**Ho Chi Minh city district land price prediction model**

Vo Khang Di

August 1, 2020

**1. Introduction**

**1.1 Business problem:**

Ho Chi Minh city is the economic capital of Vietnam. As a budding developing country, the city’s land is valuable and fluctuates often. However, it is very difficult to accurately gauge land prices to buy or sell without a loss, or to anticipate price changes in the future. It would be very useful if land prices can be estimated based on observable factors such as the number of places of interest within reasonable travel distances. This project aims to produce a model which can predictdira land prices based on these landmarks.

**1.1 Interest:**

Stakeholders who owns, or are looking to invest in, land in Ho Chi Minh city would be very interested in knowing how land prices will change when landmarks are built near their owned property, and how to properly price that land when looking to sell. Others who are looking to invest would also be interested in paying a reasonable price for a piece of land. Big investors who are looking to design entire wards would be interested in knowing what facilities should be built to raise their land prices.

**2. Data gathering and filtering**

**2.1 Data sources:**

In order to build a model measuring the relationship between land prices and the number of landmarks, we must rely on current data.

The data we require includes:

* The average land price of each district in Ho Chi Minh city
* The number and types of landmarks in an average area of each district
* The location of each district
* The boundaries of each district
* The distance of each district to the city center (district 1)

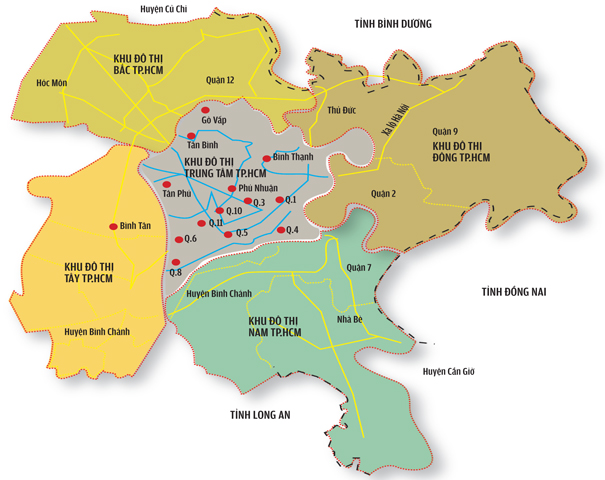
We will use a circle of suitable size, chosen to not exceed the district boundary, to define the area to gather landmark data.

The following data sources will be used to gather information:

* The average land prices in each district will be provided by a very well-known Vietnamese real estate website**: mogi.vn**
* The number and types of landmarks of the defined area will be provided by **Foursquare API**
* The coordinates of each district will be provided by **Google Maps API geocoding**
* The boundaries of each district will be gathered on **Google** and drawn onto a geojson file
* The distance to the city center can be calculated using their coordinates

**2.2 Data filtering:**

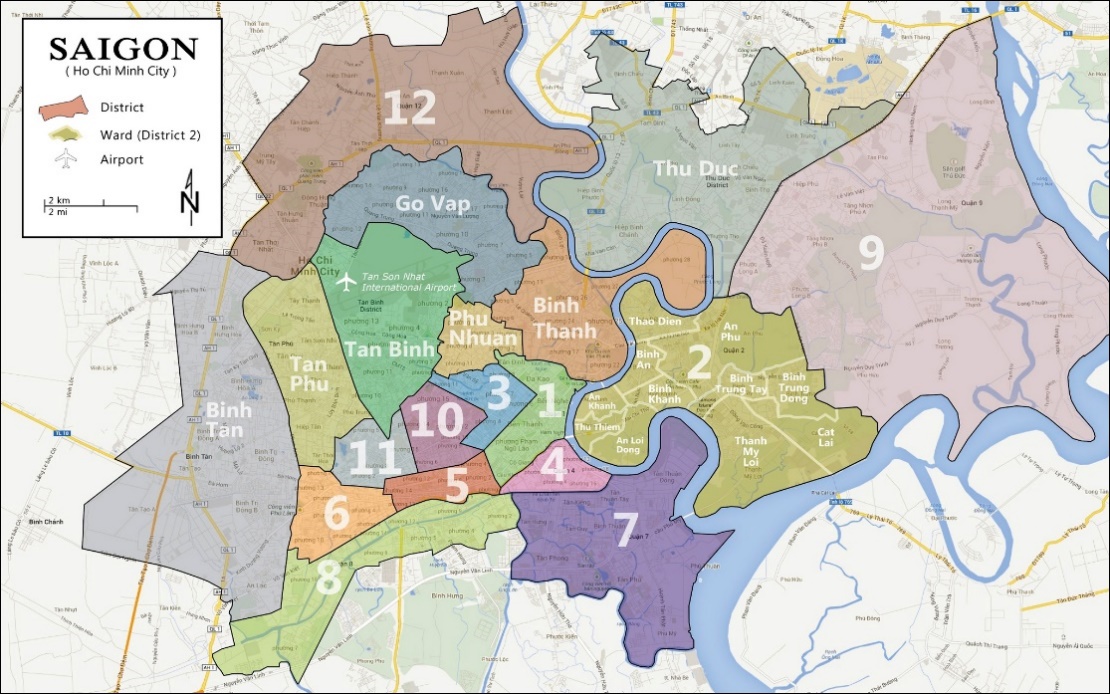
This is a rough map of every district area in Ho Chi Minh city:



