Discipline: Data Collection and Preparation

Team Name: Air Quality

Team members: Toleu Bakhauddin, Kemel Merey

Title of the research: Air Quality and Population Analysis Among Major Kazakhstan cities.

Goal: analyze the relationship between air quality and population by obtained data from two sources: the web scrapping from the latitude.to and the Air Quality API.

Data Collection:

1. Web Scraping: We collected cities name, population and coordinates using BeautifulSoap to web scraping from the latitude.to site. The scraping data consists of cities coordinates to use it on query the API. We limited the list of cities by number of populations.

2. Air Quality API: We used Air Quality API to gather information about hourly pm10 and pm2.5 data in each city. PM10 includes the dust from construction sites, landfills and agriculture. PM2.5 rate shows the emissions from combustion of gasoline, oil, diesel fuel.

The dataset is merged based on city names for future analysis.

Data Cleaning & Preparation:

1. We checked for missing values and handled missing city information in web scraping by using Unicode normalizer. For any missing pm10_avr, pm25_avr, pm10_max and pm25_max values is handled by using their median values.

2. We checked for duplicate records based on city names. This ensures that each city is represented only once on the dataset.

3. We renamed columns for better readability. We classified the cities by population and categorized the air quality based on pm2.5 average to comparisons.

Data Analysis

We conducted exploratory data analysis (EDA) to analyze the relationship between air quality and populations.

- Descriptive statistics based on population, pm10_average and pm25_average shows that in dataset contains 24 cities and the average population is 307383, while the mean value for pm2.5 concentration is 12.1μg/m³ and for pm10 is 17.9 μg/m³.
- Correlation results shows that population and pm2.5 have a strong relationship with 72%, which suggests that emissions from combustion of gasoline, oil, diesel fuel in large cities are significantly higher.

Data Visualization

1. **Population vs Air Quality**
   Scatter plot shows the relationship between city population and air quality by pm2.5 concentration. This indicates that city with higher population tend to have higher pm2.5.
2. **Top 10 cities with Worst Air Quality**
   A horizontal bar chart illustrates the 10 most polluted cities based on average pm2.5 concentration. The Almaty city has the worst air quality among the 10 cities.
3. **Air Quality by City Size**
   The box plot clearly shows that large cities with 1M+ population have worse air quality, while smaller cities with 100k-500k population have more stable and lower pm2.5 concentration.
4. **Air Quality Category Distribution**
   The vertical bar chart with categorized air quality distribution shows that the majority of cities are classified as “Moderate”, while others as “Good” air conditions. Overall, the categorized air quality shows that air quality across most cities of Kazakstan is safe for living, also none of cities classified as ‘unhealthy for sensitive groups’ or ‘unhealthy’.