Project work

Task:

The main task to analyze and make conclusions about dataset related to Kazakhstan.

Go to stats.gov.kz website to get a data.

Choose the sphere that you want to work with(example: education, medicine, transportation, etc.) Download the dataset Do main Statistical Analysis by calculating:

mean, median,mode quartiles, interquartile range variance and standard deviation

Find Correlation. Which columns could be related to each other? Try to find correlation between columns and make your conclusion.

Apply Linear Regression model. Choose two columns to analyze(choose those which have remarkable correlation). Find regression line.

Choose the platform to make analysis: Excel, Python, Java, C++.

If you want to use another platform, then come to discuss it with me. Do main Statistical Analysis by finding: mean, median,mode quartiles, interquartile range variance and standard deviation

Do Data Visualization. Here choose columns to make analysis. Plot histograms, pie chart, line chart, boxplot, scatter plot.

Find Correlation. Which columns could be related to each other? Try to find correlation between columns and make your conclusion. Apply Linear Regression model. Choose two columns to analyze(choose those which have remarkable correlation).

Find regression line. Predict the value for the future or other input.

Make a Report on your Analysis. To get high points you can use Latex platform to write your report. (Online Latex – Overleaf)

I have chosen employment of the population of Kazakhstan:

<https://stat.gov.kz/api/iblock/element/5841/file/kk/>

Part One

The first 10 records of the dataset (there are more columns in the dataset, truncated here):

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Agriculture, forestry and fisheries | mining and quarrying | manufacturing industry | supply of electricity, gas, steam and air conditioning | water supply; activities for the collection, processing and elimination of waste, elimination of pollution | Construction | Wholesale and retail sales; repair of cars and motorcycles | Transport & Assembly | Accommodation & Nutrition Services | Information and communication | Agriculture, forestry and fisheries | mining and quarrying |
| 2010.I | 2241,8 | 198,6 | 539,0 | 149,4 | 57,0 | 566,3 | 1212,0 | 505,4 | 112,2 | 130,3 | 2241,8 | 198,6 |
| 2010.2 | 2300,9 | 193,0 | 526,4 | 147,0 | 61,4 | 547,4 | 1264,0 | 517,1 | 121,2 | 114,1 | 2300,9 | 193,0 |
| 2010.3 | 2304,9 | 192,3 | 550,5 | 143,2 | 58,0 | 594,8 | 1243,6 | 512,2 | 106,2 | 109,2 | 2304,9 | 192,3 |
| 2010.4 | 2257,8 | 192,5 | 549,7 | 146,6 | 58,3 | 567,8 | 1233,0 | 517,1 | 106,4 | 111,1 | 2257,8 | 192,5 |
| 2011.1 | 2254,3 | 194,1 | 546,9 | 142,4 | 59,0 | 570,6 | 1201,9 | 511,7 | 104,1 | 110,1 | 2254,3 | 194,1 |
| 2011.2 | 2264,1 | 207,3 | 538,0 | 145,1 | 59,5 | 578,8 | 1217,0 | 523,0 | 110,7 | 113,8 | 2264,1 | 207,3 |
| 2011.3 | 2243,6 | 216,4 | 529,4 | 155,9 | 73,4 | 636,2 | 1157,7 | 565,4 | 139,2 | 141,7 | 2243,6 | 216,4 |
| 2011.4 | 2185,6 | 223,3 | 546,8 | 155,5 | 74,8 | 651,0 | 1182,1 | 550,3 | 136,1 | 143,8 | 2185,6 | 223,3 |

Main statistical analysis:

mean, median, mode:

As a data template for calculations, let's take the calculation of statistics for the industry

*water supply; activities for the collection, processing and elimination of waste, elimination of pollution* in two years and for the first 10 regions, the first quarter of 2010.

Let's present the basic formulas:

Calculate the average for a specific industry over a period

Calculation of the average for the first quarter of 2010 for 10 regions

The median is calculated by sorting the array and selectingthe middle element, if the size of the series is odd. If even, then we take the average between the two middle elements.

