

Pizza price prediction and Location analysis

Midterm Report

KunJung Lin, Roshan Kotian, Mark Trovinger

3/18/2020

Contents

1	Description of Dataset	1
2	Purpose of this project	1
3	Intended audience for project	2
4	Manipulate (tidy) the Dataset	2
5	Visualize the Data	4
6	Model details	6

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`

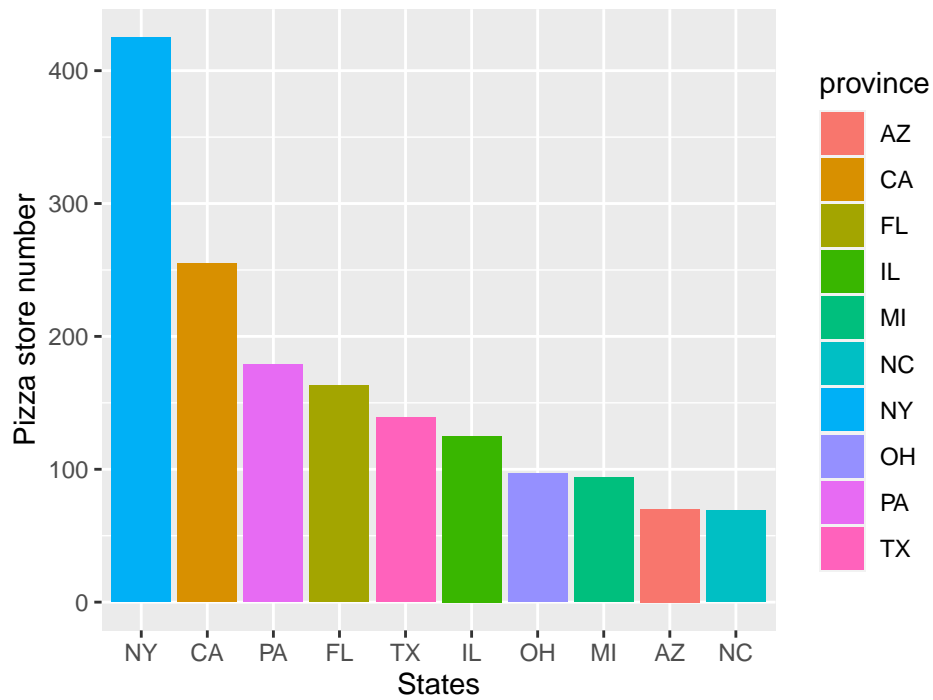
city	province	address	name
Sherwood	AR	4203 E Kiehl Ave	Shotgun Dans Pizza
Phoenix	AZ	25 E Camelback Rd	Sauce Pizza Wine
Cincinnati	OH	3703 Paxton Ave	Mios Pizzeria
Madison Heights	MI	30495 John R Rd	Hungry Howies Pizza
Baltimore	MD	3600 Eastern Ave	Spartan Pizzeria
Berkeley	CA	1834 Euclid Ave	La Vals

