Estimating Prescription Drug Costs

•••

Eleana Cabello and Emuobosa Ojoboh

Problem and Data Set Description

Problem Introduction

- The U.S. is globally known for having extremely high prescription drug prices.
- People cannot afford to purchase their prescription regularly.
  - Ration the supply they can purchase.
  - Avoid purchasing it all together in order to make ends meet elsewhere.
- Companies like GoodRx and Mark Cuban's Cost Plus Drug Company have made efforts to increase price transparency with patients.
- It is worth investigating how the manufacturing process affects the resulting cost of a drug.

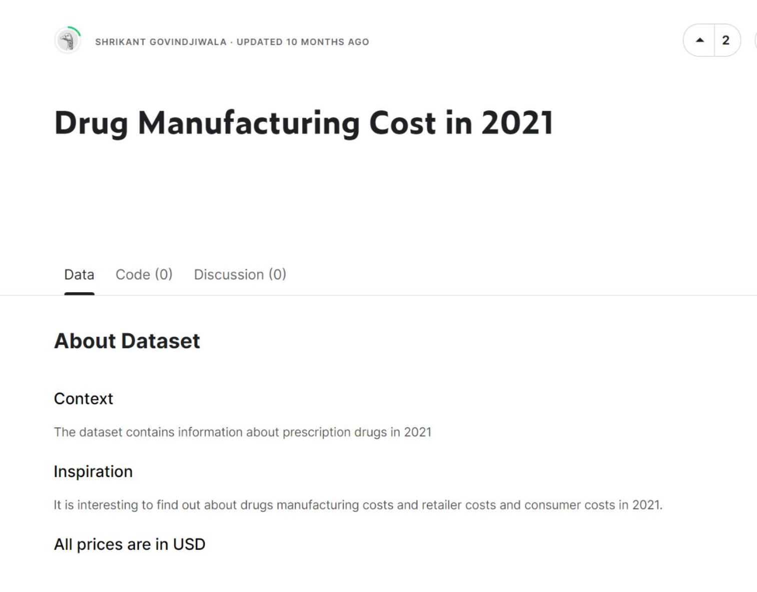
Project Goals

Our project goals are to use a publicly available dataset about prescription drug costs to:

1. Analyze patterns within the data that could explain their high prices.
2. Create a model that is able to predict drug costs with high accuracy.

Dataset Description

- 2,984 records
- Attributes:
  - Medication Name
  - Package size
  - Manufacturing cost
  - Max retailer price
  - Max consumer price
  - Max consumer VAT price
  - Year



Drug Manufacturing Cost in 2021

About Dataset

Context

The dataset contains information about prescription drugs in 2021

Inspiration

It is interesting to find out about drugs manufacturing costs and retailer costs and consumer costs in 2021.

