Philippine Junior Data Science Challenge 2024

WALKABLE: A Data Science Approach on Walkability and Safety Index (WSI) to Empower Urban Mobility and Sustainable Planning

TEAM DSENTINELS
dela Cruz, Karylle
Juanico, Hernel Niño
Macalisang, Ian Jure
Menorias, Nicole
Mugot, Chris Jallaine

EXECUTIVE SUMMARY

Urban mobility is increasingly strained by rising traffic congestion, pollution, and deteriorating public health. As cities continue to grow, the absence of data-driven solutions hinders effective decision-making in sustainable mobility planning. Policymakers and urban planners often lack the comprehensive insights needed to address these challenges, resulting in fragmented, short-term measures that not only fail to resolve the issues but frequently exacerbate them.

This project seeks to close this gap by introducing a Walkability and Safety Index (WSI), powered by advanced data analytics and predictive modeling. The WSI will assess critical factors such as infrastructure availability, safety, and environmental conditions, providing cities with reliable, actionable insights to guide infrastructure investments. By promoting active transportation options like walking and cycling, the WSI will help cities reduce car dependency, alleviate traffic congestion, and improve public health outcomes.

To further enhance the impact of the Walkability and Safety Index (WSI), the project will integrate it into an interactive data visualization platform (WALKABLE), incorporating additional unique features. This platform will transform complex data into clear, engaging visual formats, making it accessible to stakeholders across various sectors. By empowering decision-makers with intuitive data exploration, the platform will support more informed decisions, driving smarter infrastructure investments and long-term urban planning improvements, ultimately fostering more sustainable and livable cities.

KEYWORDS

Walkability, Active Transportation, Geographic Information Systems (GIS), Infrastructure Improvement, Sustainable Urban Design

BACKGROUND

In the Philippines, sustainable urban mobility is undermined by an increasing reliance on vehicles, leading to a range of societal issues that exacerbate traffic congestion, environmental degradation, and public health challenges. A prime example of this is Metro Manila, which experiences some of the worst traffic in the world, largely driven by its high population density and size (Chang et al., 2021). Addressing these issues is critical to improving mobility, reducing congestion, and mitigating environmental impacts.

The socio-economic effects involved are substantial. Helmi and Wahab (2023) indicate that traffic congestion increases fuel consumption and time wasted in traffic, raising overtime costs for commuters and businesses, with added stresses and reduced productivity. The environmental implications are also alarming, as they stem from the increased greenhouse gas emissions that contribute to climate change.

Central to a solution is the enhancement of the infrastructure for walking and bicycle riding. Outdated sidewalks, unsafe road crossings, and the absence of protected bike lanes lead people to rely on motor vehicles, worsening congestion and furthering environmental degradation (Llanto, 2016; Joy, 2024). Betterment in these infrastructures can lessen the dependency on cars, increase physical activity, and lower pollution; however, these efforts are still being hampered by the prevalence of obsolete urban planning strategies.

Data-driven urban development shall fundamentally address such prevailing challenges for the Philippines. Bondoc et al. (2018) advocate for real-time data to be integrated into traffic management systems in such a way as to optimally direct traffic flow and reduce congestion. Yuana et al. (2019) stress predictive modeling to track the impact of infrastructure developments on walkability and safety.

This project primarily aims to develop a Walkability and Safety Index (WSI) to measure urban pedestrian experience in the Philippines. This index will serve as an essential guide for urban planners and policymakers, helping to identify and evaluate areas needing improvement and supporting the creation of strategies to enhance urban mobility. The WSI focuses on five key indicators: population density, which influences pedestrian movement; night-time lighting, which impacts safety and activity levels; air quality, crucial for public health; heat index, indicating thermal comfort; and proximity to amenities, affecting access to basic services. To make the insights easily accessible and understandable, this project integrates the WSI into an intuitive data visualization platform, bridging the gap between data analysis and practical decision-making. It also ensures that even non-technical users can grasp the insights and apply them to promote sustainable urban development.

THE PROBLEM

A key challenge in sustainable urban development is the lack of a standardized, data-driven measure to assess and guide improvements in walkability and safety. Cities often lack the tools to make informed decisions on where and how to invest in infrastructure upgrades, such as pedestrian pathways, safety lighting, and environmental enhancements. Without such metrics, urban planners struggle to prioritize interventions that reduce car dependency and promote healthier lifestyles.

The Walkability and Safety Index (WSI), driven by data science, offers actionable insights into which aspects need improvement, guiding targeted investments that enhance walkability, safety, and overall urban mobility.

LITERATURE

Walkability is vital for sustainable urban planning. Several studies used GIS and street connectivity indices to assess features like sidewalk quality and public service access. These measures encourage active lifestyles and enhance urban design (Barbosa et al., 2019; Kim et al., 2020). Furthermore, the '5Ds'—density, diversity, design, destination accessibility, and distance to transit—provide a framework for understanding walkability (Kim, 2023).

Despite advancements in walkability, significant issues persist. The lack of locally gathered data and varying methodologies make it challenging for urban planners to assess walkability in real-time, hindering efforts to improve it (Hinckson et al., 2017). This inconsistency leads to fragmented strategies, as studies often include different variables, such as street connectivity and land use mix, complicating the development of a cohesive approach (Siqueira et al., 2023; Cerin et al., 2019; Portegijs et al., 2017).

The integration of thermal comfort and air quality as environmental factors into the indices of walkability is still an emerging area. Safransky (2019) and Shammas & Escobar (2019) emphasize the influence of temperature, shade, and air quality on pedestrian behavior, with few planning administrators properly incorporating these variables, While studies like Yamamura et al. (2017) illustrate the impact of polluted air on pedestrian activity, the importance of air quality in walkability indices remains underexplored.

Currently, no comprehensive walkability index integrates five key criteria: air quality, heat index, night-time lighting, distance to amenities, and population density. This presents an opportunity to create a robust tool for urban planners to enhance walkability and guide infrastructure development. This tool will empower planners to design cities for people, not just vehicles, fostering urban environments that promote physical activity, reduce emissions, and improve quality of life.

