University of Oklahoma

DSA 5103 – Intelligent Data Analytics

Dr. Charles Nicholson

Estimating Prescription Drug Costs
Final Draft

Eleana Cabello

Fall 2022

CONTENTS

EXECUTIVE SUMMARY

Problem Statement

This project aims to use a publicly available dataset on drug prescription prices to find information on what significantly influences the high cost of prescription drugs.

Summary of Findings

- Packaging size and maximum retailer price largely influenced this dataset's cost of prescription drugs.
- The following distribution methods of drugs highly influenced their price: capsules, drops, syringes, tablets, creams, gel, ointment, and solution.
- The Random Forest model predicted this dataset's maximum consumer price with the highest accuracy.

Recommendations

The dataset used for this project had no information on what condition the prescription drugs were used to treat. In future studies, these associations should be considered and included in the analysis.

PROBLEM BACKGROUND

Problem Introduction

The United States is globally known for having extremely high prescription drug prices compared to other countries. For many people, their medication is essential to continue living, leaving many stuck having no option but to pay these high prices. In some cases, people cannot afford their prescription regularly, forcing them to ration the supply they can purchase or not purchasing it all together. Either case can severely endanger one's life.

However, there have been efforts to lower these costs and provide more price transparency. One example is the app GoodRx which allows users to compare the cost of their prescription at several neighboring pharmacies. More recently created, Mark Cuban’s Cost Plus Drug Company offers generic prescription drugs at supplier pricing, typically lower than it would be at a pharmacy, with a fixed markup of 15%.

In recognition of these efforts, it is worth investigating how the manufacturing process affects the resulting cost of a drug. In this project, an open dataset about drug costs will be used to analyze patterns within the data that explain prescription drug prices.

Drug Costs Dataset

- The dataset contains information on the manufacturing cost, consumer cost, retail cost, and distribution details about a wide variety of prescription drugs in 2021.
- The dataset consists of 2,984 records.

- Predictors:
  - Medication_Name – Name of medication including dosage and distribution information.
  - Package_Size – Size, or units, of packaging . (Number of tablets, vials, capsules, etc.)
  - Manufacturing_Cost – Cost to manufacture the drug.
  - Max_Retailer_Price – Maximum retail price of the drug.
  - Max_Consumer_Price – Maximum consumer price to purchase the drug, not including taxes.
  - Max_Consumer_VAT_Price – Maximum consumer price to purchase the drug including the value-add tax. The value-add tax is a goods of service tax typically used by member states of the European Union.
  - Year – Year the record was taken.
- Dataset URL: https://www.kaggle.com/datasets/shrikantgovindjiwala/drug-manufacturing-cost-in-2021

Project Goals

- Identify underlying correlations and relationships between the different data attributes using varying analysis techniques.
- Create a regression model that can accurately estimate the consumer cost of a prescription drug based on its distribution specifics, manufacturing costs and other associated prices.

DATA QUALITY REPORT

Initial Descriptive Statistics

	MEAN	MIN	Q1	MEDIAN	Q3	MAX	SD
Package_Size	37	1	5	28	30	5000	145
Manufacturing_Cost	774	0	3	13	98	962043	18305
Max_Retailer_Price	1130	1	8	34	245	392488	12550
Max_Consumer_Price	1249	2	12	42	288	431736	13805
Max_Consumer_Vat_Price	1461	2	13	49	337	505131	16152
Year	2021	2021	2021	2021	2021	2021	0

Table 1 – Initial descriptive statistics of the numerical attributes.

Since there was no variation in the attribute year and all records were from 2021, it was not considered in any further analysis of the data set.

Outliers

Each Q-Q plot below belongs to a numerical attribute in the data. Based on their individual visualizations, there were clear outliers in each attribute. A Grubb’s test was used to confirm this as well with the highest value of each respective attribute being defined an outlier. Another trend observed was the resemblance in the Q-Q plots of the max retailer price, max consumer price, and max consumer VAT price.