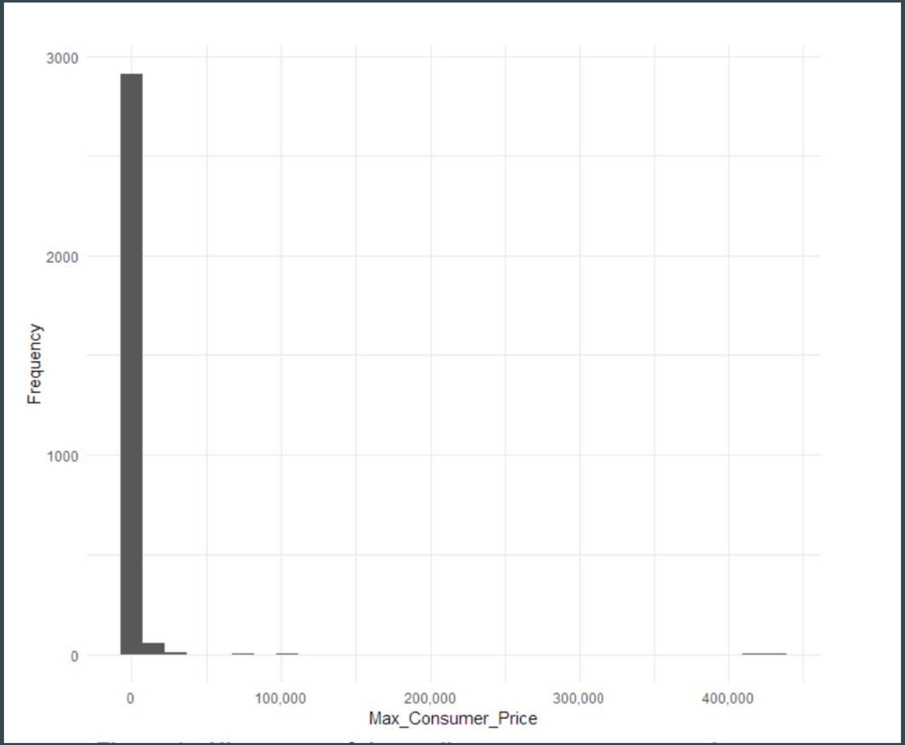
All prices are in USD

Dataset from Kaggle.com

Initial Analysis

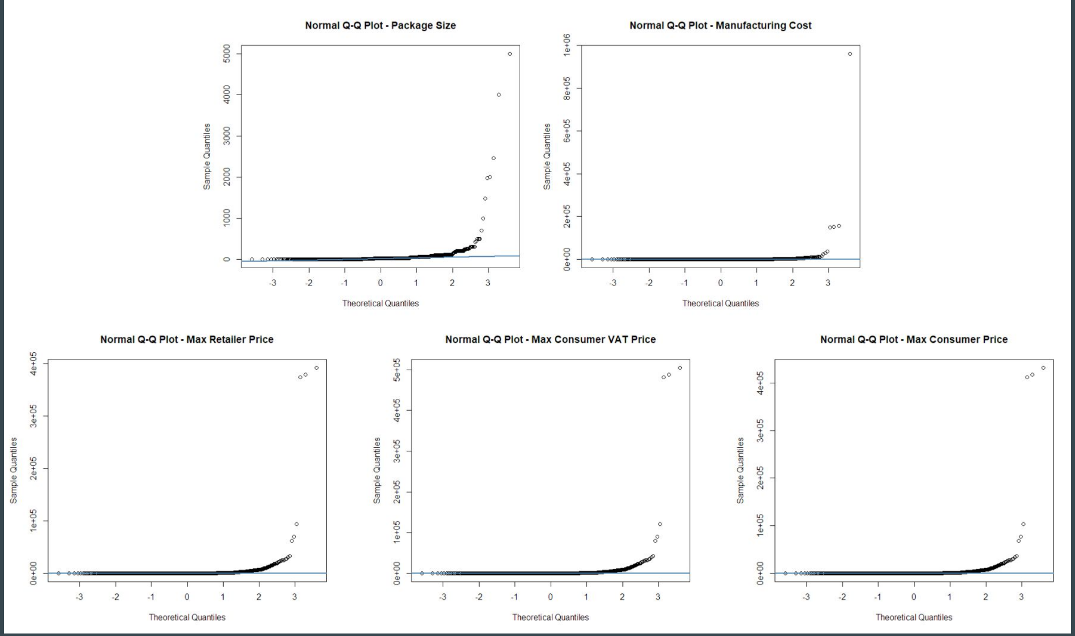
Initial Findings

- No missing values in any of the attributes.
- Skewness in the max consumer price variable.
  - ➜ To address this the variable was transformed before being used for modeling.



Initial Findings

- Outliers were visible in Q-Q plots of all the variables and confirmed using a Grubbs test for each.
  - ➜ Drug price is also dependent on the medical condition it is used for and can hold valuable information. Therefore these outliers were not removed.



