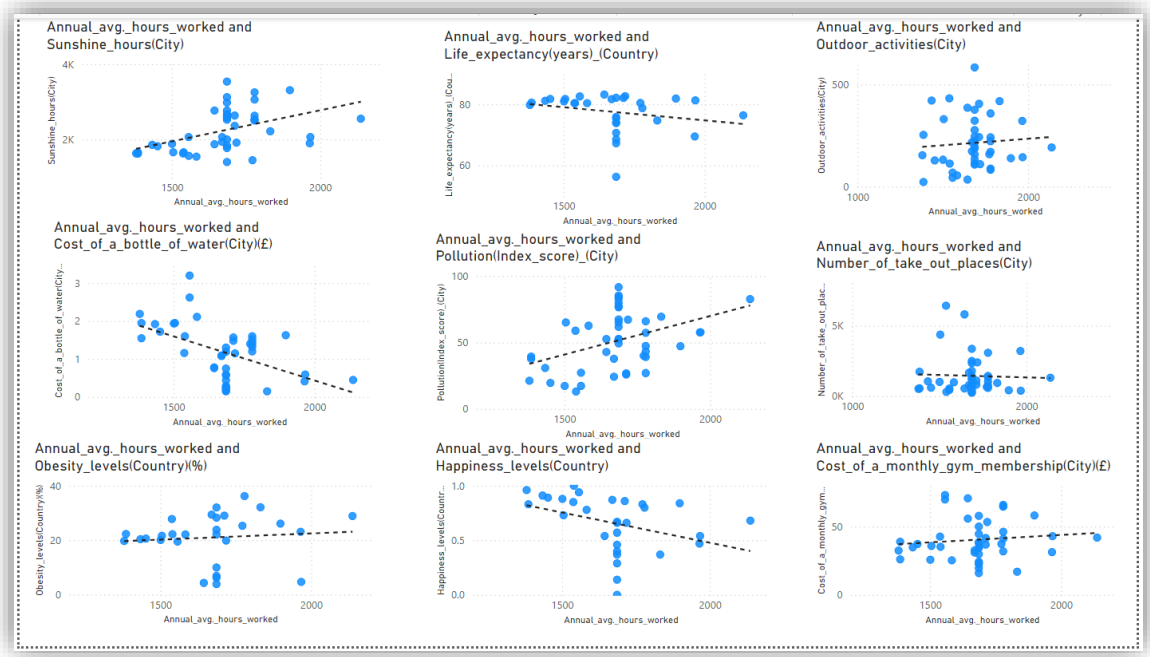
Obesity levels has no strong positive correlation with any variable and, show weak trends with most of the variables.



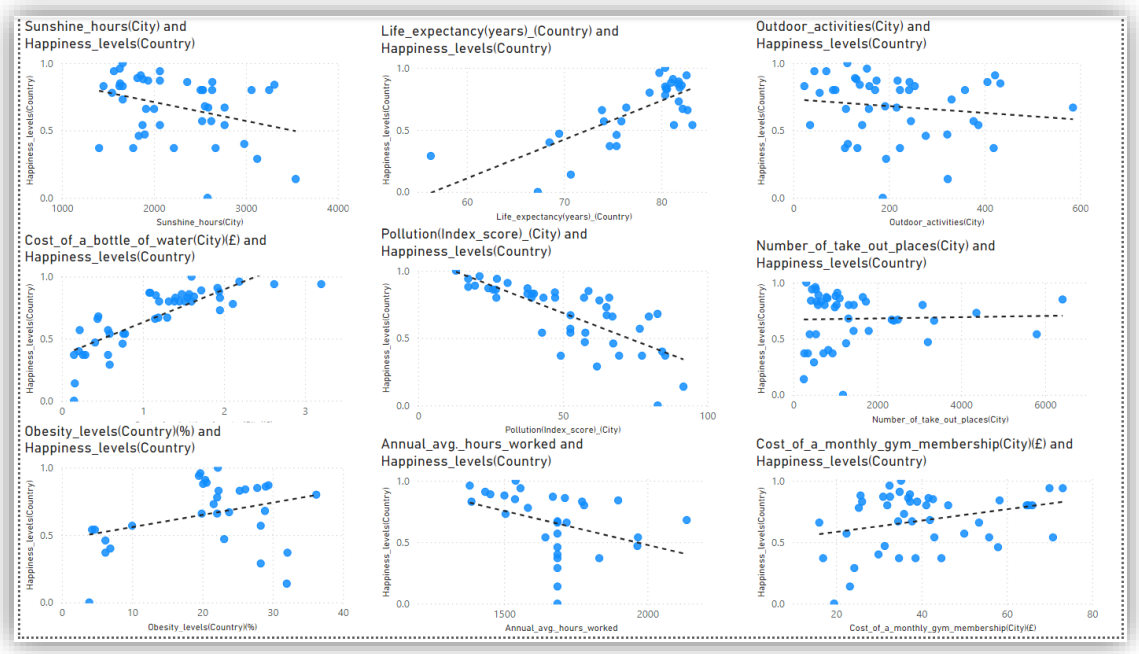
Life expectancy positively correlates with obesity level, happiness level, show weak trends with most variables and no strong negative relation.



