**Hanoi University of Science and Technology  
School of Information and Communication Technology**

A picture containing icon

Description automatically generated

**Capstone Project Report**

***Introduction to Data Science***

***Title: <put our title here 😊 >***

**Group 1**

|  |  |  |
| --- | --- | --- |
| No. | Name | Student ID |
| 1 | Bui Nguyen Bao Ngoc |  |
| 2 | Nguyen Thi Minh Chau |  |
| 3 | Nguyen Thanh Long |  |
| 4 | Nguyen Minh Tri |  |

*Hanoi, 12/2021*

Table of Contents

[Abstract 4](#_Toc91692435)

[Introduction 5](#_Toc91692436)

[Implementation 6](#_Toc91692437)

[Collecting data 6](#_Toc91692438)

[Processing data 7](#_Toc91692439)

[Data visualization 8](#_Toc91692440)

[Data analysis 16](#_Toc91692441)

[Insights and evaluation 16](#_Toc91692442)

[Conclusion & Remark 16](#_Toc91692443)

Table of figures

Table of tables

## **Abstract**

Among the fundamental needs of human beings, dining is arguably the most essential one. As the society develops, not only does the need retain its significance, but it has also become more varied. This establishes a potential market for the investors to capital into restaurants or food vendors. There are multiple data science models analyzing the possibility of opening an eatery at a particular place that has been built over the world. In this project, we make a similar design for studying restaurants and population in Hanoi. From the obtained information, we point out the insights, as well as give conclusions and evaluation about location for a food vendor of a specific category that could optimize its owners’ profits.

## **Introduction**

One of the most crucial needs of human is dining. Depending on the region, the habit of the dwellers also varies. However, with the globalization which leads to the adoption of new cultures to different areas, as well as the awareness about the importance of the introduction of various kinds of nutrients into each meal, human demands on eating are developing continuously and becoming more and more complicated. This fires the endless race of a large number of restaurants in different categories to satisfy the customers’ desires, so as to earn as much as possible. Yet, due to diversity in living condition of people, as well as the cultures between places, the choice of location to start a food business should be put under careful consideration to avoid irrational decisions, which could cause unwanted loss or failures. The evaluation depends on a set of criteria, including the population characteristics, the dwellers’ living standards, the restaurant distribution, as well as their features, and the real-estate rental price in a particular place. This information could be collected via online platforms and be analyzed with data science methodologies to produce useful conclusions supporting the judging process for opening an eatery of a kind. Currently, there are many projects that use data science to fulfill this requirement, although the target is of other regions. In this project, we gather distinct data for Hanoi, the capital city of Vietnam, in order to give appropriate assessments for starting a foody business here. Our project is divided into 5 parts:

* Collecting data
* Processing data
* Data visualization
* Data analyzation
* Insights & Evaluation

Each of which is detailly explained in corresponding sections bellow. At the end of this report, we also have a part for conclusion and comments on this project

## **Implementation**

### **Collecting data**

At the beginning of the project, we determine the needed information for serving the needs of the analysis, including data about restaurants in city, population characteristics, and real-estate current status.

a. Restaurants’ information

Since the main aim of the project is about the restaurants, data about restaurants in the region is essential. The target information about food vendors includes:

* Lists of restaurants: Name, address
* Lines: Luxury, normal, street vendor, etc.
* Types of servings
* Prices of dishes
* Customers’ rating, in terms points on a scale of 0 to 9.

This data is collected via ShoppeFood and Beamin. We combine both scrapping tools and asking for permission to use the desired data for the project’s purposes. For the sakes of ethics, we do not try to exploit the closed data from websites, but try alternatives that can fulfill our needs.

b. Population characteristic and customer services in region

Population distribution is considered one of the important factors that could affect the business of a restaurant. We collect the population density information via the Hanoi Statistical Yearbook of 2019. It is obvious that this characteristic of the city has changed after 2 years since then. However, we assume that the population growth rate of each district in the city remained over the time, and there has been no noticeable political event that affect the migration flow. Therefore, the differences are negligible, and this information could be used at current time without side effects.