Mean

Let's calculate the mean:

Mean for the entire period (for further calculations)

Median

Find the median of the fifth industry over eight quarters:

First, let's sort the array

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 57 | 58 | 58.3 | 59 | 59.5 | 61.4 | 73.4 | 74.8 |

In this case, the length is 8, so the average for the series will be:

Mode

There is no mode for the fifth industry, as all meanings are unique. Let's calculate by extending the interval to the entire interval in the dataset:

57.0, 58.0, 58.33, 59.0, 59.543, 61.4, 70.727, 73.081, 73.39, 74.406, 74.574, 74.8, 75.578, 76.026, 76.962, 77.22, 77.412, 77.6, 77.89, 78.0, 78.085, 78.451, 78.629, 78.683, 79.05, 79.5, 79.516, 79.758, 79.779, 79.938, 80.294, 80.399, 81.757, 81.762, 82.026, 82.396, 82.473, 82.552, 83.118, 83.24, 83.437, 83.458, 83.946, 83.946, 84.043, 84.141, 84.24, 84.246, 84.761, 84.97, 86.197, 87.131

For this series, there are 1 values with the highest frequency equal to two:

Mode 83.946: 2 occurrences

Quartiles

Let's calculate the quartiles for the entire interval of the fifth industry:

1. Sort the dataset in ascending or descending order
2. Determine the median (2nd quartile)
3. Determine the medians of the two halves (1st and 3rd quartile)

First, let's sort the row:

57.0, 58.0, 58.33, 59.0, 59.543, 61.4, 70.727, 73.081, 73.39, 74.406, 74.574, 74.8, 75.578, 76.026, 76.962, 77.22, 77.412, 77.6, 77.89, 78.0, 78.085, 78.451, 78.629, 78.683, 79.05, 79.5, 79.516, 79.758, 79.779, 79.938, 80.294, 80.399, 81.757, 81.762, 82.026, 82.396, 82.473, 82.552, 83.118, 83.24, 83.437, 83.458, 83.946, 83.946, 84.043, 84.141, 84.24, 84.246, 84.761, 84.97, 86.197, 87.131

Let's find the second quartile.

Q2 is the median of the entire dataset, which we calculated earlier to be 79. 5.

Let's find the first quartile

Q1 is the median of the lower half of the dataset.

And the middle will be equal to

Let's find the third quartile

Q3 is the median of the upper half of the dataset. Since the dataset has an odd number of values (57), we exclude the median itself from the upper half.

And the middle will be 83.2

Therefore, the quartiles for the given dataset are:

Interquartile

Now let's find the interquartile:

By definition, it is equal to the difference between the third and first quartile:

Variance

Let's calculate the variancefor the fifth industry:

Standard deviation

For our case:

Correlations

Two columns correlate well if they are linearly dependent or close to it.

The correlation is described by the correlation coefficient. If it is close modulo to one, then the correlation is strong, if it is close to zero, then it is not.

In this dataset, the following features correlate well with each other:

manufacturing industry and Wholesale and retail sales; repair of cars and motorcycles:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 539.0 | 526.4 | 550.5 | 549.726 | 546.9 | 537.982 | 529.367 | 546.8 | … |
| 2 | 1212.0 | 1264.0 | 1243.6 | 1232.977 | 1201.9 | 1217.006 | 1157.665 | 1182.1 | … |

Let's calculate the correlation coefficient:

Manufacturing and wholesale/retail sales, including automobile and motorcycle repairs, can have a significant correlation in employment rates. Here's why:

Supply chain: The manufacturing industry produces goods that are distributed and sold through wholesale and retail channels. Manufacturing companies rely on wholesalers and retailers to distribute their products to end consumers. This creates a demand for jobs in both sectors, as manufacturing requires workers for production, and wholesale and retail trade require workers for distribution, sales, and customer service.

