Industrial Engineering and Innovation Management (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/ieim.2023.060310
ISSN 2522-6924 Vol. 6 Num. 3

# *Analysis of Lanzhou Beef Noodle Industry Based on Linear Regression*

**Shuaihang Zhou[1], Xiangzhen He[1,*], Fucheng Wan[2], Dongchang Liu[3], Yuguang Wang[1]**

*[1]Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu, 730000, China*
*[2]Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Northwest Minzu University, Lanzhou, Gansu, 730000, China*
*[3]Yilan Big Data Technology Co., Ltd, Lanzhou, Gansu, 730000, China*
*[*]Corresponding author*

***Keywords:*** Linear Regression, Lanzhou Beef Noodles, Web Crawler, Data Preprocessing, Visualization Analysis

***Abstract:*** Lanzhou beef noodle is well known, but there is no data scientific analysis of Lanzhou beef noodle related industry research. The linear regression model has significant statistical significance and is widely used in management and economics. As a basic and widely used machine learning algorithm, linear regression plays an important role in exploring the relationship between data in different dimensions. The training of the model usually depends on a large amount of data, which can be obtained through the web crawler. This paper collected data of Lanzhou beef noodle shops through web crawler, cleaned beef noodle information and normalized data after data preprocessing, and analyzed the relationship between the number of shops and the number of population and the number of shops and the number of population density in each province by using linear regression algorithm. Finally, the descriptive statistical results and regression analysis results of "the first side of China" -- Lanzhou beef noodle industry development in the country are visualized through data visualization technology. It provides countermeasures and references for the innovation and development of Lanzhou beef noodle industry employees under the background of big data.

## 1. The Introduction

Data mining refers to the use of reasonable, appropriate and feasible methods to analyze the data to be analyzed, extract the information behind it, and form a conclusion when analyzing a large number of data studies. This is a process of data research and summary. With the development and iteration of information technology, the ability of enterprises to produce, collect, store and process data has made a qualitative leap, and the daily data throughput has reached an astonishing level. Therefore, the data analysis method is used to refine the complex data, study the development law of the data and predict the trend, and then help the management of the enterprise to make decisions. The general process of data mining includes data acquisition, data preprocessing, data storage, data analysis and data visualization. Correlation analysis is a statistical method to study the correlation

between random variables by studying whether there is a certain dependency relationship between phenomena and exploring the correlation direction and degree of the dependency phenomenon. Correlation generally exists in Lanzhou beef noodle industry. There is a mutual correlation between the number of stores, the number of population and the population density.

Since its establishment in 1915, Lanzhou beef noodle has enjoyed a worldwide reputation for its unique Northwest style of "the soup is clear for the mirror, the meat is rotten for the fragrance, and the noodles are thin for the long". It is known as "the first side of China" and has the reputation of "the name card of Lanzhou". In 2007, Lanzhou municipal government announced the first batch of intangible cultural heritage protection list of Lanzhou, and the world-famous beef noodle making skills of Lanzhou clear soup were included in it[1]. This is an industry that cannot be ignored, and the beef noodle industry plays an increasingly significant role in improving the image of Lanzhou and promoting its economy. In this paper, a complete data mining process is established, from data collection, data preprocessing, data analysis and data visualization, and finally through descriptive statistical results and regression analysis, all dimensions of Lanzhou beef noodle industry are displayed.

## 2. Research Methods and Data Sources

### 2.1. Research Methods

In the era of big data, data presents complex and huge characteristics, and it is very difficult to obtain valuable information. Data mining is a process of using massive data to explore and find meaningful relationships, rules or trends. It integrates theories and technologies from multiple fields such as artificial intelligence, machine learning, statistics and data visualization. It is the core step of knowledge discovery from massive data[2]. This study is divided into four stages: data acquisition and storage stage, data preprocessing stage, data analysis stage and data visualization stage. The specific process is shown in Figure 1:

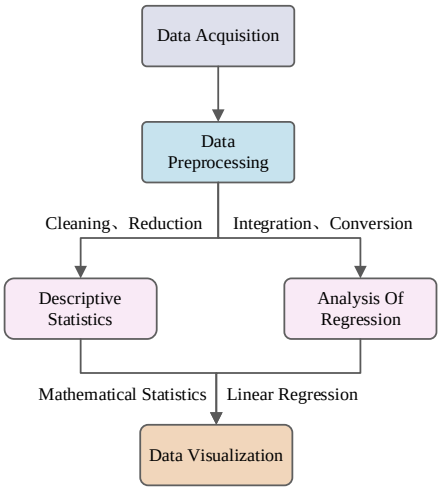


Figure 1: Data mining flow chart

1) Data acquisition and storage stage: crawler and database-related technologies are used in this stage. The web crawler captures the information of the map website, and after collecting the required data, writes it into the database according to the specific data format.

2) Data preprocessing stage: batch operation processing of data, including outlier processing, data cleaning, missing data replenishment, etc. The Python tools Pandas, Numpy, and Matplotlib provide powerful matrix operations. After the pre-processing stage, the final effective total data set will be

generated[3].

3) Data analysis stage: Descriptive statistics and regression analysis are mainly carried out on the final data set. Mathematical statistics were carried out on each dimension of Lanzhou beef noodles, and the linear regression method of machine learning was adopted to explore the relationship between the number of stores in each province and the population density.

4) Data visualization: Using tools such as Matplotlib for drawing display of the processed data.

## 2.2. Data Sources

In the era of big data, Internet data has become an important data source for industrial research. In recent years, online map service providers such as Baidu Map and Amap, as well as online lifestyle service platforms such as Meituan and EleMA have integrated a large amount of basic store data. Considering the strong real-time and coverage of the online map, this paper mainly selects Baidu Map, one of the largest online map service platforms in China, as the data source of Lanzhou beef noodles, and obtains the names of cities and streets across the country with the help of third-party websites. Use "the street name + keyword" to access the map website to obtain data including the name of the beef noodle restaurant, detailed address, score and other attribute information. In addition, after preprocessing the above data such as store location correction and error troubleshooting, the national database of Lanzhou beef noodle Restaurant was established as the support of the analysis data.

### 2.2.1. Obtaining Data

Data is obtained by Web Crawler. The basic process of web crawler data capture is as follows: first, the web page is analyzed by sending a request through a simulated browser; then get the content, the page source code for parsing; finally, the analysis is stored in the database. The Requests library is mainly used when using Python to simulate the browser to initiate requests for web pages, and the request methods are divided into GET and POST. When the request succeeds, the response content can be obtained in the form of text, binary, and JSON. At this time, the response content obtained is semi-structured data, which needs to be parsed by Python's parsing library to obtain the required data content. The web page acquisition process is shown in Figure 2:

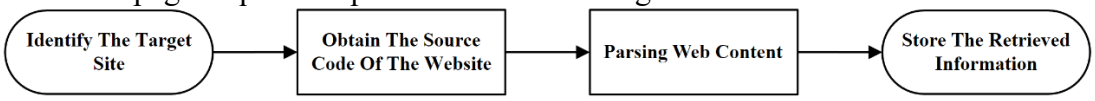


Figure 2: Web page acquisition process

The data collection of Lanzhou beef noodle shops is divided into two parts: one is to obtain the information of national roads; the second is to obtain the information of Lanzhou beef noodle shops according to the road information. The collected original data also needs data preprocessing, which mainly includes data cleaning, data integration, data conversion, data reduction and other processing methods.

### 2.2.2. Data Preprocessing

Due to the use of Baidu map search using fuzzy search, so there will be a large number of non-conforming data. Data preprocessing is to obtain relatively complete, clean and consistent data through appropriate methods without losing the original data information of the collected data, which makes the data have a higher value density and further improve the data processing efficiency. Data preprocessing can reduce the cost and difficulty of data mining, obtain the information in the data faster and more effectively through data preprocessing, improve the quality of data so that it can meet

the requirements of mining algorithms for data standards, reduce the complexity of data retrieval in subsequent mining operations, and improve the efficiency of mining operation execution. Data preprocessing mainly includes: data cleaning, data integration, data conversion and data reduction[4]. As shown in the Figure 3 below:

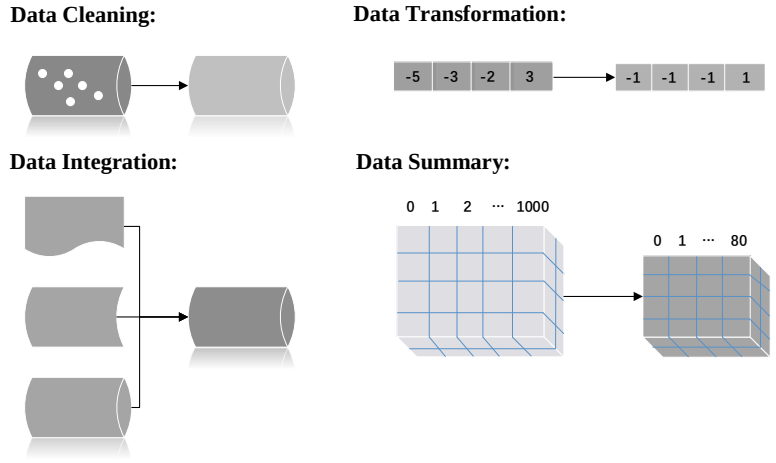


Figure 3: Data preprocessing method

Data cleaning: clearing irrelevant data and smoothing noise data in the original data set, screening irrelevant data and processing outliers[5], and standardizing "dirty data" through such processing.

Data integration: Data from different databases are integrated into the same database according to the same standard for storage. The attribute structure in the database is used to infer and identify the entity definition in real life to solve the ambiguity of data types, labels and concepts[6].

Data transformation: Used for the transformation of attribute values, such as absolute value, function and other methods can be used to transform the attribute values, dimensional transformation or transformation can be used to reduce the number of effective attribute values or find the invariant of data, including normalization, switching and projection operations, to convert the data into a form suitable for mining.

Data reduction: Scale a large amount of data to a certain extent while maintaining integrity and useful features, reduce the attribute dimension of the data itself, reduce the data capacity, so that the data can be more accurate in the later data modeling, establish an effective data model, and minimize the amount of data to save a lot of time for data mining[7].

(1) Main process of data preprocessing in beef noodle stores

Import the related libraries, read the file from the MySQL database, and save the CSV file named "whole_data.csv" to process the file. Pandas is used to read the file, and the keywords "Lanzhou beef noodles", "Lanzhou ramen" and "Lanzhou authentic Ramen" are used to select suitable store names. The obtained data is the national beef noodles store information. The name of a municipality directly under the central government will be added to the attributes of the province, and the data of business hours and prices inconsistent with the requirements will be deleted. For example, Shanghai is a municipality directly under the central government. If filling "Shanghai" into the attribute of a province in the dataset, it can facilitate unified operation later by changing the category of attribute values, converting prices and ratings to floating point data, placing hours of operation and ratings to intervals, then saving the processed data. The number of beef noodles in the country is 46,918, which is basically consistent with the official statistics.

(2) Data processing of beef noodle stores nationwide

They are grouped according to "province" and "city" respectively, and the results after grouping are counted according to the group to get the number of beef noodle shops in each province and the number of beef noodle shops in each city, which are saved as tables respectively. According to the

"score" of the group, according to the statistics, to obtain the national beef noodle shop score.

(3) The corresponding table of the number of stores and various factors in all provinces in China

The data table of national beef noodle shops is used to convert the provinces and the number of shops obtained by grouping statistics into dictionaries. The provinces and population obtained by grouping the statistics with the population table are merged, and the row where the null value is located is deleted to obtain the statistical table of the number of provinces and population. According to this method, the number of stores and the area of each province, the corresponding table of the national cities and population, and the corresponding table of the number of stores and the population density of each province are obtained.

## 3. Descriptive Statistics

Based on the previous data preprocessing work, we can obtain the dimensional information of national beef noodle stores. Using this information data, we can determine the distribution of stores in provinces and cities, their business hours, store score distribution, and more. Additionally, we can display the dimensional information of national beef noodles through data visualization.