OBJECTIVES

To guide the direction of the project, an outline is provided to help us address the challenges of urban walkability and safety.

- Design a Comprehensive Data Science Pipeline for Walkability and Safety Index (WSI)
- Develop a Dynamic and Interactive Data Visualization Platform
- Test the Usability and Performance of the Platform

SCOPE & DEFINITIONS

This project will develop a data-driven platform to enhance walkability and safety in 12 cities: *Dagupan, Palayan, Navotas, Mandaluyong, Muntinlupa, Legazpi, Iloilo, Mandaue, Tacloban, Zamboanga, Cagayan de Oro, and Davao*. These cities were selected based on available datasets from Project CCHAIN. The platform will integrate walkability and safety scores based on key dimensions such as *heat index, air quality, night-time lighting*, *proximity to amenities, and population density*. It will serve as a collaborative tool for policymakers, urban planners, and residents to make data-driven decisions and enhance infrastructure safety and accessibility.

DATASETS UTILIZED

1. **Philippines Subnational Administrative Boundaries (Barangay Level)**
[Global Administrative Areas (GADM)]
To access geospatial data for visualization and mapping.

2. **Project CCHAIN Datasets**
[Project CCHAIN]
To access historical population and infrastructure data.

3. **OpenWeatherMap API**
[OpenWeatherMap]
To access real-time heat and air quality data.

METHODOLOGY

The methodology for this project is structured around several key stages as shown in Figure 1 (see Appendices).

DATA COLLECTION

GADM Database: Geospatial data on barangay-level administrative boundaries will be acquired to visualize and map the selected areas as illustrated in Figure 2 (see Appendices).

CCHAIN Datasets: These include the urban indicators shown in Figures 3 to 6 (see Appendices). Population density (Dp = N/A, where N is the population and A is the land area) is calculated as illustrated in Figure 7 (see Appendices).

External APIs: Real-time data on environmental factors, such as the Heat Index and Air Quality Index, will be obtained to assess the safety and comfort of walking in urban spaces.

DATA PREPROCESSING

After data collection, preprocessing ensures quality and usability by cleaning the data, handling missing values, and processing key indicators to create a cohesive dataset for analysis.

DATA ANALYSIS AND PREDICTIVE MODELING

The processed dataset is analyzed to calculate the Walkability and Safety Index (WSI), derived from various factors. Statistical methods are applied to score each indicator and assess its significance to the pedestrian experience. The formula for the Walkability and Safety Index (WSI) can be expressed as:

$$WSI = \sum_{i=1}^{n} w_i S_i$$

where,

S_i is the score of each indicator

n is the total number of indicators

w_i represents the weight of each indicator

The correlation heatmap shown in Figure 8 (see Appendices) explores the relationships between indicators and WSI. This supports the modeling process by giving an initial view of variable relationships, helping identify potential issues like multicollinearity, and guiding feature selection.

The initial predictive modeling setup will use Random Forest Regressor (see Appendices, Figure 9), as a starting point, to predict WSI and evaluate the performance. The project will explore more complex models like other regression models and other tree-based models depending on performance.

The model results are evaluated to assess performance and identify which features contribute most to the predictions.

VISUALIZATION AND MAPPING

Analysis results are visualized using mapping techniques. The Walkability and Safety Index and other data will be displayed on an interactive map as shown in Figure 10 (see Appendices), offering urban planners, local governments, and residents accessible insights for informed decision-making on infrastructure investments and urban design improvements.

PLATFORM TESTING

The platform will undergo various tests, including functionality, usability, and performance assessments in selected cities, to gather real-world feedback, refine the system, and resolve any issues before its full-scale launch.

POTENTIAL CHALLENGES

1. **Data Quality and Availability**
Incomplete, outdated, or inconsistent environmental and population data may lead to inaccurate index valuation.

2. **Dynamic Urban Environments**
Urban environments are dynamic. Static datasets may not reflect recent developments in mobility infrastructure or socioeconomic conditions.

3. **Model Generalization**
The model trained may not perform well on new or unseen data, especially if there are significant differences in the data distribution or context.

MITIGATION STRATEGIES

1. **Data Quality and Availability**
The project will gather data from multiple reliable sources, conduct data cleaning and augmentation techniques

to manage missing values and outliers. Additionally, engaging local communities will provide valuable insights to enhance the dataset.

2. **Dynamic Urban Environments**
The model will be continuously updated with real-time data from open data portals, sensors, or IoT devices. Time-series analysis techniques will be employed to capture temporal changes, and flexible models will be developed to adapt to future updates.

3. **Model Generalization**
The project will perform cross-validation and use diverse datasets during training. Apply techniques like domain adaptation or transfer learning and continuously monitor model performance in production to update it as new data becomes available.

SUCCESS METRICS AND KPI'S

1. **Crowdsourcing Volume Metrics:** This metric is focused on the amount of crowdsourced data collected.
   - **Volume of Crowdsourced Data:** Measure the monthly total of user-submitted data points.

$$Monthly\ Volume = \sum (Feedback + Infrastructure\ Suggestions + Ratings + Surveys)$$

Target: Increase the volume of crowdsourced data by **20% each year**.

2. **Walkability Confidence Metrics:** This metric assesses the increase of user confidence in using walkable spaces in cities.
   - **Walkability and Safety Index:** An increasing score suggests a more walkable and trusted environment.
     Target:
     - **Low Walkability** (scores below 80): Achieve a **10% increase** within **12 months**.
     - **Moderate Walkability** (scores between 80 and 90): Maintain the score and aim for a **5% increase** within **12 months**.
     - **High Walkability** (scores above 90): Maintain the score and ensure at least **85%** of users rate their walking experience as satisfactory through periodic surveys.

