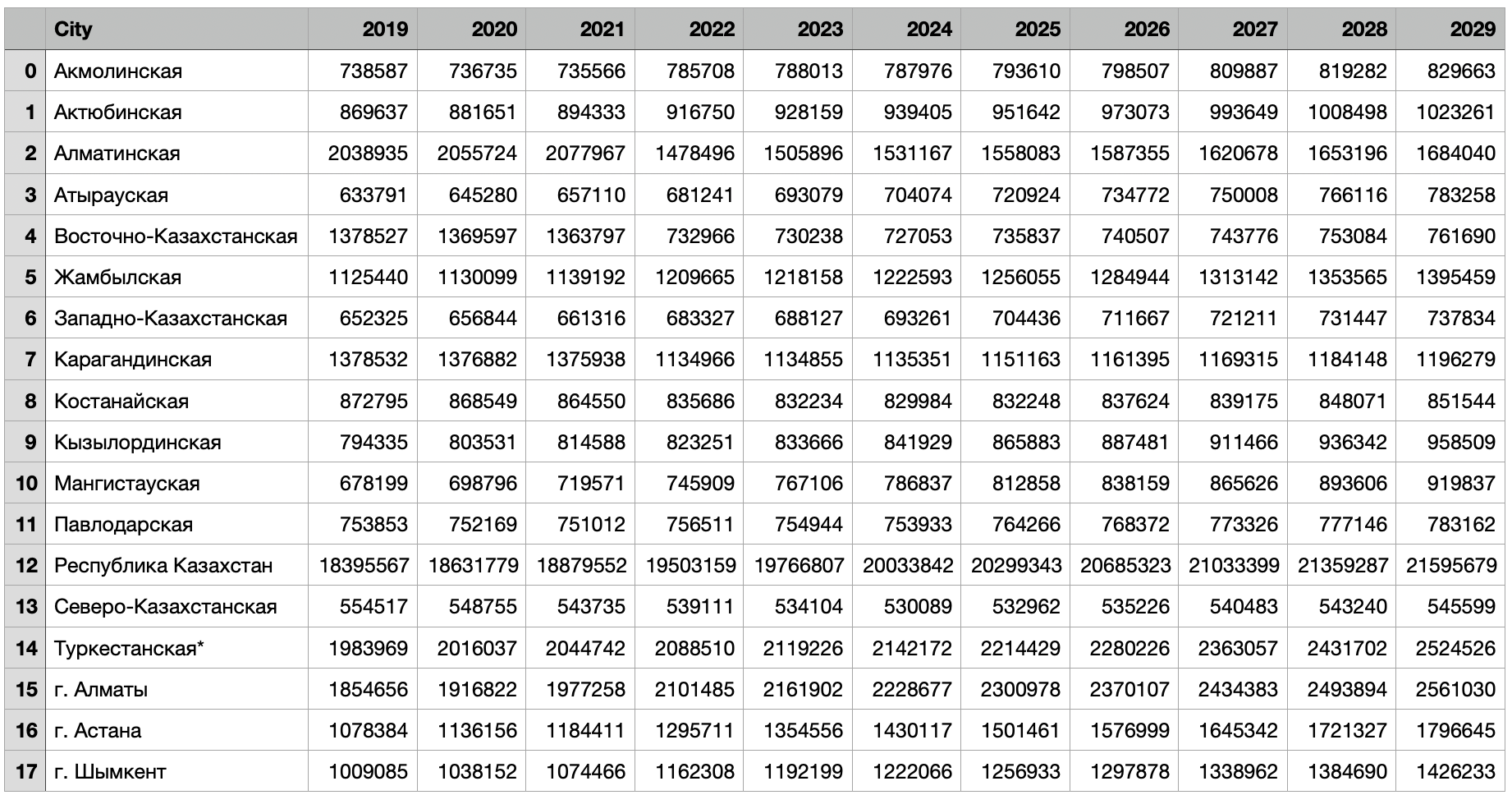
**Population Estimation**

We assumed that the investment should be considered based on the change of population in a specific region. For that purpose, the possible number of population has been estimated based on the statistics about Birth Rate, Death Rate and Migration. The equation for possible calculation:

Where,

Resultant table of population after estimate based on Linear Regression with random smoothing. We assumed that the trend will be the same based on resources from 2019 to 2024. Some of the regions have been removed due to lack of resources: Abai, Zhetisu, Ulytau



After the calculation of population, we have made another hypothesis:

*“Infrastructure should be developed based on the demand from citizens, which in case the total number of people for that particular region. All people need the same amount of resources despite their race, gender, age and social position.”*

*-DD Team*

Based on the hypothesis above, we have tried to implement the prediction models. First we identified the key sectors among all available 7 sectors. The main factors are: Energy, Economy, Transportation and Social

**Energy**

To identify the condition of energy, we have tried to calculate the total energy consumed by each region, the total number of people, the amount of stations and the capacity of each station and the price for each type of energy. The equation is following:

Where E - energy consumed by that particular year

N - population for that year

t - year and t + 1 - following year

C - capacity of each station

Where

By having those values, we can linearly project and estimate the future price for a unit of energy. By adding the inflation rate + interest of payback period of 5 years (until 2029). Each time for regular price of energy will be added another price to pay back any investments.

The same procedure has been done for all other source types, like water and electricity. So each investment can be returned by adding additional cost to the price per unit of energy source. However, due not availability of all resources and some materials not have stable price, the train of predictive model wasn’t possible.

**Transport**

For the sector of Transport, we have identified some major factors, like the number of available transportation amounts, the length of all roads, the number of population and the price for the transportation, roads of international and local road length.