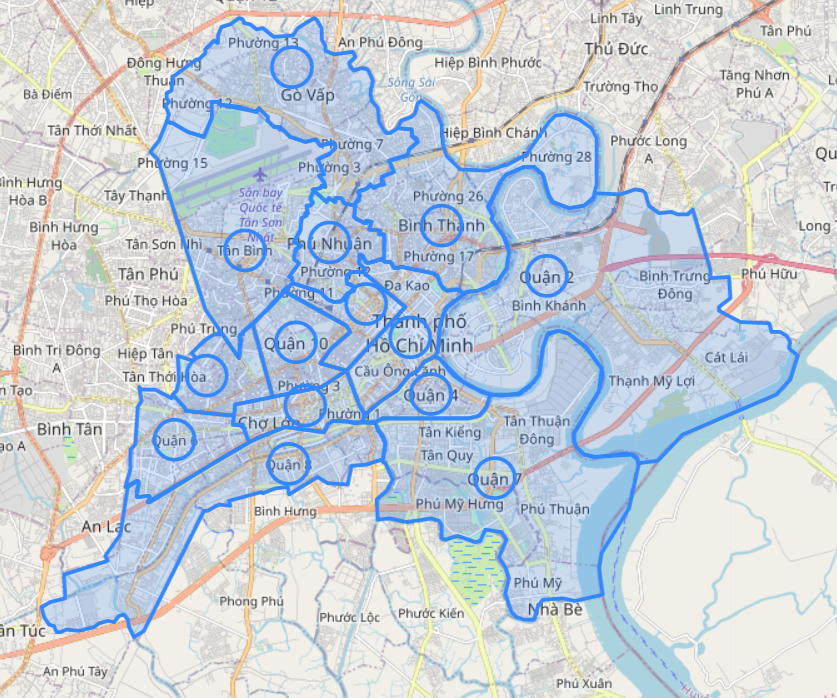
With district 1 (Q.1 on the map) as the economic center, it is mostly the case that population density and land prices drop the farther away you go. The four large areas surrounding the central one ranges from sparsely populated suburbs to uninhabited empty land. This is especially true with the southern parts which is a huge empty salt marsh.

In the interest of not collecting too many outliers and gearing the model towards predicting land prices near the city center where most investments would be made, I decide to not eliminate these areas from the dataset. Thus, the data set will include:

* District 1
* District 2
* District 3
* District 4
* District 5
* District 6
* District 7
* District 8
* District 10
* District 11
* Binh Thanh District
* Phu Nhuan District
* Go Vap District
* Tan Binh District



Here is a clearer map of the relevant area. As you can see, all of the chosen districts are in close proximity to district 1, which is the city center.

Since we are using average land prices, we must also use an average circular plot of land in each district. Due to the limitations of the data gathering method, measuring the number of landmarks within the entire area of each district proved impractical. 

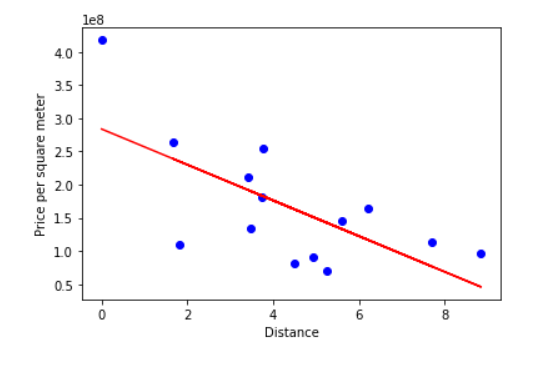
This is the map of the circular areas in each district where landmark data will be gathered. Since the location is chosen by Google Maps API geocoding, it should be a suitable average area. The circle’s size was chosen in the interest of not exceeding any district boundaries.