3. **Infrastructure Improvement Metrics:** This metric evaluates the effectiveness of infrastructure improvements in walkability and safety by counting changes and analyzing user sentiment and feedback.
   - **Implementation of Recommendation:** Maintain a count of infrastructure changes (e.g., expanded sidewalks, improved lighting, enhanced crosswalks) implemented on a monthly basis.
     Target: Facilitate **10** significant infrastructure improvements **per year**.
   - **Before-and-After Assessments:** Conduct pre-implementation and post-implementation assessments to gather baseline sentiment data from users through surveys on safety, accessibility, and overall satisfaction for consistency.

$$Improvement\ in\ Score = (PostImplementation\ Score - PreImplementation\ Score)$$

Target: Achieve a **15% increase** in **user sentiment** regarding walkability and safety.

FEASIBILITY AND INNOVATION

The project is highly feasible, utilizing established machine learning techniques, as shown in Figure 9 to get predictive power for Walkability and Safety Index. Results show that the model's Mean Squared Error (MSE) is low, indicating that the predictions are close to the actual WSI values. R-squared (R^2) also suggests an excellent fit. Also, feature importance, as shown in Figure 8, reveals which indicators are contributing most to the model's predictions, further explaining on how to improve the model.

The innovation lies in the creation of a comprehensive WSI and the integration of predictive modeling with an interactive data visualization dashboard as shown in Figure 10, enabling stakeholders to explore real-time predictions and make more informed urban planning decisions.

REFERENCES

Barbosa, J., Guerra, P., Santos, C., Nunes, A., Turrell, G., & Florindo, A. (2019). Walkability, overweight, and obesity in adults: A systematic review of observational studies. *International Journal of Environmental Research and Public Health, 16*(17), 3135. https://doi.org/10.3390/ijerph16173135

Bondoc, E., Caparas, F., Macias, J., Naculangga, V., & Estrada, J. (2018). An intelligent road traffic information system using text analysis in the most congested roads in Metro Manila. In *Proceedings of the 10th IEEE International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*. https://doi.org/10.1109/hnicem.2018.8666416

Cerin, E., Conway, T., Barnett, A., Smith, M., Veitch, J., Cain, K., & Sallis, J. (2019). Development and validation of the neighborhood environment walkability scale for youth across six continents. *International Journal of Behavioral Nutrition and Physical Activity, 16*(1). https://doi.org/10.1186/s12966-019-0802-5

Chang, Y., Jo, S., Lee, Y., & Lee, Y. (2021). Population density or population size: Which factor determines urban traffic congestion? *Sustainability, 13*(8), 4280. https://doi.org/10.3390/su13084280

Global Administrative Areas. (n.d.). *GADM database of Global Administrative Areas*. https://gadm.org/

Helmi, S., & Wahab, W. (2023). Traffic congestion effect on socio-economic of road users in Palembang City. In *Proceedings of the 4th International Conference on Economics, Management, Law and Education* (pp. 80-90). https://doi.org/10.2991/978-2-38476-072-5_9

Hinckson, E., Cerin, E., Mavoa, S., Smith, M., Badland, H., Witten, K., & Schofield, G. (2017). What are the associations between neighbourhood walkability and sedentary time in New Zealand adults? *BMJ Open, 7*(10), e016128. https://doi.org/10.1136/bmjopen-2017-016128

Joy, L. (2024). Unraveling gridlock disaster: Exploring the interplay between disaster preparedness and persistent traffic congestion in Magdiwang Noveleta, Cavite. *IJAMRS, 4*(3), 616-629. https://doi.org/10.62225/2583049x.2024.4.3.2831

Kim, E., Kim, J., & Kim, H. (2020). Neighborhood walkability and active transportation: A correlation study in leisure and shopping purposes. *International Journal of Environmental Research and Public Health, 17*(7), 2178. https://doi.org/10.3390/ijerph17072178

Kim, H. (2023). The relationship of single-family detached house prices with neighborhood walkability and disaster risk. *PLOS ONE, 18*(10), e0292657. https://doi.org/10.1371/journal.pone.0292657

Llanto, G. (2016). Philippine infrastructure and connectivity: Challenges and reforms. *Asian Economic Policy Review*, 11(2), 243-261. https://doi.org/10.1111/aepr.12141

OpenWeatherMap. (n.d.). *OpenWeatherMap API*. https://openweathermap.org/

Portegijs, E., Keskinen, K., Tsai, L., Rantanen, T., & Rantakokko, M. (2017). Physical limitations, walkability, perceived environmental facilitators and physical activity of older adults in

Finland. *International Journal of Environmental Research and Public Health*, *14*(3), 333. https://doi.org/10.3390/ijerph14030333

Safransky, S. (2019). Comfort and time-based walkability index design. *International Journal of Environmental Research and Public Health, 16*(16), 2850. https://doi.org/10.3390/ijerph16162850

Shammas, T., & Escobar, F. (2019). Comfort and time-based walkability index design: A GIS-based proposal. *International Journal of Environmental Research and Public Health, 16*(16), 2850. https://doi.org/10.3390/ijerph16162850

Siqueira, G., Mabry, R., Siyabi, H., Adeel, A., Malaj, S., & Oyeyemi, A. (2023). Construct validity of the physical activity neighborhood environment scale-Oman. *Frontiers in Public Health, 11*. https://doi.org/10.3389/fpubh.2023.1026656

Think Data Science. (2024). *Project CCHAIN* [Data set]. Kaggle. https://www.kaggle.com/datasets/thinkdatasci/project-cchain

