City Happiness Index: A Composite Indicator for Urban Well-Being

Data Analysis & Visualization

Continuous Assessment 1

BSc (Honours) in Computing in Software Development

Dundalk Institute of Technology

Student Name: Sejal Umesh Sali
Student ID: D00243448
Module Title: Data Analysis and Visualization
Submission Date: 8 May 2025

Table of Contents

1. Introduction
2. Data Cleaning and Imputation
3. Multivariate Analysis
4. Normalisation
5. Weighting and Aggregation
6. Cluster Analysis
7. Visualisation of Composite Index
8. Distribution of Sub-Indices
9. Composite Index by Region (if Region Column Available)
10. Pairplot for Sub-Indices
11. Save Cleaned and Processed Dataset
12. Composite Index Over Time for Selected Cities
13. Heatmap of City vs Month (Composite Index)
14. Boxplot Comparison of Sub-Indices
15. Scatterplot of Composite Index vs Air Quality
16. Comparison of Top and Bottom Cities by Composite Index
17. Correlation of Sub-Indices with Composite Index
18. Average Sub-Indices for Top vs Bottom Cities
19. Interactive Dashboard
20. Conclusion

Introduction

This project presents the development of a City Happiness Index, a composite metric that aggregates various environmental, social, and economic indicators to rank cities by overall liability and well-being. The purpose is to provide a robust, multi-dimensional tool for comparing and assessing urban quality of life over time and across different locations.

The dataset's initial structure and variable inspiration were sourced from the publicly available City Happiness Index 2024 dataset on Kaggle. That dataset provided a foundational schema including variables like Air Quality, Cost of Living, and Healthcare Index. From there, the dataset was expanded and refined using independent research to:

- Add new variables such as monthly granularity, Green Space Area, and Decibel Levels.
- Forecast values to the year 2030.
- Integrate missing sub-indices for a more holistic assessment.

Data Cleaning and Imputation

After integrating additional research-based variables, the dataset underwent a thorough cleaning process. This included:

Missing Value Detection:
The dataset was examined using df.isnull().sum() to identify any missing values across columns. This helped in pinpointing the specific variables requiring imputation.

Imputation Strategy:
For columns with missing values, pandas-based imputation methods were applied. Where applicable, logical replacements such as mean imputation or forward-filling based on monthly sequences were used. These approaches ensured the integrity of temporal patterns within city data.

Data Consistency Checks:

- Uniform formatting was applied to categorical fields like City, Month, and Traffic_Density.
- The Date column was created by combining the Month and Year columns to enable chronological analysis, especially for trend visualizations.

Export of Cleaned Data:
The cleaned and imputed dataset was saved as Final_Happiness_Index_Data_Processed.csv to the output directory. This version served as the reliable input for further steps such as multivariate analysis, normalization, and dashboard visualization.

The result of this step was a complete, structured dataset free from null values and inconsistencies ready for statistical operations, visualization, and machine learning tasks such as clustering.

Multivariate Analysis

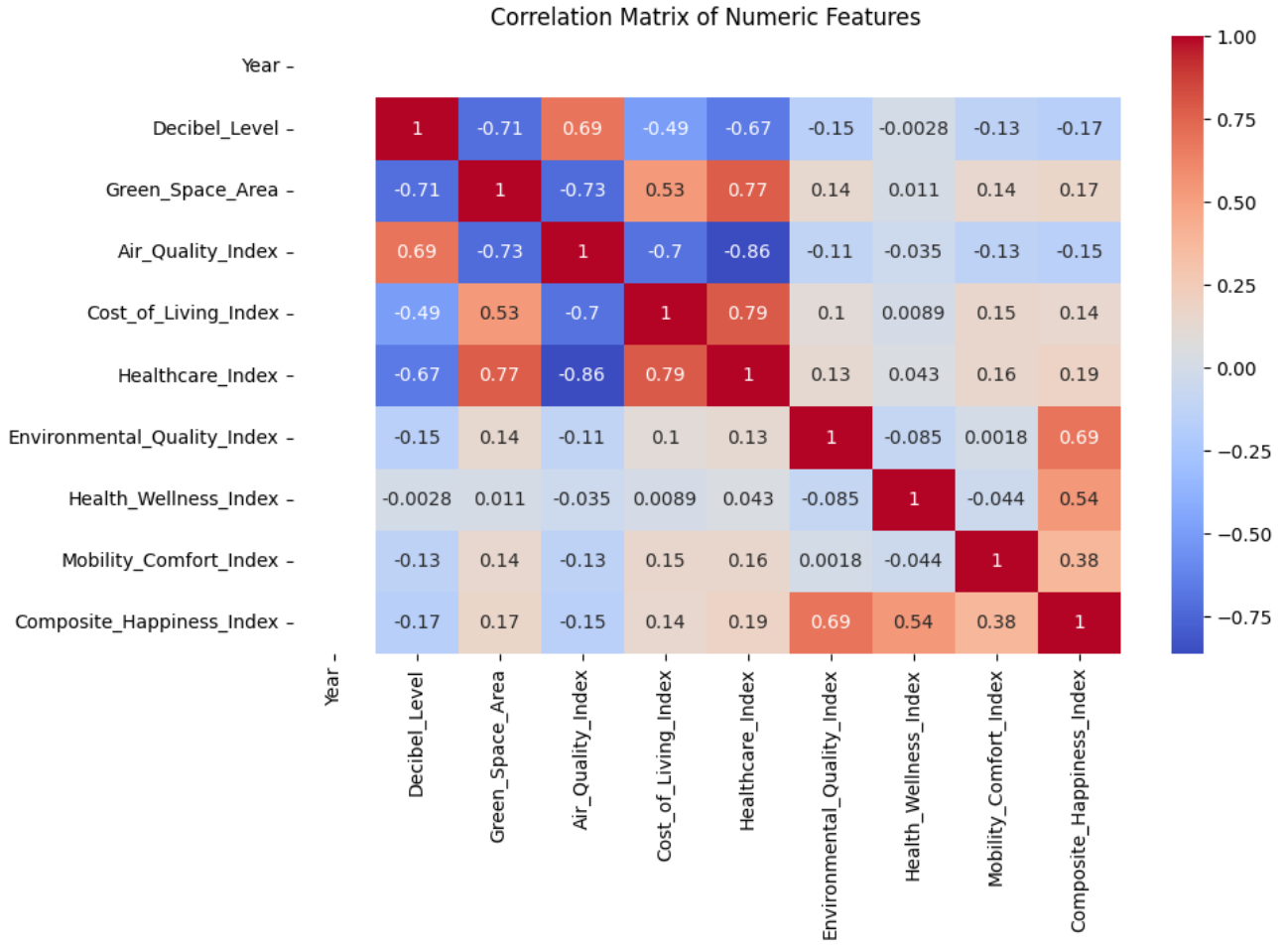
A correlation matrix and pairwise scatterplots (pairplots) were used to assess variable interdependencies. Findings:

- Strong correlations were observed between related sub-indices.
- Air Quality and Green Space strongly influenced the Environmental Quality Index.
- Cost of Living showed inverse trends with Happiness.

This step guided the feature selection process and reduced redundancy in the final index.

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Correlation matrix
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title("Correlation Matrix of Numeric Features")
plt.show()
```

Python



Normalisation

Normalization was essential to ensure all variables contributed equally to the index, regardless of their original scales. The technique used was:

- Min-Max Scaling, which transforms values to a range of [0, 1].

This allowed indicators such as Decibel Level and Healthcare Index to be directly comparable within aggregated sub-indices.

