SIS Project Report: Data Collection & Preparation

Internet Access and Higher Education in Central Asia

"Влияние доступа к интернету на развитие высшего образования в странах Центральной Азии"

The goal of the project is to analyze how the development of Internet infrastructure affects the level of higher education in Central Asian countries (Kazakhstan, Kyrgyzstan, Uzbekistan, Tajikistan, Turkmenistan). The main task is to combine data from two sources (API and Web Scraping), clean them up, analyze and visualize the relationship between the indicators.

Description of approaches

1. API: The work with the REST API (parameters, query structure, JSON format) has been studied. Used by the World Bank Open Data API (https://www.worldbank.org/ext/en/home) sources:
   - IT.NET.USER.ZS - Internet users (% of the population);
   - SE.TER.ENRR - higher education coverage (%);

   GET requests are implemented through the requests library, and data is processed using pandas.
2. Web Scraping: Using requests and BeautifulSoup, the Wikipedia page was parsed Central Asia (https://en.wikipedia.org/wiki/Central_Asia#).
   A list of Central Asian countries (Kazakhstan, Kyrgyzstan, Uzbekistan, Tajikistan, Turkmenistan) has been obtained.
3. Pandas: Creating a DataFrame table for each data source.
   Data cleaning:
   - omissions and duplicates are removed;
   - converted types (to_numeric, astype);
   - renamed columns (internet_users_pct, tertiary_enrollment_pct);

   Merging:
   - Data is combined by keys ["country", "year"].
   - Filtering by the range 2000-2025
4. Analysis: Numerical arrays and computations were handled using NumPy, and correlation, average, and descriptive statistics computations have been put into practice. For aggregation and transformation for analysis by nation and year, the groupby(), merge(), pivot(), and describe() functions were also employed.
5. Data visualization. Three types of graphs are constructed using Matplotlib:
   - Internet access growth lines by country;
   - Higher education coverage lines;
   - Scatter plot - the connection between the Internet and education;

   Axis signatures, legends, and headings are configured for graph readability.

Results from the analysis

General statistics.

- The average share of Internet users is 28% of the population.
- The average level of higher education is 31%.
- The period of active growth is after 2010.

	internet_users_pct	tertiary_enrollment_pct
count	113.00000	97.000000
mean	27.60153	31.250732
std	29.92464	16.579262
min	0.04860	7.000936
25%	2.99927	15.972096
50%	15.70000	31.701936
75%	50.60000	45.759177
max	93.39170	62.292137

Correlation Analysis (Pearson). The average correlation is 0.8, which shows a strong positive relationship between the development of the Internet and the coverage of higher education.

country		corr_internet_tertiary
Kazakhstan	internet_users_pct	0.635009
Kyrgyzstan	internet_users_pct	0.662095
Tajikistan	internet_users_pct	0.944175
Turkmenistan	internet_users_pct	NaN
Uzbekistan	internet_users_pct	0.600116

Key observations:

1. Kazakhstan and Uzbekistan are leaders in terms of digitalization and education growth rates.
2. Tajikistan and Turkmenistan are lagging behind, but they are showing gradual growth.
3. Rapid digitalization increases the involvement of the population in education and improves access to educational resources.

Conclusion

This project helped us understand the entire process of working with real data, from collecting it using API and web analysis to cleaning, analyzing, and visualizing the results. We have learned how to get information from different open sources, find meaningful relationships and visually show them using diagrams. The analysis confirmed that in Central Asia, improved access to the Internet goes hand in hand with an increase in the level of education. In general, we have acquired practical skills in working with data, statistical analysis and visualization, turning open data into real information about the development of digital technologies and education in the region.