There is also a job dependency: the manufacturingindustry relies on wholesalers and retailers to ensure they get their products to market efficiently. Wholesalers purchase goods from manufacturers in bulk and sell them to retailers, who then sell them to individual consumers. This interdependence creates a chain of jobs in the manufacturing, wholesale and retail sectors, creating employment opportunities in a variety of roles, such as manufacturing workers, supply chain managers, sales representatives, logistics specialists, and retail personnel.

We get the linear regression equation of the second column to the first:

We have already calculated the correlation coefficient, which can be used to represent the linear dependence as follows:

Part 2

Let's confirmthe calculations given above using PYTHON:

Load the dataset and select a piece of 8 by 10 from it:

df = pd.read\_excel('/content/dataset.xlsx')

df\_short = df[list(df.columns[0:11])].loc[0:7]

Next, select the tested rows for calculating statistics:

col\_sphere5 = df\_short['water supply; activities for the collection, processing and elimination of waste, elimination of pollution']

row\_time1 = df\_short.iloc[0][list(df\_short.columns)[1:]]

We calculate the statistics for the 5th domain and the first quarter, as in the first part:

Mean:

print(np.mean(col\_sphere5))

print(np.mean(row\_time1))

>>>62.682875

>>>571.2

The average coincided.

Median:

print(np.median(col\_sphere5))

>>>59.2715

Mode:

col\_sphere5\_all = df['water supply; activities for the collection, processing and elimination of waste, elimination of pollution']

col\_sphere5\_all.value\_counts().head()

>>>83.946    2

>>>57.000    1

>>>76.026    1

>>>75.578    1

>>>74.574    1

The mode is 83.946.

Quartiles:

Let's define quartiles using quantile values with given boundary values of the probability of falling into the gap to the left of this value. In our case, we set 0.25, 0.5, 0.75

print(np.quantile(col\_sphere5\_all,0.25))

print(np.quantile(col\_sphere5\_all,0.5))

print(np.quantile(col\_sphere5\_all,0.75))

>>>75.914

>>>79.50800000000001

>>>83.1485

interquartile range:

IQR = np.quantile(col\_sphere5\_all,0.75) - np.quantile(col\_sphere5\_all,0.25)

print(IQR)

>>>7.234499999999997

Variance and STD

print(np.var(col\_sphere5\_all))

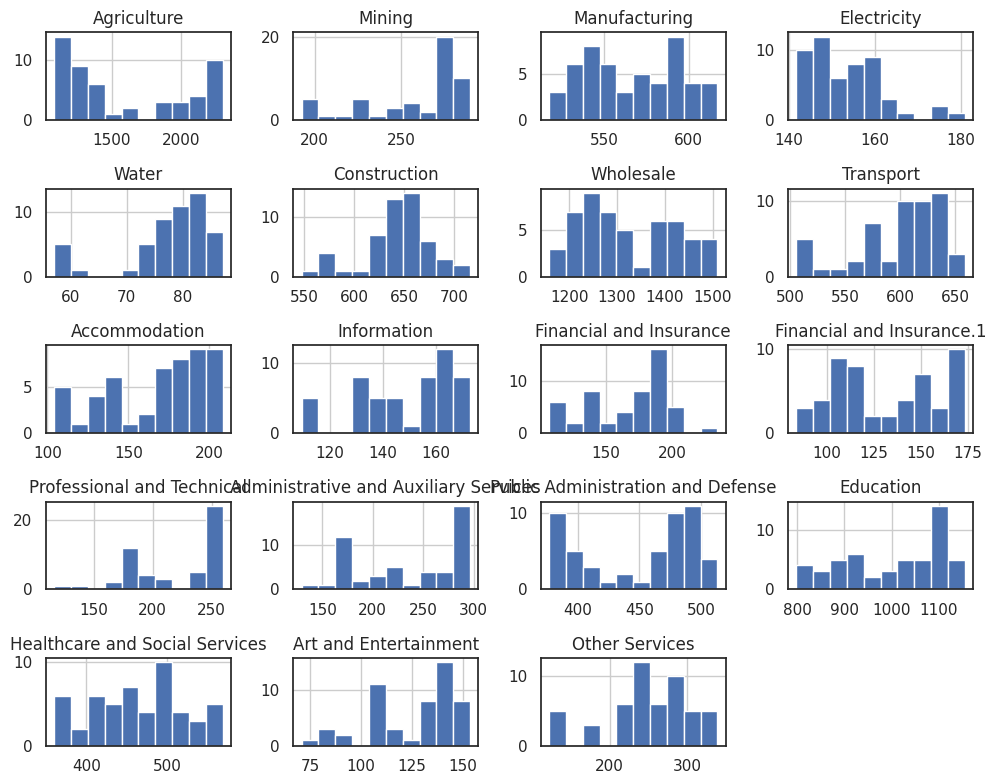
print(np.std(col\_sphere5\_all))

