Sneaker Data Mining

Zhengyi Lin

5/9/2022

Abstract

The report focuses on the premium for reselling sneakers on the popular website StockX. We are interested in what characteristics determine the resale price premium. Using the Random Forest machine learning model, we were able to accurately predict the resale price of specific limited quantities of Nike and Adidas shoes based on shoe characteristics.

Introduction

Pulling in $70 billion in 2020, the sneaker market has a powerful influence within American consumer goods. Because of the high demand for these sometimes rare and unique shoes, a powerful resale market has also emerged. The sneaker resale market was worth as much as $2 billion in 2019, a figure that has only increased as more and more players try to get in on the sometimes over 2000% profit margin earned from the rarest of sneakers.

Why is this relevant? Premiums are a quick and simple benchmark to measure the profitability and desirability of a specific sneaker. Many characteristics, such as colorway, brand, size, and material can make or break a shoe sale. The physical characteristics of shoes are not the only determining factors for premiums, much like other retail goods, shoe sales have a seasonality component as well. This makes understanding the timing of a sale crucial. Premiums can demonstrate to resellers which characteristics make a shoe more profitable. Premiums can also be useful to buyers: based on characteristics, what price is a good deal and what prices border on irrational?

Methodology

Part I: Data Descriptions

The final dataset used in this project is located in shoe_final.csv

Scripts used to merge variables and clean data are located in cleaning.R

The main data for this project was sourced from the popular online sneaker marketplace, StockX. The dataset contains the details of 99,956 orders of *Adidas Yeezy* and *Nike x Off-White* shoes on StockX from September 2017 to February 2019. Each row represents a unique sale on the website. The variables associated with each sale are: *Buyer Region* (State), *Order Date*, *Brand*, *Sneaker Name*, *Retail Price*, *Sale Price*, *Release Date*, and *Size* (StockX lists shoes in mens' sizing).

Premium was created from this initial dataset using the difference between resale and retail price, and *Relative Premium* is the relative change in price from retail to the eventual order price.

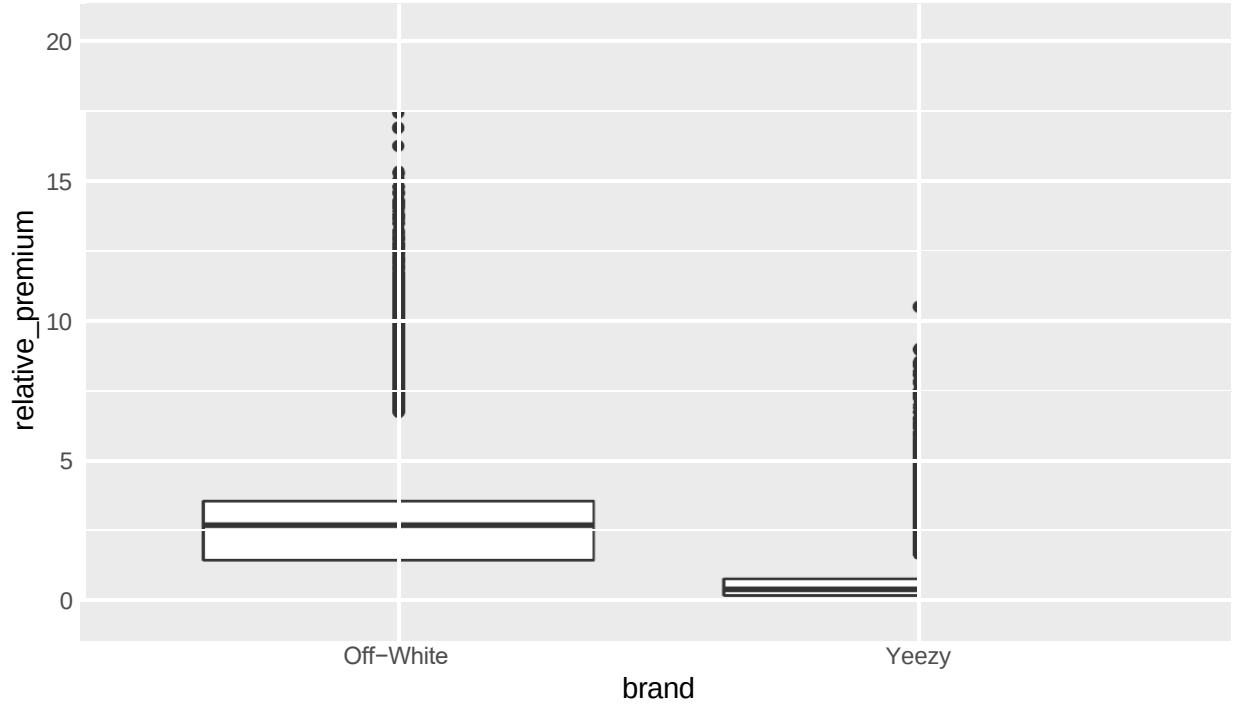
We collected additional variables regarding characteristics of each shoe including: *Material*, *Lace Type*. *Primary Color*, *Secondary Color*, and *Tertiary Color*. Primary color represents the dominant shade while secondary and tertiary colors are extra accent or trim colors associated with the sneaker.

