Oct,9,2020

Predicting the GDP of China Cities

1. Introduction

Background

Gross domestic product (GDP) is a monetary measure of the market value of all the final goods and services produced in a specific time period. GDP (nominal) per capita does not, however, reflect differences in the cost of living and the inflation rates of the countries; therefore, using a basis of GDP per capita at purchasing power parity (PPP) is arguably more useful when comparing living standards between nations, while nominal GDP is more useful comparing national economies on the international market. Gross Domestic Product (GDP) per capita shows a country's GDP divided by its total population. The table below lists countries in the world ranked by GDP at Purchasing Power Parity (PPP) per capita, along with the Nominal GDP per capita. PPP takes into account the relative cost of living, rather than using only exchange rates, therefore providing a more accurate picture of the real differences in income.

1. Introduction

Problem

The GDP of a region can be inflected in many aspects, including the urbanization rate and the average life span of the region. In a word, the higher the urbanization rate is, the higher the average life span is, more likely is that the GDP of the region is higher. So, as we have learn to explore the neighborhoods of a city, I wonder if the vary type of the venue in the neighborhood can be a reflection. To be explicitly, if the rate that cafe takes of the total venues is higher in a same cultural environment, the more likely the GDP tends to be higher, because the need for unnecessaries imply that people are more healthy. In my research, I'm going to verify my hypothesis.

Interest

If hypothesis proves true, I believe there will be two group of audience will be interested in my research. First, the city management official, if they are to improve the GDP of the city, will set some policy which encourage the opening of some venues that may have positive effect to the city's GDP. Secondly, some entrepreneurs may be interested. After the build and calculate the predicted GDP, the will know if the venue they are t open is existing enough in the city.