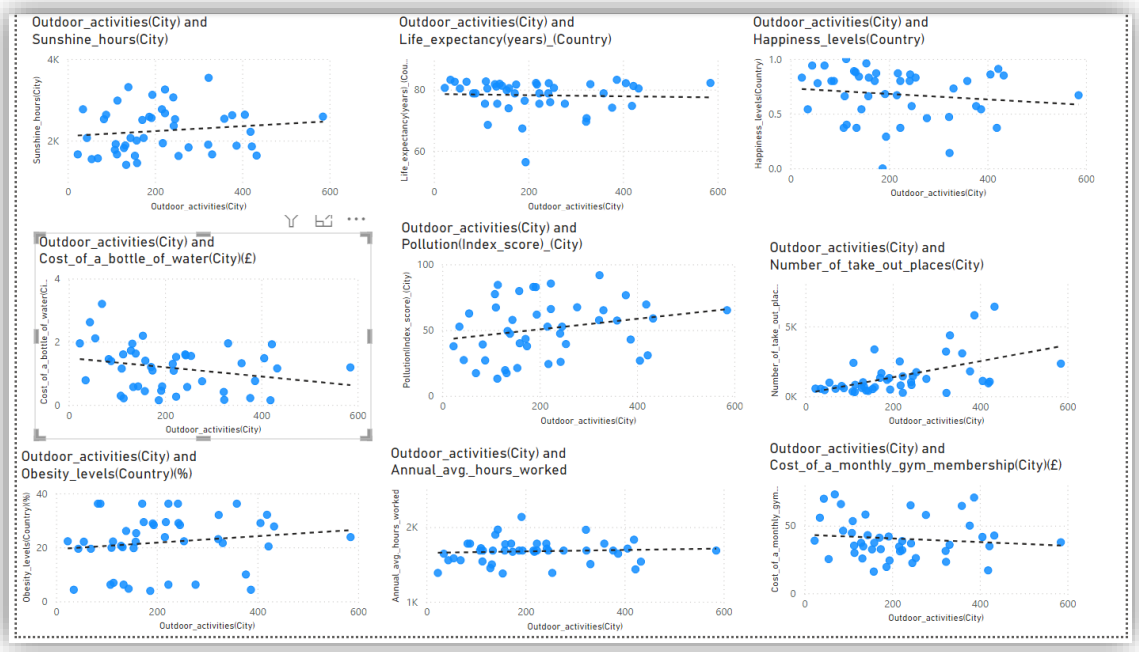
Pollution scores have no strong positive correlation and show weak trends with most variables and negatively relate to water bottle costs and happiness level.