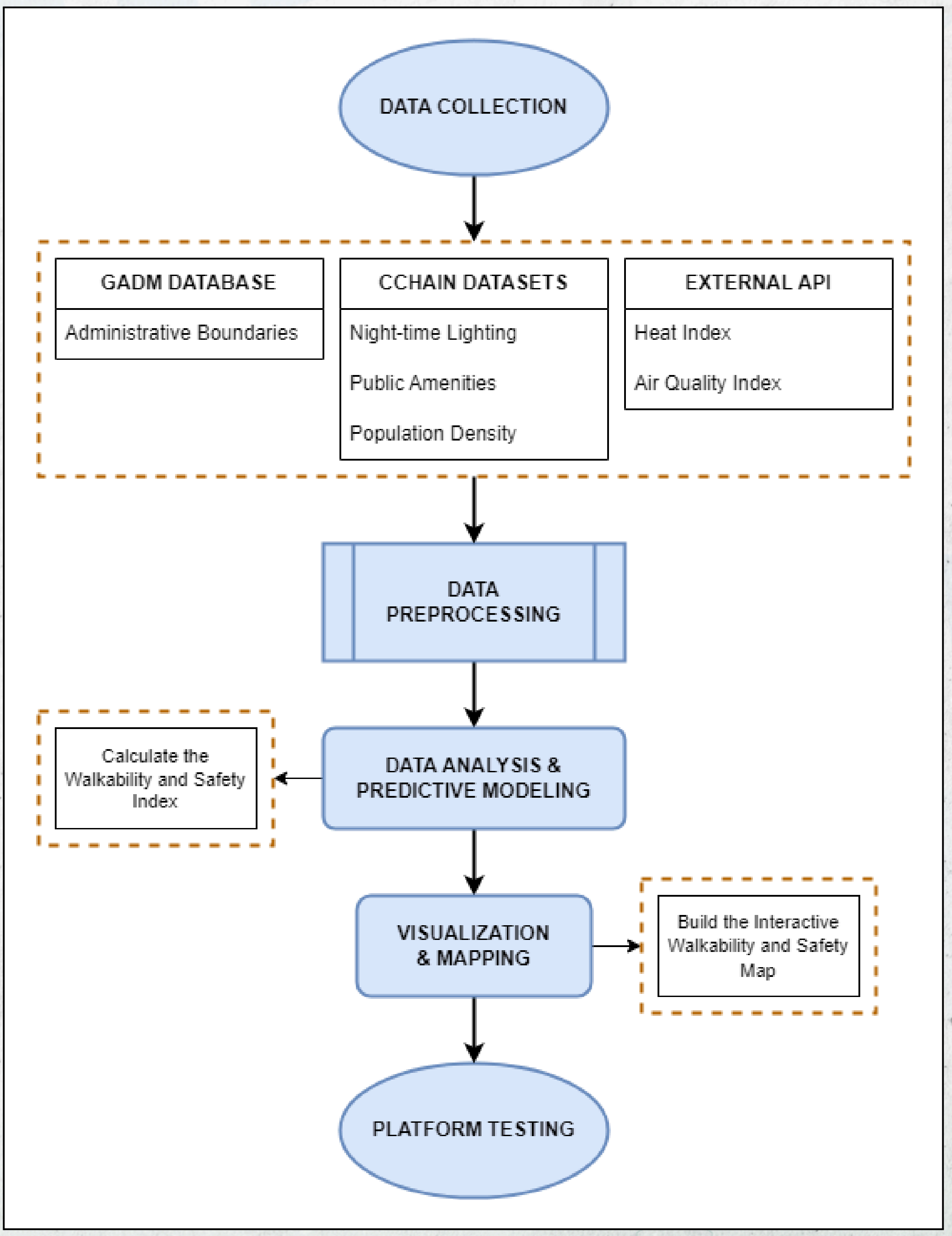
Yamamura, S., Fan, L., & Suzuki, Y. (2017). Assessment of urban energy performance through integration of BIM and GIS for smart city planning. *Procedia Engineering*, 180, 1462-1472. https://doi.org/10.1016/j.proeng.2017.04.309

Yuana, S., Sengers, F., Boon, W., & Raven, R. (2019). Framing the sharing economy: A media analysis of ridesharing platforms in Indonesia and the Philippines. *Journal of Cleaner Production*, 212, 1154-1165. https://doi.org/10.1016/j.jclepro.2018.12.073

APPENDICES

Figure 1

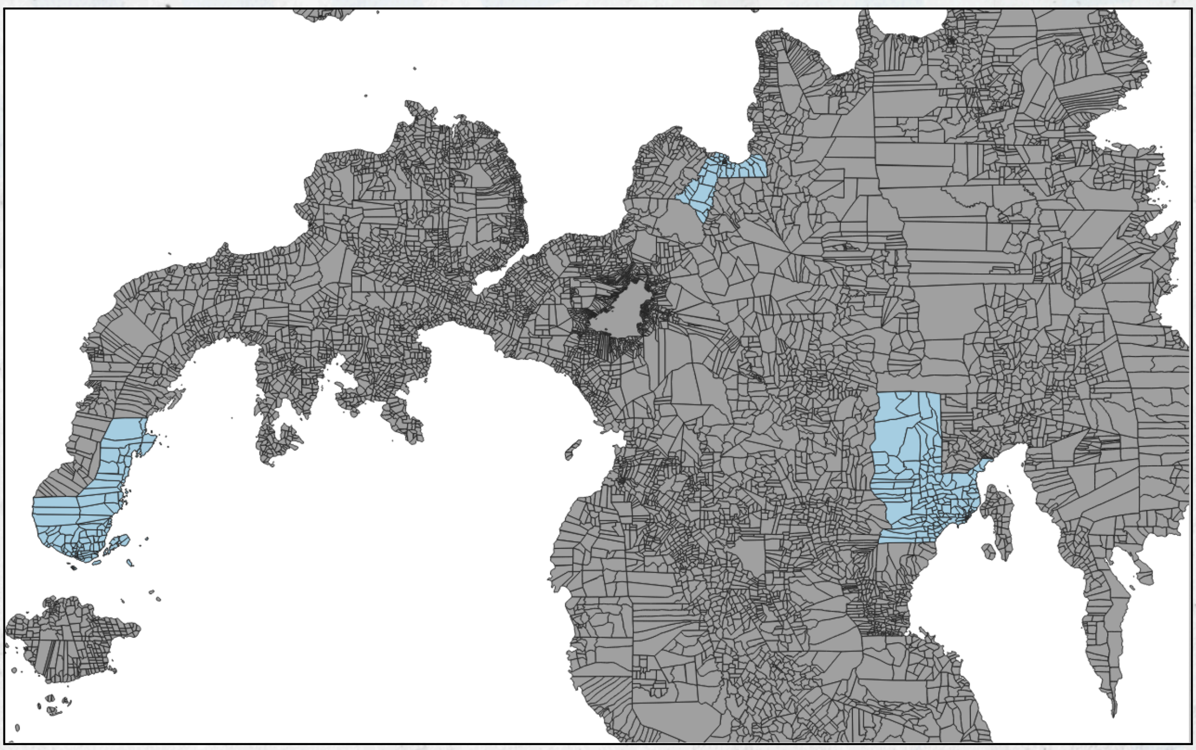
Project Methodology Flowchart



Note. The methodology involves data collection, preprocessing, analysis, modeling, visualization, and testing. This data-driven approach aims to improve urban walkability and support sustainable city planning.