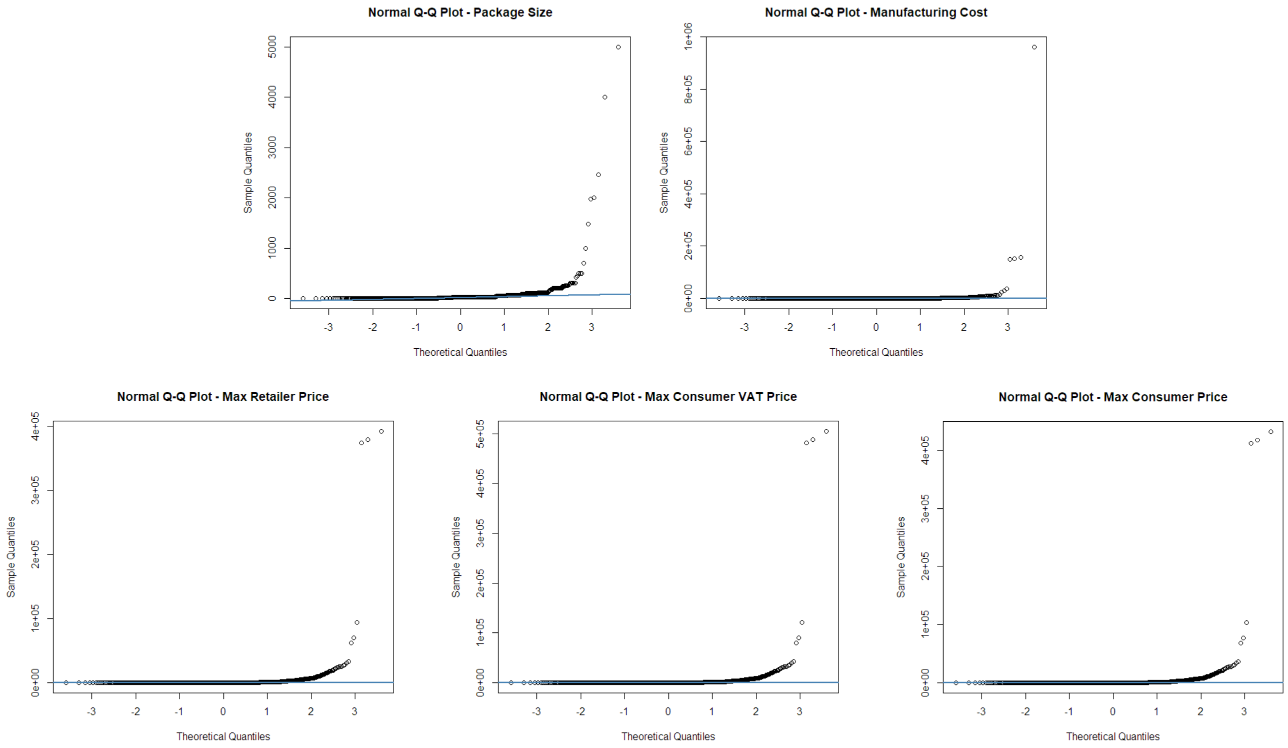

Figure 1 - Q-Q plots of numeric attributes.

Missing Values

No missing values were found in any of the attributes.

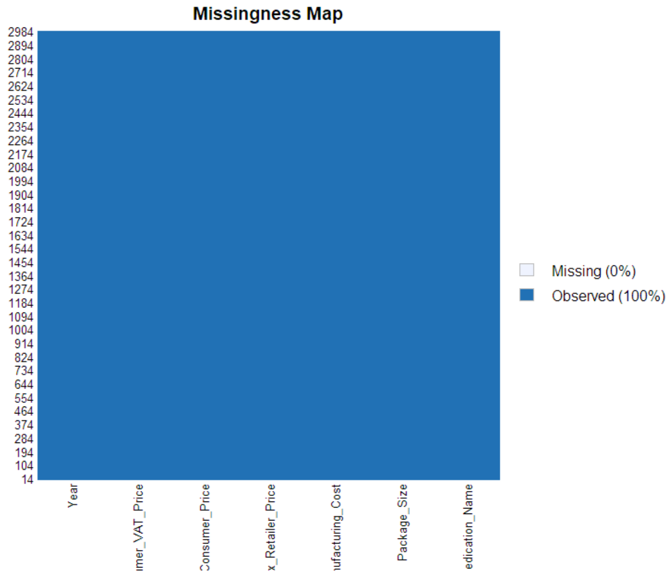

Figure 2 - Missingness map of all attributes.

Skewness

After initial analysis of the max consumer price variable, there is clear skewness in the values.

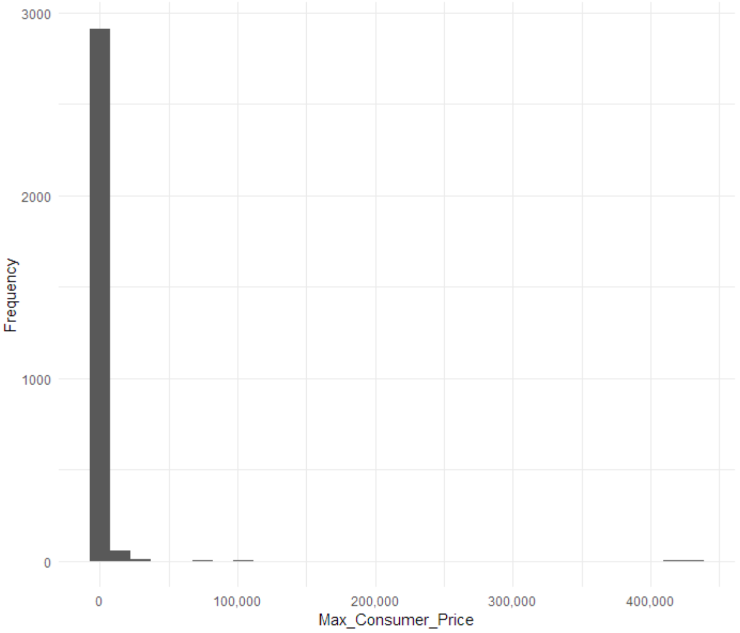

Figure 3 - Histogram of the attribute max_consumer_price.