>>>58.77544970821006

>>>7.666514834539229

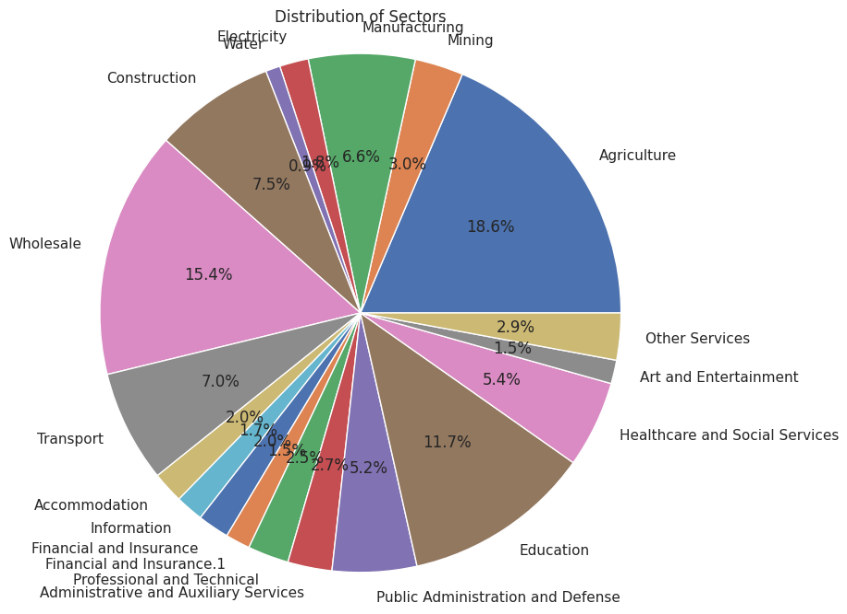
Data visualization

First, let's build histograms.

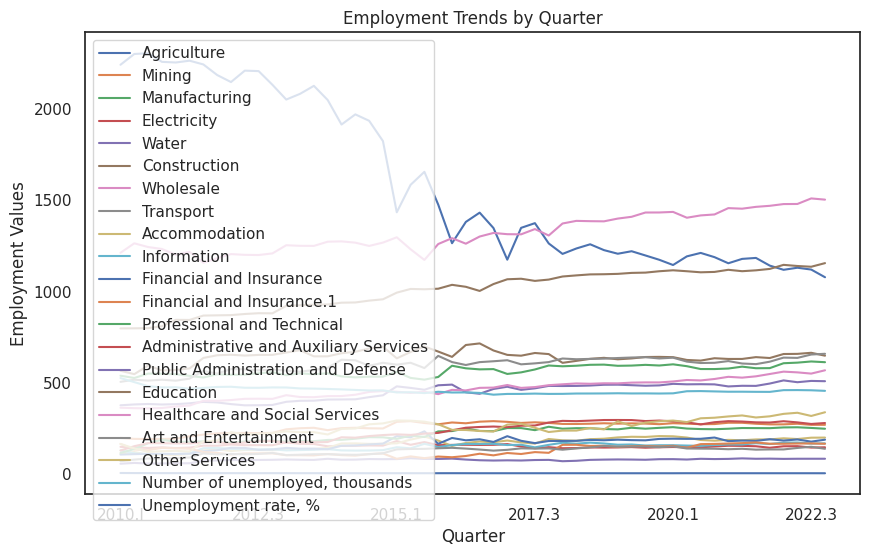


Signs construction, water suply have a distribution close to normal. Some features have a bimodal distribution: agriculture, manufacturing, wholesale, accomadation. And bimodality indicates a change in the demand for workers from one stable value to another