```python
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
normalized_cols = ['Air_Quality_Index', 'Decibel_Level', 'Green_Space_Area', 'Cost_of_Living_Index', 'Healthcare_Index']
df[[col + '_Norm' for col in normalized_cols]] = scaler.fit_transform(df[normalized_cols])
df.head()
```

Python

Weighting and Aggregation

Sub-indices were created from thematically grouped indicators:

- Environmental Quality Index
- Health & Wellness Index
- Mobility Comfort Index

These were aggregated using a weighted sum to produce the Composite Happiness Index. While weights were not explicitly detailed, the structure implies equal or proportional weighting based on the relative importance of each domain.

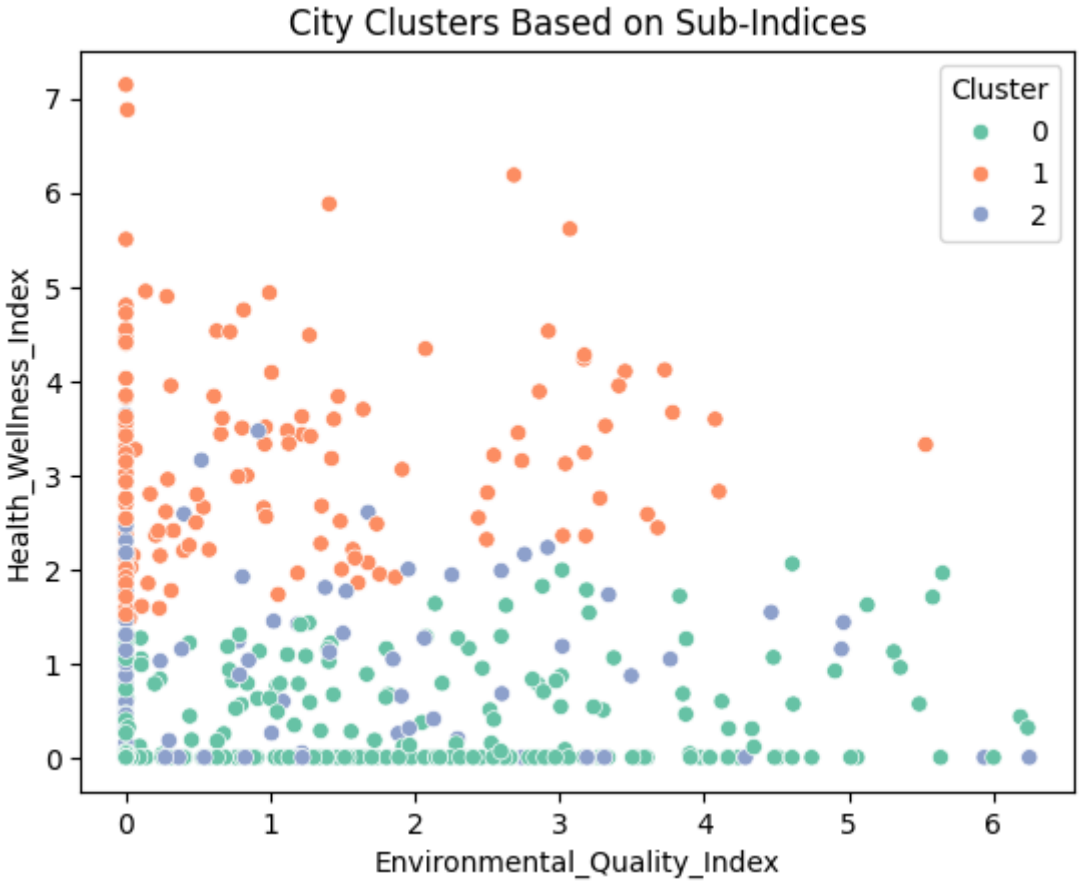
```python
# Used provided sub-indices and create final composite index
df['Composite_Happiness_Index'] = (
    0.4 * df['Environmental_Quality_Index'] +
    0.35 * df['Health_Wellness_Index'] +
    0.25 * df['Mobility_Comfort_Index']
)
df[['City', 'Composite_Happiness_Index']].sort_values(by='Composite_Happiness_Index', ascending=False).head(10)
```

	City	Composite_Happiness_Index
261	Vienna	3.783236
573	Wellington	3.607431
171	Oslo	3.544929
256	Vienna	3.372604
136	Lisbon	3.251533
36	Denver	3.243002
469	Ottawa	3.174471
295	Yerevan	3.165678
547	Ulaanbaatar	3.117675
367	Florence	3.116303

Cluster Analysis

K-Means clustering was applied to segment cities into distinct happiness tiers. This unsupervised learning approach identified cities with similar happiness profiles, helping visualize disparity and group trends.

```
from sklearn.cluster import KMeans

features = ['Environmental_Quality_Index', 'Health_Wellness_Index', 'Mobility_Comfort_Index']
kmeans = KMeans(n_clusters=3, random_state=42)
df['Cluster'] = kmeans.fit_predict(df[features])