Figure 2

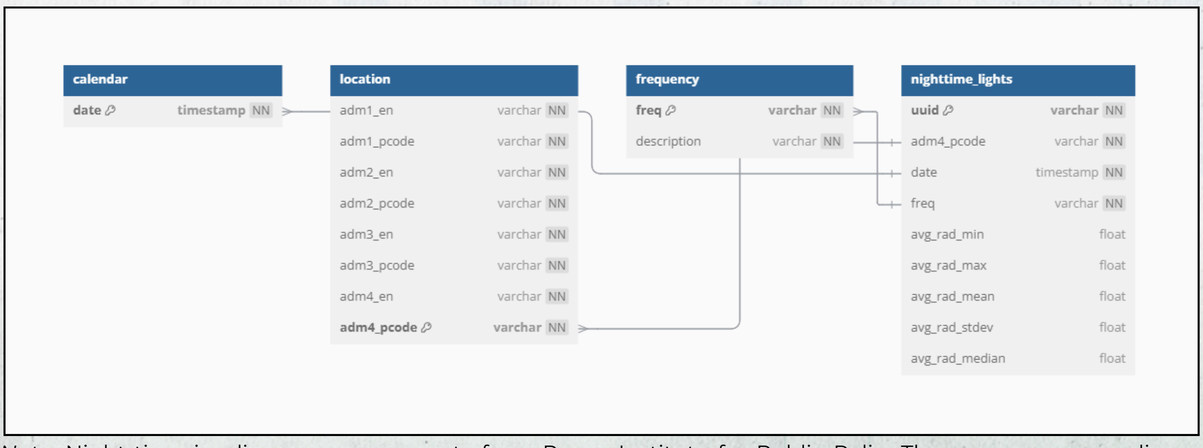
Map of Mindanao with the Selected Cities



Note. Regions highlighted represent the selected cities in Mindanao (Zamboanga, Cagayan de Oro, and Davao) which are part of the scope. From *GADM database of Global Administrative Areas,* Global Administrative Areas (https://gadm.org/download_country.html)

Figure 3

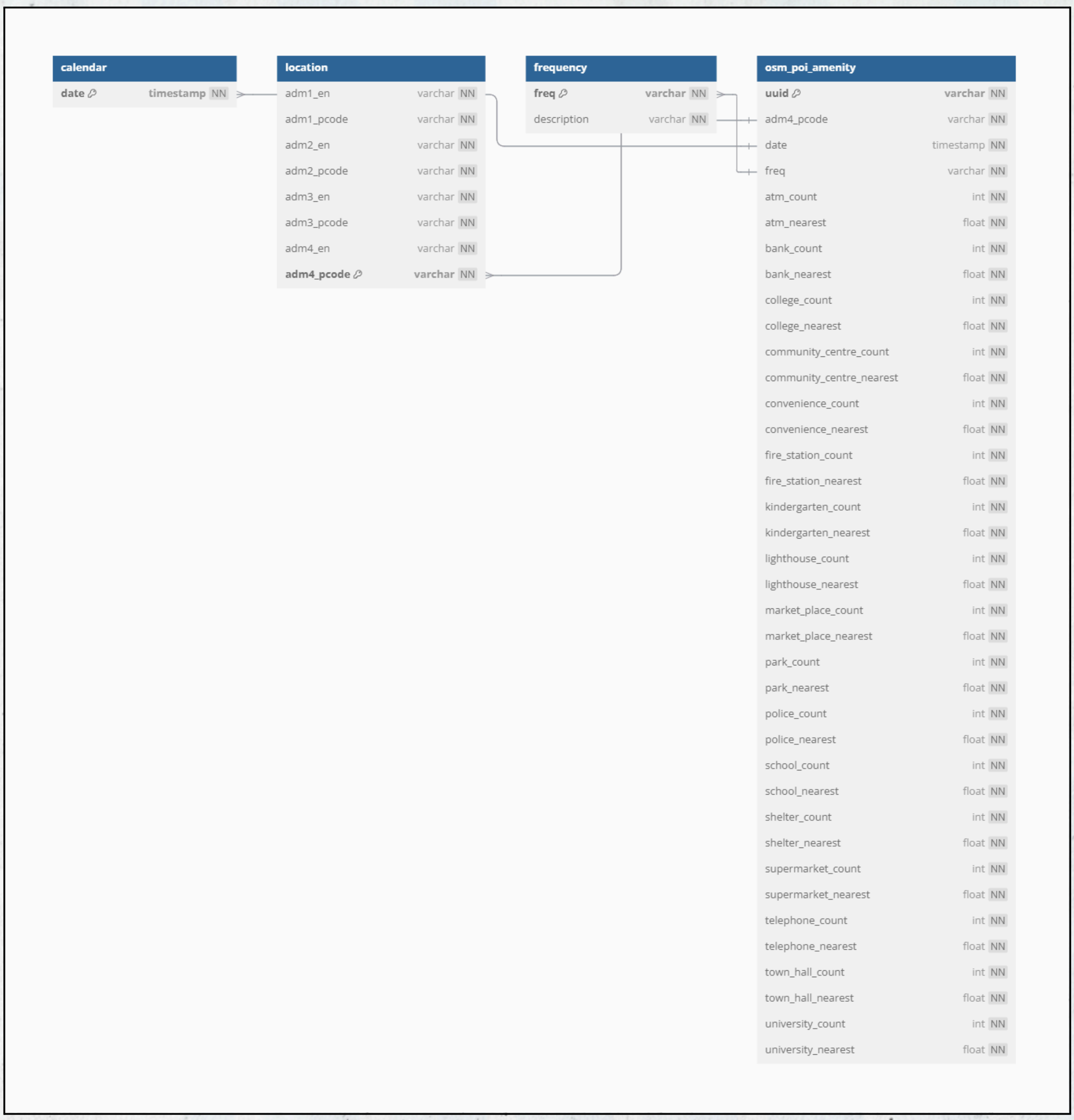
Database Schema for Night-time Lights Dataset