Where T - number of available transports

N - number of people in that particular region

t - year and t + 1 - for following year

By the equation above, the possible amount of public transport is calculated. A similar equation was used to calculate the amount of potential number of passengers of public transport.

Where R - number of passengers

N - number of people in that particular region

After the application of each equations, the total investment is calculated for following changes:

Where P - price for public transport

Investment - planned investment to cover new demands

N - number of people in that particular region

First, the price for transportation is estimated without application of new investment by Linear Regression with random smoothing. After that new price estimation, we add another cost to payback all the investments done for that region. Again, due to lack of available information, we couldn’t calculate the possible investment needed in the region and the possible price, but still calculated the possible amount of passengers as well as needed bus amount for each particular region.

**Social**

Social factors include the amount of medical and education buildings. To estimate the need for future number of building, we used the following equations:

Where M - number of medical buildings in a particular region;

N - amount of people in a specific region;

By estimating the amount of needed hospitals, there is a probable need in the amount of hospitals. The same methods have been used for estimating the number of education buildings.

Where E - number of educational buildings in a particular region;

N - amount of people in a specific region;

By above calculations, we have tried to estimate the possible needs among each region to possibly assess the level of priority.

**Technical Solution**

We used some libraries to work with the dataset that we gathered, for visualization of tables and graphs and a numpy library to work with an array of information. The libraries are following:

**import pandas as pd**

**import matplotlib.pyplot as plt**

**import numpy as np**

We used datasets about Current Population, Birth Rate, Death Rate and Migration Rate

**population = pd.read\_excel(dir + 'population.xlsx')**

**death = pd.read\_excel(dir+ 'death.xlsx')**

**birth = pd.read\_excel(dir + 'birth.xlsx')**

**migration = pd.read\_excel(dir + 'migration.xlsx')**

We removed some of the cities due to lack of information

**cities\_to\_remove = ['Абай', 'Жетісу' , 'Ұлытау']**

**# Use np.isin() to filter out the cities to remove**

**cities = cities[~np.isin(cities, cities\_to\_remove)]**

**print(cities)**

We cleaned

**# Refine the data cleaning function to handle numbers with spaces and convert them to integers**

**def clean\_numeric(data):**

**# Remove any spaces and replace missing or placeholder values ("-") with 0, then convert to integer**

**return data.apply(lambda x: x.astype(str).str.replace(" ", "").str.replace("-", "0").str.replace(",", "").astype(int))**

**population = clean\_numeric(population)**

**death = clean\_numeric(death)**

**birth = clean\_numeric(birth)**

**migration = clean\_numeric(migration)**

**population['City'] = cities**

**birth['City'] = cities**

**death['City'] = cities**

**migration['City'] = cities**

We then created a new population table to store the values for years from 2025 to 2029. To make a Linear Regression a bit realistic, we added some smoothing.

**# Prepare a dictionary to store projections for each city**

**np.random.seed(42)**

**years = [2025, 2026, 2027, 2028, 2029]**

**# Create an empty DataFrame to store the new projections**

**new\_populations = pd.DataFrame(columns=['City', 2025, 2026, 2027, 2028, 2029])**

**new\_populations['City'] = cities**

**new\_populations**

Then we started our calculations to estimate the future population

for city in cities:

# Extract data for the current city

population\_city = population[population['City'] == city].iloc[:, 1:-1].values.flatten()

birth\_city = birth[birth['City'] == city].iloc[:, 1:-1].values.flatten()

death\_city = death[death['City'] == city].iloc[:, 1:-1].values.flatten()

migration\_city = migration[migration['City'] == city].iloc[:, 1:-1].values.flatten()

# Calculate the average annual growth rate

annual\_growth = (birth\_city - death\_city + migration\_city) / population\_city[:-1]

avg\_growth\_rate = annual\_growth.mean()

# Start the projection with the last known population value

population\_projection = [population\_city[-1]]

# Project the population for the next five years

for year in years:

noise = np.random.uniform(-0.005, 0.005) # Add noise for smoothing

adjusted\_growth\_rate = avg\_growth\_rate + noise

next\_population = round(population\_projection[-1] \* (1 + adjusted\_growth\_rate))

population\_projection.append(next\_population)

new\_populations.loc[(new\_populations['City'] == city), year] = next\_population

new\_populations

The final results were concatenated with old values to make a new Dataset which is called ‘All Population’.

All other calculations were done by Excel, where we implemented the equations that have been described above. Then the calculations have been rounded to the nearest whole number after. All calculations were used to estimate the need based on the change in the population of each region.

**Conclusions**

Based on the change in the population, we have tried to estimate the need in the next 5 years. We haven’t been able to estimate the most priority, but we have attempted to estimate the need to make a choice. For example, in 2025 there will be a need for additional universities in Karaganda due to an increase in the population of that city. By those changes, we can estimate the needed amount of infrastructure in different parts of Kazakhstan. If there were additional days, more precise and efficient results would be obtained and reported.

**Limitations**

1. **Dataset Limitation:** some data is missing, irrelevant, incorrect. The more structured dataset creation takes a huge amount of time;
2. **Too optimistic**: the calculations were calculated based on the theory that the increase in each variable might be constant in the future. So our estimation is made for best case situations, so for some worst case scenarios, the calculations might be wrong.
3. **Overgeneralized**: the assumption that there is no difference in the need regarding the age, race, ethnicity and gender is not correct, however due to limited amount of time and resources, that assumption was the key to make analysis and calculations;
4. **Simple**: simple mathematical equations and Machine learning algorithms have been used due to limitation in computational resource, to have more complex models and calculations, more computational resources are needed;