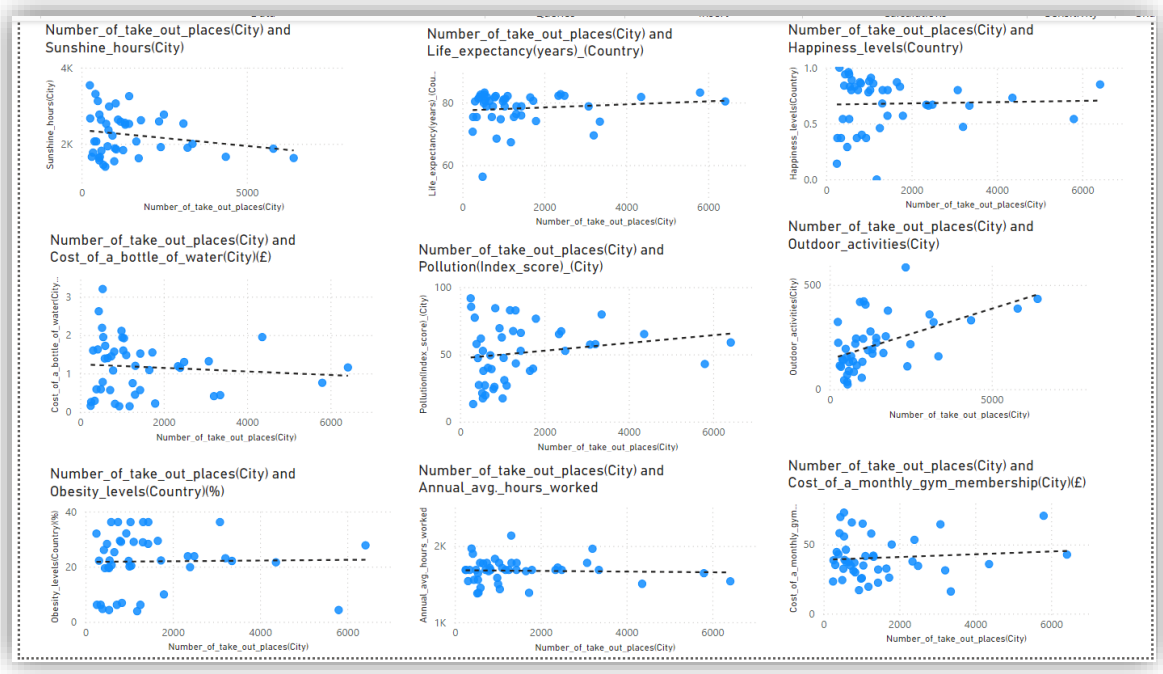
Annual average worked hours positively correlate with life pollution and show weak trends with most variables and negatively relate to water bottle costs.



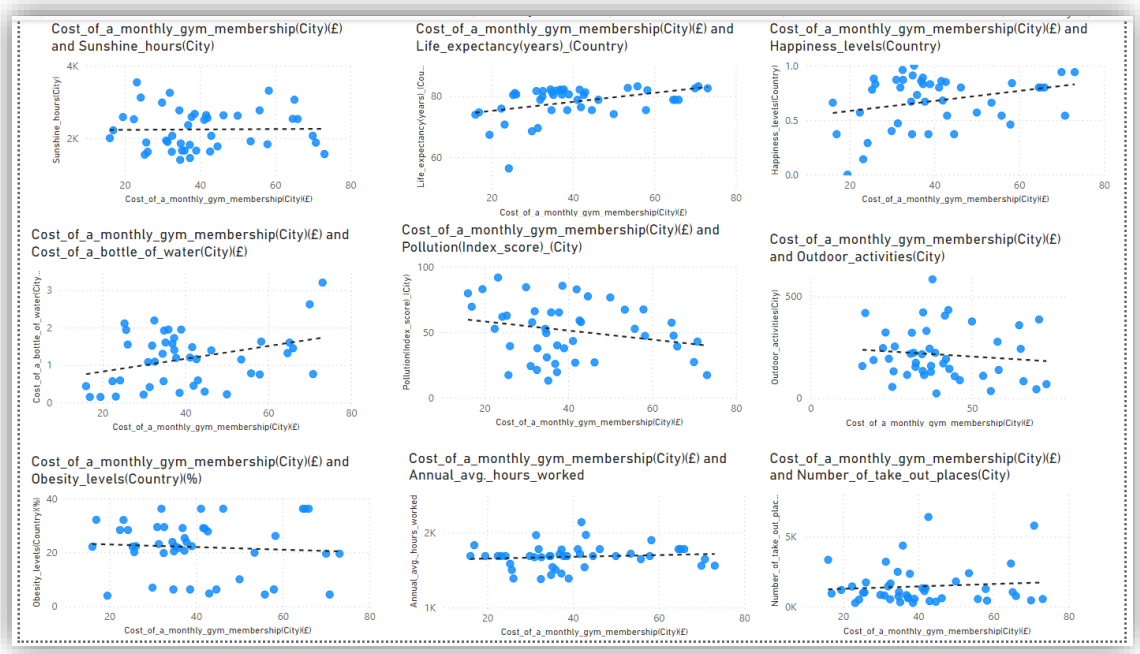
Happiness level positively correlate with obesity level and life expectancy, show weak trends with most variables and negatively relate to pollution index.



Outdoor activities have no strong positive or negative correlation and show weak trends with most variables.



Number of takeout places positively correlate with outdoor activities, show weak trends with most variables and no strong negative relation.