Experimental tool: Matplotlib library. Matplotlib library is a multi-platform data visualization library based on NumPy array. It has good operating system compatibility, graphical display interface and output format, and cross-platform is the most important feature of it.

### 3.1. Number of Beef Noodle Shops in Each Province

According to the result table obtained from data preprocessing, we should import the Matplotlib library and use Pandas to open the number of beef noodle shops in each province. Next, select "province" for frequency statistics and draw the number of beef noodle shops in each province, as shown in Figure 4.

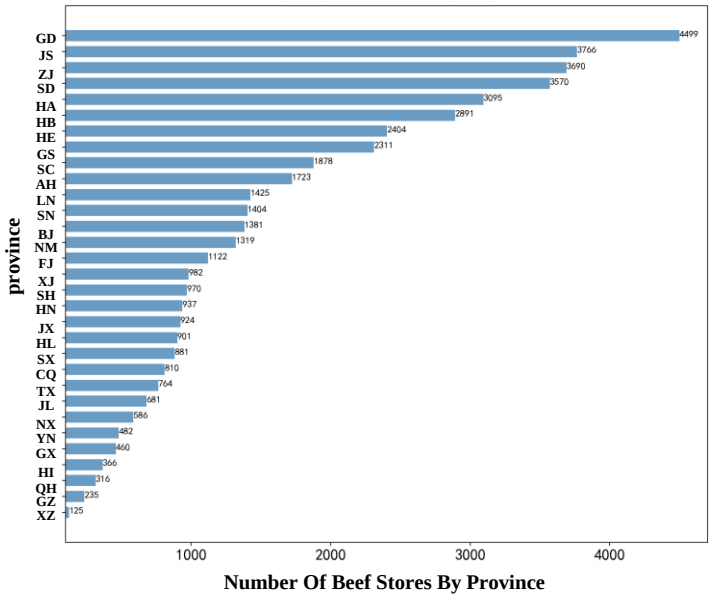


Figure 4: Number of beef noodle shops in each province

There are a total of 46,918 beef noodle shops in China, among which Guangdong, Jiangsu and Zhejiang provinces, the three southern provinces, have the largest number of beef noodle shops, and Guangdong Province has 4,499 Lanzhou beef noodle shops. The number of shops in Jiangsu, Zhejiang and Shandong provinces was almost flat, with an average of 3,650. The Tibet Autonomous Region and Guizhou Province have the least number of beef noodle shops. It could be the small

population or diet. On the whole, in provinces with large population and better economic development, Lanzhou has the majority of beef noodle shops.

### 3.2. Number of Beef Noodle Shops in Each City

In order to obtain the number of shops in various cities in the country, tabulated data was obtained by Pandas, which was then sorted to obtain the ranking of the top 15 Pandas, as shown in Figure 5.

