**Data Analysis Plan for Event (Heatwave and High Air Pollution) and Non-Event (Average Temperature and Moderate/Good Air Quality) Survey**

**1. Research Objective**

The primary objective of this research is to collect evidence on how events like heatwaves and high air pollution affect the health, daily routines, and overall well-being of urban youth, and to understand their priorities for improving the healthiness and sustainability of their cities.

**2. Variables and Data Categorization**

**Demographic Variables:**

• Age Group: 13–17, 18–24, 25–34, >35
• Parent Status: Yes/No
• Gender: Female, Male, Other/Prefer not to say
• Household Income (USD): <$100, $100–$499, $500–$1,499, $1,500–$4,000, >$4,000

**Health and Wellbeing Indicators:**

• Self-Perceived General wellbeing: Great, Good, Okay, Not good
• Sleep Quality: Scale from Terrible to Great
• Physical Symptoms: (e.g., heat exhaustion, respiratory difficulties, headaches)
• Mental Health Indicators: Low mood, anxiety/stress, difficulty concentrating

**Physical Activity Data:**

• Physical Activity Levels (minutes/day): 0 to >60 minutes
• Disruptions to Daily Life: (e.g., missed work/school, late arrivals, canceled social events)

**Perception and Attitudes:**

• Concern About Climate Events: Scale from Not at all concerned to Extremely concerned
• Satisfaction with Government Response: Scale from Very satisfied to Very dissatisfied
• Urban Improvement Ideas: Open-ended responses categorized into key themes

**Survey Metadata:**

• Previous Participation: Yes/No

#### 2.1 Variable Coding

1. **Demographic Variables:**
	* Age (continuous variable)
	* Gender (categorical: Female, Male, Other/Prefer not to say)
	* Parental Status (categorical: Yes, No)
	* Youngest Child’s Age (continuous)
	* Household Income (ordinal: 5 levels)
2. **Health and Well-being:**
	* Self-reported general wellbeing (ordinal: Very good to Very bad)
	* Sleep quality (ordinal: Very good to Very bad)
	* Physical symptoms (binary: Yes, No for each symptom)
3. **Physical Activity:**
	* Duration of physical activity (ordinal: 6 levels)
4. **Impact on Daily Activities:**
	* Disruption indicators (binary: Yes, No for each disruption)
5. **Community Preparedness and Heatwave Concern:**
	* Concern about heatwaves (ordinal scale: 1 to 10)
	* Satisfaction with community response (ordinal: 5 levels)

**3. Data Cleaning and Preparation**

Data will be collected from heatwave and high air pollution events as well as non-events (days with moderate/good air quality and average or below-average temperatures for that month).

**Data Cleaning Steps:**

1. Translation and Standardization: Convert survey responses to English.
2. Structural Review: Assess completeness, identify missing values, and check for duplicate entries.
3. Formatting Consistency: Ensure uniform text and numerical formatting.
4. Error Correction: Fix spelling, naming inconsistencies, and logical errors.
5. De-identification: Remove any personally identifiable information from open-ended responses.

**Handling missing data:**

In this study, missing data is expected to be minimal, as all quantitative questions were compulsory. Potential gaps may arise for the gender question (where “I’d prefer not to say” is an option), the income questions (where “I don’t know” is allowed), and the optional qualitative question. Responses such as “I’d prefer not to say” or “I don’t know” will be treated as valid data reflecting participant choices rather than as missing data.

Missing data, such as from participants who began the survey but were later excluded (e.g., due to age i.e, older than 29 years and had no children), were reviewed to confirm they met exclusion criteria. These responses were not included in the final analysis to maintain data quality and relevance.

For the qualitative question, non-responses will be considered a normal part of voluntary participation, and the analysis will focus solely on the responses provided without any imputation. The extent of missing data will be documented, and any observable patterns will be noted to ensure transparency in the analysis.

**Handling duplicate data:**

Repeated data (duplicates) were identified in SPSS using the “Identify Duplicate Cases” function. This was accessed via *Data - Identify Duplicate Cases*, where a unique identifier i.e., such as participant ID was selected to check for exact duplicate entries. Cases were flagged, and the dataset was reviewed to confirm whether these repetitions were true duplicates or intentional repeated measures. Duplicates were then removed using the “Select Cases” option, choosing to delete duplicates permanently after they have been reviewed and confirmed as true duplicates. If necessary, aggregated values (e.g., means or sums) were calculated for non-exact duplicates before proceeding with further analysis.