- Renaming column names of `pizza_info` dataset for better readability

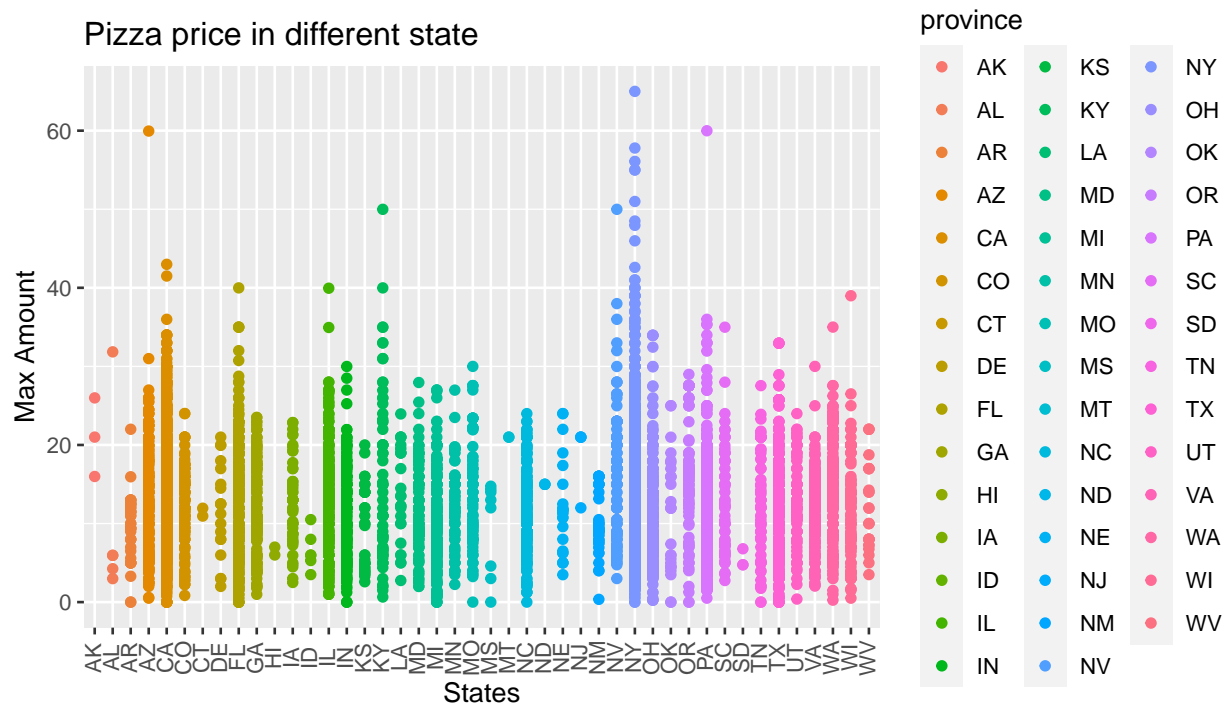
`pizza_info` data set consists of raw column names which can be modified to more informational names for better readability

