Population of Kazakhstan

Made: Olzhas Sanatbek, Dauren Rakhmetolla, Adilkhan Zamanbek

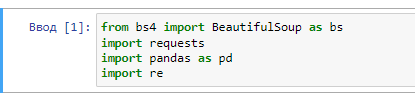
# Introduction

We analyzed the population of Kazakhstan, and showed this data in plot, scatter plot, histogram. Also we make a machine learning model for prediction for the population.

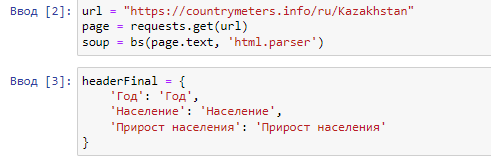
# I. Data Collection

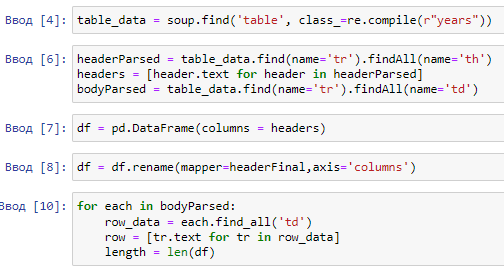
The Country Meters resource was chosen as the data source. We wanted to know the dynamics of population growth in our country, what was the increase in percentage terms, how the situation changed depending on time, era etc.

Every work starts from importing libraries.



After that I added constant like ‘url’, and method soup. Added names to final headers.

  
This resource had table data, so I wrote it in my code, found the element named “table” in html code. For headers I chose ‘th’ in ‘tr’ tag, and the main data was in ‘td’ tag.



As a result I had the following data and saved it in csv file ‘populkz.csv’



# II. Data Visualization (5pts)

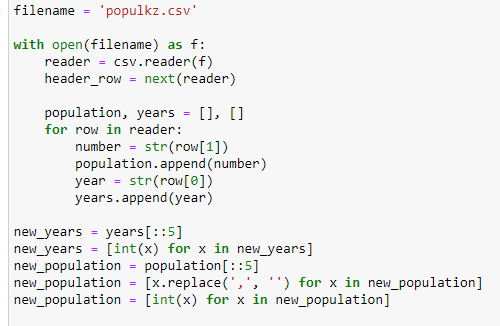
◦ Visualize the data statistics: histograms, plots, scatterplots, diagrams. ◦ Use matplotlib or seaborn

Data visualization refers to the exploration of data through its visual representation. Visualization is closely related to data mining, which uses code to explore patterns and relationships in a data set. A dataset can be either a small list of numbers that fits in one line of code, or an array of many gigabytes. High-quality presentation of data is not limited to a beautiful picture. When a simple, visually appealing representation is chosen for a data set, its meaning becomes clear to the viewer. People notice patterns in a dataset that they weren't aware of. Fortunately, you don't need a supercomputer to visualize complex data. Thanks to the efficiency of Python, you will be able to quickly explore datasets of millions of individual data items (data points) on a simple laptop.

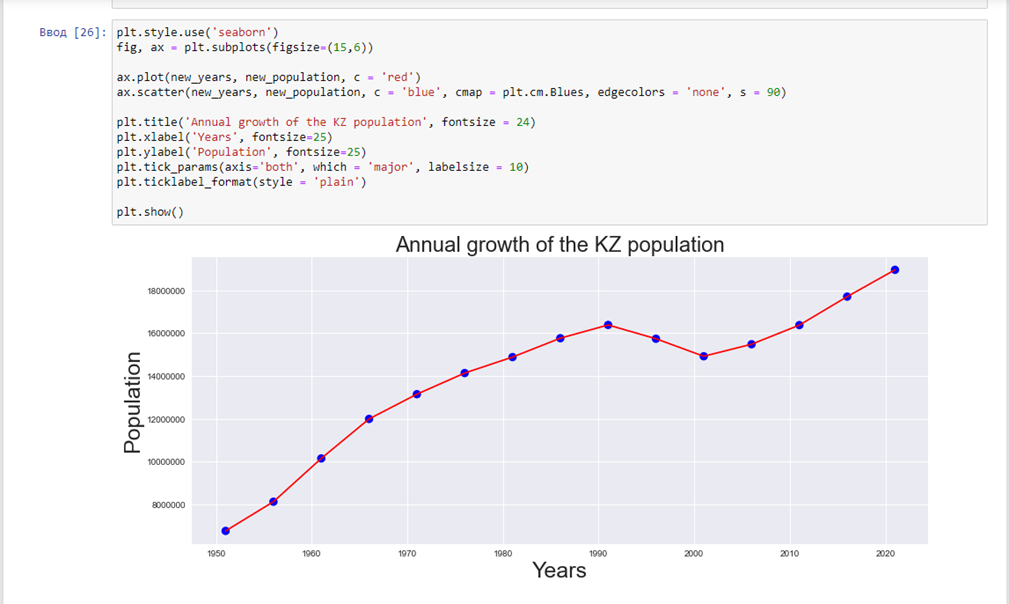
Since our group chose the topic Population of Kazakhstan, we will visualize the data taken from the site<https://countrymeters.info/en/Kazakhstan>.