EXPLORATORY DATA ANALYSIS

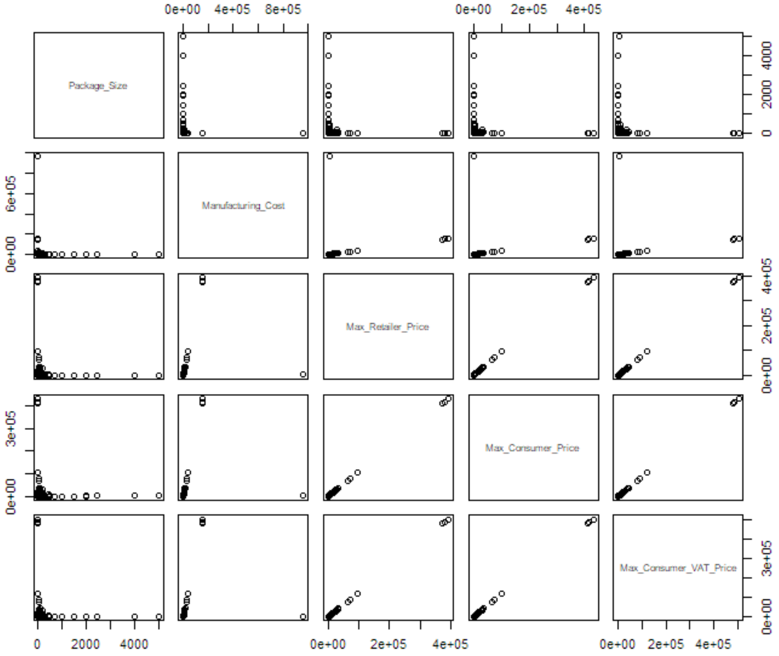

Figure 4 – Scatterplots of each of the attributes in comparison with one another.

Plots from Fig. 4 suggest many of the attributes have linear relationship with one another like max retailer price, max consumer price, and max consumer VAT price. 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.

The correlation plot shows light positive correlation between manufacturing cost, max retailer price, max consumer price, and max consumer VAT price.

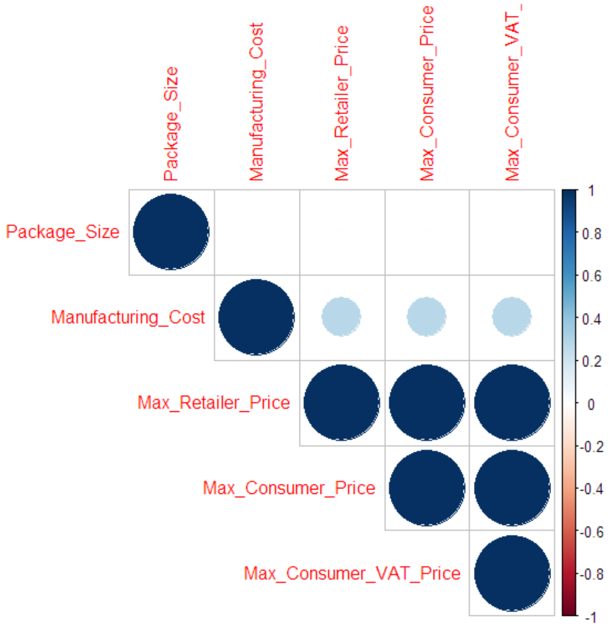

Figure 5 – Correlation plot of all numeric attributes.

Fig. 6 is further examination of the relationship between manufacturing cost and max consumer price. The data was split into two sets for easier observation, records with manufacturing cost less than 50,000 and those equal to or higher than 50,000. For most there is a clear positively linear relationship between manufacturing cost and max consumer price. However, there are one or more observations that do not fit this observed relationship. Although these drugs cost more to manufacturing than most of the observations, they are significantly less expensive than the rest.

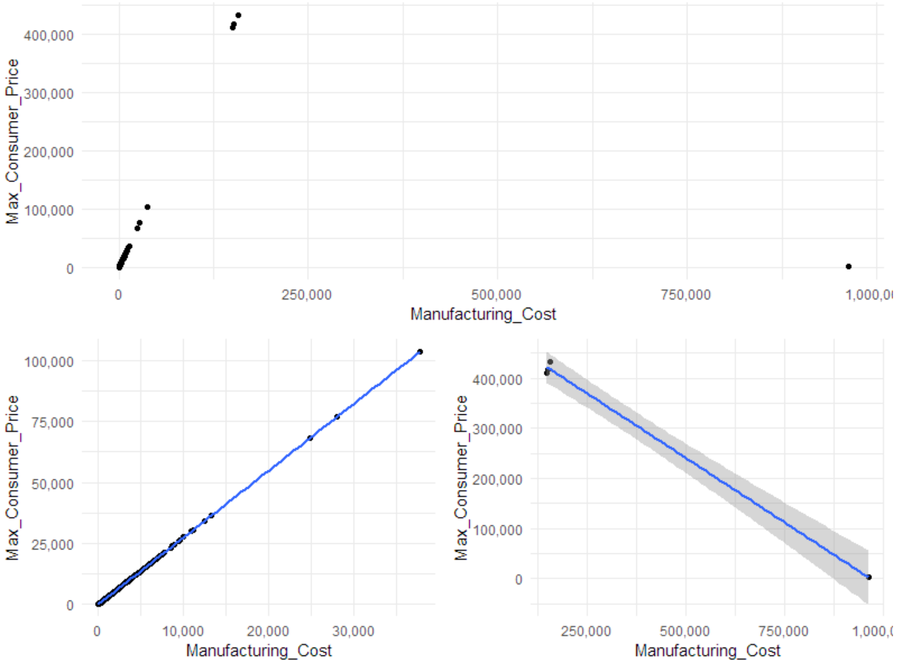

Figure 6 – Scatter plots of max_consumer_price vs manufacturing_cost. (Lower left) Records with manufacturing costs less than 50,000. (Lower right) Records with manufacturing cost equal to or greater than 50,000.