Note. Night-time irradiance measurements from Payne Institute for Public Policy.The mean average radiance from this dataset will be used in calculating the walkability and safety index. From *Project CCHAIN*, 2024, Kaggle (kaggle.com/datasets/thinkdatasci/project-cchain)

Figure 4

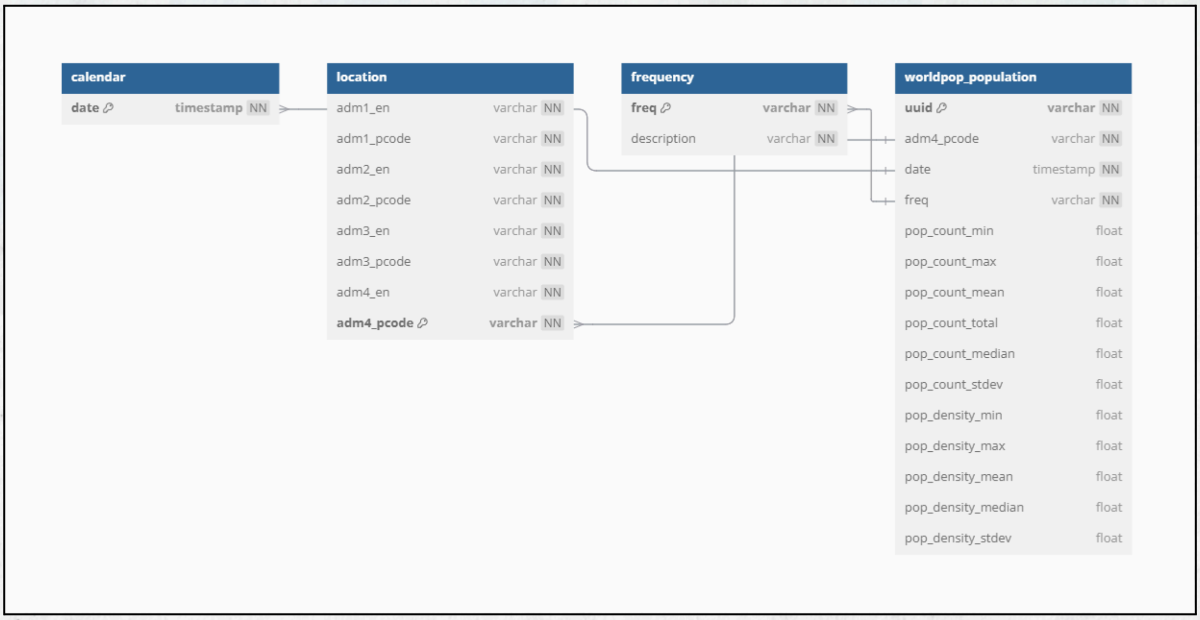
Database Schema for Amenities Dataset



Note. Basic public and private service facilities were analyzed, with statistics for each point of interest (POI) type—including sums and nearest distances—calculated for every barangay using vector zonal statistics. From this dataset, distance to amenities will be used in calculating the walkability and safety index. From *Project CCHAIN*, 2024, Kaggle (kaggle.com/datasets/thinkdatasci/project-cchain)

Figure 5

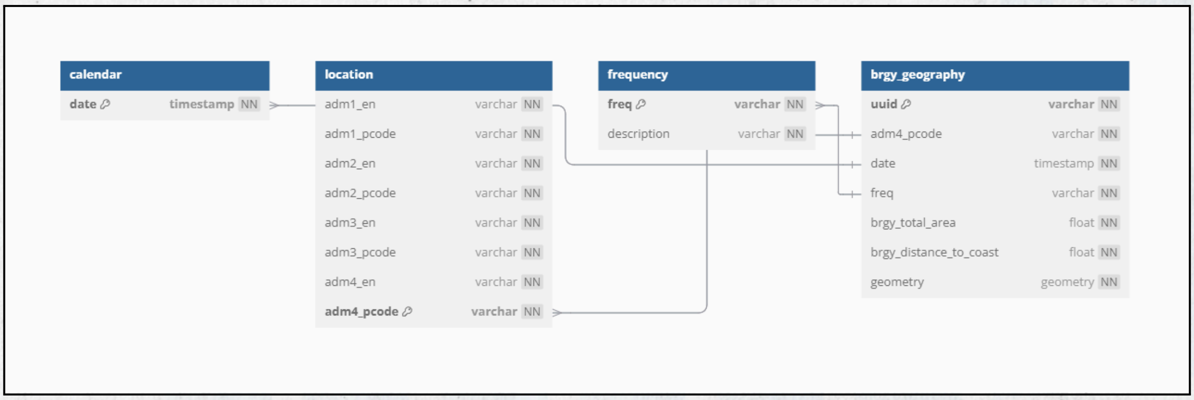
Database Schema for Population Dataset



Note. Gridded population estimates are sourced from WorldPop, covering population counts and density estimates. From this dataset, population count will be used for population density calculation. From *Project CCHAIN*, 2024, Kaggle (kaggle.com/datasets/thinkdatasci/project-cchain)