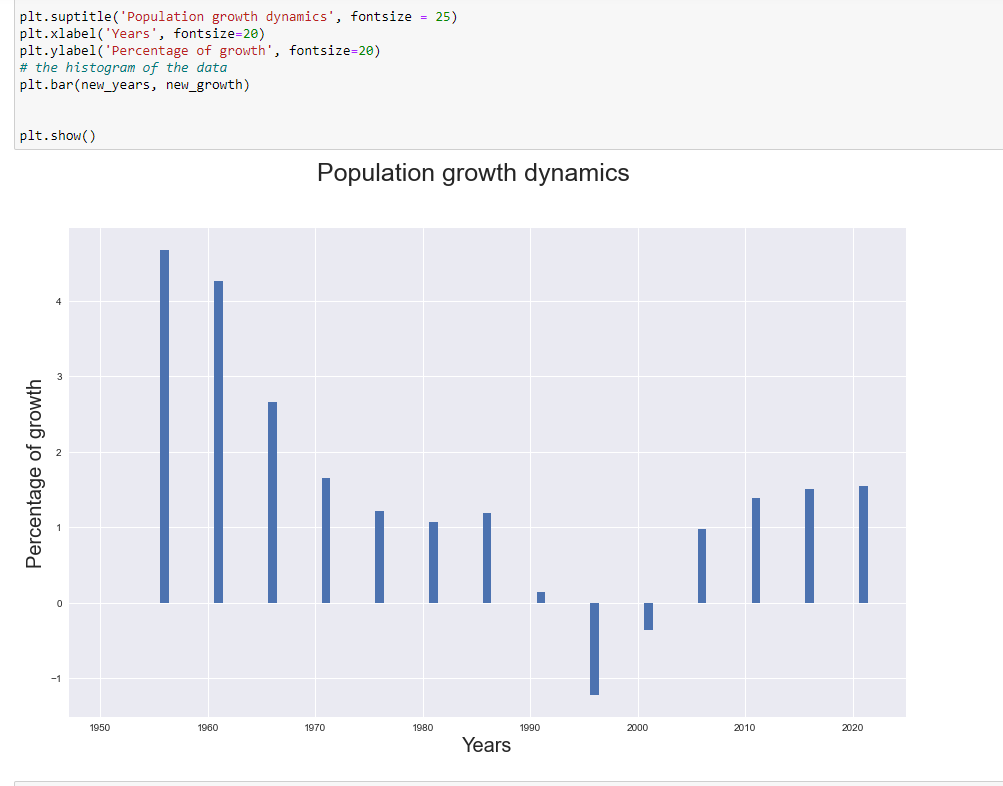
First of all, I would like to build a simple graph where the user can see data that describes the ratio of the population of Kazakhstan to years, that is, in what year how many people lived in our country. The main library which we are going to use is *matplotlib.*

Via the module *csv* we can easily read the data which we parsed earlier. Then we will sort the data by the columns in lists. Since there are too much information to visualize, we need to cut some data. That is why we take every 5th element of the lists.



Then some magic from Python and matplotlib library, and graph is done.



Almost the same we used to do for other types of graphs. Only some specific differences.