Fig. 7 was used to better analyze the relationship between consumer cost and packaging size. The data was split into three sets for easier observation, records with packaging size less than 200, those equal to or greater than 200 but less than 1000, and those equal to or greater than 1000. For most packaged drugs with less than 200 units or greater than 1000 units there seems to be a strong linear relationship between their size and the max consumer cost. However, for drugs with packaging sizes between these two values there does not seem to be a clear linear relationship.

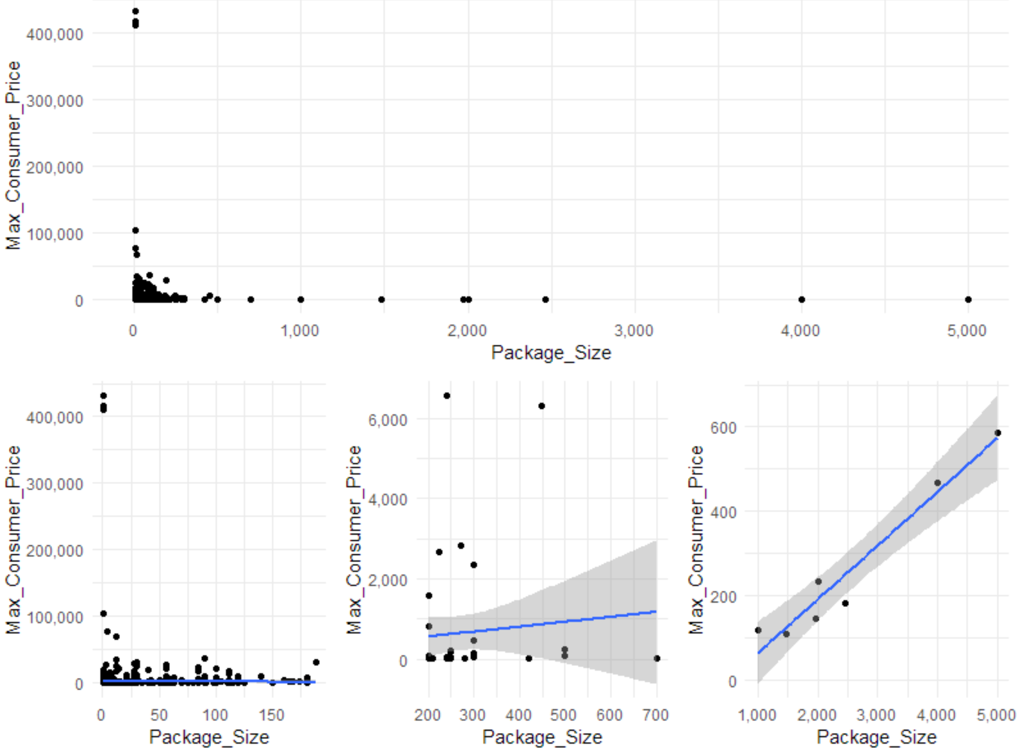

Figure 7 – Scatter plots of max_consumer_price vs package_size. (Lower left) Records with packaging size under 200. (Lower middle) Records with packaging size equal to or over 200 but lower than 1000. (Lower right) Records with packaging size equal to or greater than 1000.

An additional attribute was created to record the distribution method used by each drug named *method*. This information was initially recorded in the medication_name variable. String values for this attribute were cleaned by removing common patterns and numbers not associated with the method used for the drug. Then the word cloud below in Fig. 8 was created to obtain the most common distribution methods used for the drugs in the data set. The top 15 methods used were chosen and are listed below. If a record was not associated with any of the selected methods, its method was recorded as “unknown”.

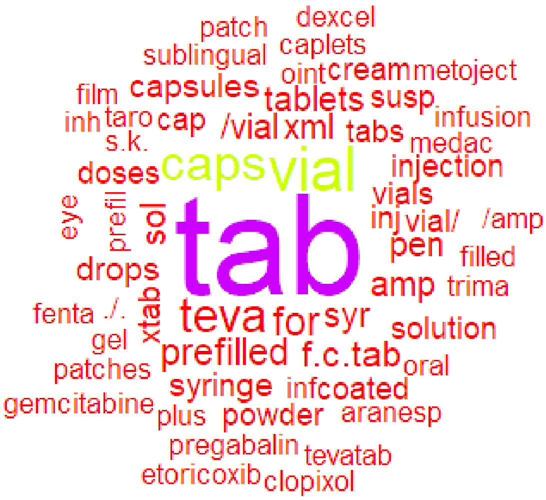

Figure 8 – Word cloud of frequent words in the attribute medication_name.