# Visualize clusters
sns.scatterplot(data=df, x='Environmental_Quality_Index', y='Health_Wellness_Index', hue='Cluster', palette='Set2')
plt.title("City Clusters Based on Sub-Indices")
plt.show()
```



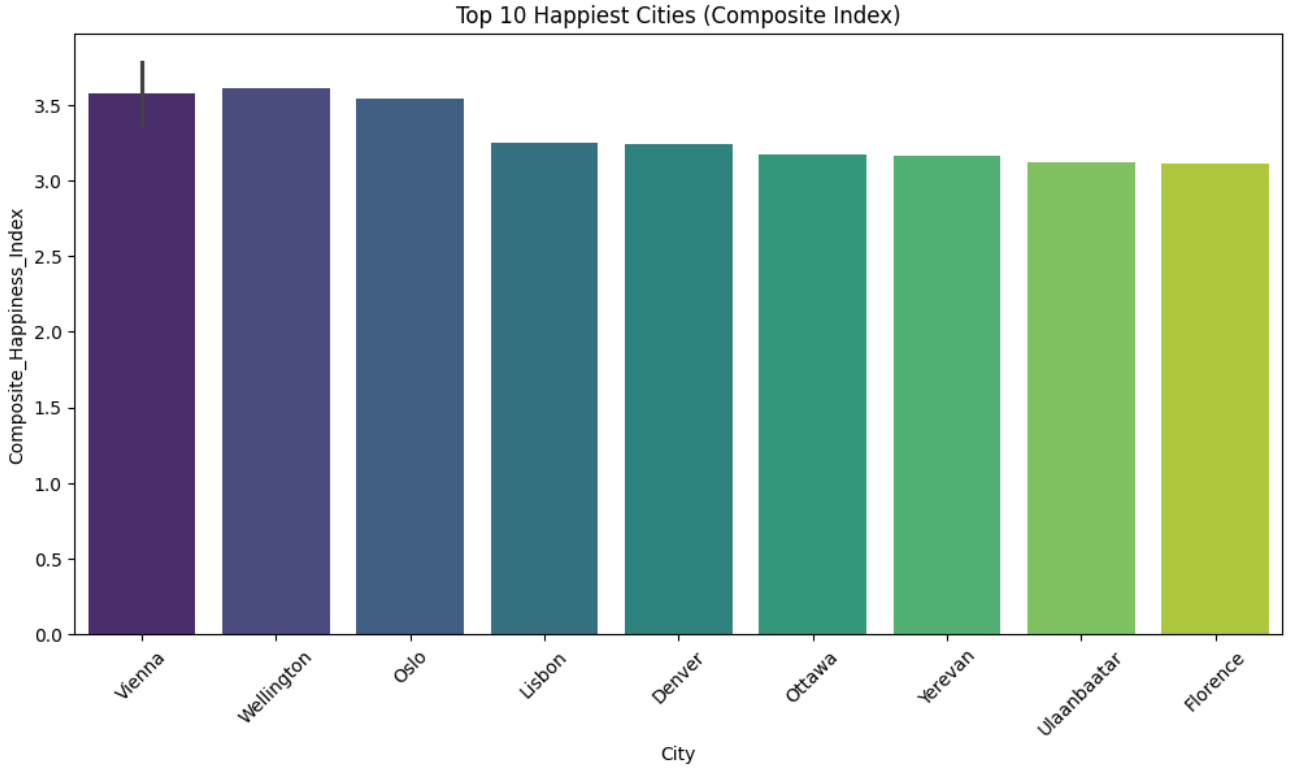
Visualisation of Composite Index

Bar plots and line graphs were generated to show:

- Happiness scores across cities.
- Changes over months.
- Ranking distributions.

These visualizations provided intuitive comparisons of urban performance.

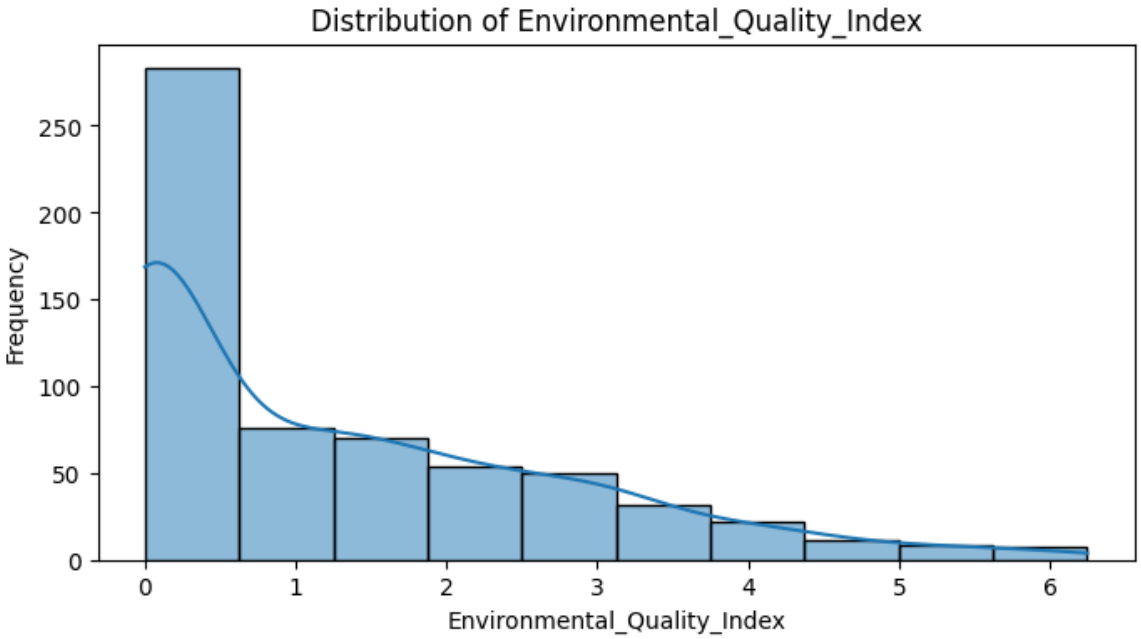
```python
# Bar plot of top 10 happiest cities
top_cities = df.sort_values(by='Composite_Happiness_Index', ascending=False).head(10)
plt.figure(figsize=(12, 6))
sns.barplot(data=top_cities, x='City', y='Composite_Happiness_Index', palette='viridis')
plt.title("Top 10 Happiest Cities (Composite Index)")
plt.xticks(rotation=45)
plt.show()
```

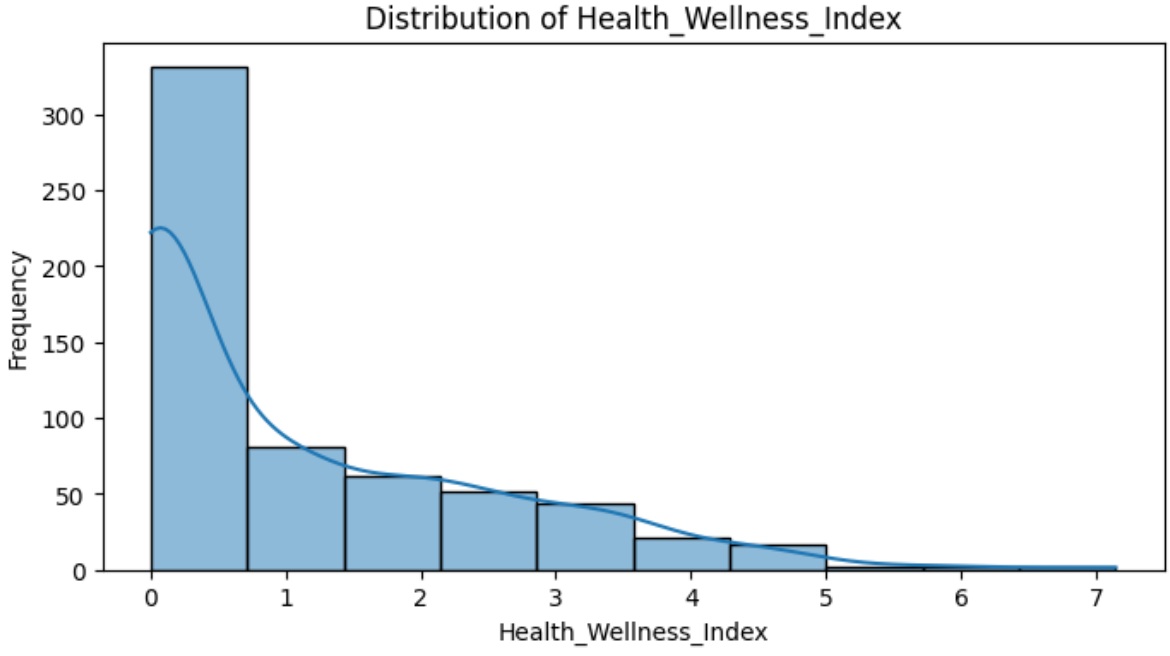


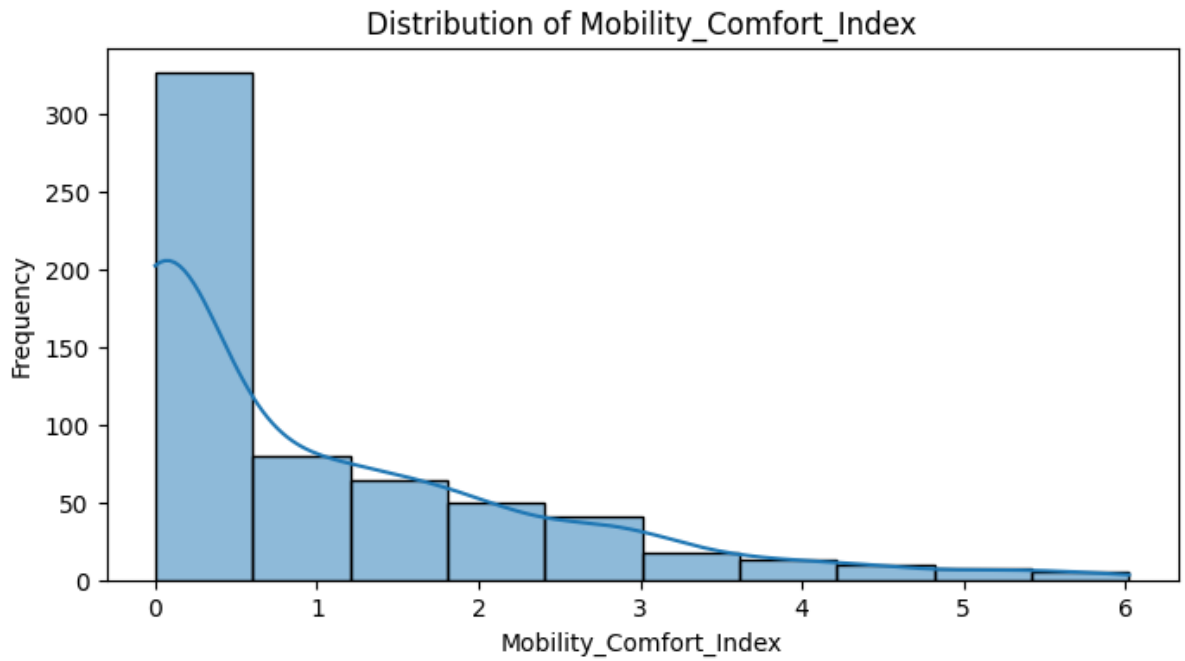
Distribution of Sub-Indices

Histograms and density plots illustrated how sub-indices like Environmental Quality and Mobility Comfort were distributed. This helped identify outliers and skewness in certain cities.

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Plot distribution of each sub-index
sub_indices = ['Environmental_Quality_Index', 'Health_Wellness_Index', 'Mobility_Comfort_Index']
for col in sub_indices:
    plt.figure(figsize=(8, 4))
    sns.histplot(df[col], kde=True, bins=10)
    plt.title(f'Distribution of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.show()
```