Because preferences for shoes could depend on economic or personal financial conditions, we added the variables: *Sporting Goods Index* (Monthly), *USA Monthly Retail Sales Index* (Monthly), *State Disposable Income per Capita* (Yearly), and *State Population* (Yearly). These demographic variables were collected from the U.S. Census Bureau, the Federal Reserve, and the Bureau of Economic Analysis.

To address the geographical component of our data, we figured that there is a cultural component to preferences in sneaker purchases. The emergence of a big resale market for exclusive Nike and Adidas shoes may be associated with interest in athletics and basketball. After all, the designs of all of these shoes were either created with a purpose to either run of play basketball or derived from other sneakers made for that purpose. Included in our data is also the historic performance of each NCAA Division I basketball team aggregated by state. The variables used are: *Overall Win Loss Percentage Win Percentage 2019*, *Win Percentage 2018*, *NCAA Championships*, *AP Final Poll* (Number of appearances on final AP rankings), *AP Rank 2019* (Average ranking on final AP poll of each team), and *AP Rank 2018*.

Part II: Summary Statistics

brand	min	first_quantile	median	mean	third_quantile	max	sample_size	sd
Off-White	0.015	1.432	2.684	2.83	3.546	20.3	27794	1.89
Yeezy	-0.155	0.218	0.436	0.64	0.814	10.5	72162	0.67



Compared to Adidas' Yeezy shoes, the Nike x Off-White shoes across the board require a higher resale premium and price volatility. In 557 Yeezy orders, we saw shoes sold for less than retail. There is no such thing in Nike's order.

Across the data set, the average relative premium for Nike X Off-White was about 284 percent, and for Yeezy it was about 64 percent. In our data set, Nike x Off-White shoes accounted for 27,794 of 99,956 orders (about 28%), while Adidas Yeezy shoes accounted for 72,162 orders (72%).