Normal Q-Q Plot - Package Size

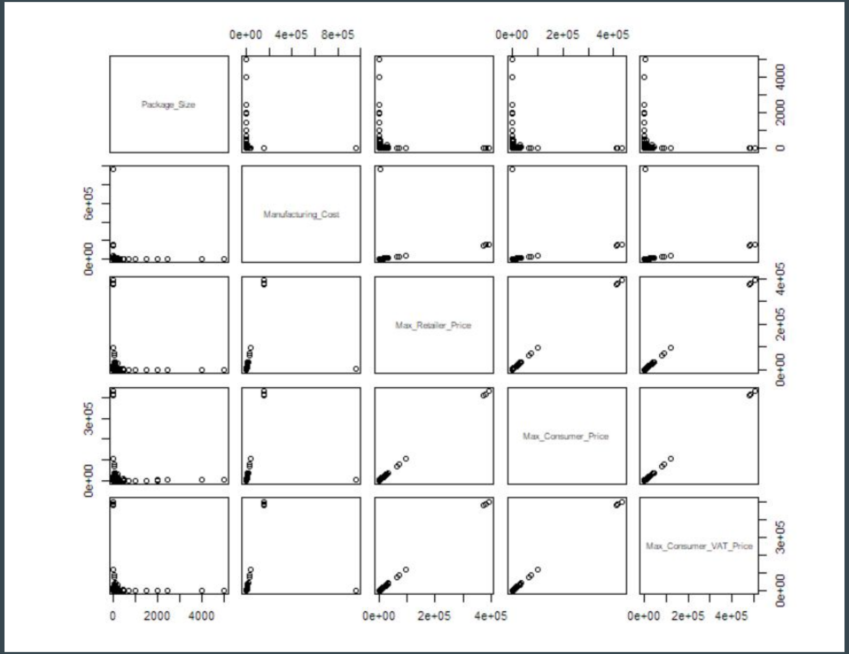
Normal Q-Q Plot - Manufacturing Cost

Normal Q-Q Plot - Max Retailer Price

Normal Q-Q Plot - Max Consumer VAT Price

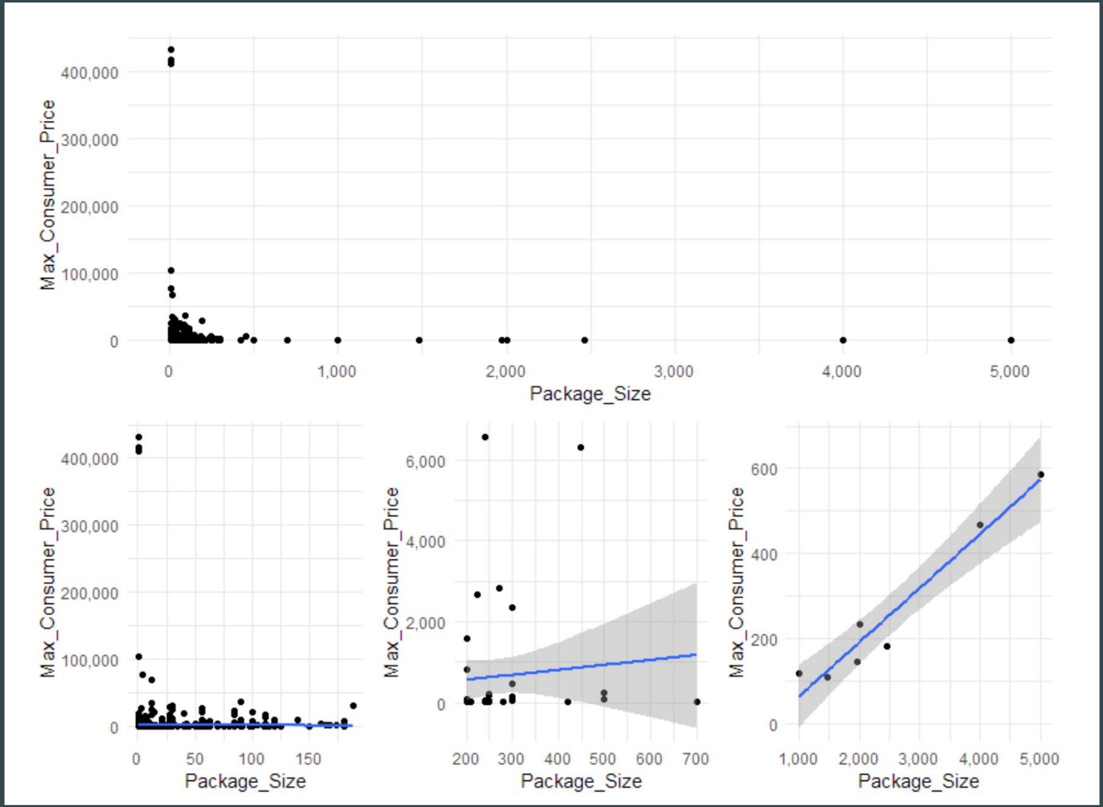
Normal Q-Q Plot - Max Consumer Price

Relationships Between Attributes

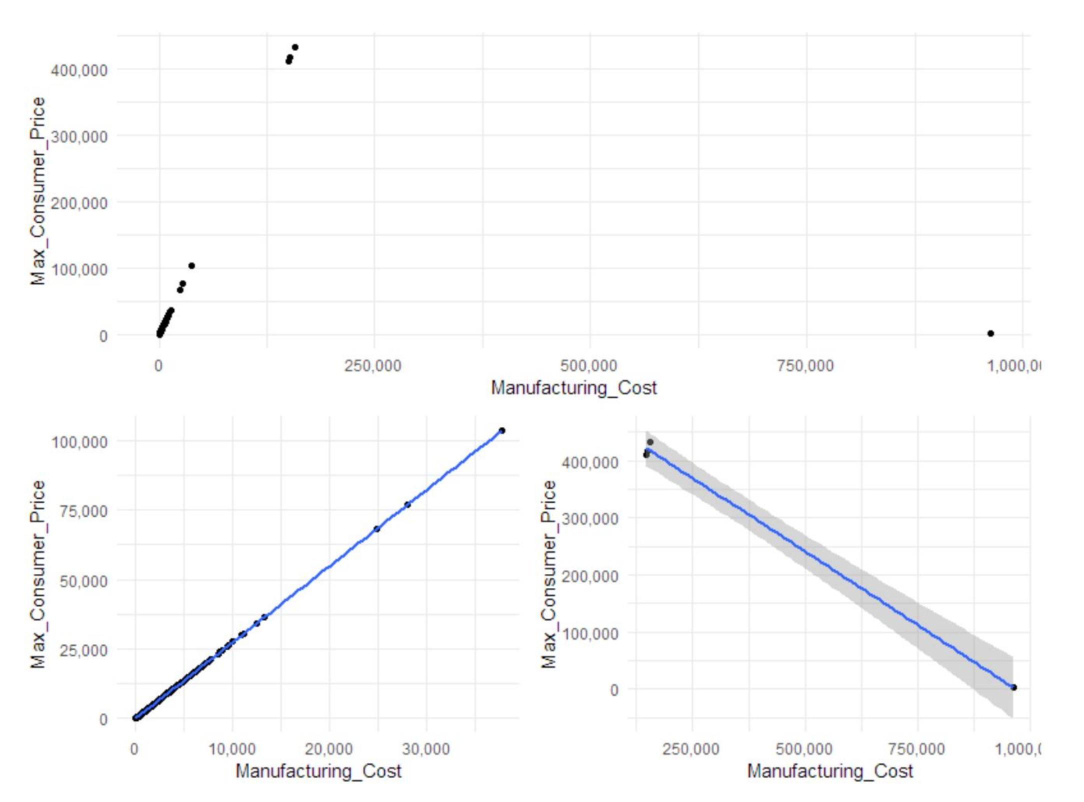


- Max retailer price, max consumer price, and max consumer VAT price displayed a linear relationship.