Composite Index by Region (if Region Column Available)

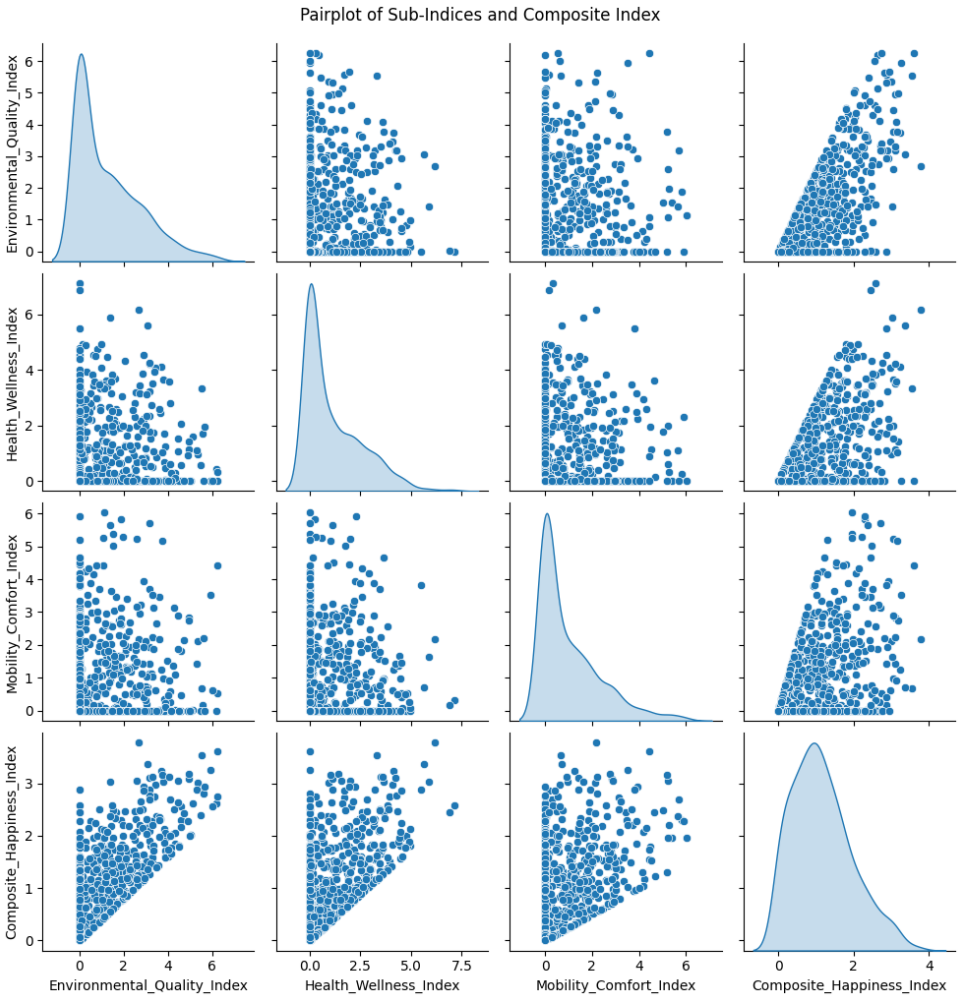
This step was conditionally designed for regional comparisons. If regional groupings (e.g., continents or zones) are available, average scores can be computed to benchmark broader geographical performance.

```
if 'Region' in df.columns:
    region_avg = df.groupby('Region')['Composite_Happiness_Index'].mean().sort_values(ascending=False)
    region_avg.plot(kind='bar', figsize=(10, 6), title='Average Composite Index by Region')
    plt.ylabel('Composite Happiness Index')
    plt.show()
else:
    print("Region column not available in dataset.")
```

Pairplot for Sub-Indices

A Seaborn pairplot was generated to visualize the relationships between all sub-indices. This matrix view further clarified which indicators were interrelated and validated the dimensionality of the final index.

```
# Pairplot for visual relationship
sns.pairplot(df[['Environmental_Quality_Index', 'Health_Wellness_Index', 'Mobility_Comfort_Index', 'Composite_Happiness_Index']], diag_kind='kde')
plt.suptitle('Pairplot of Sub-Indices and Composite Index', y=1.02)
plt.show()
```



Save Cleaned and Processed Dataset

The cleaned and enriched dataset, with normalized and imputed values, along with calculated sub-indices and the final composite score, was saved to CSV for reproducibility and downstream analysis.