Figure 3: Monthly Average Relative Premium, Over Time

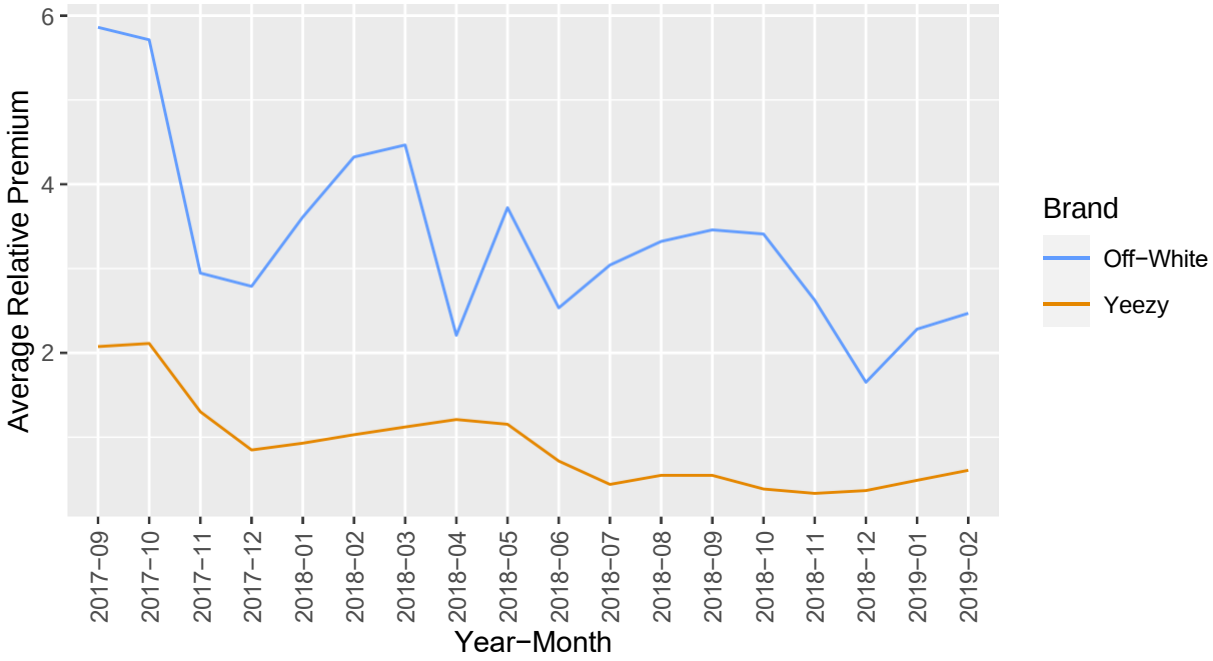
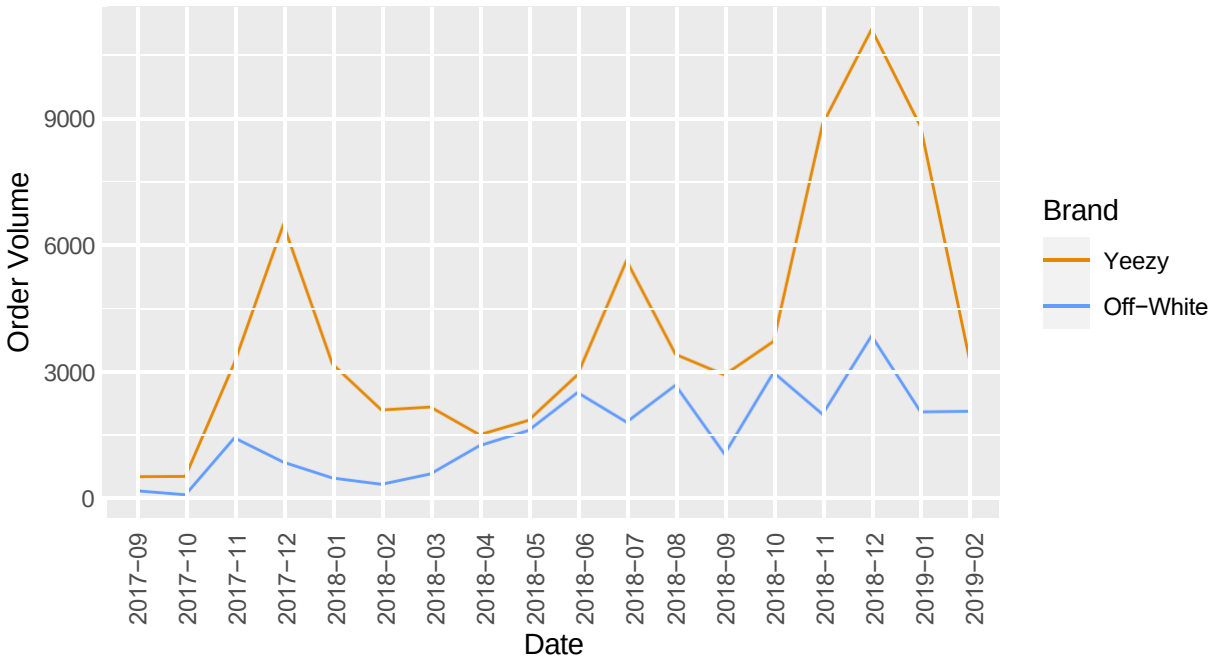


