Pizza price prediction and Location analysis

Final Report

KunJung Lin, Roshan Kotian, Mark Trovinger

3/18/2020

Contents

1 Description of Dataset

The dataset in question is from Datafiniti, a company that provides businesses with a wide variety of information on retail products, properties, and companies. The data is a list of over 3500 pizzas from multiple restaurants across the United States.

The dataset is 4.7 MB unzipped, which does not make the big dataset *data* by any definition, but has 10,000 rows, which makes it large enough for our purposes.

2 Purpose of this project

The purpose of this project is to present a theoretical business owner with a workup of what the likely price should be for new pizzas. To do so, we will be using two different predictive models, one that is simpler and one that is more complex. The purpose of having two different models is to illustrate how much value could be derived from a more sophisticated approach to solving the problem versus a more straightforward, cheaper solution.

3 Intended audience for project

The intended audience for this project would be the owner of a pizza restaurant that is looking to potentially expand its business or a new owner looking to get started in the pizza business. The final report should not be overly technical, as making a report too technical can often turn off stakeholders and make it more likely that a proposed solution will not be implemented.

4 Manipulate (tidy) the Dataset

- Remove the Columns Have the Same Value
- Select Specific Columns to Data Frame

  We split some columns to another table because for some predictions, we do not need to use such a wide data frame. If we can split them into some small data frame, it can make the data frame more readable and useable. At this point, we split it into two tables, which are `pizza_store_info` and `pizza_info`.

  - `pizza_store_info` stores the information about the store.
  - `pizza_info` stores the information about pizzas and using `id` to connect with `pizza_store_info`.

  After we do some research, the `menus.amountMax` and `menus.amountMin` might mean one specific flavor of pizza that has the price for a whole pizza or a slice if two values are different. To use the data correctly, we need to check the `menus.description` and `menus.name` to see if there has some detail of the price of not. At the same time, we also need to remove the data that the `menus.amountMax` or `menus.amountMin` is `0`.

For the `pizza`, we still keep this variable. In case we need to use the data that has been removed.

- Correct the Data Type of Data Frame

  After some manipulate, then we need to do correct some data type of columns. For example, `province` should be a factor instead of character.

- Summary of the Data Frame

Table 1: Summary of `pizza`

city	province	address	name	menus.amountMax	menus.name
Sherwood	AR	4203 E Kiehl Ave	Shotgun Dans Pizza	7.98	Cheese Pizza
Phoenix	AZ	25 E Camelback Rd	Sauce Pizza Wine	6.00	Pizza Cookie
Cincinnati	OH	3703 Paxton Ave	Mios Pizzeria	6.49	Pizza Blanca
Madison Heights	MI	30495 John R Rd	Hungry Howies Pizza	5.99	Small Pizza
Baltimore	MD	3600 Eastern Ave	Spartan Pizzeria	5.49	Pizza Sub
Baltimore	MD	3600 Eastern Ave	Spartan Pizzeria	10.99	White Pizza

Table 2: Summary of `pizza_info`

menus.amountMax	menus.name
7.98	Cheese Pizza
6.00	Pizza Cookie
6.49	Pizza Blanca
5.99	Small Pizza
5.49	Pizza Sub
10.99	White Pizza

