Target Income Analysis

Dylan Vener

Introduction

Terre Haute is a small college town, housing two universities, Rose-Hulman Institute of Technology and Indiana State University, as well as a population of approximately 61,040 individuals. The town is home to several large chains, prominently Walmart, and after the local KMart closed its doors in the summer of 2015, there was speculation about who would fill the space. One popular theory was that Target might open, as theirs was a notable absence, but a long-held belief in the populace is that Target would not come to Terre Haute because its median household income of $33,317 was too low for Target to be interested.

My goal was to investigate the plausability of this rumor, as well as analyzing how local income affects Target's strategy for choosing store locations in general.

Methodology

In order to answer these questions, I was provided with median household income data for zipcodes in the United States, with the rest of the data used scraped from Target's website, the Federal Reserve Bank of St. Louis Economic Research website, and flowingdata.com.

```r
#Code to scrape Target's website, with the input being a valid state abbreviation
grab_stores <- function(state){
  url <-paste(c('http://gam.target.com/store-locator/state-result?lnk=statelisting_stateresult&stateCode
  temp.pg <- read_html(url)
  stores.df <- as_data_frame(fromJSON(html_text(html_nodes(temp.pg,'#primaryJsonResponse'))))$storeList)
  return(stores.df)
}
stores <- lapply(state.abb,grab_stores)
stores.df <- bind_rows(stores)
stores.df$zipCode <- strtrim(stores.df$zipCode,5)
```

Once this data was collected, I took in the provided data source, from the American Community Survey, and did some minimal cleaning

```r
income.df <- read_csv('~/Statistical Programming/ACS_14_5YR_B19013_with_ann.csv',col_names=c('ID','zipCo
  filter(med.inc!='-') %>%
  transform(med.inc=as.integer(med.inc))
income.df <- as_tibble(income.df[complete.cases(income.df),])
income.df$target <- ifelse(income.df$zipCode %in% stores.df$zipCode,'Yes','No')
```

To answer the latter question, I strove to look at how Target's strategy for choosing locations has changed over time. To do this I downloaded historical data from the St. Louis website, which contained the Per Capita Income(PCI) of fifty large cities, annually, from 1965 to 2014.

```r
city.df <- read_csv('~/Statistical Programming/city_names_t.csv')
city.df <- gather(city.df,year,pci,`1969-01-01`:`2014-01-01`)
city.df$year <- parse_date(city.df$year)
city.df <- city.df %>% group_by(year) %>% mutate(z.score = scale(pci))
```

I did some minimal cleaning, as well as calculating each cities PCI Z Score within each year.

As well, I found a dataset published by FlowingData.com, by Nathan Yau, which includes three columns, Target store opening date, and the City and State it was built in.

```
dates.df <- read_csv('~/Statistical Programming/opendate.csv')[,c('Open Date','City','State')]
dates.df$`Open Date` <- parse_date(dates.df$`Open Date`,'%m/%d/%Y')
#Nearest new year to the opening date, for easier joining with city.df
dates.df$year <- round_date(dates.df$`Open Date`,unit = 'year')
```

I then did an inner join of these two data sets, leaving me with 207 cities pci and Z Score, and date of opening.

```
final.df <- inner_join(dates.df,city.df,by=c('City','year'))
```