Another feature was added in order to minimize noise and improve the model’s accuracy which is the distance between each district and district 1. Prices naturally rise as we get closer to district 1. Therefore, to accurately measure the impact of each type of landmark on the price of land, we need to take other factors into account. Population density was also considered, but I was unable to obtain accurate enough population density data to improve the model. This is due to the other data being collected in a circle within each district, while population density is measured throughout the entire district. Obtaining population density within only the circle also proved extremely impractical. Incorporating the inaccurate population density data lowered the model’s accuracy, so it was removed.

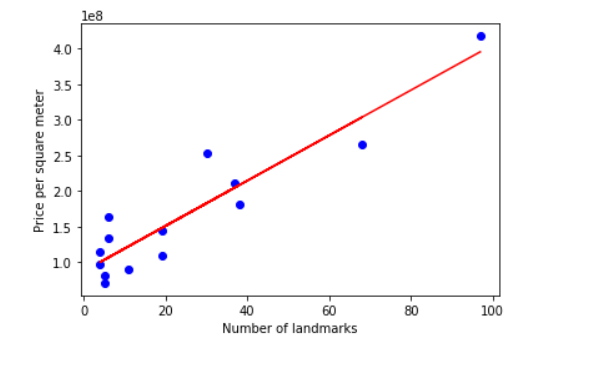
**3. Data analysis:**

**3.1 Model selection:**

Let us look at some of the data collected and their relationship with land prices, starting with distance from the city center:



This plot shows a rather linear relationship. Next is the total number of landmarks in each district:



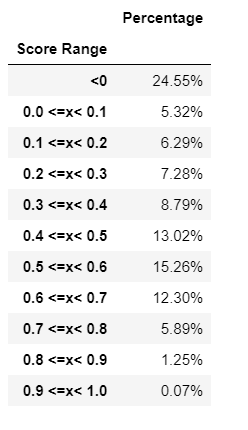
This is an even clearer linear relationship.

In the main model, we will use each individual type of landmarks rather than the total landmarks, but the same relationship should apply to a decent enough degree. Therefore, we should select Multiple Linear Regression.

**3.2 Model and accuracy:**

Due to the model being Multiple Linear Regression with too many features, it is impossible to properly visualize. The model itself is presented inside the Notebook. However, we can still assess its accuracy.

Due to the relatively low amount of data samples. The model’s accuracy, namely its variance score, vary wildly depending on which districts are chosen for the data or testing set. Thus, for now, we can only evaluate how often the it produces an accurate model by calculating all possible combinations of training and testing sets, we arrive at this dataframe:



We can see that around 25% of the time, a completely inaccurate model would be built. The majority of the models tend towards the 0.4-0.7 variance score range.

**3. Results and Discussion:**

Through Data analysis we were able to construct a moderately accurate model capable of predicting land prices in the central districts of Ho Chi Minh city, Vietnam. We also produced a dataframe which will show us how the accuracy will change as more data is collected.

Based on the coefficients of the model, we can also safely say most of these landmarks, as well as the distance of the district to the city center, have a fairly significant effect on the price of the land.

The model is far from complete, however, and requires more data.

This is because there were some difficulties:

Firstly, I was unable to identify all landmarks in every district. This is because Foursquare API does not always give full Vietnamese addresses which would include information on which district the landmark belongs to, making it difficult to sort landmarks by district. Furthermore, it is also unable to locate every landmark in a district. For example, a massive and famous landmark of district 1 is Ben Thanh Market, but Foursquare API did not list it.

Vietnamese districts are also distributed and has boundaries drawn in no particular order, there is no logic to their size and placements other than history. As such it was too difficult to try covering the entirety of the area in each district.

These difficulties resulted in the data gathering area being confined to circles in which not all landmarks were collected. This is a massive loss in data.

If these problems were somehow solved, or if enough data were otherwise amassed through different means, the method above would produce a significantly more accurate and useful model. Not only that, other features such as average population density, which was impossible to measure in a circle, would become available for use to further enhance accuracy.

Currently, the model above has proven to be at least somewhat accurate in predicting test set prices, further testing on real data sets is required for refinement, but this model would suffice for now. The dataframe measuring accuracy will help assess the model’s improvement as more data are collected.

**4. Conclusion:**

The purpose of this project was to identify the different types of landmarks in the central districts of Ho Chi Minh city, Vietnam and gauge their effects on land prices while using their distance to the city center as a factor to improve accuracy. This is to aid stakeholders who have, or are wishing to invest, in real estate in Ho Chi Minh city.

With this model, they will be able to predict the change in land prices as landmarks are built and/or demolished in each district and make more accurate judgements on land prices when they are anticipating buying or selling.

More data is required to make the model accurate.