2. Data acquisition and cleaning

Sources

1. Location data: the location data of the various cities selected as the training group.

   From website: https://bbs.pinggu.org/plugin.php?id=dsu_paulsign:sign

2. GDP data: the GDP data from some official websites.

   From website: https://bbs.pinggu.org/plugin.php?id=dsu_paulsign:sign

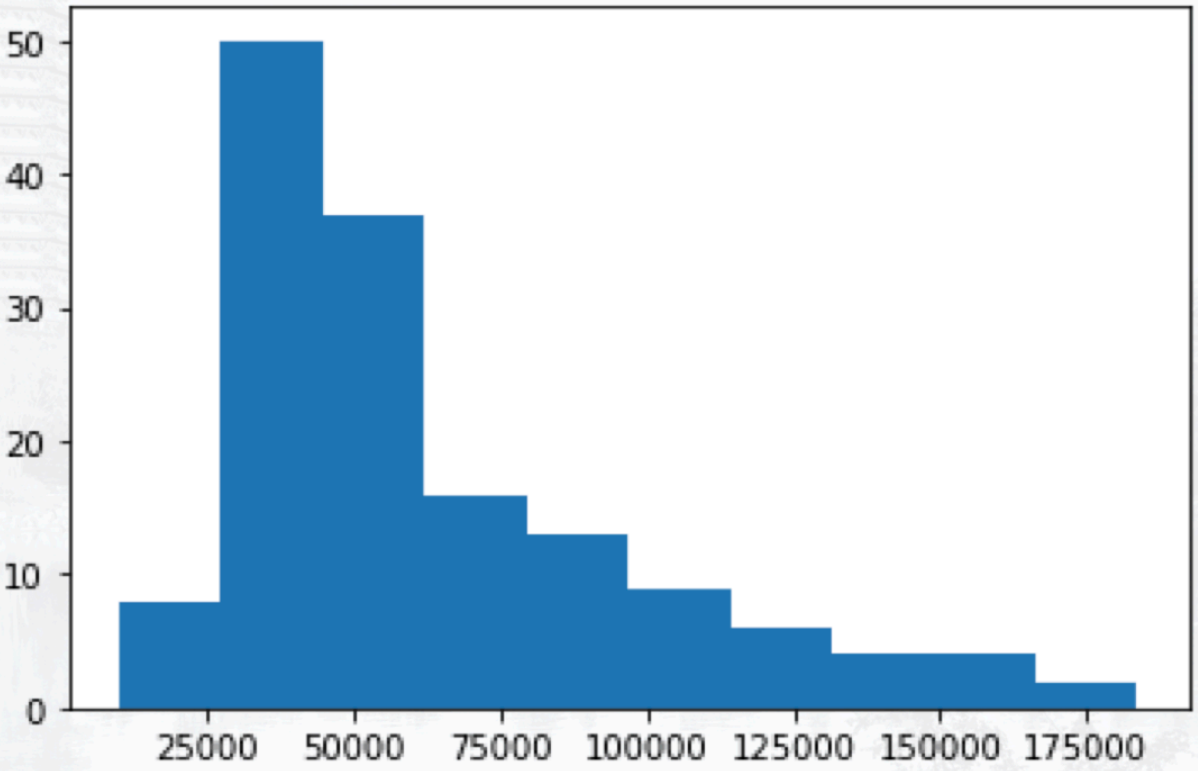
3. Foursquare API

2. Data cleaning and feature selection

Simple feature selection during data cleaning.

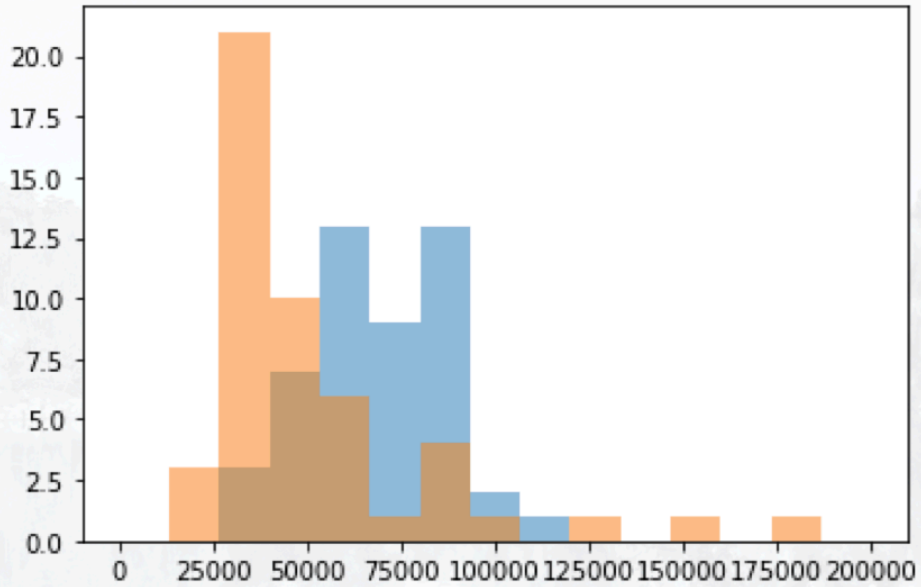
Kept Features	Dropped features
Hotel, Coffee Shop, **Fast Food Restaurant,** **Shopping Mall,** **Train Station,** **Park,** **Chinese Restaurant,** **Pizza Place,** **Bus station,** **Historic Site** ``````	Afghan Restaurant Mediterranean Restaurant Bike Rental / Bike Share Huaiyang Restaurant Taiwanese Restaurant Shanxi Restaurant Boarding House Indonesian Restaurant Pastry Shop ```````