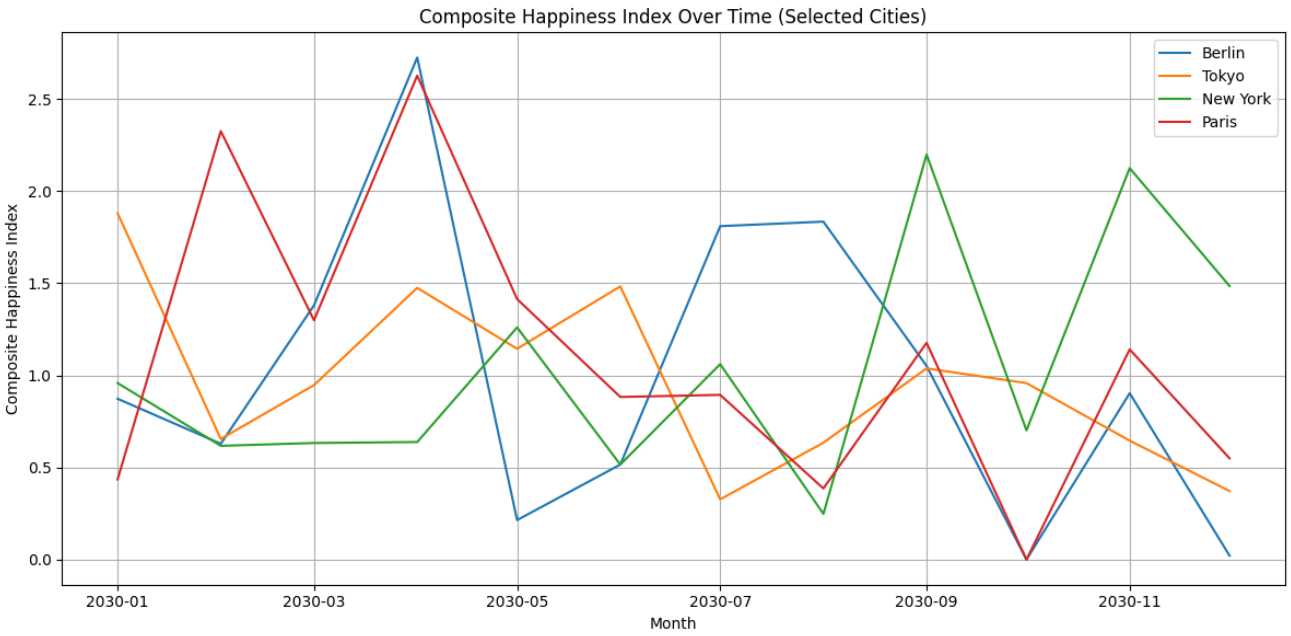
```
import os
output_dir = 'output'
os.makedirs(output_dir, exist_ok=True)
df.to_csv(os.path.join(output_dir, 'Final_Happiness_Index_Data_Processed.csv'), index=False)
print('Dataset saved to output/Final_Happiness_Index_Data_Processed.csv')
```

```
Dataset saved to output/Final_Happiness_Index_Data_Processed.csv
```

Composite Index Over Time for Selected Cities

Time-series line plots were used to visualize how happiness evolved month-to-month for selected cities. This allowed the identification of seasonal effects or irregular trends.

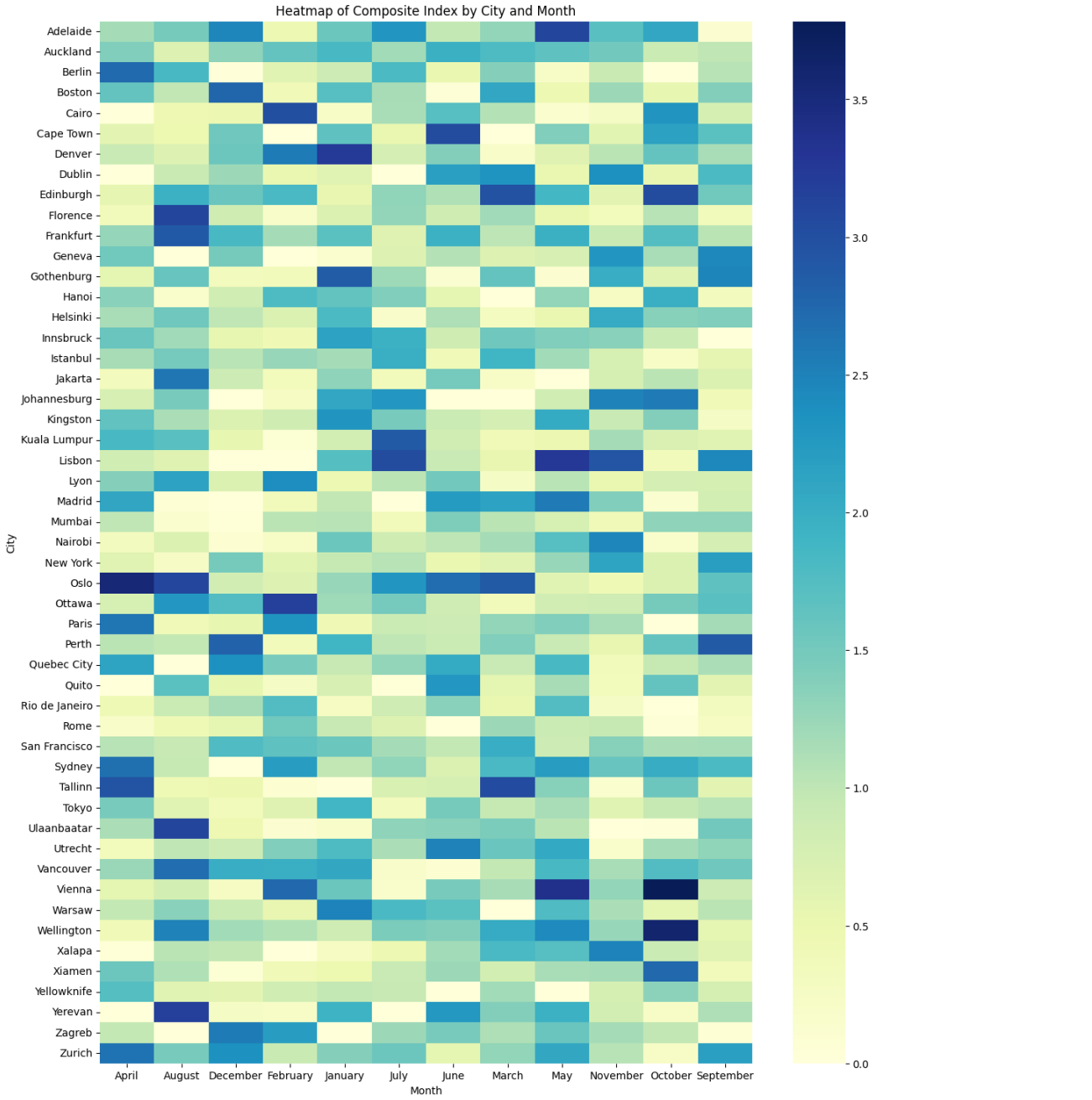
```
# Line plot of Composite Index over time for a few cities
selected_cities = ['Berlin', 'Tokyo', 'New York', 'Paris']
df['Date'] = pd.to_datetime(df['Month'] + ' ' + df['Year'].astype(str))
plt.figure(figsize=(12, 6))
for city in selected_cities:
    city_data = df[df['City'] == city]
    plt.plot(city_data['Date'], city_data['Composite_Happiness_Index'], label=city)

plt.title("Composite Happiness Index Over Time (Selected Cities)")
plt.xlabel("Month")
plt.ylabel("Composite Happiness Index")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```



Heatmap of City vs Month (Composite Index)

A heatmap matrix displayed the Composite Happiness Index for each city across all months. This visual format easily highlighted fluctuations and high/low scoring patterns over time.