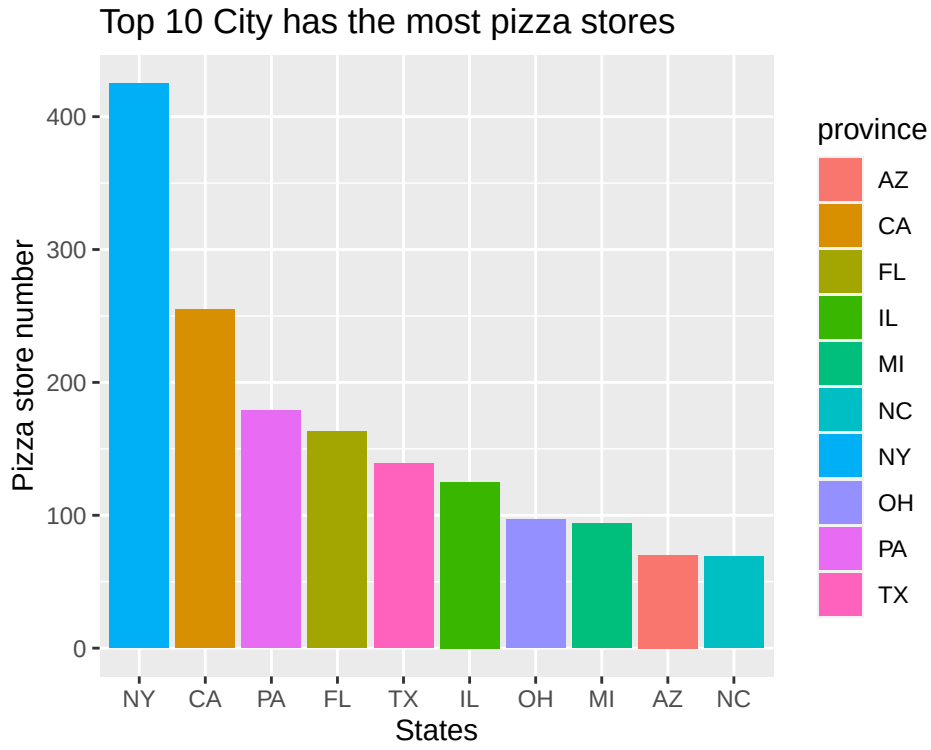
Table 3: Summary of `pizza_store_info`

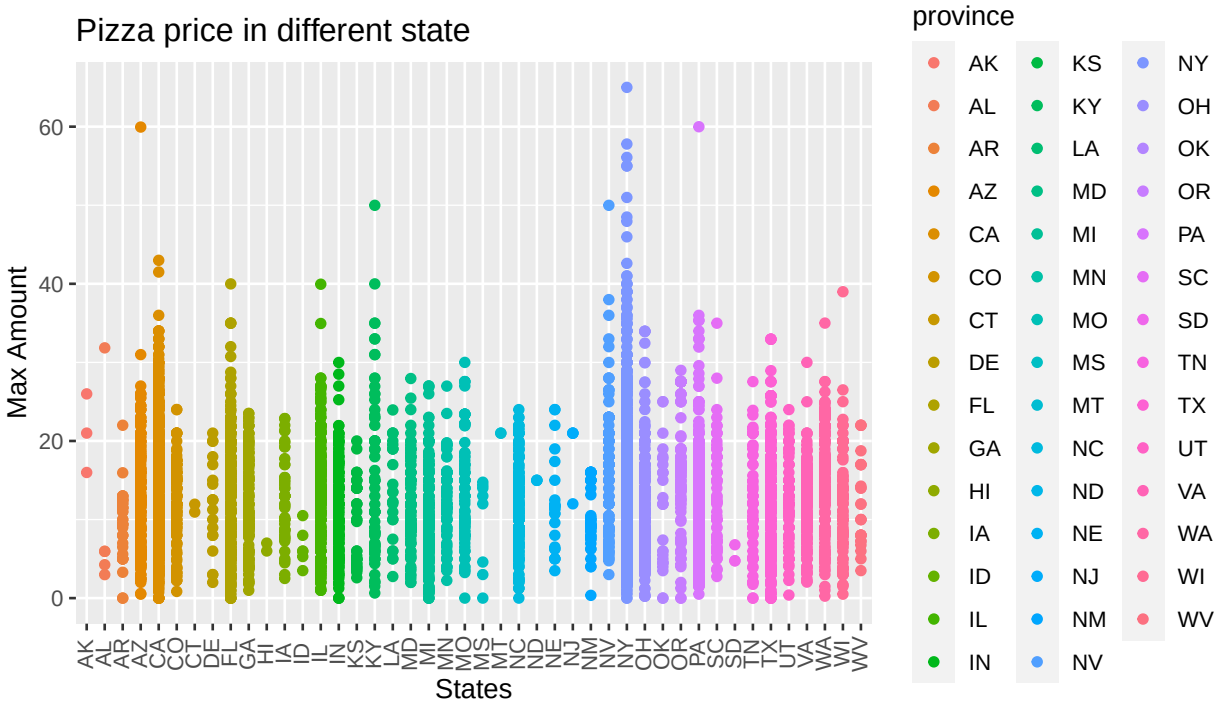
city	province	address	name
Sherwood	AR	4203 E Kiehl Ave	Shotgun Dans Pizza
Phoenix	AZ	25 E Camelback Rd	Sauce Pizza Wine
Cincinnati	OH	3703 Paxton Ave	Mios Pizzeria
Madison Heights	MI	30495 John R Rd	Hungry Howies Pizza
Baltimore	MD	3600 Eastern Ave	Spartan Pizzeria
Berkeley	CA	1834 Euclid Ave	La Vals