Conclusion

Using the scraped Target data and the ACS, I created the following density chart, wherein locations where a Target exists is shown in blue, and location where Target does not have a location is show in red, with the color only beginning after Terre Haute's median income of $33,137.
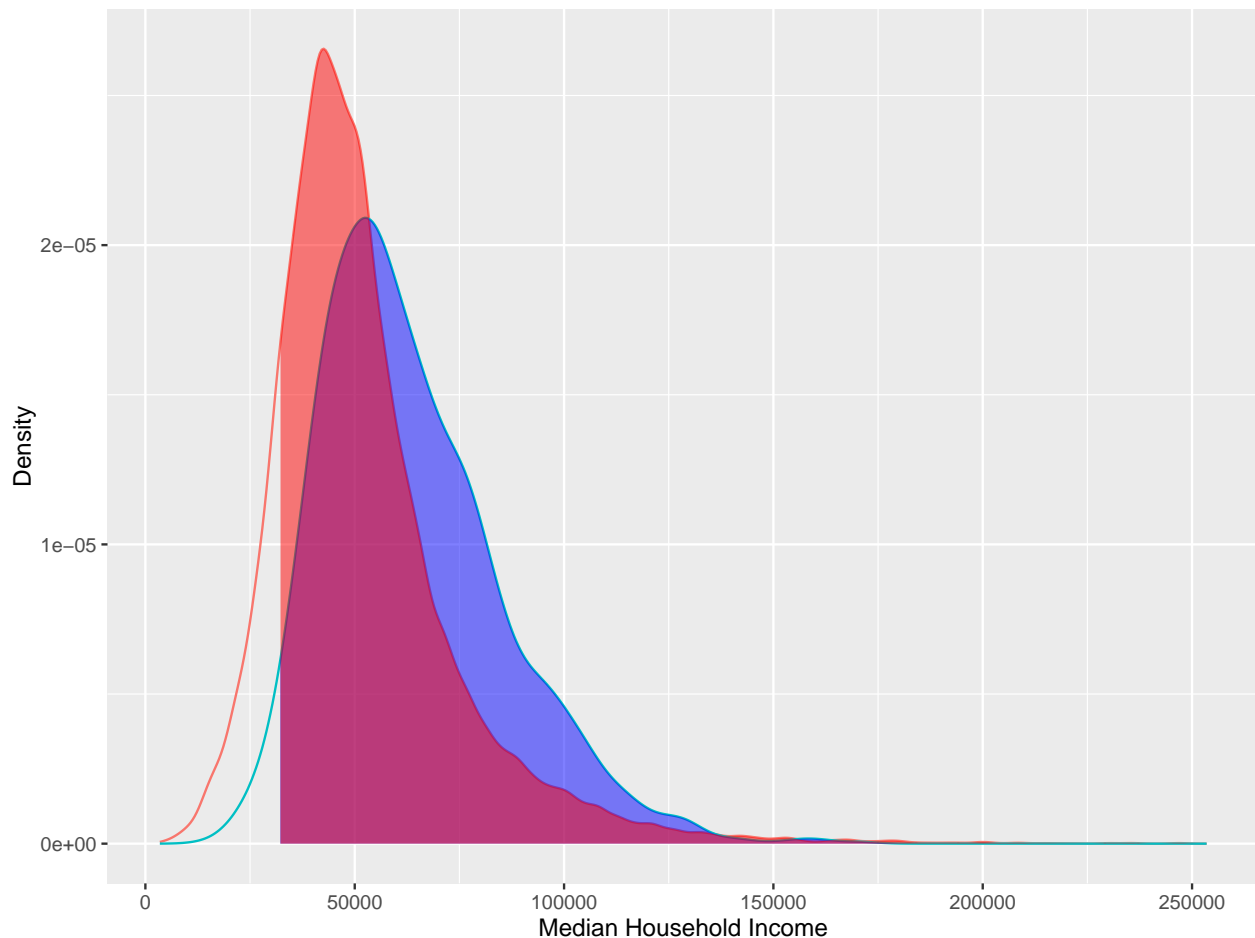
As one can see from this graph, the places where Target builds stores have a higher mean income, $63,707.33 vs $51,923.9, though they have a very similar standard deviations, $21,758.56 vs $22,284.78. Importantly, one can clearly see that Terre Haute, with an income of $33,317, is well below both of these, and while there are Targets in locations with lower incomes, notably, the lowest income location is only $15,405, with a Z Score of -1.39, it would be definitively on the lower end of the spectrum. From this, we can say that Terre

Haute is explicity not below a strong threshold for there to be a Target, but it is likely going to be less likely chosen as a location.