Figure 6

Database Schema for Barangay Geography Dataset



Note: From this dataset, barangay total area will be used for population density calculation. *From Project CCHAIN, 2024, Kaggle (*[*kaggle.com/datasets/thinkdatasci/project-cchain*](kaggle.com/datasets/thinkdatasci/project-cchain)*)*

Figure 7

Code Snippet for Data Analysis on Indicators - Calculating Population Densities

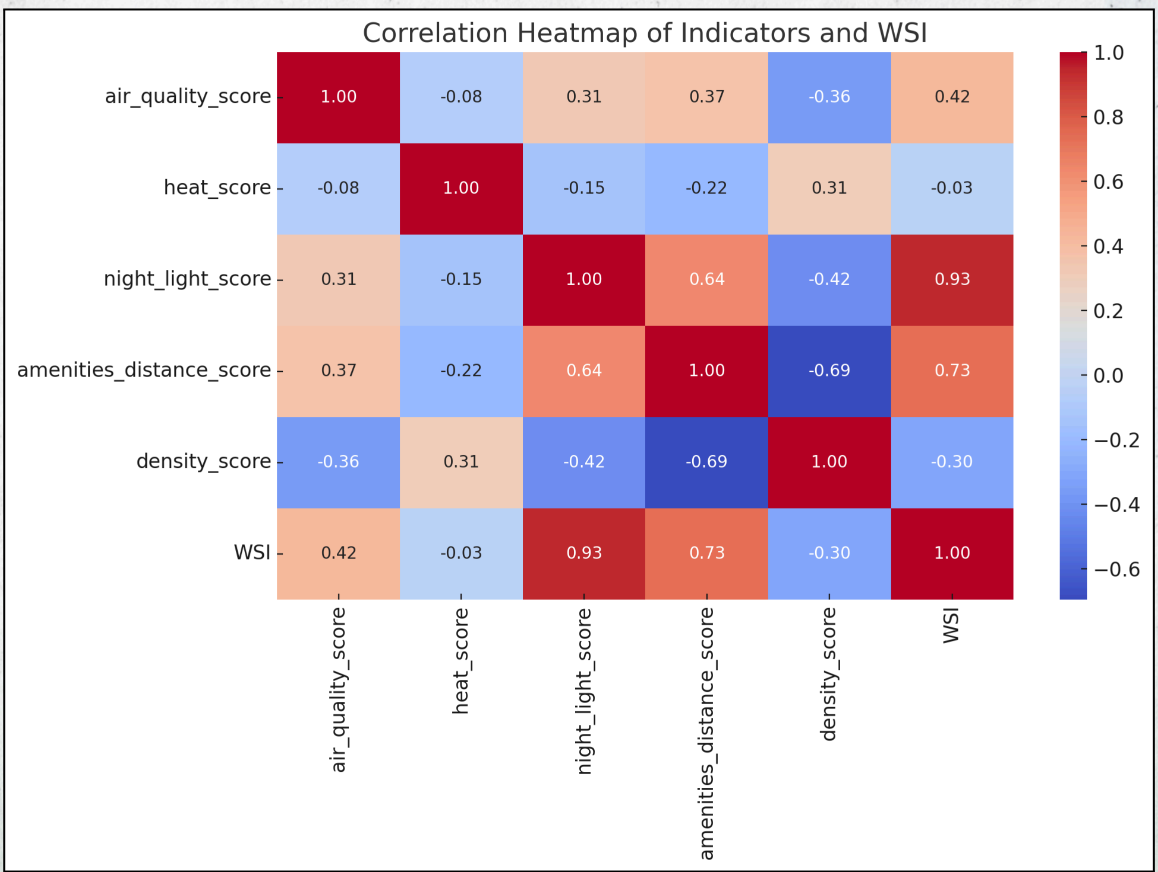
```python
density_data = pd.merge(pop_data, land_data, on="adm4_pcode", how="inner")
density_data = density_data[['pop_count_total', 'brgy_total_area']]

# Calculate the population density (population / square kilometer)
density_data['population_density'] = (density_data['pop_count_total'] / density_data['brgy_total_area'])
density_data.head(10)
```

adm4_pcode	pop_count_total	brgy_total_area	population_density
PH050506037	3198.0	0.4738	6749.683411
PH050506019	355.0	0.0763	4652.686763
PH063022118	1997.0	0.1259	15861.795075
PH063022015	785.0	0.1261	6225.218081
PH063022110	9718.0	0.2428	40024.711697
PH063022040	2017.0	0.2656	7594.126506
PH063022020	2118.0	0.0626	33833.865815
PH063022109	761.0	0.1506	5053.120850
PH034919018	285.0	13.7074	20.791689
PH063022152	1367.0	0.8800	1553.409091

Note. Preliminary data exploration in calculating population densities for various barangays.

Figure 8

Correlation Heatmap of Indicators and WSI



Note. The correlation heatmap reveals how each of these indicators relates not only to WSI but also to each other. Notably, Night Light Score and Amenities Distance Score shows a strong positive correlation with WSI. Air Quality Score also contributes positively, but to varying extents. Density Score and Heat Score show moderate negative correlations.