```
# Pivot table and heatmap
pivot = df.pivot_table(index='City', columns='Month', values='Composite_Happiness_Index')
plt.figure(figsize=(14, 18))
sns.heatmap(pivot, cmap='YlGnBu', annot=False)
plt.title("Heatmap of Composite Index by City and Month")
plt.xlabel("Month")
plt.ylabel("City")
plt.show()
```

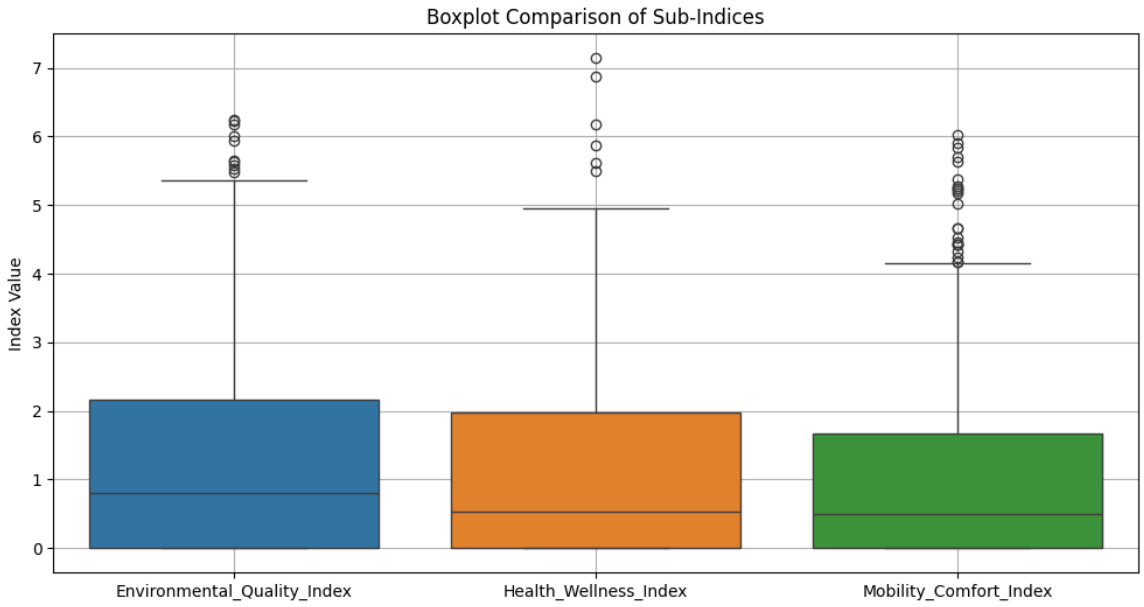


Boxplot Comparison of Sub-Indices

Boxplots were used to compare the spread and central tendency of different sub-indices. This was useful in:

- Identifying which sub-indices were more variable.
- Detecting outliers in city performances.

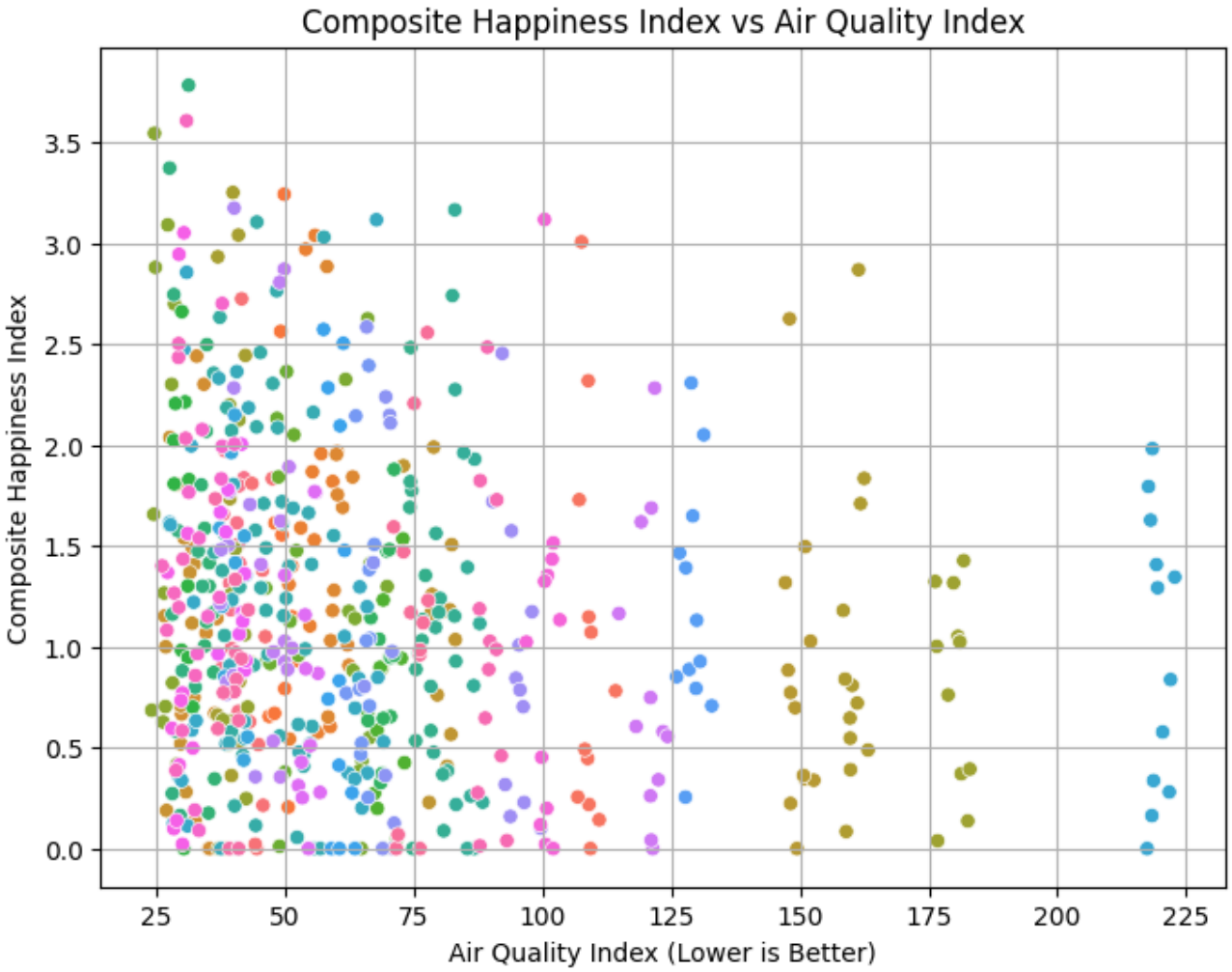
```
plt.figure(figsize=(12, 6))
sns.boxplot(data=df[['Environmental_Quality_Index', 'Health_Wellness_Index', 'Mobility_Comfort_Index']])
plt.title("Boxplot Comparison of Sub-Indices")
plt.ylabel("Index Value")
plt.grid(True)
plt.show()
```



Scatterplot of Composite Index vs Air Quality

