**Report: Air Quality & City Population Analysis**

**Students: Kalikyzy Daniya and Manatkyzy Zhanel**

**The goal of this project**

To analyze city-level air quality (PM2.5) and its association with city population using two heterogeneous sources:

1. an open **API** for structured air-quality measurements, and
2. **web scraping** of a static webpage for population figures.  
   We study whether larger cities tend to exhibit higher fine particulate concentrations and discuss caveats that may affect the relationship.

**Data Collection**

1. **OpenAQ API (API source).**  
   We queried the **OpenAQ Measurements** endpoint for parameter pm25 (µg/m³) across a selected set of cities over the **last 60 days**. For each city we aggregated all available measurements to compute a city-level mean (pm25\_mean). (Pagination handled until all results were retrieved.)
2. **Wikipedia (Web scraping source).**  
   Using requests + BeautifulSoup, we scraped each city’s **Wikipedia** page and parsed the **infobox** to extract the most recent **population** value for the city proper. We respected robots.txt, targeted rows containing the keyword “Population,” and implemented heuristics to parse formats such as 2,054,000 or 15.5 million.

**Cities (example set for reproducibility):** Almaty (KZ), Astana (KZ), Tashkent (UZ), Bishkek (KG), Istanbul (TR), Moscow (RU).  
**Time window:** last 60 days relative to the date of execution.

**Data Cleaning**

1. **Type handling.**
   * Cast pm25\_mean to numeric (float, µg/m³).
   * Cast population to integer (people), removing commas/spaces and converting “million/mln/млн” to absolute counts.
2. **Missing values & duplicates.**
   * Drop duplicate city entries arising from multiple pages/aliases.
   * Exclude cities with missing pm25\_mean or population from correlation analysis (kept in raw for transparency).
3. **Normalization & harmonization.**
   * Standardize city and country identifiers (e.g., “Astana,” “KZ”).
   * Ensure consistent units and column names: city, country, pm25\_mean, population, wiki.
4. **Outputs.**
   * Save the **raw merged** join for auditability: air\_quality\_raw\_join.csv.
   * Save the **clean** analysis table: air\_quality\_clean.csv.

**Data Analysis**

1. **Descriptive statistics.**  
   Compute summary stats (min/mean/median/max, IQR) for pm25\_mean and population. Inspect city-level spread and check for obvious outliers (e.g., exceptionally large population city or unusually high PM2.5 city).
2. **Correlation.**  
   Estimate **Pearson’s r** between population and pm25\_mean and report **p-value**.
   * **Interpretation guide:**
     + |r| < 0.2: very weak; 0.2–0.4: weak; 0.4–0.6: moderate; 0.6–0.8: strong; >0.8: very strong.
     + Statistical significance judged by p-value (e.g., α = 0.05).
3. **Sensitivity/robustness.**  
   Recompute r after removing a potential high-leverage city (largest population or highest PM2.5) to ensure the relationship is not driven by a single observation.
4. **Contextual considerations.**  
   Interpret results in light of likely confounders: industrial activity, heating season, valley‐basin topography, dust events, monitoring network density and siting.

**Data Visualization**

1. **Scatter plot (Population vs. PM2.5)** with a fitted **trendline** (polyfit). Label points by city to show relative positioning. Axes labeled with units (people, µg/m³), note the time window (“last 60 days”).
2. **Bar chart: mean PM2.5 by city** (sorted descending) to spotlight the most/least polluted among the sample.
3. **Bar chart: population by city** (sorted descending) to show scale differences that may bias perception in the scatter.  
   **Optional (if you extend analysis):**

* **Boxplot** of daily PM2.5 distributions by city (if daily aggregation is computed) to display variability.
* **Correlation heatmap** if adding more variables (e.g., PM10, NO2) in an extension.

**Results**

Replace placeholders after running your notebook.

* **Average PM2.5 (µg/m³):** mean ≈ **<insert>**, median ≈ **<insert>**; range: **<min>–<max>** across cities.
* **Population:** mean ≈ **<insert>**, median ≈ **<insert>**; range: **<min>–<max>**.
* **Correlation:** **r(population, PM2.5) = <insert>**, **p = <insert>** → (e.g., “moderate positive and statistically significant” / “weak and not significant”).
* **City ranking by PM2.5:** highest — **<A>**, **<B>**; lowest — **<C>**.
* **Notable mismatches:** e.g., **<City>** has high population but moderate PM2.5 (possible strong winds/mitigation policies), while **<City>** has lower population but higher PM2.5 (industrial sources/local stagnation).

**Conclusion**

This project shows how to **combine an API** (OpenAQ) and **static web data** (Wikipedia) into a single analytical dataset, perform transparent cleaning, and evaluate a policy-relevant question. The **sign and strength of r** clarify whether **population size** is a meaningful proxy for air-quality pressure in our sample; however, **population alone is insufficient** to capture the full drivers of PM2.5. Further variables (emissions inventory, traffic counts, meteorology) would improve explanatory power.

**Recommendations**

* **Augment variables.** Add meteorological indicators (wind speed, temperature inversions), traffic intensity, and industrial activity proxies.
* **Extend the window.** Compare 60-day means with seasonal/annual averages to reduce seasonal bias.
* **Improve population comparability.** Align definitions (city proper vs. metropolitan area) and ensure similar reference years.
* **Assess monitoring coverage.** Document station count and siting per city; weight averages if necessary.