- Manufacturing cost vs max retailer price, max consumer price, and max consumer VAT price seem to show an increasing exponential relationship.
- Packaging size vs manufacturing cost, max retailer price, max consumer price, and max consumer VAT price seem to show a type of decreasing exponential relationship.

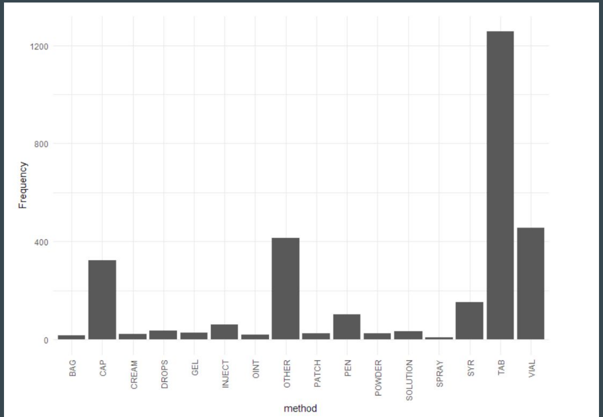
Max Consumer Price vs Package Size

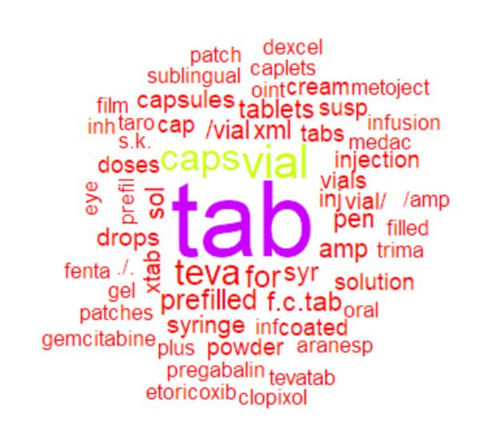


Max Consumer Price vs Manufacturing Cost



Text Mining Approach





Word cloud of the most come distribution methods of the prescription drugs.