We also predict that the region having many important centers of services such as schools or hospitals should be put on more attention, since the dinning demand of dwellers living in these places are usually higher compared to others. In the scope of this project, we gather list of schools and hospitals in Hanoi via directory websites, as well as the official portals of Hanoi Department of Education and Training and Hanoi Department of Health

c. Real-estate information

The price for renting business places also contributes to deciding to start and maintaining the restaurants. Therefore, we collect data about rental price for business places over the districts in the city. This information could be found on the real-estate agency websites.

### **Processing data**

Because of the non-regulated and non-verified data over the internet, the collected dataset contains multiple noises. In order to have appropriate information for the next stages, we put much time and effort on cleaning progress.

a. Restaurant information

Of all the data categories mentioned, restaurant data is the most varied. This is also the set of data having fewest errors. However, the vendors’ categories must be put correctly. Based on the list of dishes sold in these stores, we use tagging methods to classify the restaurants. Since each of which may serve more than one kind of food, we decided to do this manually instead of using tools, in order to get the most accurate dataset.

After having analyzed this data, we obtain a set containing the fields:

* < list all the fields here>

One vendor can be put under more than one category, but no more than 3. In this project, we pointed out and work on 14 different types of restaurants:

* <list the categories here>

Note that, the “restaurant” category targets at normal restaurant serving normal savories of Vietnam, maybe also of other regions, but not on a specific theme. The “snacks, street snacks” category aims at small street vendors with non-identical dishes such as mixed desserts, grills. In addition, “breakfast, street breakfast” is the type of vendors selling identical dishes breakfasts, such as sticky rice, banh mi, etc.

b. Population and services

The information about population density of each district in Hanoi had been calculated and presented in detail in Hanoi Statistical Yearbook 2019; therefore, there is no need to perform any further process on this data.

In addition, the list of schools and hospitals over the city is extracted from different online sources, including the portals of the city governments, so as to get the most appropriate addresses. For the records having unclear destination, we tried to fill in with the nearest location, to Ward level, found on Google Map.

c. Real-estate information

This is the data set required most effort for cleaning and correcting. Because of the spontaneity of the data publishers, the obtained data are mainly non-monitored and verified. There is a considerable proportion of results that has blank data, or unreasonable information. In this project, we eliminate all the records with no data about area for rent, and use mean value for the price for renting each area meter per month. This value is estimated from other reliable records in the same region.

### **Data visualization**

In order to visualize data in a proper way that support the analysis stage, we need to understand the aim of visualization. We put questions related to the dataset and the aim of the project so as to determine the suitable kinds of graphs to demonstrate the information.

a. Restaurant data

For the restaurant data, we put attention on these questions:

- How do the restaurants distribute over the districts?

- What is the distribution of restaurants with respect to their type over the districts?

- What is the price distribution with respect to each type of cuisine?

- What is the price distribution with respect to districts?

- How is the rating of restaurants in districts?

- What is the relationship between restaurants’ ratings and their corresponding price range?

From these questions, we decided to use stacked bar-chart and histogram to easily illustrate and compare different factors considered.

Chart, bar chart, waterfall chart

Description automatically generated

From the above figure, we can see that, there are 3 types of restaurants being focused at this moment: (normal) restaurant, cafeteria and office meals. There are few food vendors serving only one specific kind of cuisine, eg. Japanese or Korean food only. The reason for this is that these meals require high-quality chiefs so as to deliver the original tastes of the dishes. Moreover, the investor also has to put more budgets on specific ingredients for making the meals. The labor and material cost outweighs the profits that could be obtained, but these types of restaurants are also picky: customers may prefer a place that has a wider range of dishes rather than which serves only one outstanding type of cuisine. There are too many risks to inject capacity in this one-theme kind of food store. Varying the food served helps target more class of customers, and thus, improve the potential of it earning more profits.

Dong Da is the district with the highest number of restaurants in all types considered. The second and third favorable places for opening a food vendor is Hoan Kiem and Hai Ba Trung district. These are all regions in the inner city, which has a high population, leading to high and diverse demand of dining. Moreover, these districts also have many schools, hospitals and other services, which results in people needed places to eat to fulfill their basic need.