Most Frequent
Tablets
Vial
Capsules
Syringe
Injection
Pen
Drops
Powder
Solution
Cream
Patch
Gel
Ointment
Spray
Bag

From Fig. 9 it is visually noticeable that most drug recorded in the dataset are given in tablets, vials, or capsules.

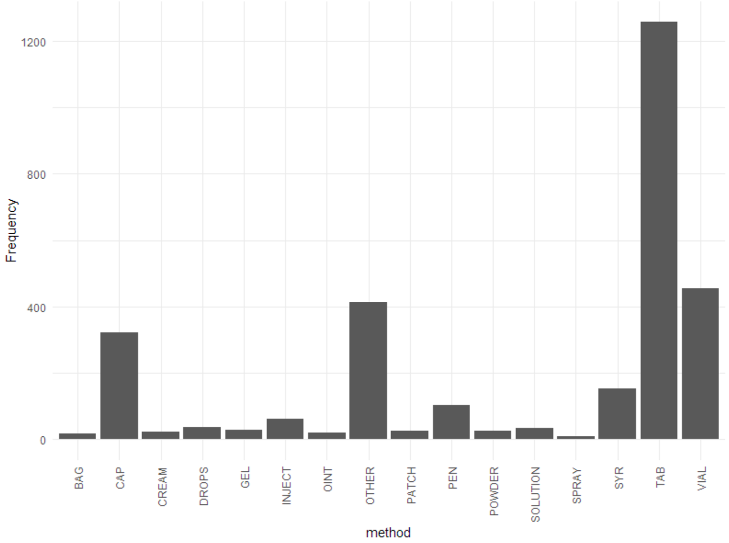

Figure 9 – Bar graph of the different drug distribution methods used and their frequency.

The histogram plots below show that the least costly drugs to manufacture are typically those distributed as tablets, capsules, injections, and creams. The drugs that cost the most to manufacture were those distributed in vials.

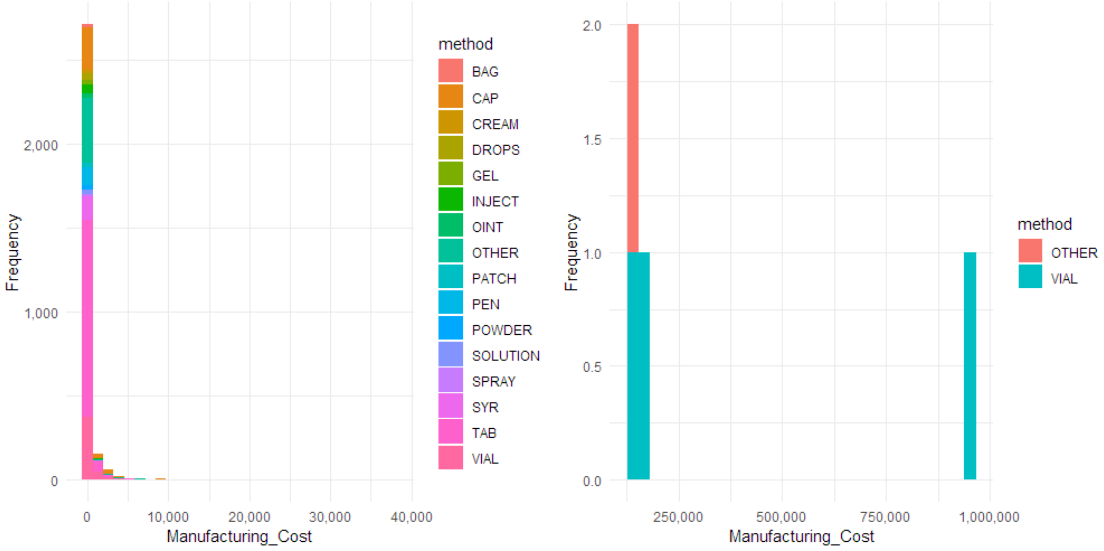

Figure 10 – Histograms of manufacturing_cost with color comparison of the different drug distribution methods.

Fig. 11 shows that the least expensive drug for consumers are those distributed as tablets, syringes, capsules, and vials.

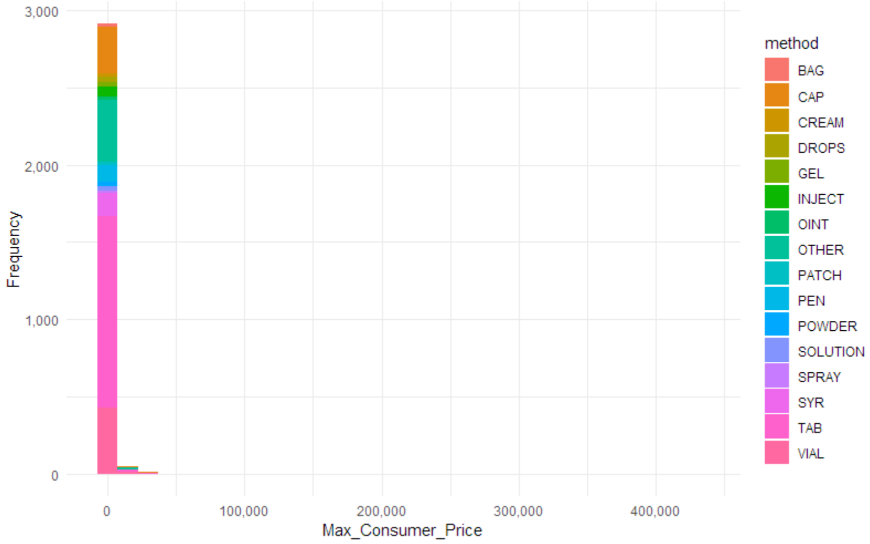

Figure 11 – Histograms of max_consumer_price with color comparison of the different drug distribution methods.