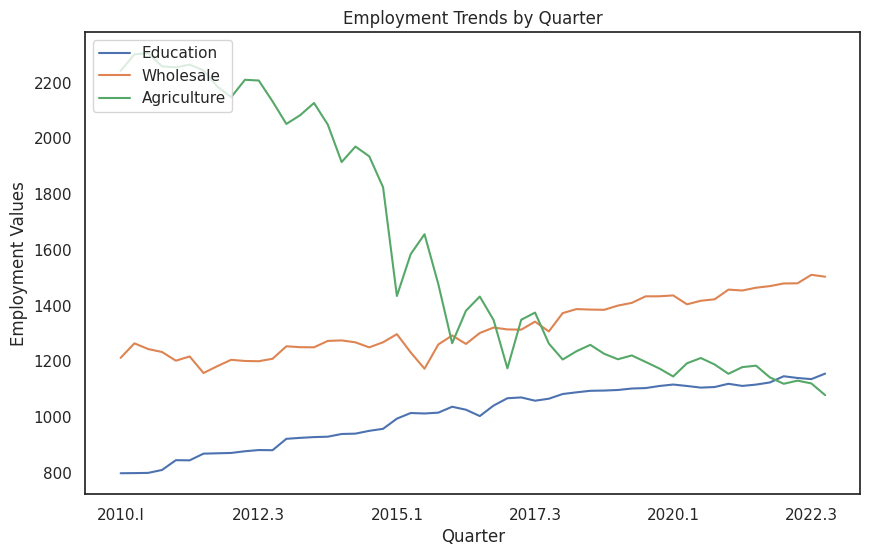
Next we can plot a pie chart



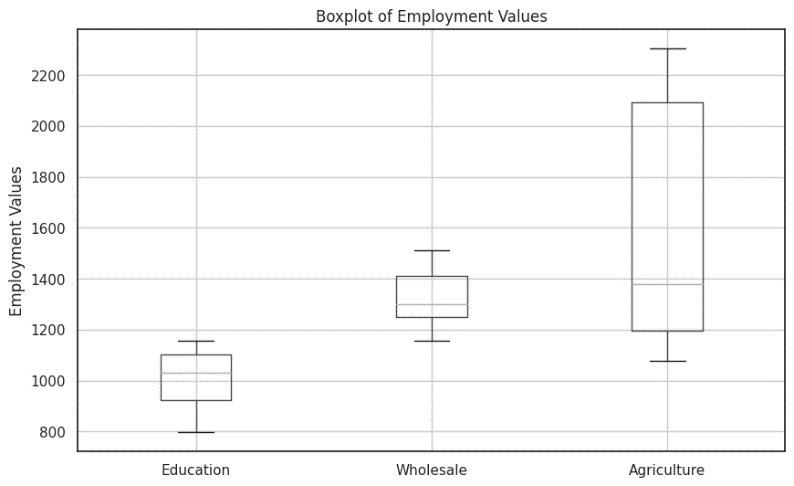
High demand for working staff in threefold coverage: Wholescale, agriculture, education.