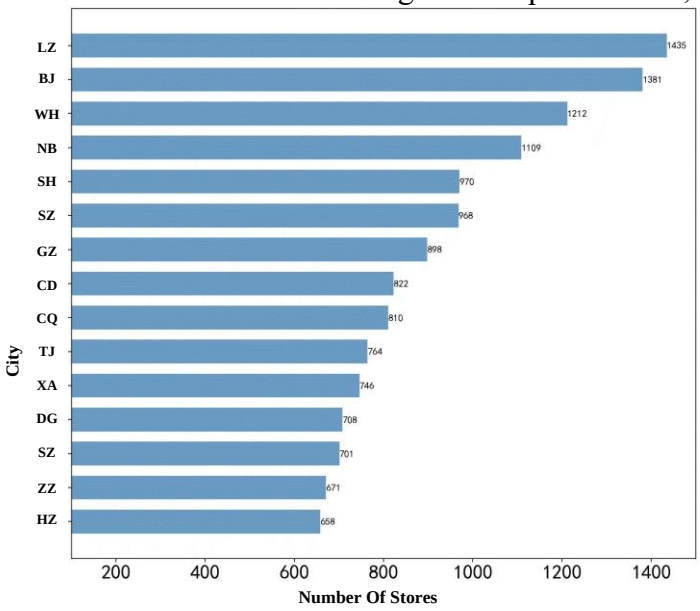


Figure 5: Number of beef noodle shops in each city

The figure visually displays the number of beef noodle shops in cities across the country, among which Lanzhou, Beijing, Wuhan and Ningbo have more than 1,000 shops, and Lanzhou has the largest number of shops. The number of shops in Shanghai and Shenzhen is basically the same. In addition, southern cities account for a large proportion among the top 15 cities, which proves that although Lanzhou beef noodle is a representative specialty of northern pasta, it is widely accepted and popular in the whole country.

### 3.3. Business hours Distribution of Beef Noodle Stores Nationwide

After processing, the operation hours of national beef noodle shops were analyzed. The operation hours data of beef noodle shops were obtained by using Pandas. Matplotlib drew the operation hours distribution maps of national beef noodle shops and those of Gansu Province, as shown in Figure 6 and 7.

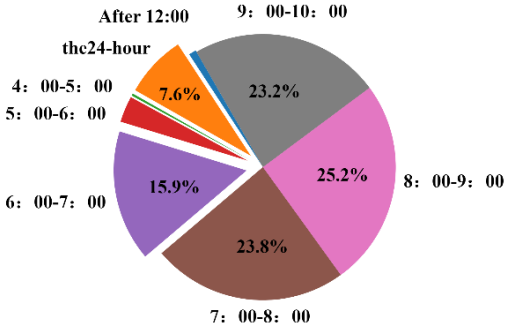


Figure 6: National operating hours

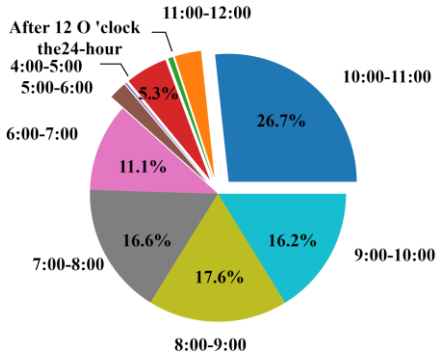


Figure 7 Business hours chart of Gansu

It can be intuitively seen that the business hours are concentrated between 7:00 and 10:00, among which 8:00 and 9:00 account for the highest proportion. In the same way, the distribution of business hours of beef noodle shops in Gansu Province is drawn, as shown in Figure 3.4. The business hours are concentrated from 7:00 to 11:00. Therefore, whether in Gansu Province or the whole country, anyone who wants to eat beef noodles can always find a shop open at any time of the day.

### 3.4. Rating Distribution of Beef Noodle Stores Nationwide

Scoring distribution map was drawn for the processed national beef noodle shop information. Matplotlib was used to read the scoring data of national beef noodle shops and obtained the scoring distribution map for national beef noodle shops. It was concluded that 71.8% of the national beef noodle shops were scored between 4-5 points, and the score of 1-3 points was no more than 5.4%. Thus it can be seen that Lanzhou beef noodles are widely praised in the whole country. As shown in Figure 8.

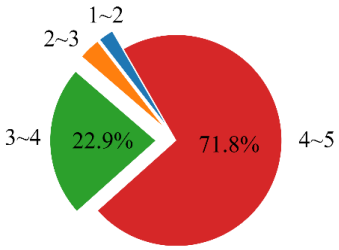


Figure 8: Scoring distribution of beef noodle stores across the country

Through the descriptive analysis of the relevant data of all provinces and cities in China, it can be seen that the number of stores in each region presents a step-type gap, and the business hours are basically in line with the public eating time. Basically, in each region, as long as you want to eat beef noodles, you can find open stores, and the number of stores above the pass line is more than 90%. This shows that the taste of Lanzhou beef noodles can be accepted by the public in different eating habits across the country.