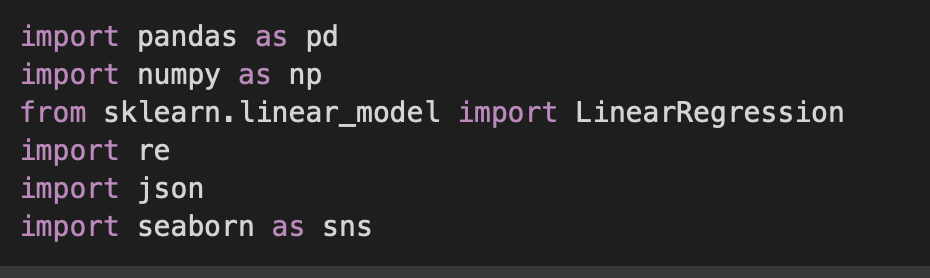
# III.Data Analytics

We make a population prediction. To realize this we used a linear regression model.

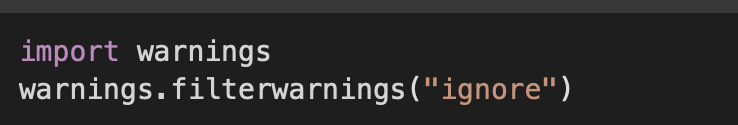
Goals:

* To train the model
* Make the prediction
* Show these in scatter plot

The following libraries we used:

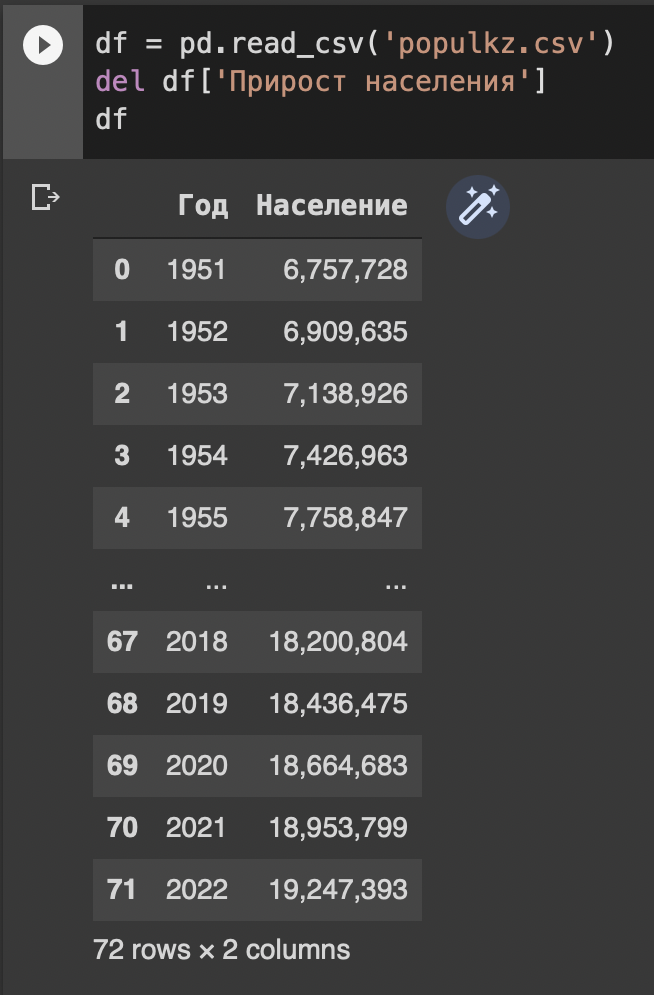


To ignore warnings in code output, we used warnings module

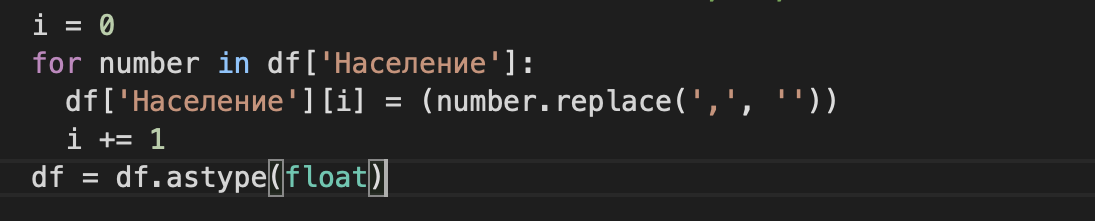


### Read data and cleaning

After we read the populkz.csv file, we used pandas.read\_csv() and deleted the 3rd column.