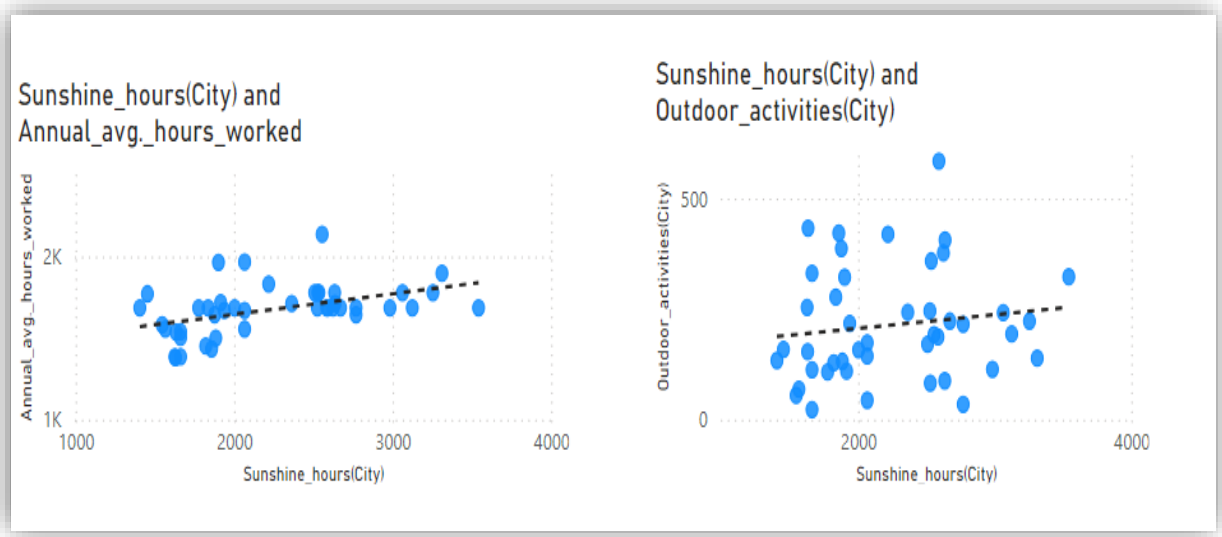


Cost of monthly gym membership have no strong positive or negative correlation and show weak trends with most of variables.

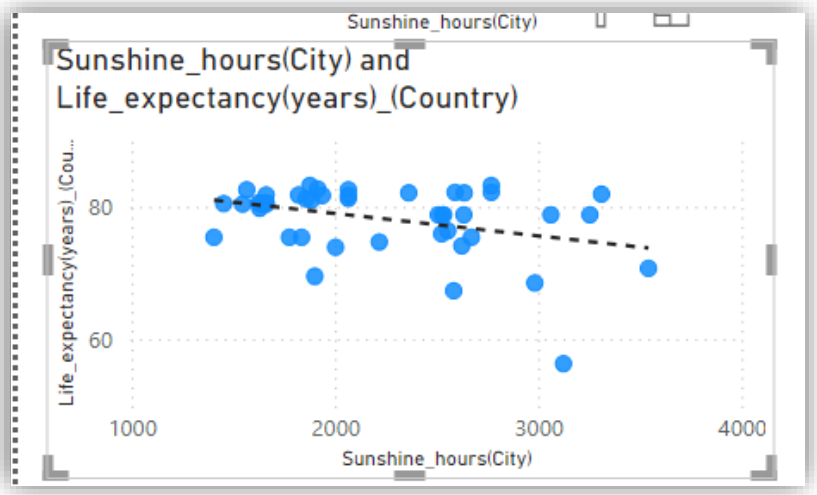
Data storytelling

Have you ever thought about what makes living in an urban area so enjoyable? Important information can be obtained by looking at indicators like sunshine hours, water bottle prices, pollution levels, and annual working hours. Examining the connections and patterns between these various factors offers a realistic picture of urban living and aids in identifying the fundamental forces influencing wellbeing.