METHODOLOGY

Data Preparation and Splitting

- The attribute *year* was dropped as all the records belonged to the same year.
- The attribute *medication_name* was dropped as the number of unique values would have made the analysis of it difficult.
- The attribute *max_consumer_price* was transformed using the equation below to address the skewed distribution of its values.

$$\log\left(max\_consumer\_price \;\; + 1\right)$$

- 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 for the following models.
- All following models discussed were trained using 10-fold cross validation.

GLM Model

A generalized linear model was created using all the attributes. Based on the results, attributes with p-values less than 0.05 were taken to create a subset of significant attributes, listed below.

Package_Size
Max_Retailer_Price
method.CAP
method.DROPS
method.OTHER
method.SYR
method.TAB
method.CREAM
method.GEL
method.OINT
method.SOLUTION

Another generalized linear model was created and trained using the subset, its performance throughout training when using 10-fold cross validation can be seen below in Fig. 12.

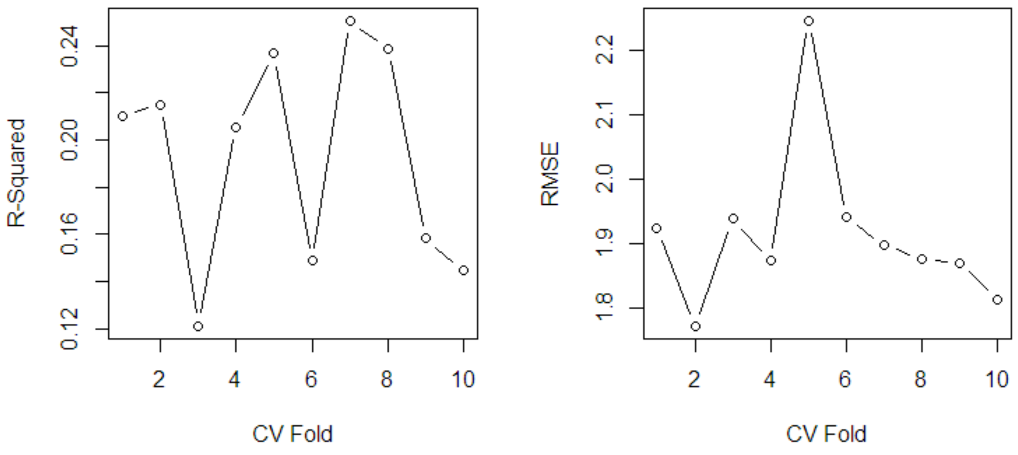

Figure 12 – R-Squared and RMSE of GLM model throughout training using 10-fold CV.

LASSO Regression Model

A LASSO regression model was created using the subset of attributes previously used for the generalized linear model. Hyperparameter tuning along with 10-fold cross validation was then to find the optimal value for lambda of 0.0066 (Fig. 13).

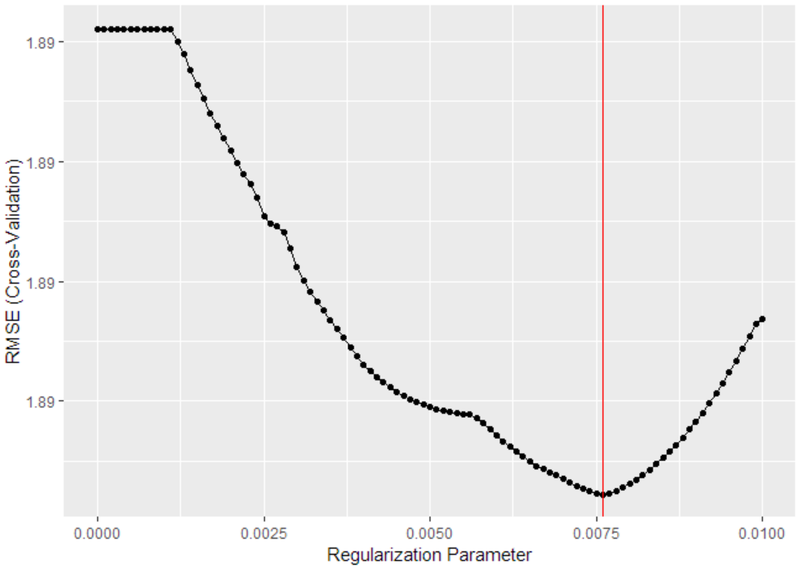

Figure 13 - Hyperparameter tuning results of lambda for LASSO Regression model.