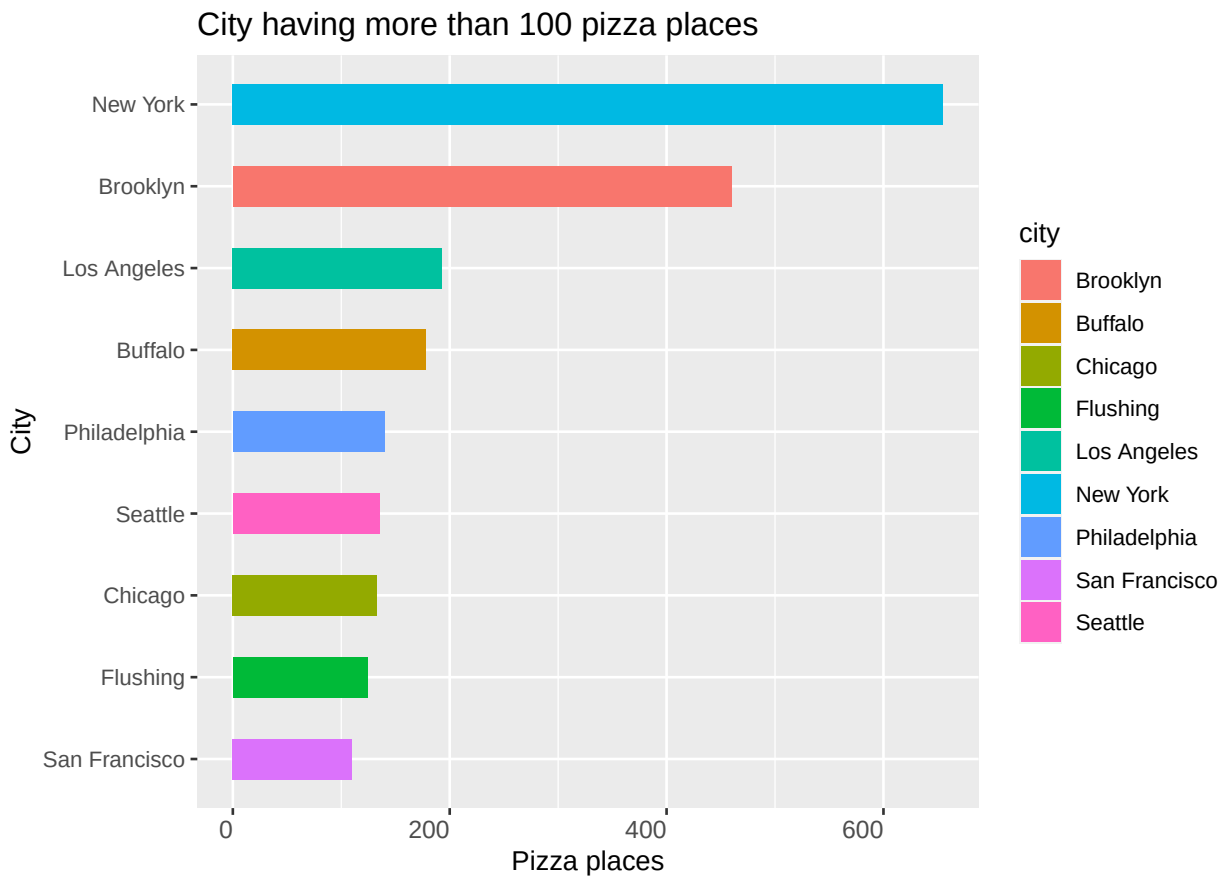
- Renaming column names of pizza_info dataset for better readability

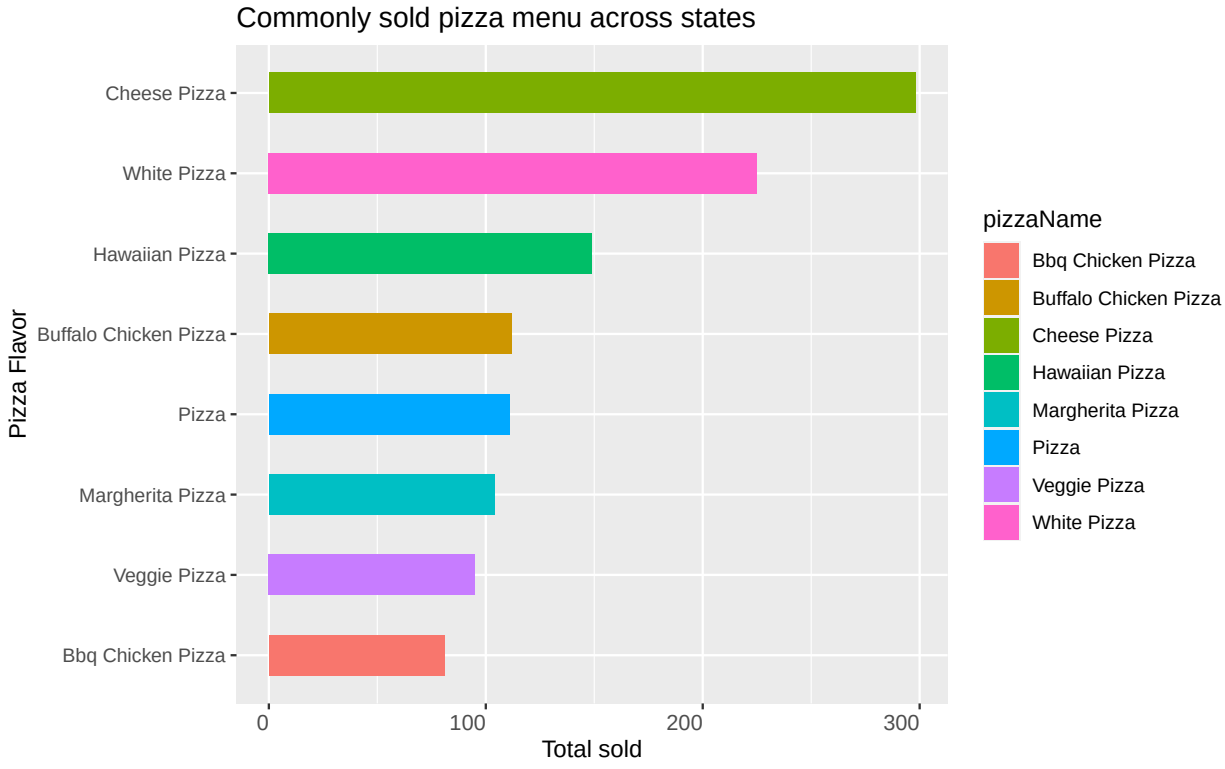
  `pizza_info` data set consists of raw column names which can be modified to more informational names for better readability