According to the dataset, outdoor activities and the average number of working hours per year are all positively impacted by an abundance of sunshine hours. This can be seen from the below graphs,

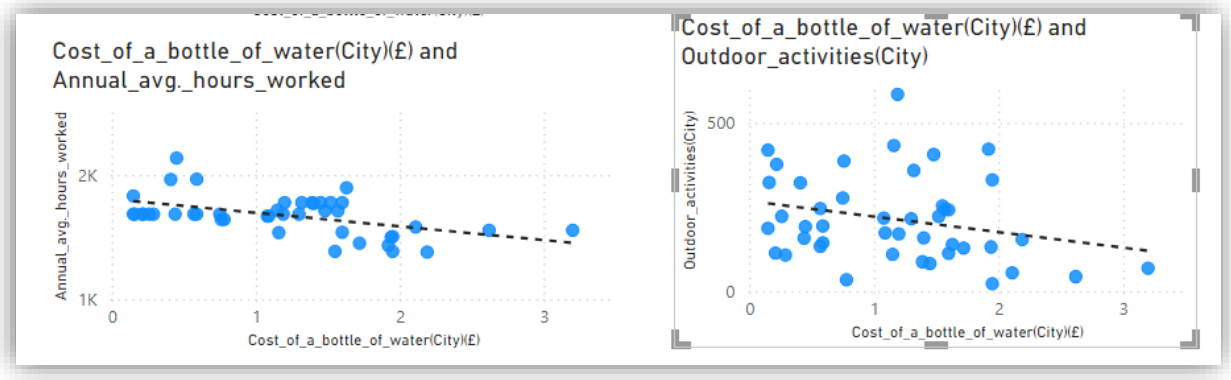


These results indicate that bright days encourage a more active and productive lifestyle. But there is a drawback as well. Life expectancy is negatively correlated with sunshine hours. This might be brought on by less sleep, excessive UV exposure, or other health issues.

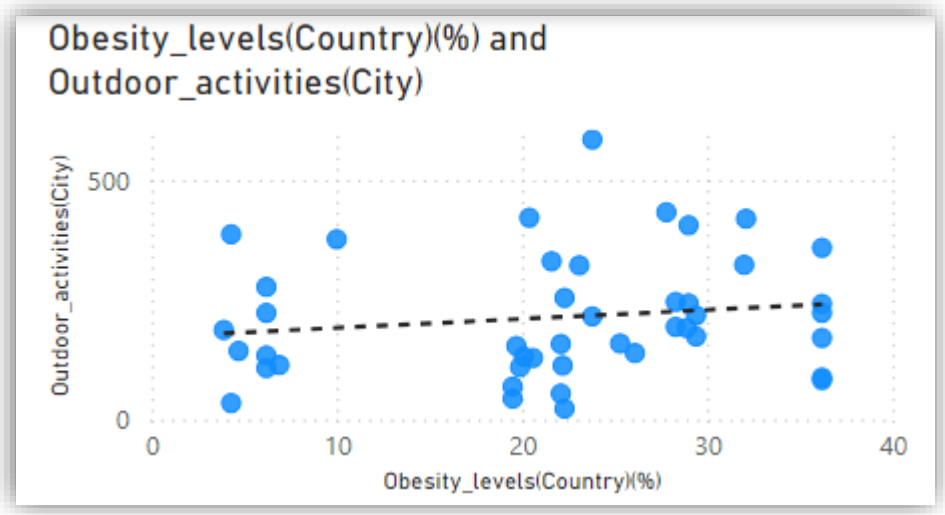


The increase in water bottle prices demonstrates a negative relationship with annual working hours and outdoor activities. This correlation indicates that the affordability of basic

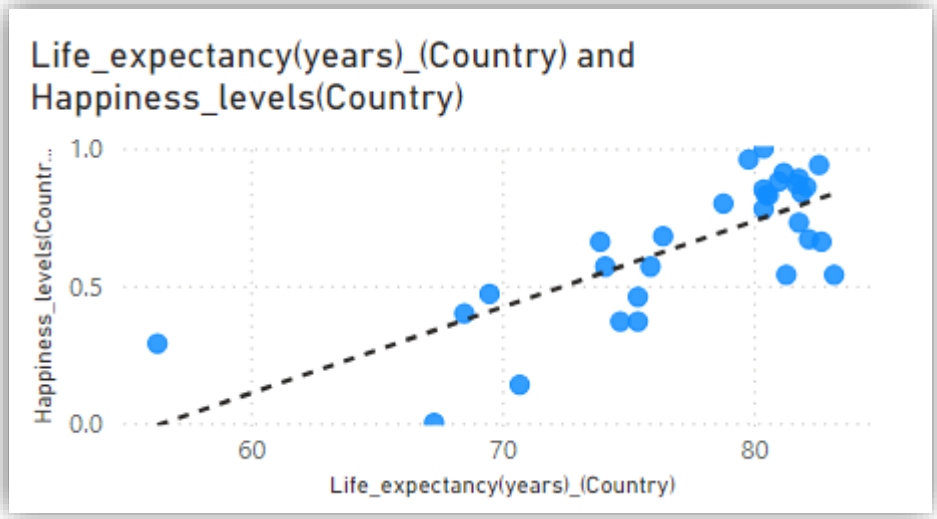
resources like water is critical for fostering active lifestyles. Access to clean and affordable water is an essential element of well-being, particularly in cities where outdoor activities are a key contributor to happiness.