Figure 1: Monthly Order Volume



Plotting the average premium over time, we can see that there again appears to be a seasonality effect. *Figure 3* displays the average premium by brand over time. Interestingly, the average premium seems to dip

for each brand around the holiday season. This could be because the orders for Yeezys spike up much more significantly compared to Nikes as shown in *Figure 1*. The downward trend of premiums over time could be due to a variety of factors: possibly more people are selling on StockX over time, driving premiums down as sellers compete for consumers. Another factor driving down premiums could be that there is more stock of shoes being put out by Nike and especially Adidas that eventually make their way into the resale market.

A look at monthly order volume by brand (*Figure 1*), reveals a definite seasonal pattern, with orders spiking for both brands around the holiday season in both 2017 and 2018. The data also exhibit non-seasonal spikes in order numbers that appear to be linked to specific product release dates and restocks. For example, we believe the July 2018 spike in *Yeezy* orders could be associated with the late June release of the *350 V2 "Butter"*. It should be noted that the steep decline in orders around February 2019 is due to the data ending in the middle of the month.

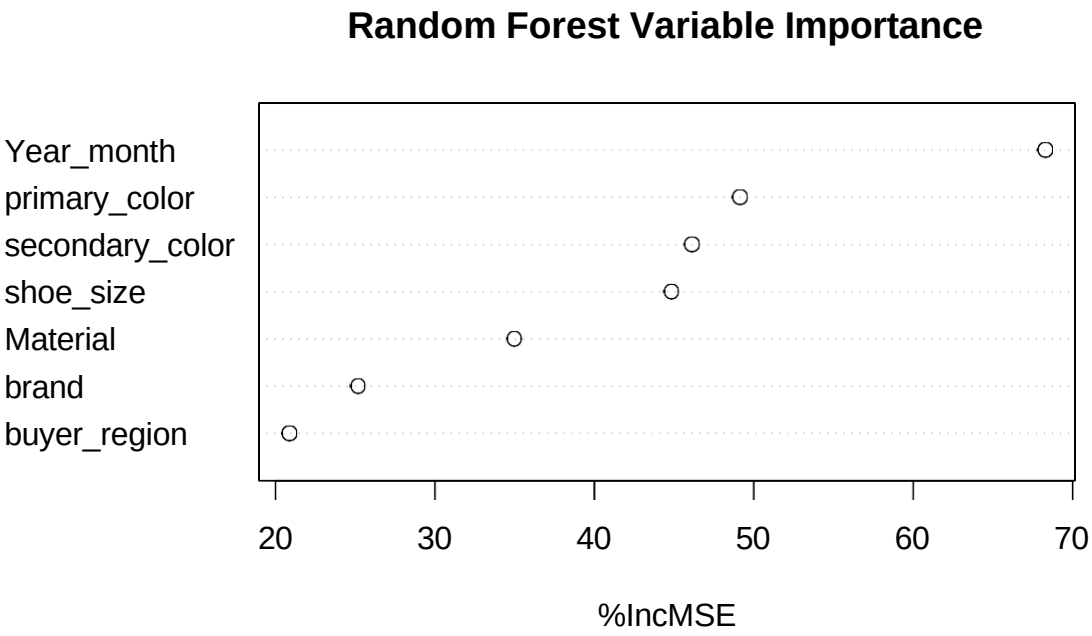
Part III: Random Forest Model

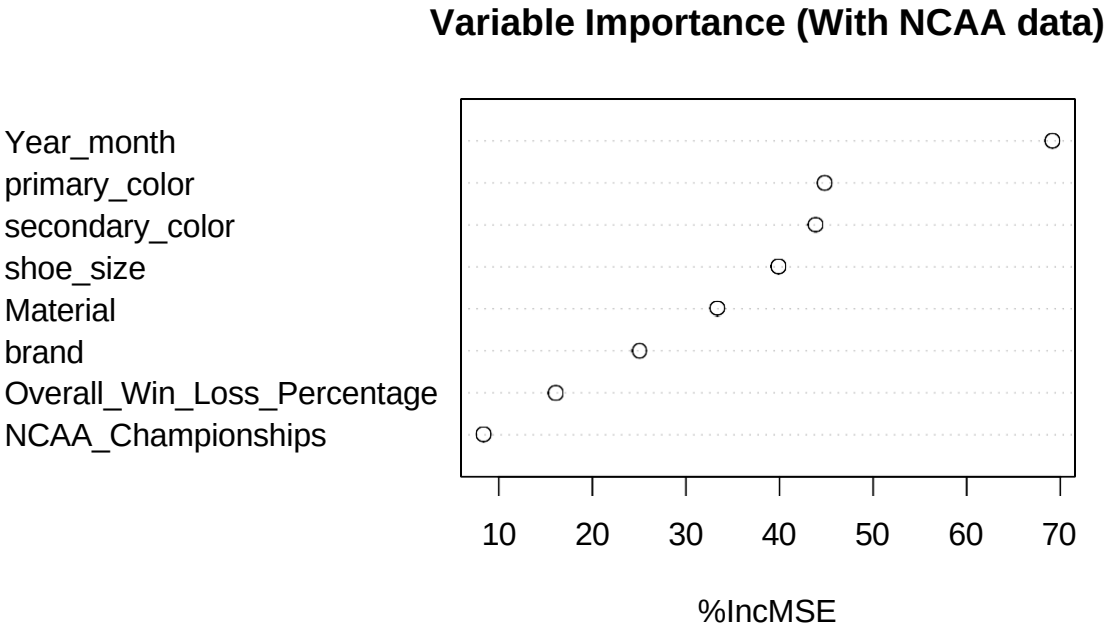
Our goal is to identify the most important predictors of resale premiums. Due to the large amount of data and heterogeneity of variable types, we decided to adopt random forest model. Since it is difficult to quantify individual preferences for buying sneakers, Random Forest will be able to find and consider every combination of interactions and sneaker characteristics. For our pattern, our complexity parameter is set to 0.002 and we use 300 trees. For cross-validation, our data was divided into a training set and a test set, with 20% of the data reserved for testing.

Relative Premium is this case is our dependent variable. With *Sneaker Name*, *Size*, *Buyer Region* (State), *Order Date* (grouped by month), *Primary Color*, *Secondary Color*, and *Material* as our independent variables. The complexity parameter for our decision tree model was placed at .02, minimum observations for split at 300, and max depth at 4. For random forest, our complexity parameter was set at .002 and the number of trees set to 300. For cross-validation, our data was split into testing and training sets, with 20% of the data reserved for testing.