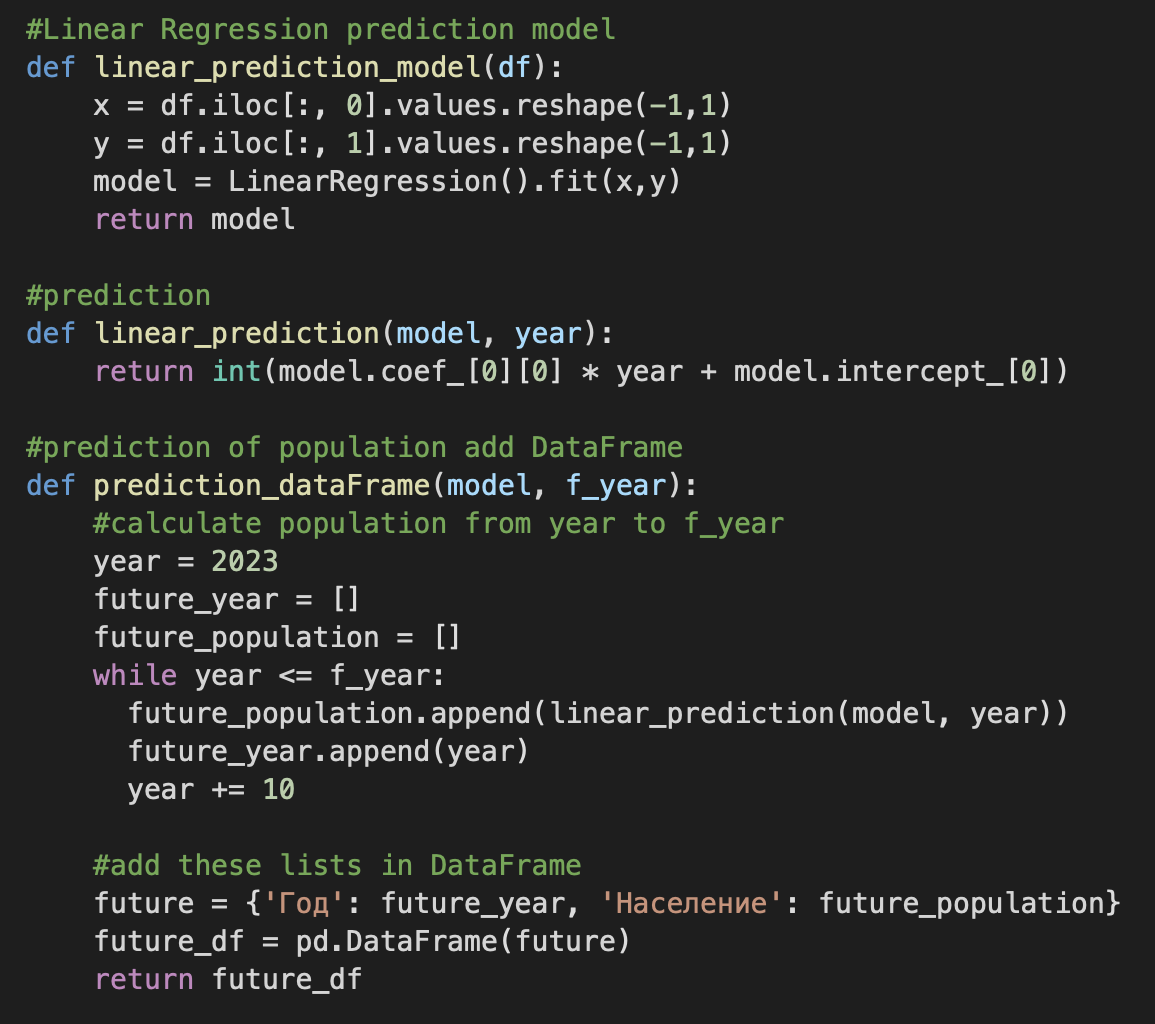


But in this data frame in population column not read like numbers, to solve this problem we delete “,” in numbers and convert to float.



### Make prediction of population

For prediction we used the model LinearRegression() from sklearn library. To train this model we used this dataframe.



This linear\_prediction function simply takes the coefficient and intercept from the model and uses ‘y=mx+c’ to derive the predicted population. And the prediction\_dataFrame function calculates the previous year's predictions and adds the data frame.

All these added the main function.