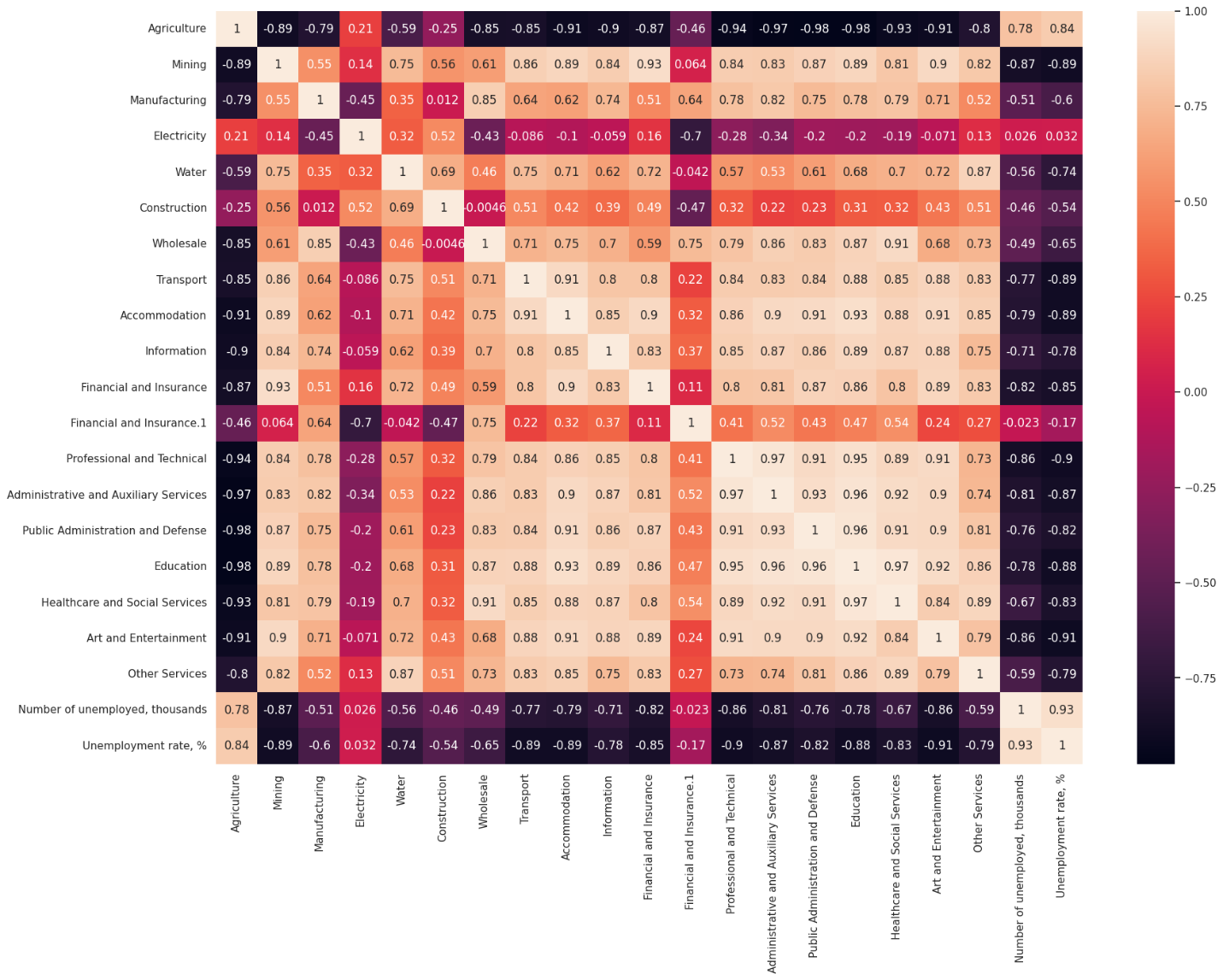
Here you can see a trend of declining demand for workers in the field of agriculture and a gradual increase in demand for workers in the field of education and wholesale trade.



Here you can clearly see what a big decline in agriculture and good growth in education and wholesale trade.



Correlation

Let's build a correlation matrix.

Here we can see the relationship between wholesale trade and manufacturing, as we have noted. Also good columns that correlates is agriculture, with almost all the columns. Because when agriculture fell, other signs, on the contrary, grew.

Also well positively correlated with many features of the technical area and administration. This happens because with an increase in demand for workers, an increase in demand for administration and technical support of products for analyzing and systematizing work is growing.