Results

Overall Data

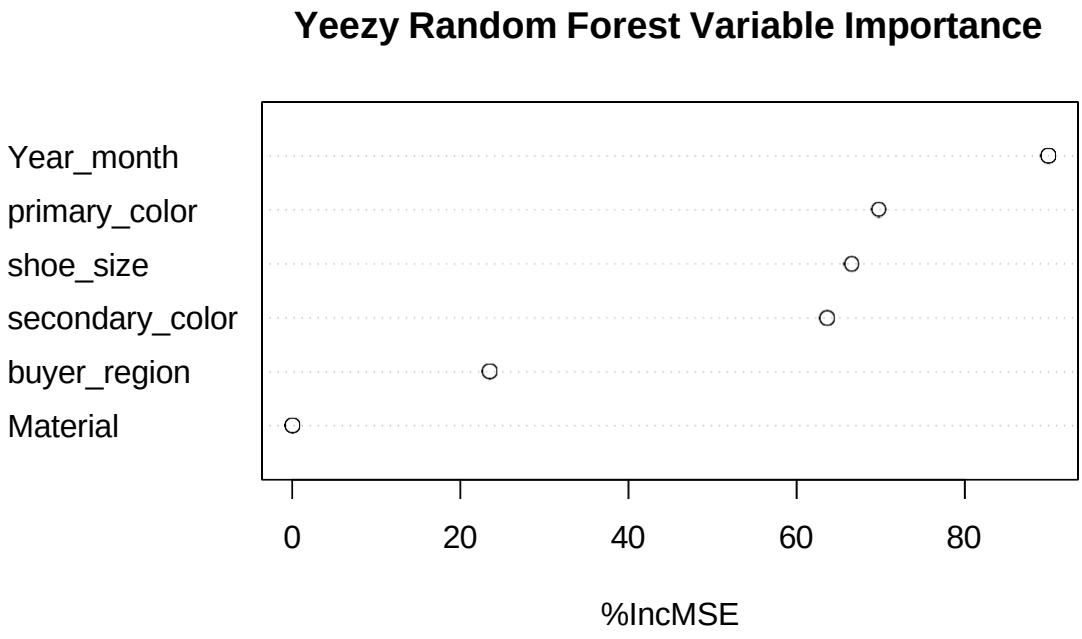


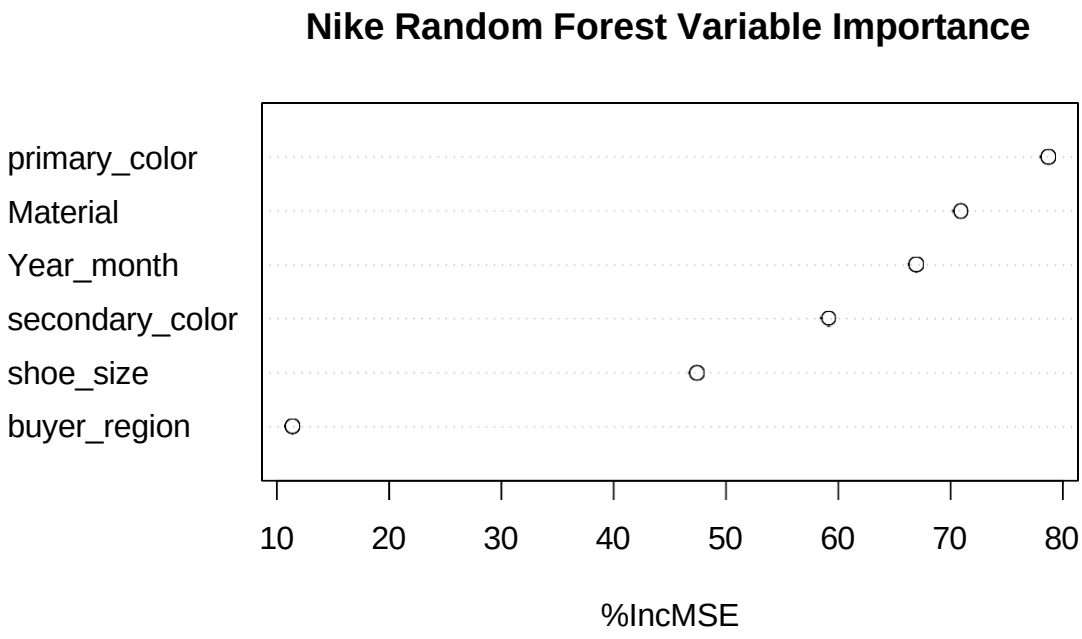


Model	RMSE
Tree	0.835
Forest	0.425
Forest with NCAA Data	0.431

We obtain an out of sample root mean square error (RMSE) of 0.425, but we suspect that the high level of the buyer's region variable bias the model in favor of New York or California prices, whereas the first variable importance plot shows that buyer's region is the least important variable for the accuracy of our model. The second model incorporates NCAA data (aggregated by state) rather than using the states themselves.

Brand Specific





Model	RMSE
Yeezy Forest	0.276
Yeezy Forest with NCAA Data	0.284
Nike Forest	0.539
Nike Forest with NCAA Data	0.540

This model is used to control the overall data for brands, but we get more reliable findings when we separate the data. Looking at the two brands separately, we're looking at shoes that seem to be aimed at two completely different markets, as Nike shoes are in significantly higher demand and have a wider range of prices. Focusing on each brand individually allows the model to more accurately see the importance of each predictor. The RMSE results show that we can predict Adidas' sales premium more accurately than Nike.

