Grocery Store Analysis

Database Management

Jimmy Anglim, Raima Jamil,
Shengke Lian, Nesar DSouza

Introduction

- H Mart is an Asian grocery store chain looking to expand to new counties in Virginia. The business wants to look at publicly available data to find counties where there is potential to open new stores.

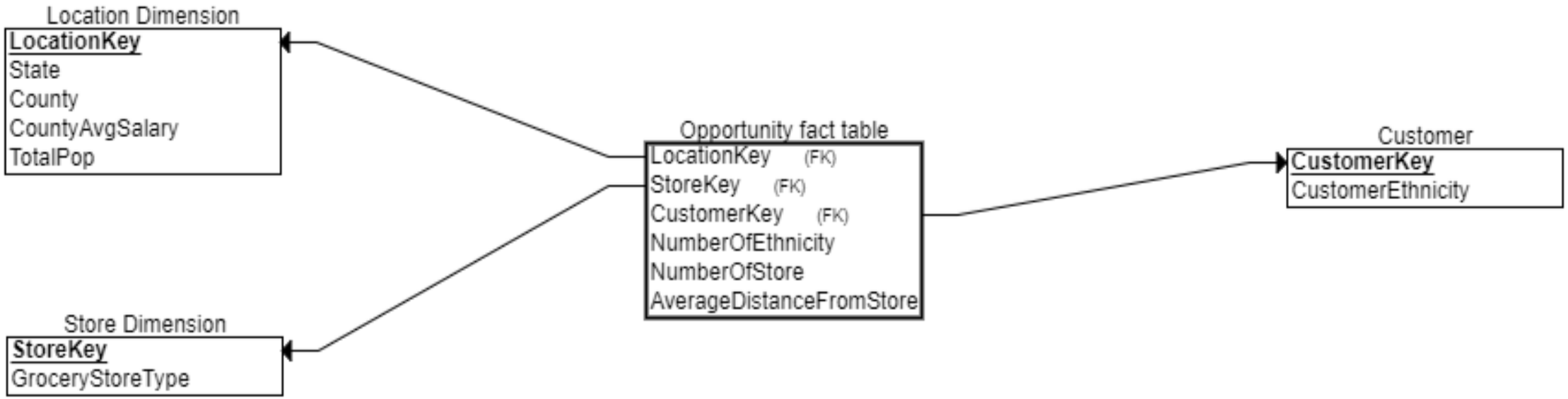
- H Mart will look at the following factors:
  - Household income
  - Number of existing specialty stores
  - Ethnic population distribution