From the historical data, we can see that Target has mantained a similar strategy when choosing locations over the past fifty years, staying very uncorrelated with respect to income in the major cities for which I have date.
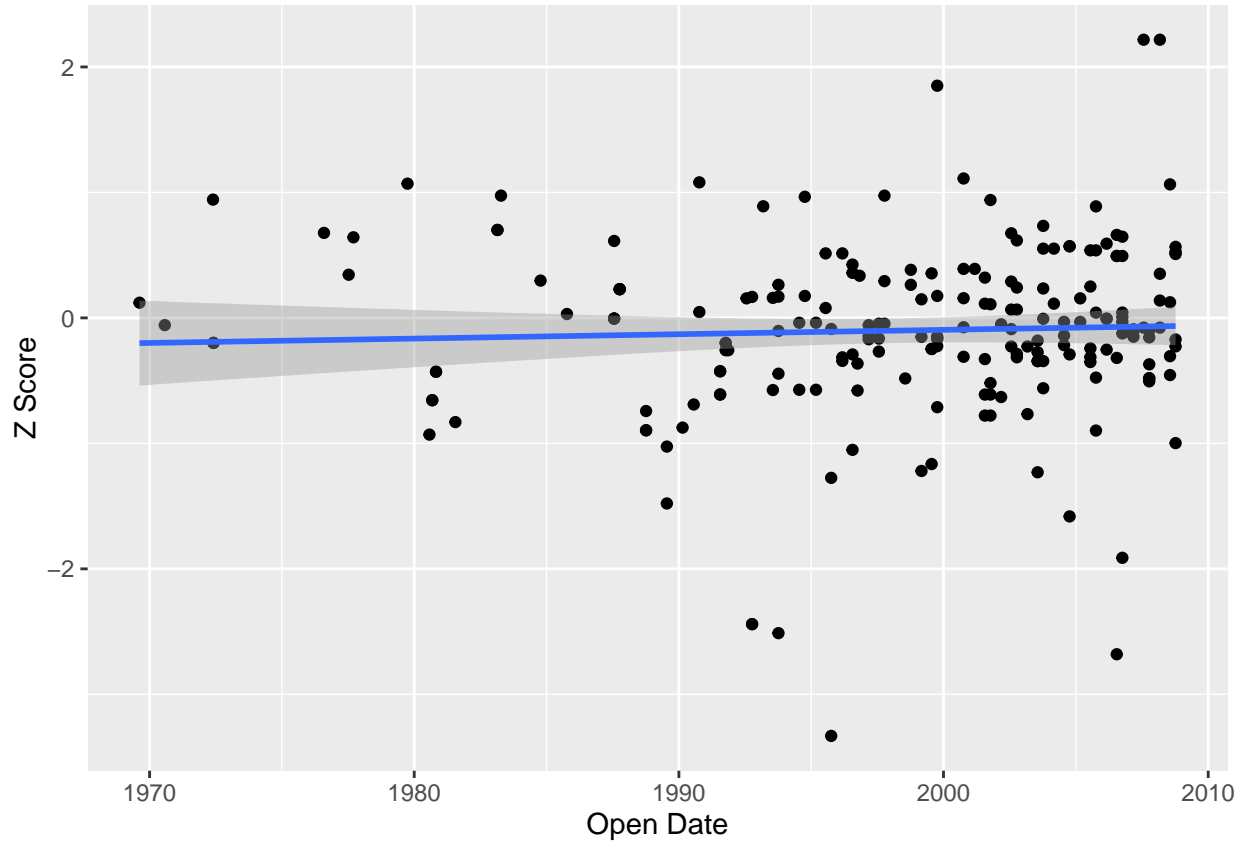
This is a scatter plot of the Z scores of the cities where Target has opened, from the year they opened.