Tablets
Vial
Capsules
Syringe
Injection
Pen
Drops
Powder

Solution
Cream
Patch
Gel
Ointment
Spray
Bag

Modeling Approach and Results

Data Preparation

- Year was dropped as all the records belonged to the same year.
- Medication name was dropped as the number of unique values would have made the analysis of it difficult.
- Max consumer price was transformed using the equation below to address the skewed distribution of its values.

$$\log(max\_consumer\_price + 1)$$

- One-Hot encoding of the attribute method was conducted to create several new variables indicating the method used for the drug.
- 85% of the dataset was used for training and 15% was used for testing.

Initial Results

Generalized Linear Model

Attributes that significantly influenced max consumer price:

Package_Size

Max_Retailer_Price

method.CAP

method.DROPS

method.OTHER

method.SYR

method.TAB

method.CREAM

method.GEL

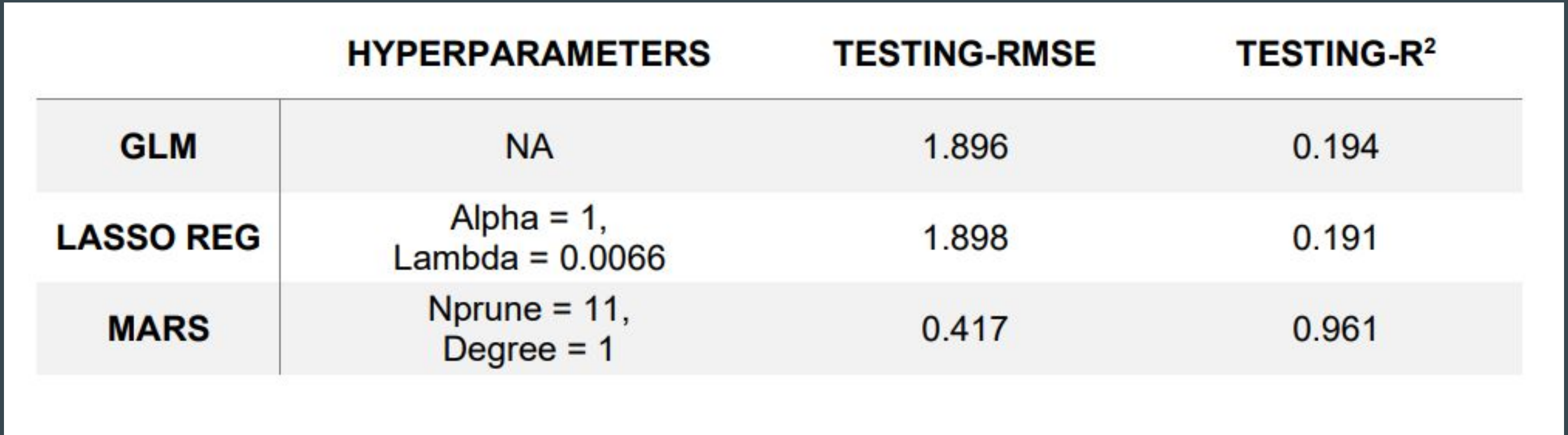
method.OINT

method.SOLUTION

Modeling Results

- A LASSO Regression model and MARS model were trained using this set of optimal features.

- 10-fold cross validation was used to tune their hyper parameters.



	HYPERPARAMETERS	TESTING-RMSE	TESTING-R^2
GLM	NA	1.896	0.194

LASSO REG	Alpha = 1,
	Lambda = 0.0066
		1.898	0.191

MARS	Nprune = 11,
	Degree = 1
		0.417	0.961

Limitations and Future Work

- No information in the data set on the medical conditions the prescription drugs treat.
- Manually gather data or combine with another publicly available data set.

Thank you!