Datasets Used

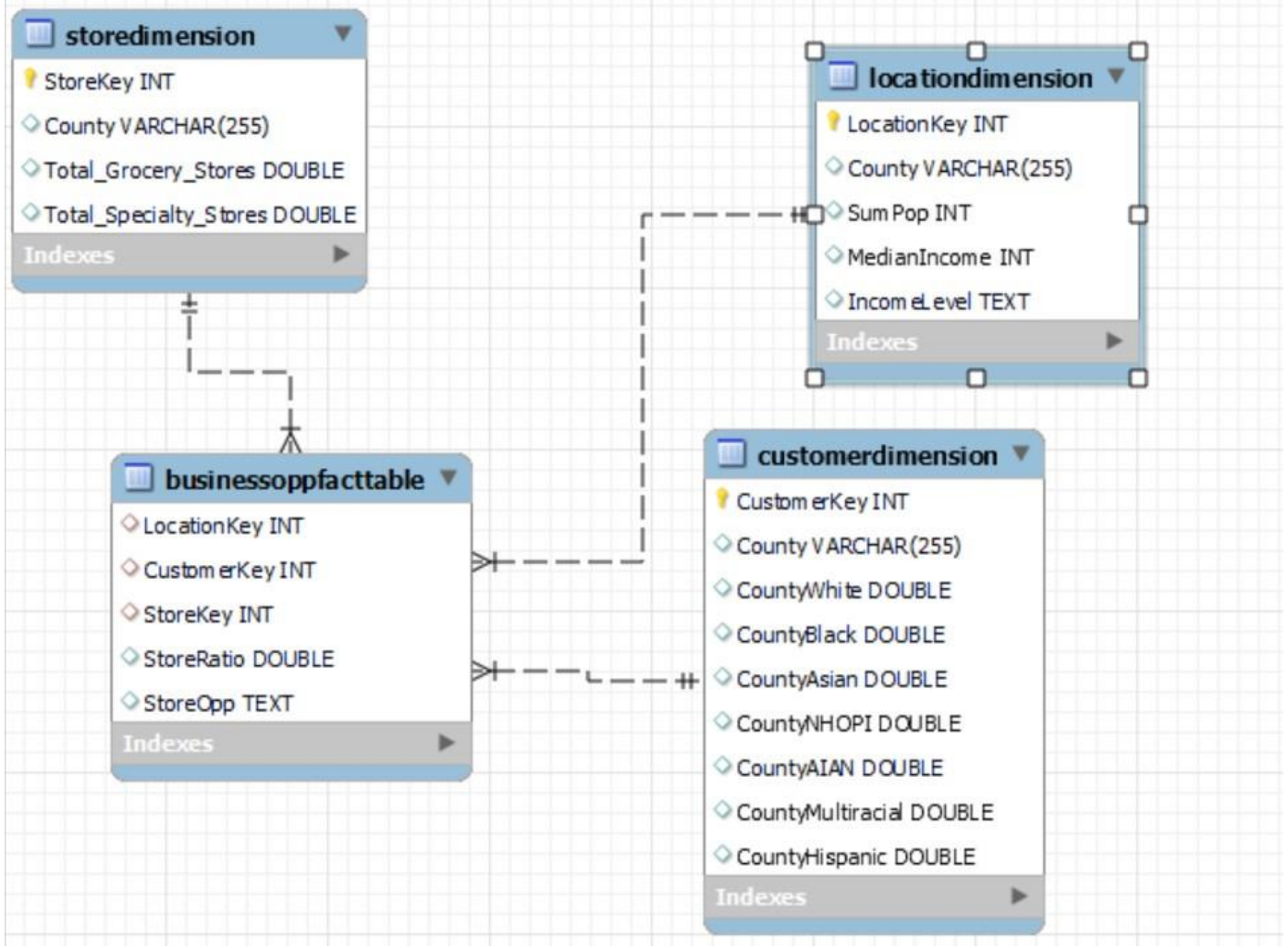
- USDA FoodEnvironmentAtlas - Store Availability (2016)
  - Contains store availability metrics like number of Grocery stores, supermarkets, specialty stores, club stores, convenience stores, etc.

- 2015 Food Access Research Atlas (2015)
  - Contains population, income, and food access data