5 Visualize the Data

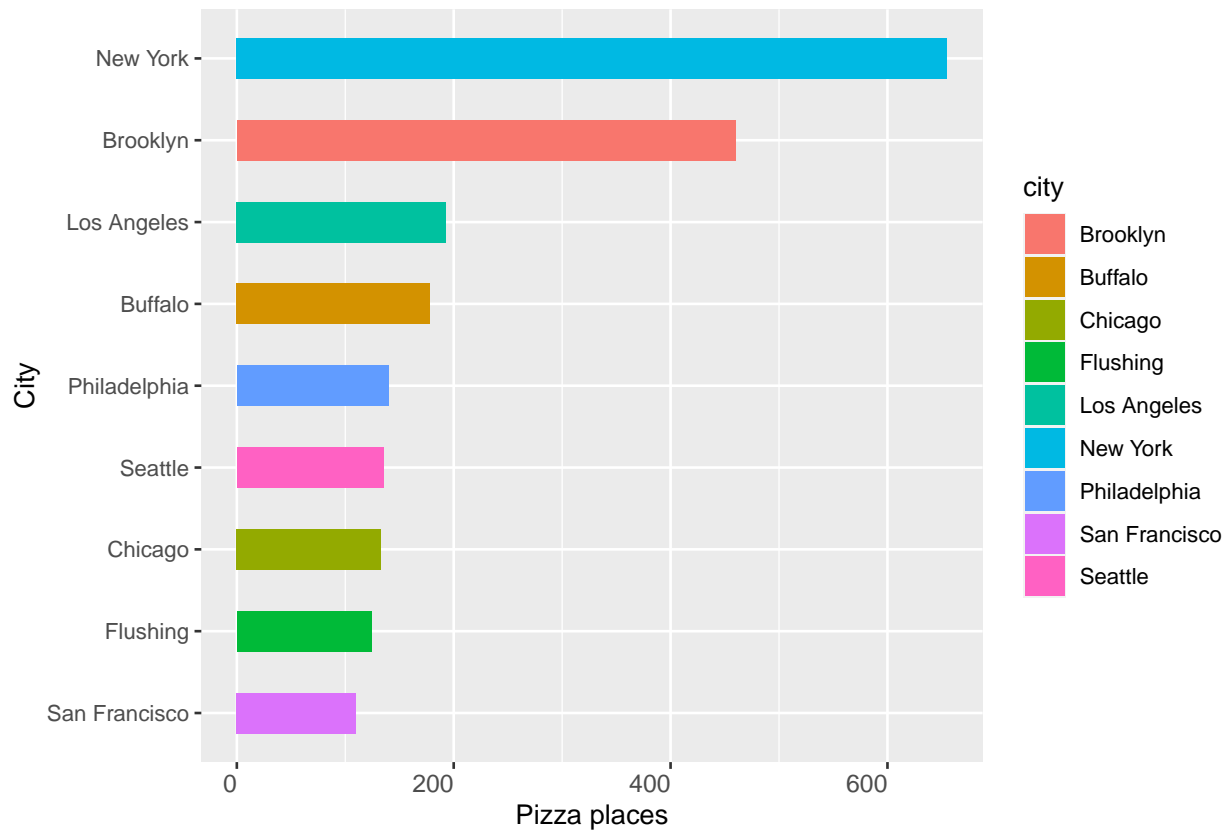
Top 10 City has the most pizza stores



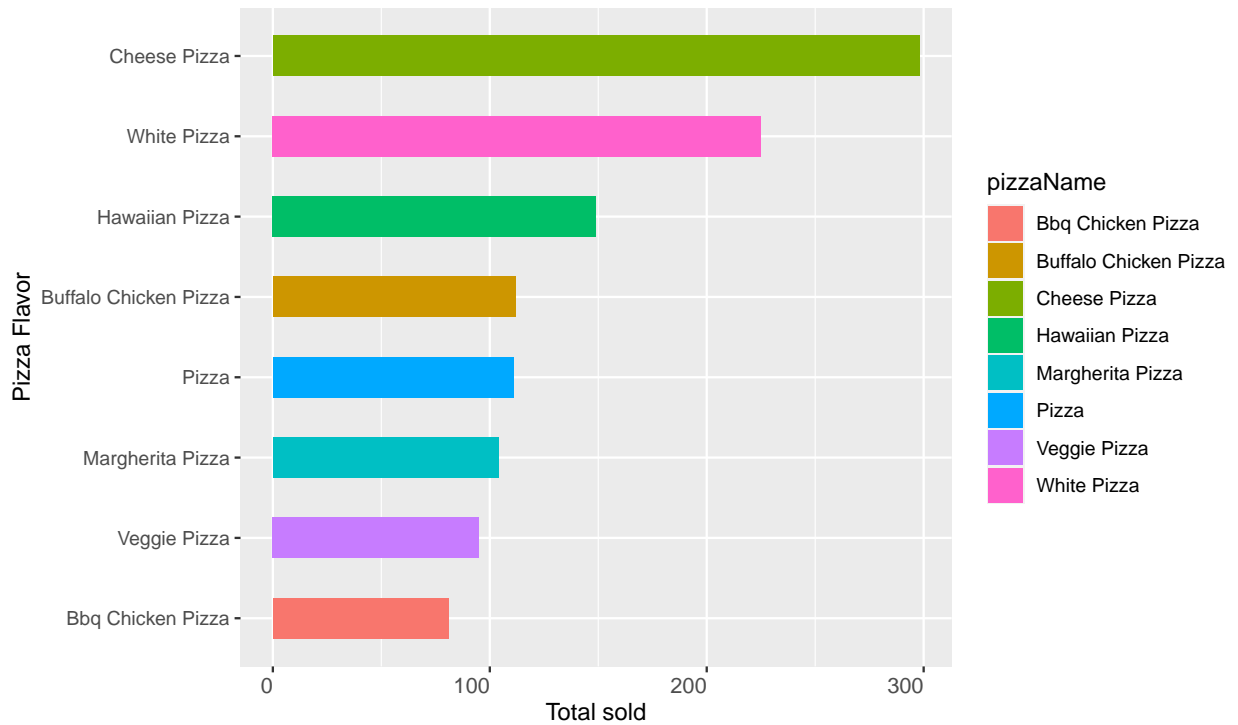
Pizza price in different state



City having more than 100 pizza places



Commonly sold pizza menu across states



6 Model details

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 is not an impossible obstacle to overcome, since we will be learning about modelling packages in R later in the course.