A scatterplot showed the relationship between air quality and the composite index. As expected, better air quality positively correlated with overall happiness, highlighting its critical role in urban satisfaction.

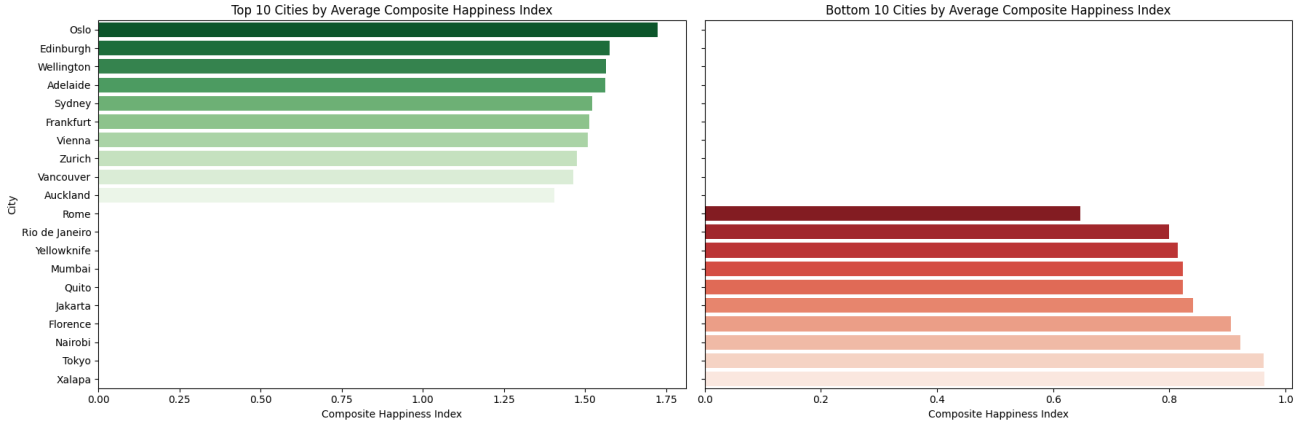
```
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df, x='Air_Quality_Index', y='Composite_Happiness_Index', hue='City', legend=False)
plt.title("Composite Happiness Index vs Air Quality Index")
plt.xlabel("Air Quality Index (Lower is Better)")
plt.ylabel("Composite Happiness Index")
plt.grid(True)
plt.show()
```



Comparison of Top and Bottom Cities by Composite Index

Cities were sorted and compared based on their happiness scores. Tables and bar graphs showcased the top-performing and lowest-ranking cities, enabling easy benchmarking and insights into leading practices.

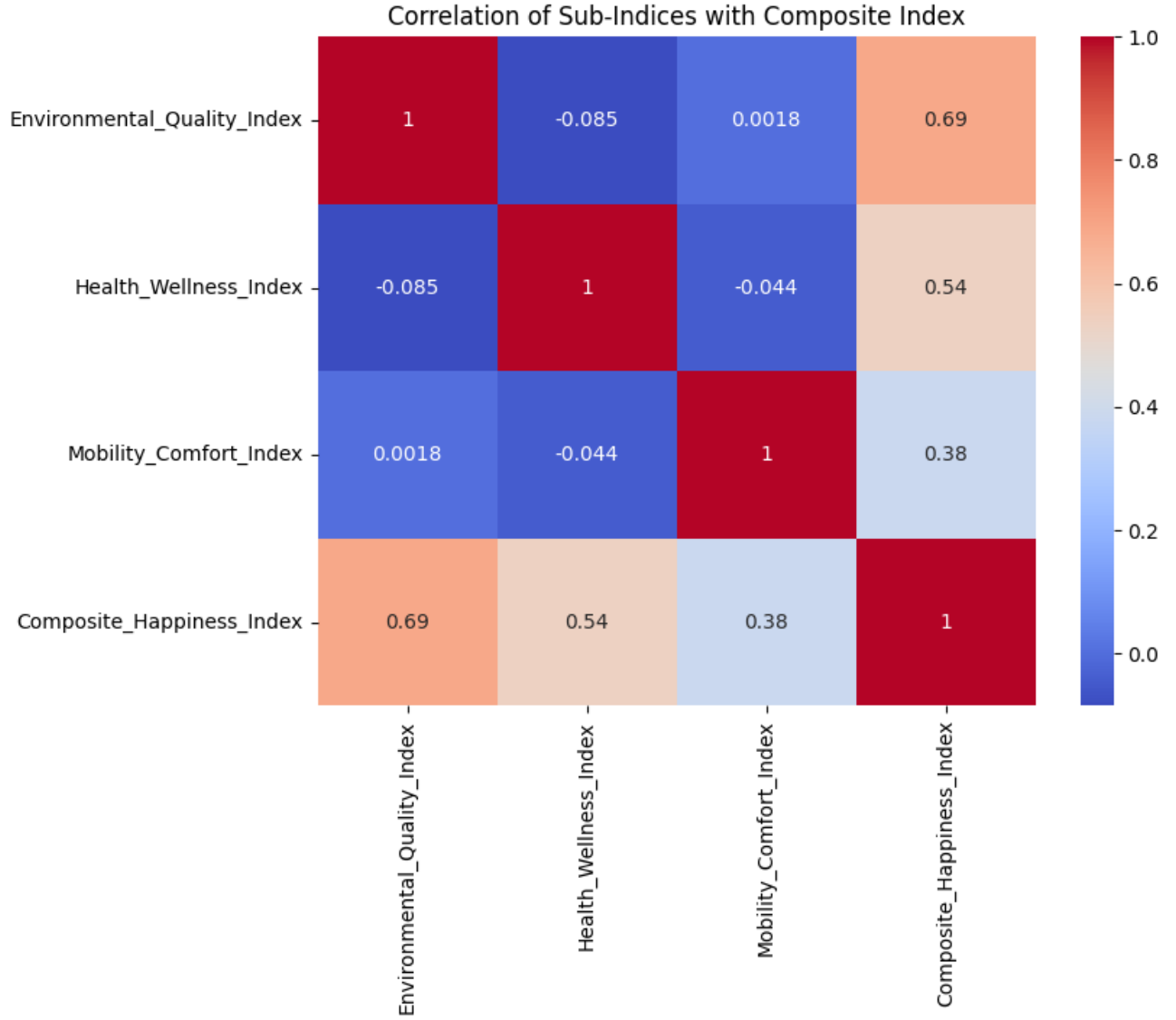
```
# Top 10 cities
top_10 = df.groupby('City')['Composite_Happiness_Index'].mean().sort_values(ascending=False).head(10)