Initial Schema



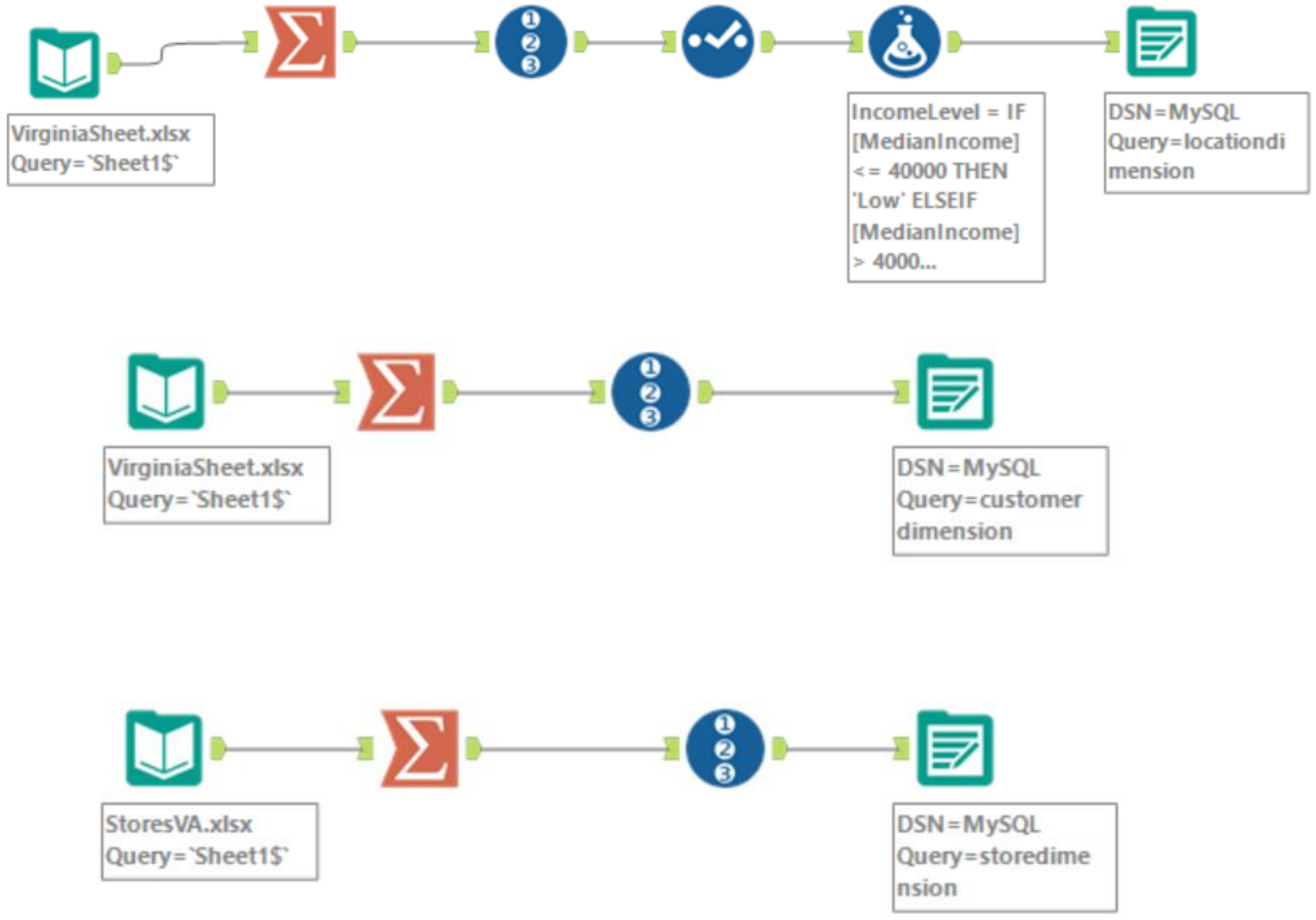
Final Schema



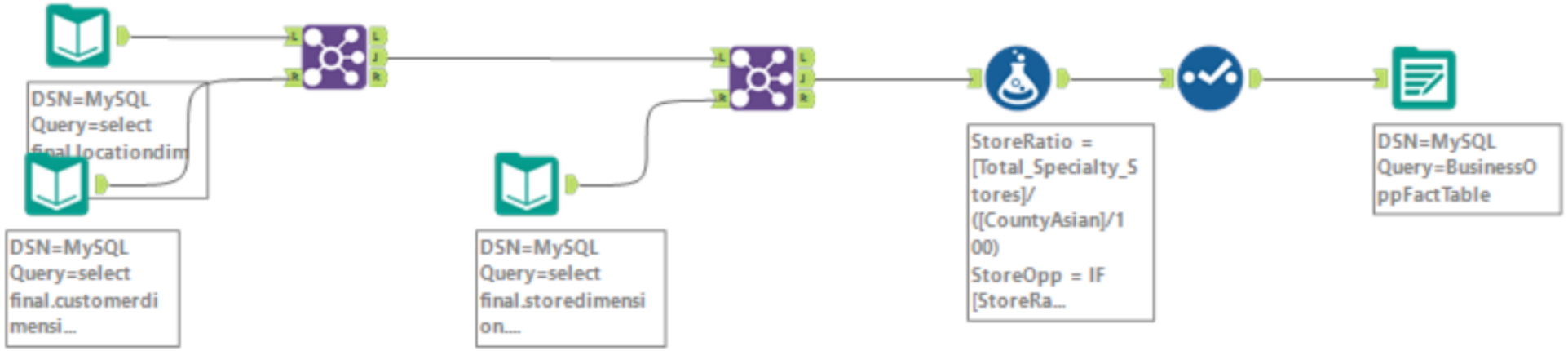
ETL Processes

- We used Alteryx to create the Location, Customer, and Store dimensions
- The Location dimension is composed of the columns State, County, Income, and Total Population