The performance of the model throughout training using a lambda = 0.0066 and an alpha = 1 can be seen below in Fig. 14.

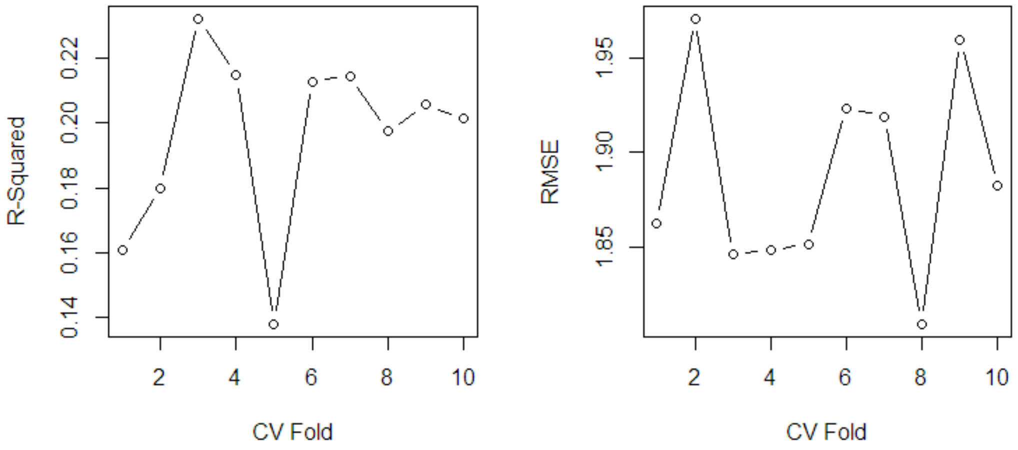

Figure 14 – R-Squared and RMSE of LASSO Regression model throughout training using 10-fold CV.

MARS Model

A MARS model was created using the same subset of attributes used by the previous two models. Hyperparameter tuning was done to find the optimal number of terms to retain and interaction effects to include (Fig. 15).

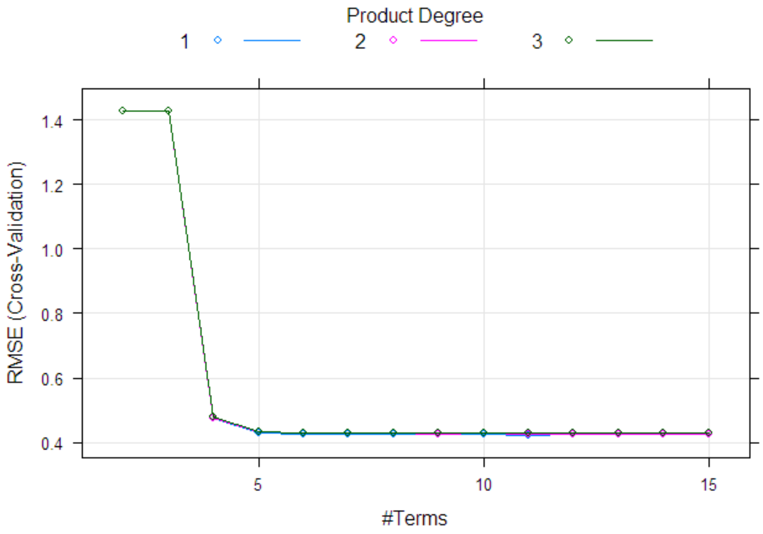

Figure 15 - Hyperparameter tuning results of degree and nprune for MARS model.

The MARS model was then trained using an nprune = 11 and degree = 1 with 10-fold cross validation.

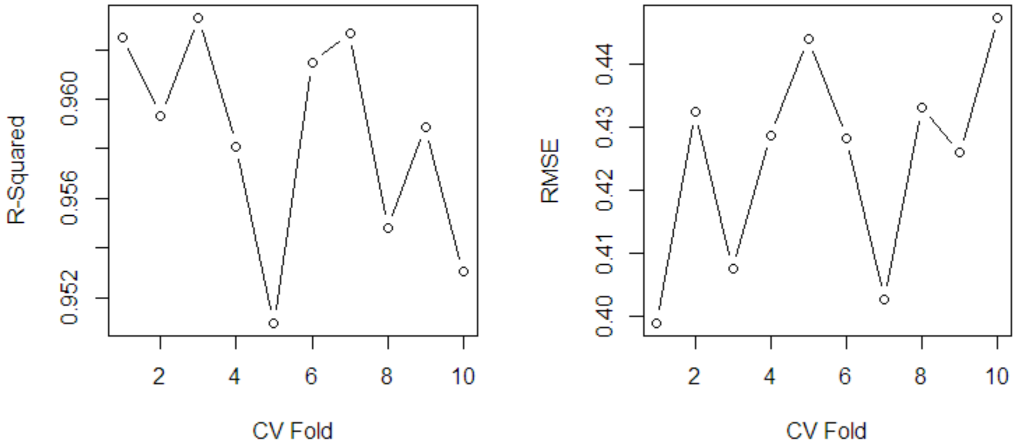

Figure 16 – R-Squared and RMSE of MARS model throughout training using 10-fold CV.