Chart

Description automatically generated

Figure… compares the price distribution with respect to 14 categories. From the figure, we can have a conclusion about the average price of cuisines, as shown in the bellow table

|  |  |
| --- | --- |
| **Phân loại đồ ăn** | **Phân khúc giá phổ biến (đơn vị: nghìn đồng)** |
| Đồ ăn vặt, đồ ăn vỉa hè | < 200 |
| Đồ ăn sáng | 50-100 |
| Cơm văn phòng | 45-75 |
| Cơm bình dân | ~50 |
| Pizza | 100-400 |
| Cơm, phở | 100 |
| Nhà hàng | Rất phong phú, chủ yếu dưới 500 |
| Ẩm thực Hàn Quốc | < 150 (\*) |
| Ẩm thực Nhật Bản | < 300 |
| Trà sữa | < 100 |
| Café | 35-70 |
| Bánh ngọt | 100 – 600 |

The dishes in normal restaurants are usually in the price range of 100 – 400 thousand VND. For the street vendors, dishes cost less than 100 thousand VND. This figure for the cafeteria varies from 35 to 75 thousand VND for a cup of dessert.

Note that there are few data collected for the Korean restaurant type, therefore the information for this kind may be not reliable.

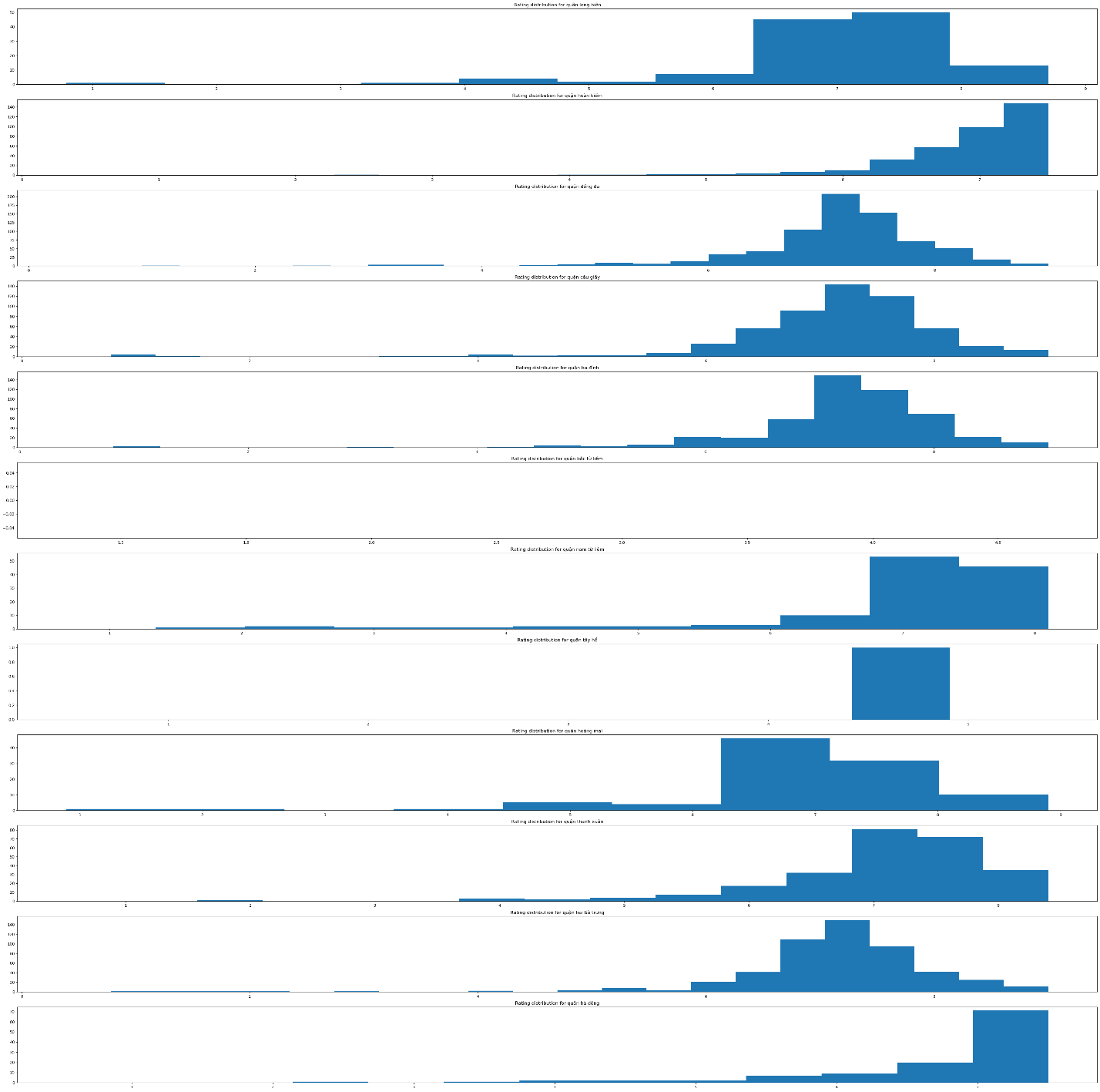
Chart, timeline

Description automatically generated

The above diagram demonstrates the price distribution according to different regions of the city. This price, to some extends, can reflect the living standard of the dwellers in the area. Normally, people living and working in a particular place will use that location’s services, which include the dinning. The price of dishes in each district, therefore, may illustrate the average budget one might put in order to have a meal.

|  |  |
| --- | --- |
| **Quận** | **Giá cả trung bình (đơn vị: nghìn đồng)** |
| Long biên | Tập trung mạnh ở mức dưới 100, có nhiều ở mức 100 – 200 |
| Hoàn Kiếm | < 350 |
| Đống Đa | < 300 |
| Cầu Giấy | < 300 |
| Ba Đình | < 250 |
| Bắc Từ Liêm | < 300 (\*) |
| Nam Từ Liêm | < 300 |
| Tây Hồ | < 300 |
| Hoàng Mai | < 100, dao động mạnh quanh mức 200 |
| Thanh Xuân | < 300 |
| Hai Bà Trưng | < 300 |
| Hà Đông | < 200 |