- The Customer dimension is composed of the ethnic distribution of the counties
- The Store dimension is composed of the total number of specialty stores in the counties

ETL Processes



ETL Processes



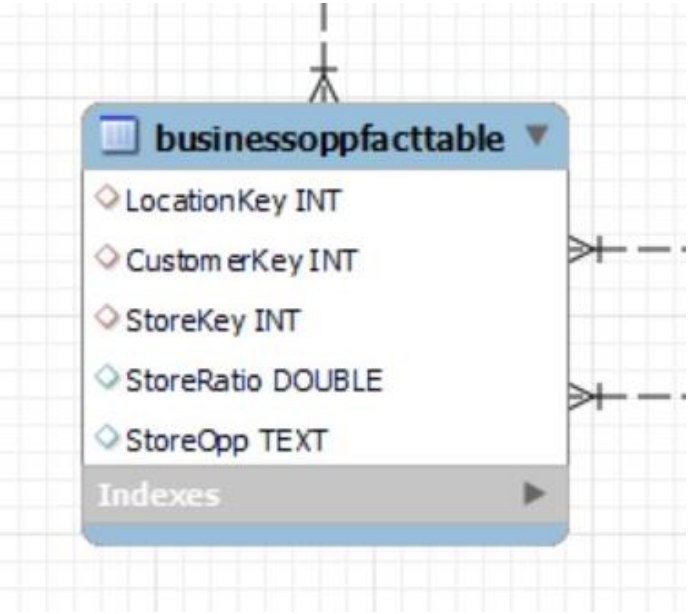
Fact Table

StoreRatio:

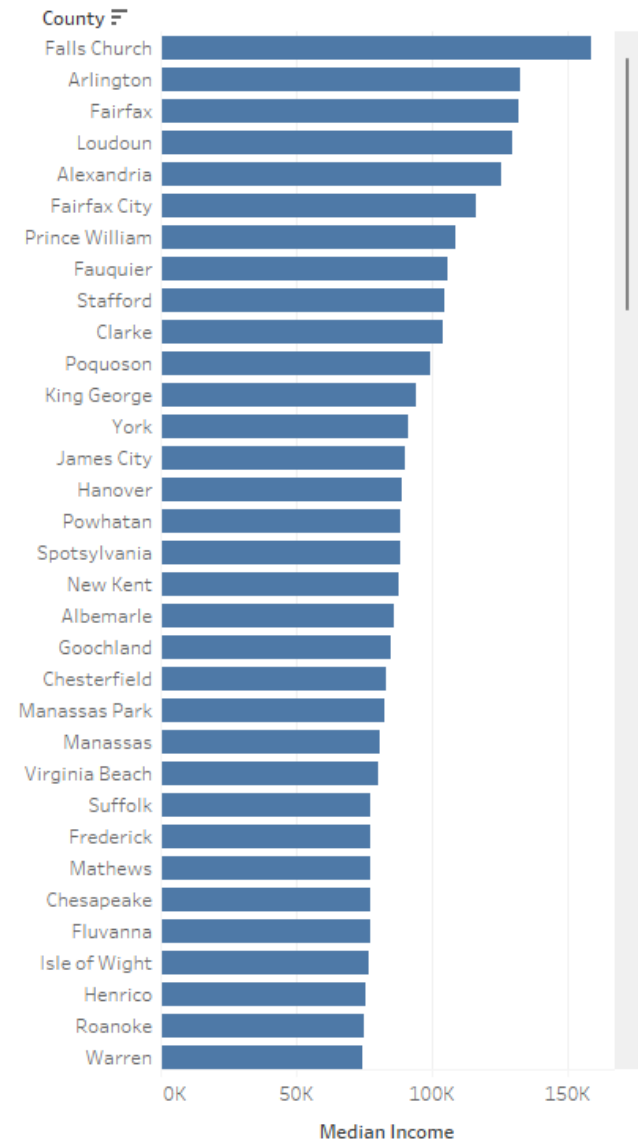
- Calculated as number of stores per 100 Asian people