## 4. Correlation Analysis

Correlation analysis method analyzes the correlation between two or more variables under natural conditions without human control or intervention. Through the numerical table of the number of stores and various factors in each province obtained by data preprocessing, the linear regression method[8] of machine learning is intended to explore the relationship between the number of stores in each province and various factors. The linear model f(x)=w*x+b is adopted to explore the relationship between the number of stores and the provincial area, For example, if the provincial area is assigned to y, the target value f(x,θ)=w*x+b≈y is obtained[9]. In order to evaluate the model performance of the training results, the sum of squares of the gap between the target value and the predicted value is calculated. Therefore, the formula of quantified loss function is as follows:

$$\text{Loss}(y,\hat{y}) = sum((y - f(x,\theta))^2) \tag{1}$$

The task of linear regression is to find the θ value when the Loss is minimum. The Stochastic Gradient Descent (SGD)[10] is adopted. The formula of random gradient descent is:

$$\theta = \theta - \alpha \frac{d(loss(\theta,x,y))}{d(\theta)} \tag{2}$$

The parameter θ is constantly updated iteratively. When the iteration times are sufficient, SGD can converge to the local optimal solution, making the objective function constantly close to the minimum value.

## 4.1. Relationship between the Number of Stores and Population in all Provinces

According to the relationship between the number of stores and the number of population, the value of w is 1.97, the value of b is 1520.82, and the value of Loss is 26300.83. The number of stores is proportional to the number of people, and the model obtained is: y=1.97*x+1520.98. The relationship between the number of stores and the area of provinces is shown in Figure 9.

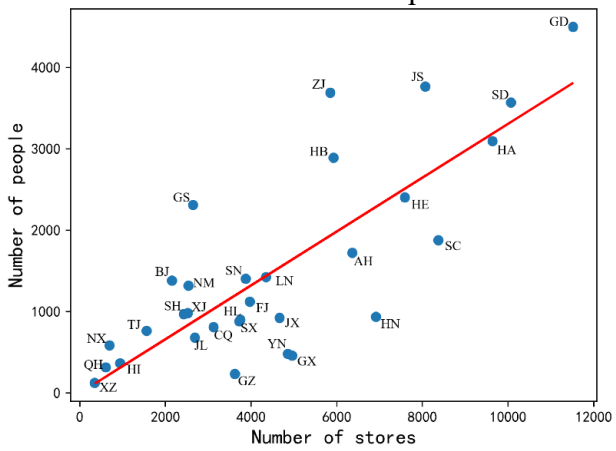


Figure 9: Relationship between the number of stores and the number of population

The other cities that deviate far can be analyzed separately. Take Lanzhou as an example, the actual number of beef noodle shops is much higher than the value expressed in the expression. The intuitive reason is that it is related to the regional food culture. More detailed reasons can be analyzed in a deeper level.

## 4.2. The Relationship between the Number of Stores and Population Density in All Provinces

The scatter plot is drawn for the relationship between the number of stores and population density in these provinces, as shown in Figure 10.

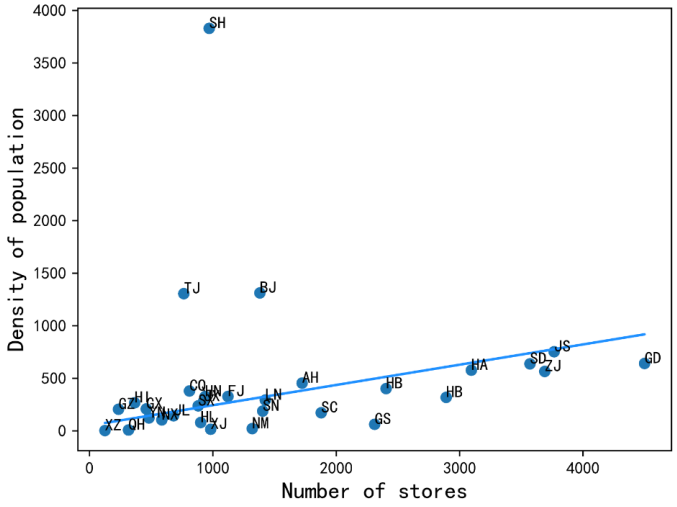


Figure 10: Number of stores and population density

Obviously, there are three municipalities directly under the central Government are discrete points, and the common point of these three cities is that the population density is too large. The linear relationship between other provinces is as follows: y= 0.14*x+47.57, where w is 0.14, b is 47.57, and loss is 230872.55. Except for discrete points, the number of stores is positively correlated with population density. Among them, Shanghai has the furthest deviation, with 972 stores and a population density of 3,830, which is relatively small compared with other cities with the same population density. More detailed analysis can be made of provinces where the population diverges further.