From the histogram and the summary table, the citizens have a tendency of giving approximately 300 thousand VND for a normal meal in food stores over the districts. In the border regions such as Hoang mai, Long Bien district, the price is a little bit lower. Remark that there are little data about restaurant in North Tu Liem, therefore the information for this region might be not reliable.



This figure illustrates the restaurant rating in different districts. The customers’ evaluation for a restaurant are usually based on their experiences with the food, as well as the services delivered. On a scale of 0 to 9, most restaurants are rated 7-8. There is no significant differences over the districts, except from Tay Ho, which has the restaurant rating around 4 to 5.

|  |  |
| --- | --- |
| **Quận** | **Rating** |
| Long biên | 7-8 |
| Hoàn Kiếm | 6-8 |
| Đống Đa | Tập trung mạnh ở 7-7.5 |
| Cầu Giấy | 7-8 |
| Ba Đình | 7-8 |
| Bắc Từ Liêm | Không có dữ liệu |
| Nam Từ Liêm | 7-8 |
| Tây Hồ | 4 - 5 |
| Hoàng Mai | 7-8 |
| Thanh Xuân | 7-8 |
| Hai Bà Trưng | Tập trung mạnh ở 7-7.5 |
| Hà Đông | Tập trung mạnh ở 7 |

Chart

Description automatically generated

Finally, for the relationship between rating and price range, we see that most higher ratings are for the stores with the dishes cost from 100 to 400. This range is also the preferred by most of the citizens over different kinds of restaurant in all districts. The higher the price, the lower the rating, since there are not many people wanted to put high budget for normal quality services, or the services that just satisfy their basic needs.

|  |  |
| --- | --- |
| Mức giá trần | Rating |
| 100 | 6.5-7.5 |
| 200 | 7-7.5 |
| 300 | 7-8 |
| 400 | 7-7.5 |
| 500 | 4-5 |
| 600 | 4-5.5 |
| >600 | 4.5-5 |

Therefore, the price range of 100 to 400 thousand VND attracts more customers, and costs for dishes should be estimated to be assessed in this range.

### **Data analysis**

### **Insights and evaluation**

## **Conclusion & Remark**