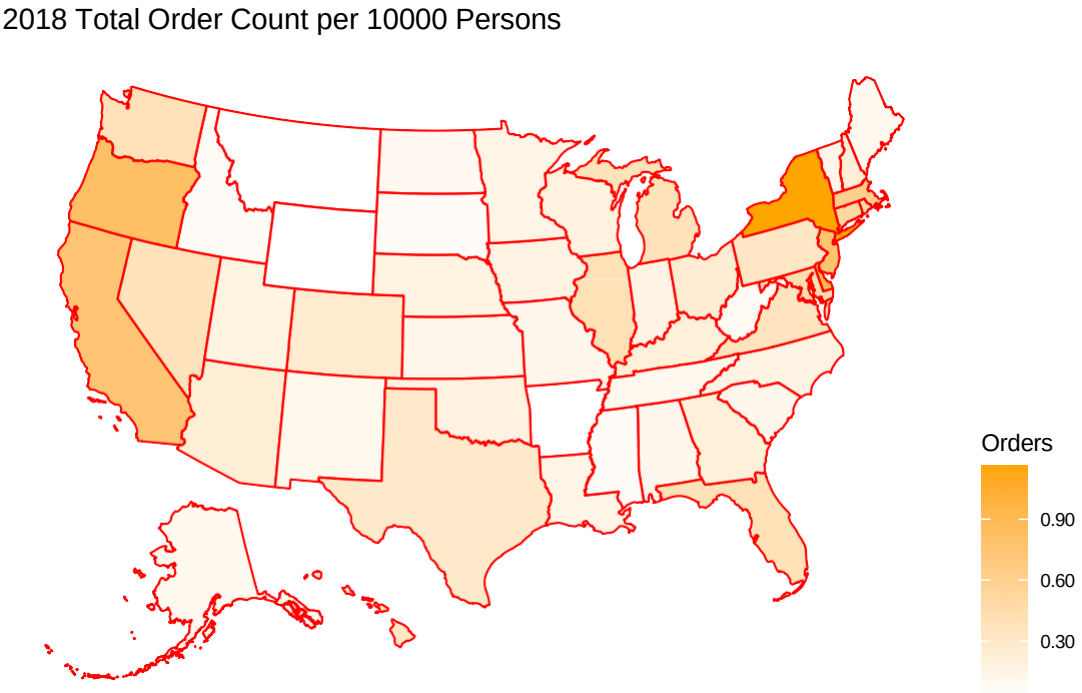
Appendix

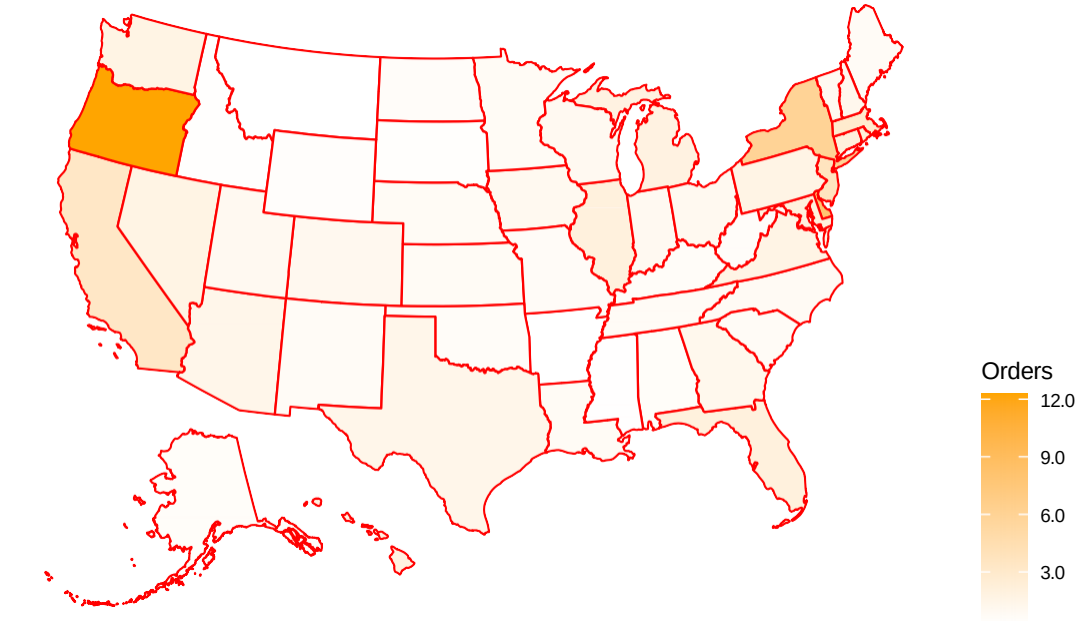
A.1 Linear Model

Model	RMSE
Overall	0.826
Yeezy	0.365
Nike	1.185

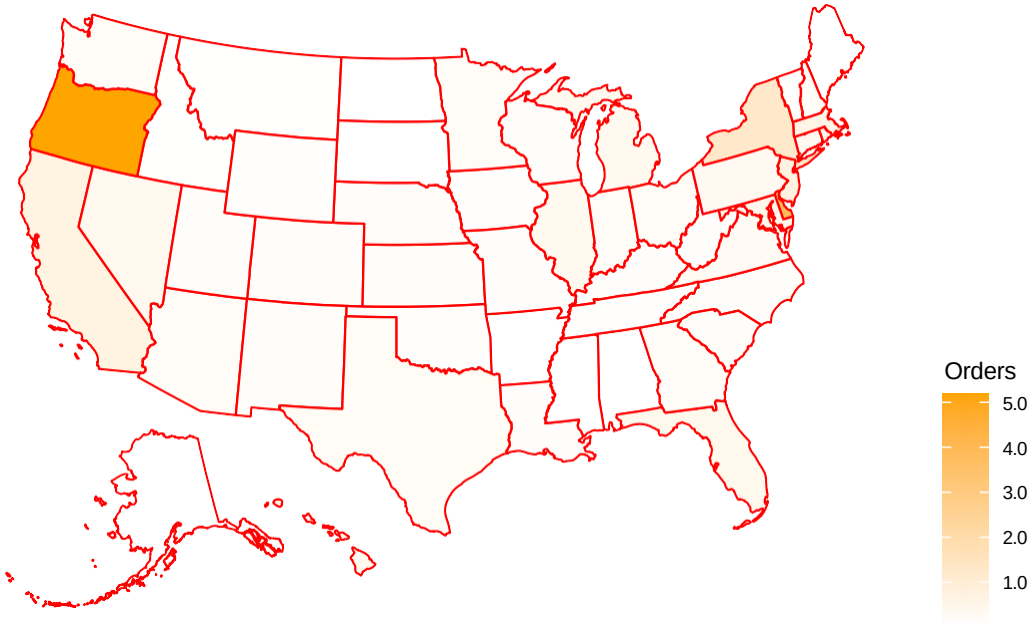
We tried to use the OLS regression model to predict premiums because there are a large number of categorical variables in our data and the random forest model requires a lot of computing power, which makes it difficult to run on some of our machines. The results are in the same direction as those we get with random forests, but the error is higher. Using the same shoe characteristic variables, NCAA data and market data, we ran three models: one for overall data and two for each brand.

A.2 Oregon's high sales volume per capita





2019 Total Order Count per 10000 Persons



One interesting fact we found in our analysis was that Oregon's order volume was significantly higher than other states. We controlled for population and disposable income to see if there was anything in the data that could explain the odd placement of Oregon's order numbers.

Oregon orders more sneakers per capita than any other state even when controlling for income. While the internet has made sneaker culture a global phenomenon, Oregon specifically likely has a high share of sneaker enthusiasts, perhaps related to Portland being home to the headquarters of Nike, LaCrosse, Dr. Martens, and the North American headquarters of Adidas, Li-Ning, and more. Furthermore, Portland has found itself to be a [id5][major hub] for the outdoors and shoe industry.