To find correlation between wholesale and manufacturing we can use this code:

df['Wholesale'].corr(df['Manufacturing'])

>>>0.8501917870668514

The value coincided with the calculated

Linear regression

Let's find the regression line using the numpy library:

from sklearn.linear\_model import LinearRegression

X = np.array(df['Manufacturing'])

Y = np.array(df['Wholesale'])

coefficients = np.polyfit(X, Y, 1)

# Extract the slope and intercept from the coefficients

slope = coefficients[0]

intercept = coefficients[1]

# Print the coefficients

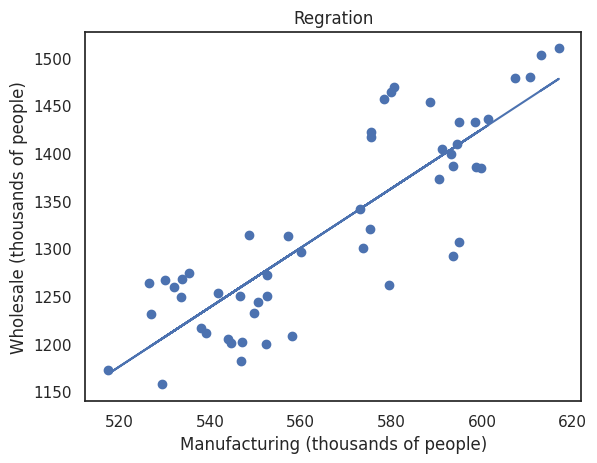
print("Slope:", slope)

print("Intercept:", intercept)

>>>Slope: 3.109589643721574

>>>Intercept: -440.4183119694752

Values, I coincided with the calculated



Let's also train a linear regression model to predict values:

model = LinearRegression()

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X,Y,test\_size=1/3,random\_state=42)

model.fit(X\_train.reshape(-1,1),Y\_train)

Ypred\_train = model.predict(X\_train.reshape(-1,1))

Ypred\_test = model.predict(X\_test.reshape(-1,1))

error\_train = mean\_squared\_error(Y\_train,Ypred\_train)

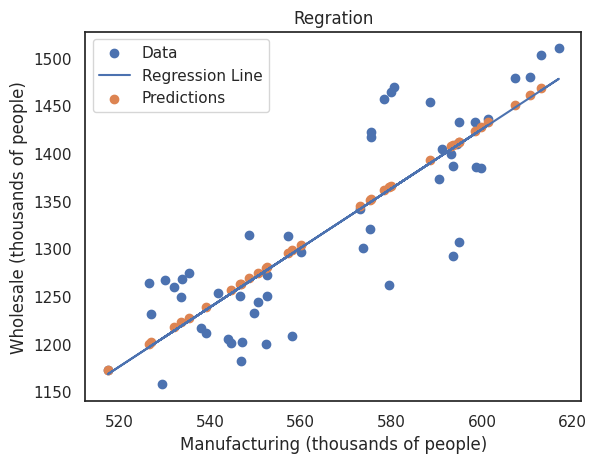
error\_test = mean\_squared\_error(Y\_test,Ypred\_test)

print(error\_train)

print(error\_test)

>>>2722.5644694769762

>>>2825.9665620733617



It can be seen that the predicted values fell on the regression line, which shows a good predictive function.

*Conclusion*

The analysis focused on employment in Kazakhstan using a set of data obtained from the official statistics website. The analysis included statistical analysis, data visualization, correlation analysis, and linear regression modeling.

The calculated statistical indicators give an idea of the central trend, variability and distribution of employment in various industries.

Several data visualization methods have been used. They helped to understand the distribution, trends and relationships in the employment dataset.

Correlation analysis was done to identify relationships between different columns in the dataset. It was noted that manufacturing and wholesale/retail sales showed a good correlation in employment rates. This correlation can be explained by the interdependence between the manufacturing, distribution and retail sectors in the supply chain.

A linear regression model was built using static methods and machine learning. The model that was used to analyze the correlation between manufacturing and wholesale/retail sales. The regression line is obtained and its equation is calculated.