# Bottom 10 cities
bottom_10 = df.groupby('City')['Composite_Happiness_Index'].mean().sort_values().head(10)

# Plot side by side
fig, axes = plt.subplots(1, 2, figsize=(18, 6), sharey=True)

sns.barplot(x=top_10.values, y=top_10.index, ax=axes[0], palette='Greens_r')
axes[0].set_title("Top 10 Cities by Average Composite Happiness Index")
axes[0].set_xlabel("Composite Happiness Index")

sns.barplot(x=bottom_10.values, y=bottom_10.index, ax=axes[1], palette='Reds_r')
axes[1].set_title("Bottom 10 Cities by Average Composite Happiness Index")
axes[1].set_xlabel("Composite Happiness Index")

plt.tight_layout()
plt.show()
```



Correlation of Sub-Indices with Composite Index

This analysis quantified how strongly each sub-index (e.g., Health, Environment) influenced the overall index. It validated the importance of each component and guided possible re-weighting considerations for future iterations.

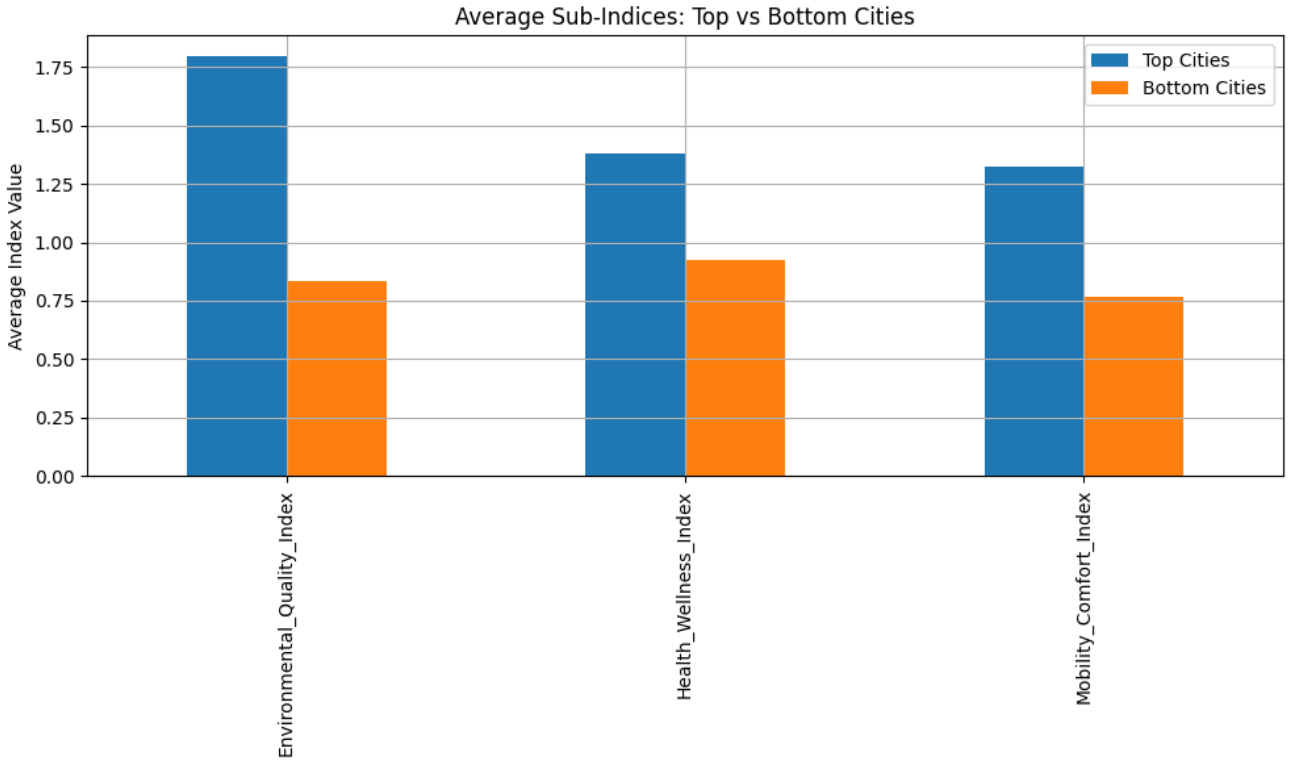
```
correlation = df[[
    'Environmental_Quality_Index',
    'Health_Wellness_Index',
    'Mobility_Comfort_Index',
    'Composite_Happiness_Index'
]].corr()

plt.figure(figsize=(8, 6))
sns.heatmap(correlation, annot=True, cmap='coolwarm')
plt.title("Correlation of Sub-Indices with Composite Index")
plt.show()
```



Average Sub-Indices for Top vs Bottom Cities

The average values for top vs. bottom-tier cities were compared across sub-indices. This highlighted which factors (e.g., Green Space, Healthcare) most consistently distinguished happy cities from less happy ones.

```
# Select top and bottom cities
top_cities = top_10.index.tolist()
bottom_cities = bottom_10.index.tolist()

# Computed average sub-indices
avg_top = df[df['City'].isin(top_cities)][['Environmental_Quality_Index', 'Health_Wellness_Index', 'Mobility_Comfort_Index']].mean()
avg_bottom = df[df['City'].isin(bottom_cities)][['Environmental_Quality_Index', 'Health_Wellness_Index', 'Mobility_Comfort_Index']].mean()

# Combined into one DataFrame
comparison_df = pd.DataFrame({'Top Cities': avg_top, 'Bottom Cities': avg_bottom})
comparison_df.plot(kind='bar', figsize=(10, 6), title='Average Sub-Indices: Top vs Bottom Cities')
plt.ylabel("Average Index Value")
plt.grid(True)
plt.tight_layout()
plt.show()
```



Interactive Dashboard

To complement the data analysis and static visualizations, an **interactive dashboard** was developed using **HTML**, **CSS**, and **JavaScript** (with **Chart.js**). This dashboard provides a dynamic and visually engaging way to explore the City Happiness Index.

Features and Implementation:

- **HTML (index.html)**: Structured the layout, headings, and sections of the dashboard. It includes responsive design elements, card-style UI containers, and embedded charts.
- **JavaScript (script.js)**: Handles data visualization logic. It dynamically renders:
    - A **bar chart** of the top 10 cities by composite index
    - A **radar chart** comparing sub-indexes for leading cities
    - A **traffic density pie chart**
    - A **monthly trend line chart** for user-selected cities
    - A **heatmap-style bar chart** comparing environmental, economic, and health indicators across cities
- **CSS (style.css)**: Styled the page to enhance readability, layout, and visual appeal, ensuring mobile responsiveness and consistent theming.

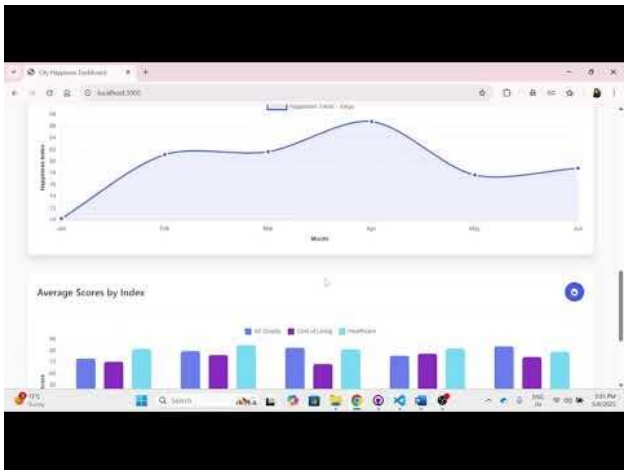
This dashboard allows users to:

- Select cities and observe monthly trends
- Compare sub-indices visually
- Instantly grasp traffic conditions and average scores

Purpose:

The dashboard enhances the accessibility and interpretability of the index by enabling non-technical stakeholders—such as city planners or citizens—to interact with the data without needing to run Python code.

The dashboard files are included in the submission and can be opened in any modern browser.



Conclusion

The City Happiness Index developed in this project provides a comprehensive and data-driven framework for assessing urban well-being. By integrating environmental, health, economic, and comfort-related variables into a single composite indicator, it offers a holistic measure of livability across global cities.

The project began with a structured dataset sourced from Kaggle's City Happiness Index 2024, which was then thoughtfully expanded through additional research to include new variables, forecasted values for the year 2030, and monthly granularity. These enhancements significantly increased the dataset's relevance and depth.

Through methodical data cleaning, imputation, normalization, and aggregation, the index was constructed with attention to statistical soundness and practical interpretability. Visualizations such as heatmaps, scatterplots, and time-series analyses brought clarity to city-wise patterns and highlighted key differences between high- and low-performing regions.

While the weighting scheme and clustering results provided actionable insights, future iterations of the index could benefit from:

- Expert validation of weighting structures
- Inclusion of more real-time or crowd-sourced metrics
- Broader benchmarking against external indices such as HDI or the OECD Better Life Index

Overall, this work demonstrates how multidimensional, well-structured data can be transformed into meaningful indicators that aid decision-making in urban planning and public policy.