StoreOpp:

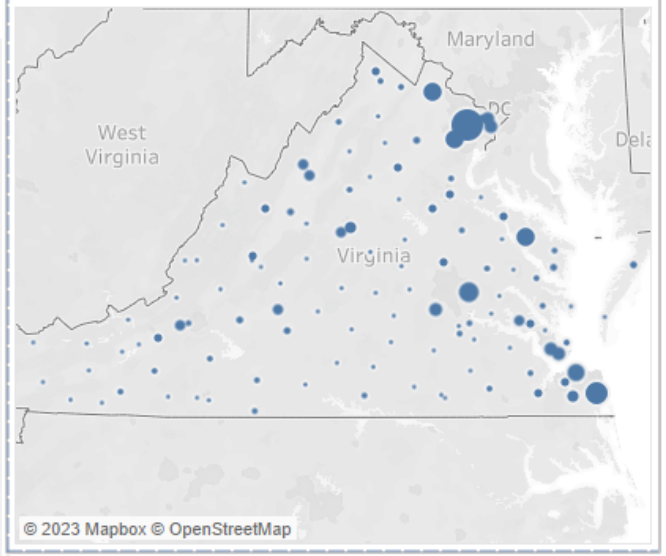
```
IF [StoreRatio] < 1 and [CountyAsian] > 100
and [IncomeLevel] == 'High' or 'Middle'
THEN "Yes" ELSE "No" ENDIF
```