5 Visualize the Data






City having more than 100 pizza places
New York
Brooklyn
Los Angeles
Buffalo
Philadelphia
Seattle
Chicago
Flushing
San Francisco
City
0
200
400
600
Pizza places
city
Brooklyn
Buffalo
Chicago
Flushing
Los Angeles
New York
Philadelphia
San Francisco
Seattle


Commonly sold pizza menu across states
Cheese Pizza
White Pizza
Hawaiian Pizza
Buffalo Chicken Pizza
Pizza
Margherita Pizza
Veggie Pizza
Bbq Chicken Pizza
Pizza Flavor
0
100
200
300
Total sold
pizzaName
Bbq Chicken Pizza
Buffalo Chicken Pizza
Cheese Pizza
Hawaiian Pizza
Margherita Pizza
Pizza
Veggie Pizza
White Pizza

6 Model Proposal

The goal of the predictive model will be used in pricing pizza. We hypothesize that a more complex model, such as Random Forest Regression, will have a greater accuracy in predicting pricing than a simpler model, such as Linear Regression.

While a more complicated model tends to require more compute resources to train, the trade-off is that the more complex model performs better, frequently dramatically. In this case, the difference in required compute resources is probably fairly small, but it would be interesting to see how well the two models compared.

We will be testing our hypothesis by looking at the accuracy rates for both the simpler model, and the more complex model. The model with a higher accuracy score will be the one recommended for use in the final report.

The main challenge that could cause issues is our relative lack of experience using R for machine learning. While some in our group have experience using machine learning models, that experience is in Python. This isn't an impossible obstacle to overcome, since we will be learning about modelling packages in R later in the course.

7 Model Implementation

In this section, we will be discussing implementing the two models chosen for this project, Linear Regression and Random Forest Regression.

7.1 Data Preprocessing

Before we can start building models, we need to prepare our data for the model building process. To start we create a data.frame that contains a subset of the dataset that includes both the pizza information and the store information. In order to have a response variable to build a model with, we will create a new column, avgPrice, that adds the minimum and maximum prices and divides by the number of menu items. Since we don't have the number of menu items, we can make an educated guess that it will be between 5 and 10 menu items.

7.2 Linear Regression

Now that we have a response variable to test against, we can look at the first model we tested for the project, Linear Regression. Linear Regression is a model that can be trained and deployed quickly, requiring far less compute resources than other, more complicated models. Because we are looking at the geographic location for how to price an average item from a pizza restraunt, it makes sense to examine the three location based variables in our data; city, state, and ZIP code.

7.3 Random Forest Regression

Much like the Linear Regression model, the Random Forest will examine the relationship between the response variable and the dependent variables. Unlike with linear regression, random forests won't work with factors that have greater than 52 levels. For the purposes of the project, it means we are unable to use the city parameter from our data.

7.4 Model Comparison

For the purposes of comparing the two models, we will look at the Mean Squared Error for both the Linear Regression model, and the Random Forest model. In the case of the Random Forest model, we need to look at the final element in the array mse, as it contains the cumulative MSE for the entire forest.

As we can see from the plot, the Linear Model has a lower Mean Squared Error than the Random Forest. However, we cannot necessarily infer that the Linear Model would be the better choice, as we were not able to use all of the data. This isn't necessarily a weakness in the model, but rather would require a greater investment in data engineering, in order to parse the city column into a form that the model would accept.