Pharmaceutical Pricing and Market Changes Over Time

Deep Adusumalli, Alex May, Emily Painter

1 Introduction

Medical care costs and rising drug prices have been dark topics looming over the minds of Americans for many years now, with the Stanford Institute for Economic Policy Research estimating nationwide collective medical debt of at least $140 billion [1]. We took this project as an opportunity to see if there was any validity behind the claims of increased drug prices, as well as to gain better insight into the federal government's involvement in the pharmaceutical market. Given the vastness of the medical market, we hoped that a subset of the market, the Department of Veteran Affairs (VA), could act as a microcosm of the medical market as a whole. More than just seeking confirmation of commonly held conceptions about the consumer pharmaceutical market, we wanted to investigate the relationship between vendors and the federal government specifically.

The VA regulates the Federal Supply Schedule, which governs the majority of federal pharmaceutical purchases [2]. Therefore, in our analysis of the federal government's involvement in the pharmaceutical market, we chose to focus on the VA specifically. Our primary dataset documents all active contracts between the VA and pharmaceutical vendors and is updated on the 2nd and 16th of each month. We took a snapshot of the data on 09/15/2021 and all of our conclusions and graphs have come from the data as of that date. There are 20,226 rows over 17 variables, each observation representing a unique negotiation between the VA and the given vendor [3]. The specifics of precisely which variables we decided to include in our analyses is expanded upon in Section 2.1. We also chose to use a secondary dataset from 2016 so that we could isolate trends and make comparisons across a longer period of time. Our secondary data has 17,312 rows over 16 of the same variables (the 17th variable included in the 2021 dataset was never used in analysis) [4].

Upon initial familiarization with our data, our primary focus was on price, how price had changed over time, and how different types of products had influenced price change. While this did prove an interesting avenue worthy of pursuit, we immediately saw that the price type variable provided some of the strongest relationships amongst all the data and gave us substantial insight into the nature of federal contracts themselves. In combination with price, the classification of a product, and whether or not a product was covered (definition provided in Section 3.3), a clear picture began to form of some of the most interesting questions to ask to illuminate the nature of the VA's contracts, the relationship between vendors and the federal government, and even the trends of the pharmaceutical market at large.

Before examining price or the nature of the products themselves (through the classification and covered variables), it is important to first establish some common characteristics of pharmaceutical contracts themselves, primarily through analysis of the price type variable; this will be the focus of the first section of our analysis, Section 3.1. From there, we will continue to investigate vendor relationships with the VA that were previously touched upon in Section 3.1, introducing the classification variable and beginning to see how different types of products change in price over time; this is Section 3.2. Finally, we will fully explore how the nature of products themselves influences pharmaceutical contracts, introducing the covered variable in Section 3.3. First, though, we will see how we performed the initial data cleaning.

2 Data Cleaning

2.1 General Cleaning and Set-Up

Upon our initial cleaning of the data, we did not yet know which variables would be useful and which would be included in our final analyses. As such, we cleaned every variable, including those variables that didn't prove altogether useful or interesting. We also added two variables to the primary data frames with information we had derived from the initial data, one of which was never used in our final graphs.

As we became more familiar with the dataset we decided to exclude `contract_number`, `NDC`, `sub_item_identifier`, `package`, `generic_drug` and `trade_name_drug` from our analysis. All of these variables proved either impossible to analyze due to lack of standardization and having too many unique values (`generic_drug`, `trade_name_drug`, `package`, and `contract_number`), or provided little useful information (`sub_item_identifer` was almost entirely `NA` and `NDC` contained much of the same information as other variables provided). We additionally chose not to use `prime_vendor` or `compliant` in our final analysis, simply because they didn't offer any particularly interesting trends or conclusions.

In the end, we used nine variables at least once in our analyses: `vendor`, `contract_start`, `contract_end`, `class`, `covered`, `price`, `price_start`, `price_type`, and `contract_length`, with `contract_length` being a variable we created using the `contract_start` and `contract_end` variables.