Random Forest

Lastly, a random forest model was created using the previously mentioned subset. After hyperparameter tuning was conducted, using five randomly selected predictors were seen to produce the model with the lowest RMSE.

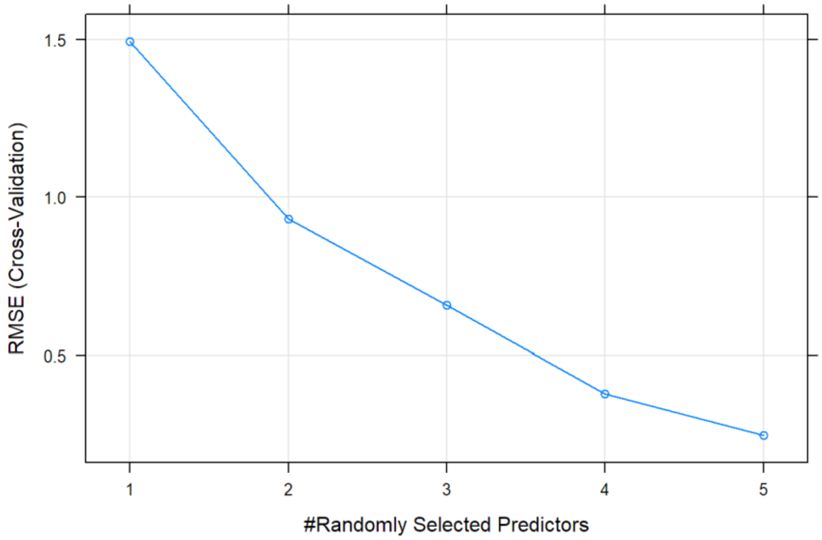

Figure 17 - Hyperparameter tuning results of Random Forest model.

The model was then trained using 5 randomly selected predictors with 10-fold cross validation, results can be seen in Fig. 18.

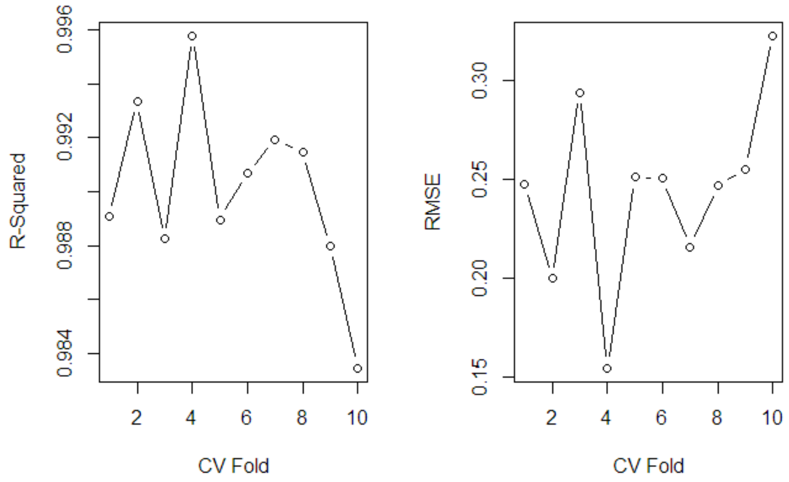

Figure 18 - R-Squared and RMSE of Random Forest model throughout training using 10-fold CV.

RESULTS

The average performance of the all the models throughout training using 10-fold cross validation can be seen below in Table 2.

Table 2 - Average training performance of the models.

	HYPERPARAMETERS	CV-RMSE	CV-R^2
GLM	NA	1.92	0.193

LASSO REG	Alpha = 1,
	Lambda = 0.0066
		1.890	0.196

MARS	Nprune = 11,
	Degree = 1
		0.425	0.958

RANDOM FOREST	Mtry = 5	0.244	0.990

The R-Squared and RMSE of each of the models when tested with the testing data set can be seen in Table 3.

Table 3- Testing performance of the models.

	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

RANDOM FOREST	Mtry = 5	0.221	0.989

Overall, the Random Forest model outperformed all the other models. The MARS model performed similarly to the Random Forest model. However, the Random Forest model had a significantly lower RMSE than the MARS model. The GLM and LASSO regression models performed very poorly compared to the other models.

CONCLUSION

The aim of this project was to analyze patterns within the data that could explain high prescription drug prices and create a model that could predict maximum consumer drug prices with high accuracy. Four models were created to achieve these goals: a GLM, a LASSO regression model, a MARS model, and a Random Forest model. From the initial model developed, the GLM, it was found that 11 attributes from the 20 that were kept after data preparation statistically significantly influenced the maximum consumer price of a prescription drug. From these results, it was concluded that packaging size, maximum retailer price, and 9 of the top 15 methods of distribution of the medicines highly dictated the maximum consumer price of the drugs.

These 11 attributes were then used to create the other three models discussed, of which two performed with very high accuracy. Although the Random Forest and MARS models performed very well, more data would need to be used to test these models to ensure overfitting did not occur.

Prescription drug costs are also related to the health condition they are designed to treat. Therefore, the most significant limitation of the dataset selected for this project was that it needed to include information on what the medications in the dataset were used to treat. In future studies, information should be gathered manually on what conditions these medications are linked to, or this dataset should be combined with another preexisting dataset that offers this information.