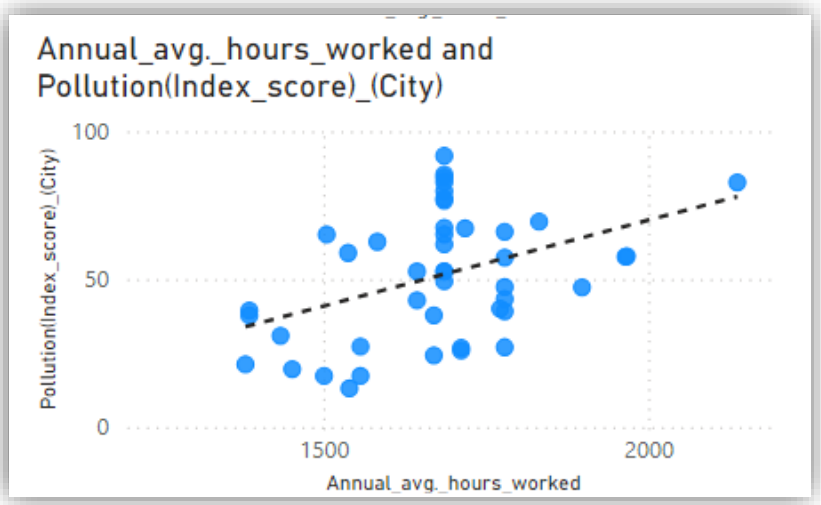
Interestingly, there is a slight positive correlation between obesity levels and outdoor activities. This trend may reflect a growing health consciousness, where individuals engage in outdoor activities to combat weight-related issues. It highlights the importance of promoting active lifestyles to address obesity while ensuring access to supportive resources like gyms and public spaces.

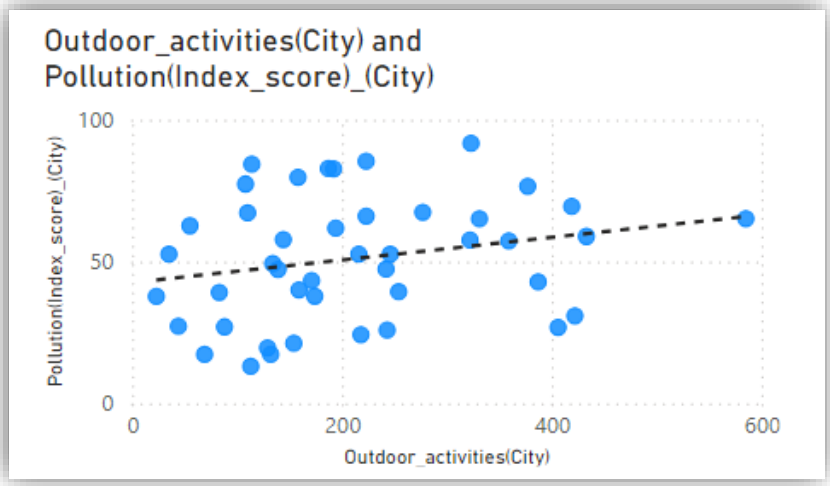


Life expectancy levels show a strong positive correlation with happiness levels, suggesting that longer lifespans directly enhance overall well-being. Conversely, increased pollution levels have a significant negative effect on life expectancy, underscoring the importance of environmental quality in urban settings. Poor air quality and exposure to pollutants diminish the health of city residents, shortening lifespans and reducing happiness.



Annual working hours also contribute to higher pollution levels, revealing the environmental toll of economic activity. Furthermore, increased outdoor activities lead to rising pollution levels, indicating the unintended consequences of urban development and industrialization. These findings emphasize the critical need to balance economic growth with environmental sustainability to protect the health and happiness of city dwellers.





Ultimately, this analysis highlights the complex interplay between urban development, environmental sustainability and individual well-being. By addressing these factors, policymakers and city planners can create urban environments that promote happiness, health and sustainability, ensuring that urban living continues to thrive for future generations.