**4. Statistical Analysis Plan**

**Descriptive Analysis (Comparing Event vs. Non-Event Periods): Using IBM SPSS Version 29**

***Demographic i.e., Age, Parent, Child’s Age, Gender and Income:***

* **Stratification by crosstabulation:** Split the basic **demographic** data by *event (heatwave and high air pollution)* vs. *non-event* (normal temperatures and healthy levels of air quality) categories.
* **Comparisons:** Use appropriate statistical tests based on data distribution to understand the significant difference during heatwave / high air pollution and normal temperature / health air quality level periods. The statistical tests may include:
	+ **Normally distributed data:** Independent t-tests.
	+ **Skewed data:** Mann-Whitney U tests.
	+ **Categorical variables:** Chi-square tests (or Fisher’s exact test if expected cell counts are low).

***Proposed Statistical tests***

| Demographic | Statistical Test | Reason |
| --- | --- | --- |
| Age | Independent Samples t-Test | compare the mean age of participants between heatwaves and non-heatwaves across all cities (Normal distributed data) |
| Parent | Pearson Chi-Square Tests | Compares categorical data |
| Child’s Age | Independent Samples t-Test | compare the mean age of participants between heatwaves and non-heatwaves across all cities (Normal distributed data) |
| Gender | Pearson Chi-Square Tests | Compares categorical data |
| Income | Pearson Chi-Square Tests | Compares categorical data |

***Health & Wellbeing and Daily Activity Disruptions:***

* **Stratification:** Split the data by *event* vs. *non-event* categories (air pollution, and heat waves).

**Inferential Analysis:**

| Research Question (RQs) and Key Variables | Analysis |
| --- | --- |
| **RQ 1:** How do high levels of air pollution and heat wave events affect the self-perceived health and wellbeing of children, young people, and parents (CYPP) living in cities?**Independent:** Heatwave /High Air Pollution Impact (Yes/No)**Dependent:** Self-reported Health and Wellbeing- General Wellbeing Perception (Feeling) -5 scale ordinal)- Sleep Quality - 5 Scale ordinal- Physical and Mental Health Symptoms - yes/no (binary) | **- Descriptive Statistics:**Summarize the distribution of self-perceived health & general wellbeing, sleep quality, and reported symptoms.**- Chi-Square Test of Independence:**To assess the significant association between event i.e., **heatwave / non-heatwave** and **reported symptoms** (e.g., headache, anxiety).**-Ordinal Logistic Regression:Outcome:** Self-perceived general wellbeing (feeling) and sleep quality (ordinal)**Predictors:** Heatwave / non-heatwave, demographic variables, air pollution /normal air quality levels.**-Binary Logistic Regression:Outcome:** Presence or absence of specific health symptoms (e.g., respiratory difficulties, headache)**Predictors:** Heatwave and High air pollution, age, gender, income. |
| RQ 2: How do heat waves affect the daily activities of CYPP?**Independent variables:** Event / Non-Event**Dependent variables:** Daily Activity Disruptions and Physical Activity | **Descriptive Statistics:**Summarize the distribution of daily activity disruptions and physical activities during both event and non-event.**- Chi-Square Test of Independence:**To assess the significant association between event i.e., **heatwave / non-heatwave** and **daily activities** disruptions.**-Ordinal Logistic Regression:Outcome:** Physical activity**Predictors:** Heatwave / non-heatwave, demographic variables, air pollution /normal air quality levels.**-Binary Logistic Regression:Outcome:** daily activity disruptions |

**5. Thematic Analysis for Open-Ended Responses (Urban Improvement Ideas): answers RQ 3:** What ideas do young people have for how their cities can become more healthy environments to live in?

Analytical Approach (Using NVIVO Software):

1. Familiarization:
 - Read responses for overall insights.
 - Identify recurring terms using Word Frequency & Text Search Queries.
2. Initial Coding:
 - Create Nodes (Codes) for key ideas.
 - Use In Vivo Coding (direct participant quotes) for authenticity.
3. Identifying Themes:
 - Group similar codes into broader themes.
 - Example: “Plant more trees,” “better parks,” and “urban cooling” as Green Infrastructure.
4. Reviewing Themes:
 - Validate coding consistency using Coding Comparison Queries.
 - Ensure themes are well-supported by data.
5. Defining & Naming Themes:
 - Write descriptions for each theme and explain its significance.

**6. Data Visualization Plan**

• Bar Charts: Prevalence of health symptoms, disruptions, and concern levels.
• Stacked Bar Charts: Overlapping impacts (e.g., missed school/work across age groups).
• Box Plots: Comparison of wellbeing indicators (e.g., sleep quality, mental health) by income level.
• Word Clouds: Visual representation of qualitative themes.