3. Exploratory Data Analysis



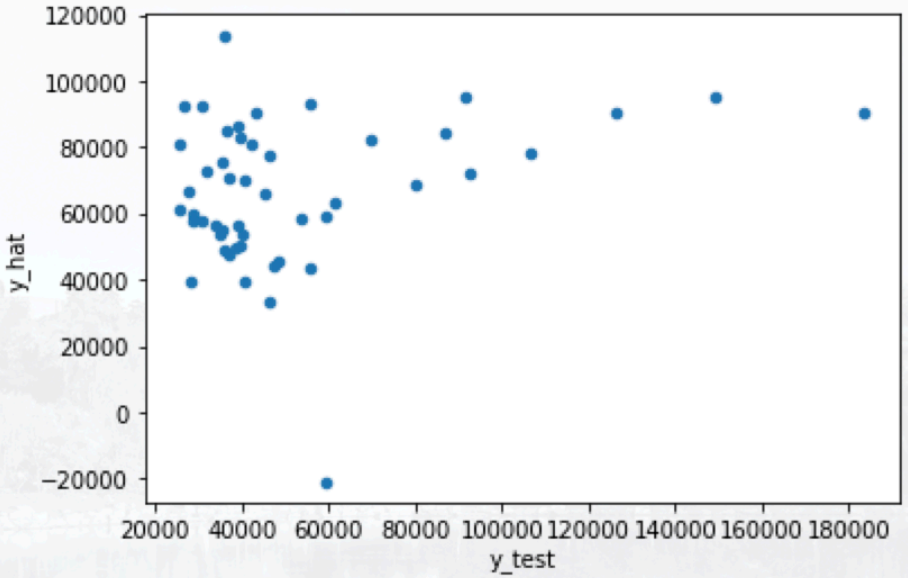
Frequency histogram of GDP per capita of cities in China

4. Predicting GDP(linear regression)



Distribution of actual and predicted GDP per capita using linear regression with equal weights of samples.

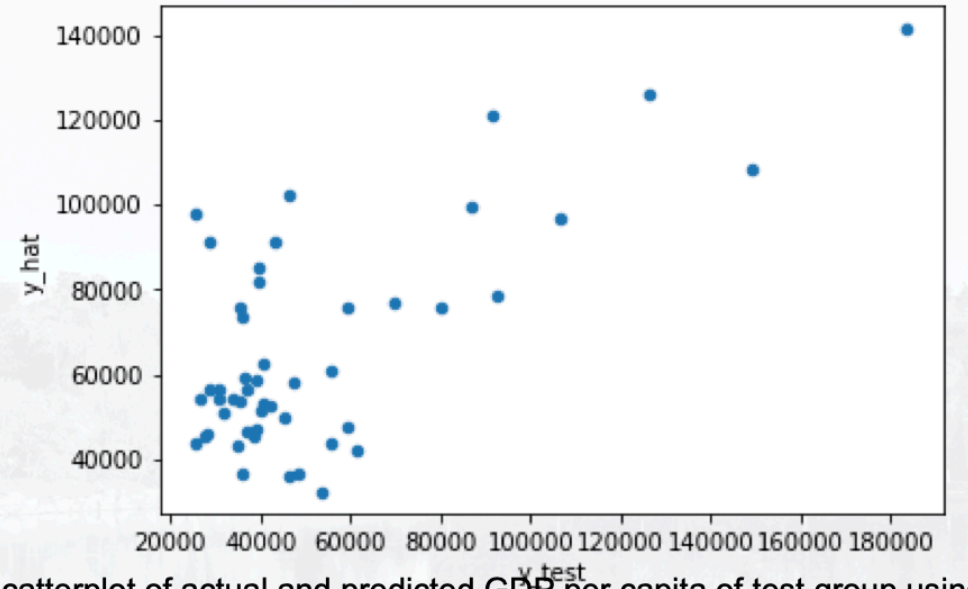
4. Predicting GDP(linear regression)



Scatterplot of actual and predicted GDP per capita of test group using linear regression.

```
Residual sum of squares: 1316241803.28
Variance score: -0.28
```

4. Predicting GDP(GBDT regression)



Scatterplot of actual and predicted GDP per capita of test group using GBDT regression.

```
Residual sum of squares: 724404106.42
Variance score: 0.30
```