For further demonstration of this fact, we have the following graph, which shows how the PCI for cities have grown over time, and how Target has chosen stores over time.
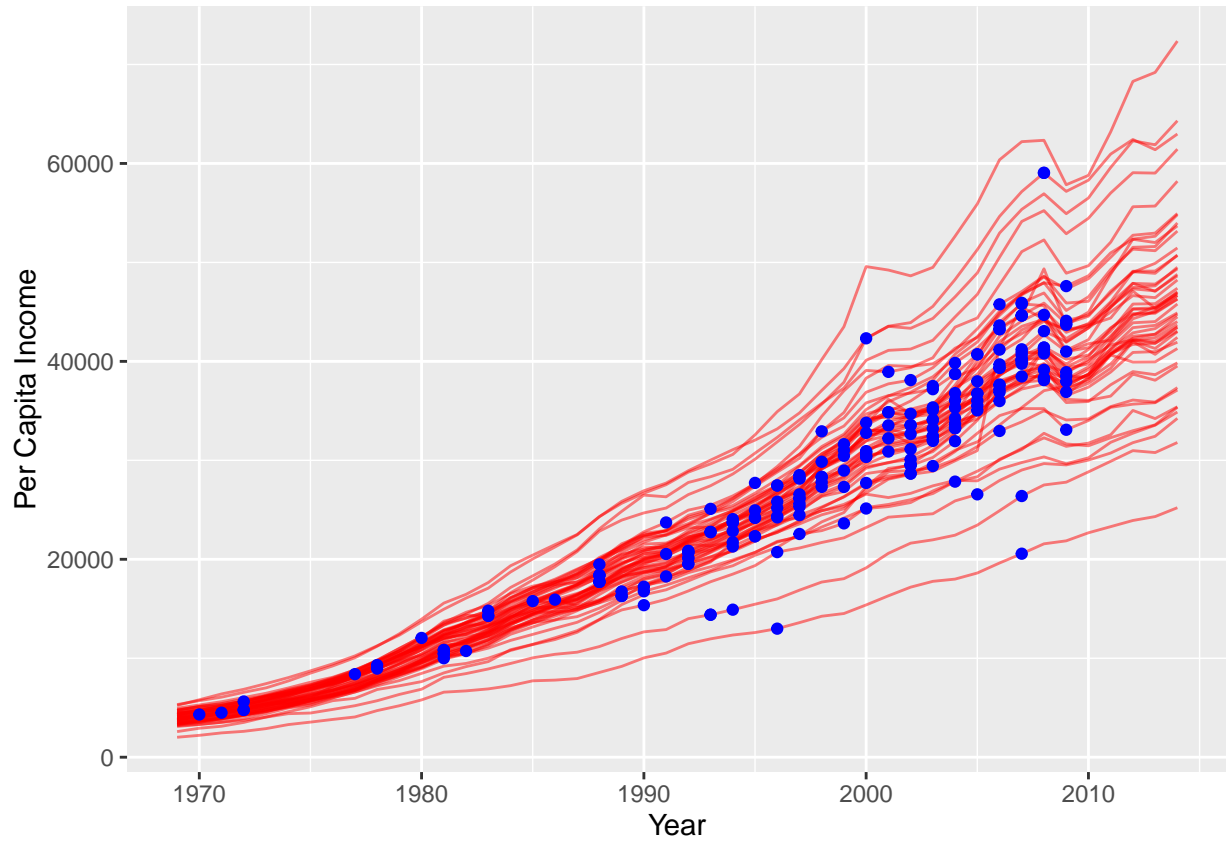
The major shortcomings of this analysis arise from two features. Firstly, analysis at zipcode level might overloop larger trends in an area, for example, a store might be openend in an Industrial area next to a large population center, leading to a lower family income in that zipcode, but a significantly higher income if one looks at the population which shops at that store. Secondly, for the analysis of the time series for cities, the analysis suffers from the population drawing only from large cities, which, because of their large size, all are likely to be larger than any cutoff Target would have, while smaller town's, where a stronger effect might be observed, are absent from the dataset obtained and analyzed.