## 5. Conclusion

Data mining technology in big data affects the development of all walks of life, and catering is no exception. This paper takes Lanzhou beef noodle shops as the research object, analyzes the information of national beef noodle shops, and further deepens the understanding of Lanzhou beef noodle, a unique geographical food culture phenomenon. Lanzhou beef noodles, as a unique symbol of Lanzhou's food culture, has witnessed the hundred-year urban development process of Lanzhou. It vividly depicts the unique food culture of Lanzhou people and has become a lens to understand the life of Lanzhou people. However, like other big data problems, the lack of detailed attribute data restricts the further study of Lanzhou beef noodle Restaurant. If it can match the daily sales volume of beef noodle restaurant, detailed population data and precise occupation division, the social attributes of Lanzhou beef noodle can be further studied.

In this paper, Python language is used to acquire, pre-process, data mining and analysis of the national beef noodle data. Finally, the information of the national beef noodle shops and the relationship between them and various factors are analyzed, and the algorithm model based on linear regression is obtained. But this data mining implementation process is mainly to explain how to use Python data mining work, as well as the implementation of the data mining process. Details such as optimal parameter adjustment and model optimization have not been carried out on the final generated model. Based on the work in this paper, the following research work can be carried out:

1) Divide time periods to obtain data, build a prediction model, and predict the changing trend of stores in the future.

2) Due to the limited scale of this data source, the obtained results are only data analysis of Lanzhou beef noodle industry. More industry data can be collected for analysis and more secret information can be obtained.

3) This mining method can be applied to different fields for research and data analysis results of different industries can be obtained.

## Acknowledgments

This work is supported by Lanzhou Chengguan District Science and Technology Planning Project (2021RCCX0016), Gansu Provincial University Young Doctoral Fund Project (2022QB-016) and the Fundamental Research Funds for the Central Universities (31920220010).

## References

*[1] Peng Cheng, Su Su. Ma Li Min has a great development of beef noodle industry in Lanzhou. Food and Beverage World, 2021 (04):12-15.*

*[2] Wu Yong, He Changtian, Fang Jun, Zhang Chao. Fraud auditing of Financial Statements based on Big Data Mining analysis. Finance and Accounting Monthly, 2021 (03):90-98.*

*[3] Fang Ji, Xie Huimin. Application research of Python in Big Data Mining and analysis. Digital Technology and Applications, 2020, 38 (09):75-76+81.*

*[4] Guo Lirong. Web crawler program Design based on Python. Electronic Technology and Software Engineering, 2017 (23):248-249. (in Chinese)*
*[5] Ren Zhiwei. Research on Data Preprocessing for Data-driven Modeling. Henan University of Science and Technology, 2013.*
*[6] Li Xuelong, Gong Haigang. Overview of Big data systems. Science China Information Science, 2015, 45(01):1-44.*
*[7] Wu Yong, Chen Hui, Zhu Weidong. Management Accounting System Reconstruction based on big data analysis technology. Finance and Accounting Monthly, 2019(07):61-68. (in Chinese)*
*[8] Li xueyang. Shao Xigao. Research on the Influencing Factors of College Enrollment Quality Based on Multiple linear Regression and Lasso Regression. Journal of Ludong University (Natural Science Edition), 2022, 3 8(04):350-356.*
*[9] Li Mengsi. Research on Linear Regression Algorithm for Data Privacy Protection based on two-way computation. Shanghai Ocean University, 2020.*
*[10] Qian Ying, Fang Xiunan. Multiple linear regression Model and its application. China Science and Technology Information, 2022 (04):73-74. (in Chinese)*