Median Income

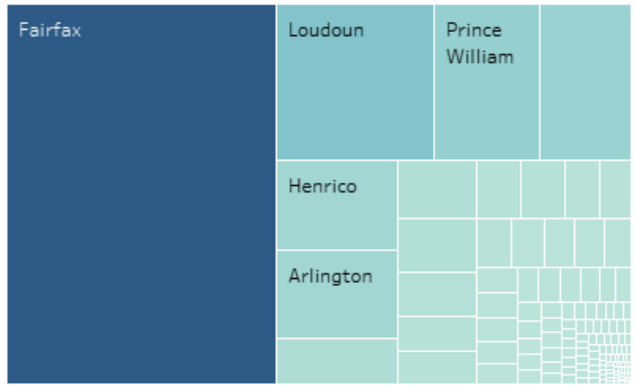


Store Count

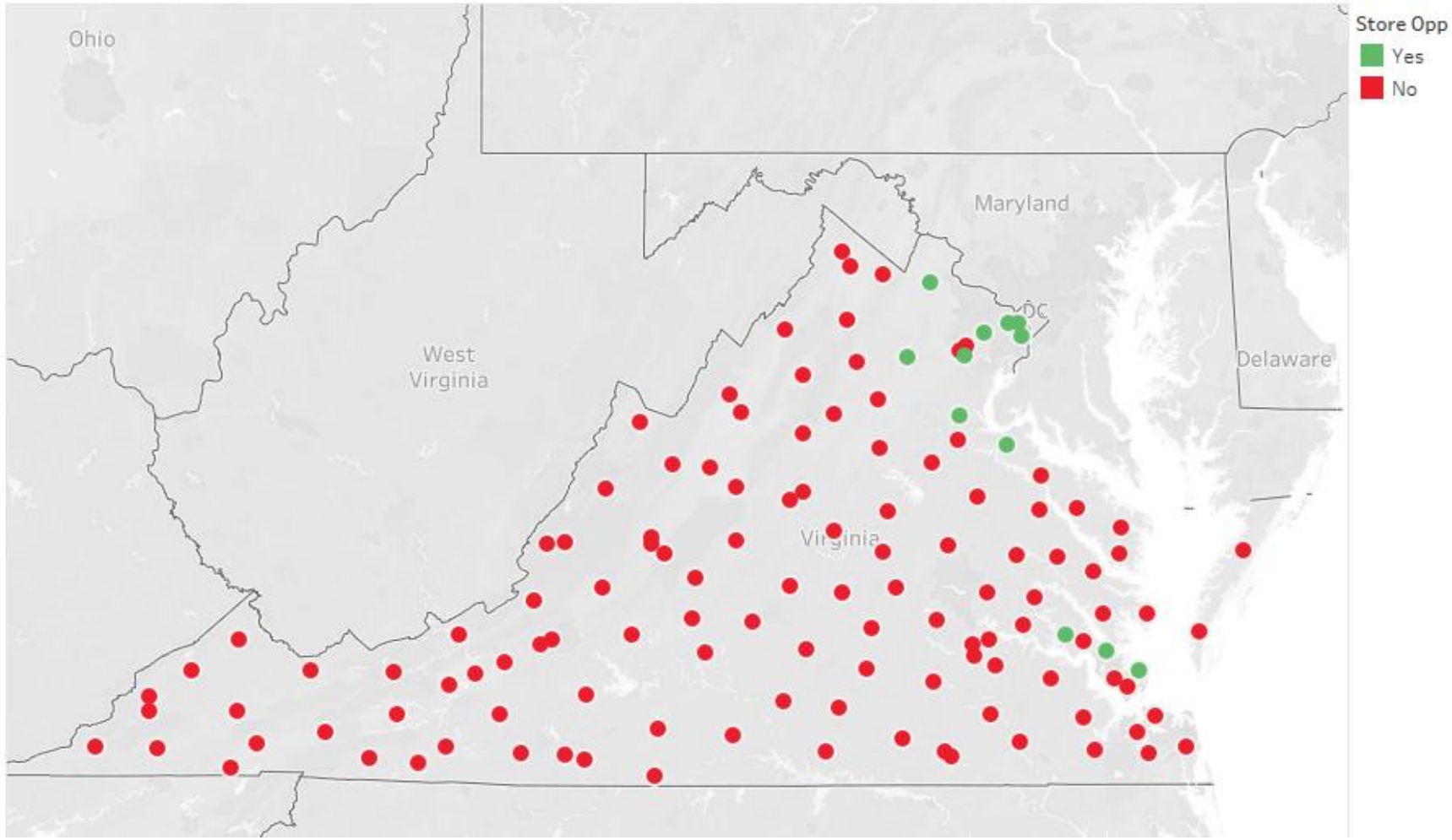


Total Specialty Stores
0.00
20.00
40.00
64.00

Population Map



County Asian
4
189,661


Ohio
West Virginia
Maryland
DC
Delaware
Virginia
Store Opp
Yes
No

Conclusion

Using our model exclusively, H Mart would choose to open stores in 12 counties. However, there are many other factors that influence the decision to open a store. These might include:

- Costs
- Nearby H Mart locations
- Competitors

Our model should be used as an exploratory tool for H Mart

Final Thoughts

Difficulties:

- Understanding the practical differences between dimensions and fact tables
- Creating the dimensional models

Not-so-difficult parts:

- ETL processes
- Creating the tableau dashboard

What did we learn? What would we have done differently?