Figure 9

Code Snippet for Predictive Model Setup

```
# Define features (indicators) and target (WSI)
X = data[['air_quality_score', 'heat_score', 'night_light_score', 'amenities_distance_score', 'density_score']]
y = data['WSI']

# Split the dataset into training and testing sets (80% train, 20% test)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
#Train a Random Forest Regressor
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
```

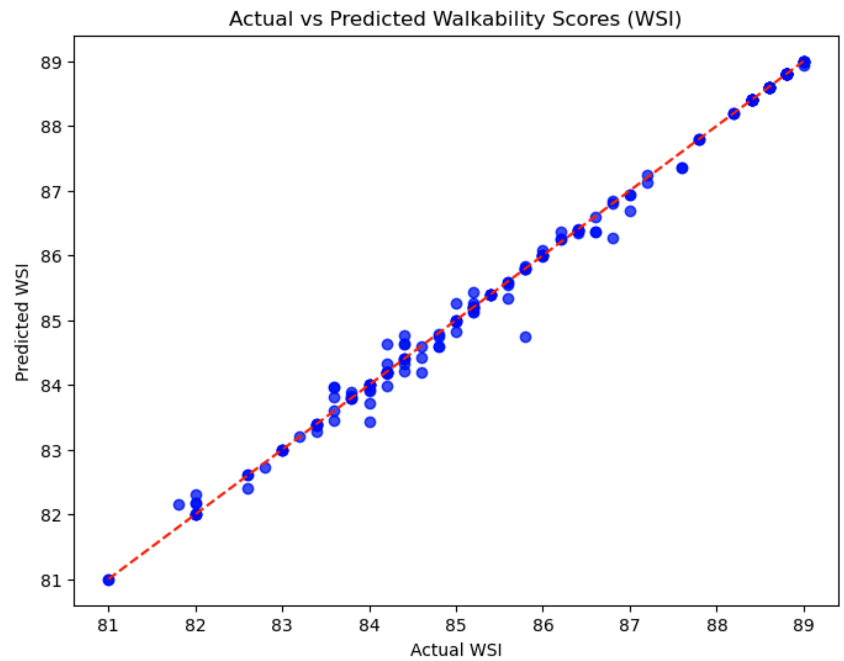
```
RandomForestRegressor
RandomForestRegressor(random_state=42)
```

```
#Make predictions on the test set
y_pred = rf_model.predict(X_test)
```

```
#Evaluate the model performance
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
print(f"R-squared: {r2}")
```

```
Mean Squared Error: 0.022329772727273534
R-squared: 0.9952282267987581
```

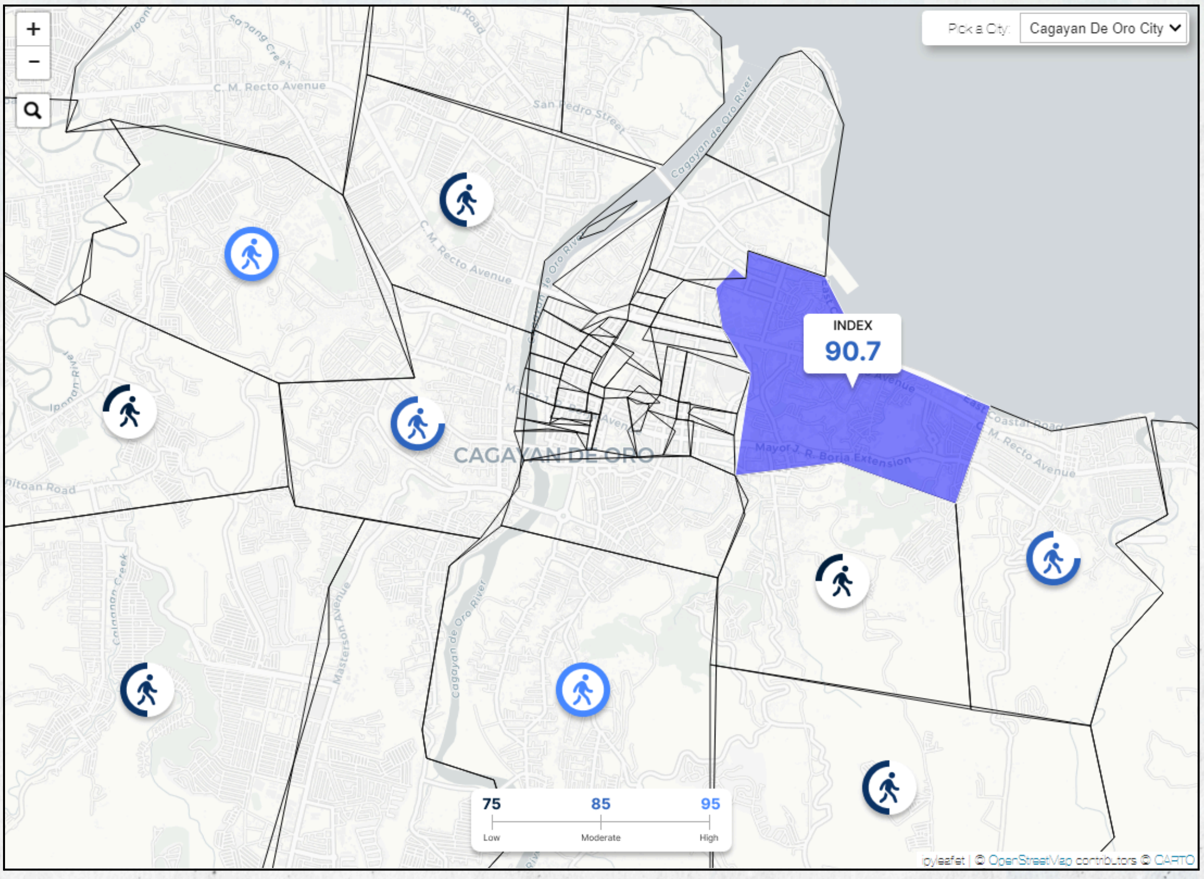
```
#Plot actual vs predicted values
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_pred, alpha=0.7, color='blue')
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red', linestyle='--')
plt.title('Actual vs Predicted Walkability Scores (WSI)')
plt.xlabel('Actual WSI')
plt.ylabel('Predicted WSI')
plt.show()
```



Note. Model performance metrics for the Random Forest Regressor shows Mean Squared Error (MSE): 0.0223, which is quite low, indicating that the predictions are close to the actual WSI values. R-squared (R^2): 0.995 suggests that the model explains 99.5% of the variance in WSI, which is an excellent fit. The Actual vs. Predicted Walkability Scores (WSI) plot shows that the predicted values align closely with the actual WSI values, which indicates that the model is performing well.

Figure 10

Project WALKABLE's Initial Interface Design



Note. Initial interface design of the platform integrating the Walkability and Safety Index (WSI).