Initial data cleaning of most variables was simple: coercing variables into appropriate data types—`contract_start`, `contract_end`, and `price_start` into Date objects, `price` into numeric type, `price_type` into a factor, and `covered` into a logical value. Other variables that were never used were also coerced into their corresponding data types. The `contract_length` variable was found by subtracting `contract_start` from `contract_end` (Date objects support basic addition and subtraction). Additionally, though it affected very few observations, some date entries were incorrect (in the far future, or a stop date was chronologically before a start date) and were thus reassigned to `NA`.

The `class` variable was the first variable that required significant additional work in order to be used effectively. In its original form, the classification column contained a sequence of two letters followed by three numbers, representing the precise categorization of the given product according to the VA's classification system [5]. To further generalize the classification (from groupings as specific as "Amphetamine-like Stimulants" to the more broad "Central Nervous System"), we first took the substrings of the classifications to only keep the first two letters, which represent these more general families. From there, we factored the `class` variable into these 31 families, renaming them from their abbreviated letters to their full names. Though the vast majority of the data was categorized correctly from this step alone, there were some rows in which the classification code had not matched any VA standard code. For these cases, we went into the data itself and looked at what had been categorized incorrectly. If the inconsistency was as simple as a letter that was not capitalized, or a "0" entered instead of an "O", we reassigned the classification to the appropriate code. If the inconsistency was less clear, we looked at the specific case and determined if the classification could be determined with confidence in another way—namely, by looking to see if that same product had been entered elsewhere in the data with an appropriate classification. These two methods for handling fringe cases allowed us to properly classify almost all entries, and the few remaining that could not be easily sorted with confidence were assigned `NA`.

2.2 Summary Statistics vs Raw Data

Something that proved challenging about our dataset was not only its number of entries but its spread. More specifically, one of the variables of the most interest, `price`, was very hard to work with in its initial state. Values ranged from under ¢10 to over $2,000,000. This made it exceptionally difficult to work with in its raw form. Initially, to combat this problem, we created a new variable, called `factored_price`, which categorized each product into logarithmic price categories: $10 - $99, $100 - $999, et cetera. While this initially proved effective, as our graphs evolved, we found using a summary statistic to be a more useful way of representing price. Given price was our only naturally continuous variable, we were also reluctant to turn our only continuous variable into yet another discrete quantity, making using a summary statistic advantageous not only in its best representation of the data itself but also in providing us the opportunity to create graphs that were not only discrete.

After we determined that a summary statistic would potentially be an effective way to visualize the price variable, we then had to consider which statistic to use: mean or median. Given the previously described immense spread, we found in pricing, mean prices, especially mean prices by classification, were often incredibly skewed. For example, for the immunological classification, the mean price is $100,467.50, but the median price was only $714.00. As such, we determined that, when working with the price variable, the median price would typically be a better measurement of the true nature of the price distribution than the mean. The natural sacrifice when using median price is losing information from the extreme cases (the immunological classification, though it contained the most extreme price disparities, never stood out in our final graphs given our usage of median), but since our goal in using the price variable was to see the typical case, we found the sacrifice of the extreme cases worth the improved picture that median price provided in comparison with the mean price.

3 Analysis

3.1 Price Typing and Contract Negotiations

Coming into our dataset, we believed that our primary focus would be on the relationship between price and time, with classification providing additional information as to how specific types of products had evolved over time. While we certainly conducted this analysis as well (Figure 3.2.1), we quickly realized that price type was a broad variable worth analyzing from multiple angles, and it quickly became one of our primary focuses in our analysis. Particularly, the VA National Contract price type's differences from the FSS and Big 4 stood out in several of the coming graphs and provided better insight into the VA's priorities and pharmaceutical negotiations. Before moving to interpret the results, some important definitions are included to be used in this and further sections [2].

FSS: Federal Supply Schedule. The broadest pricing, available to all direct federal purchasers.

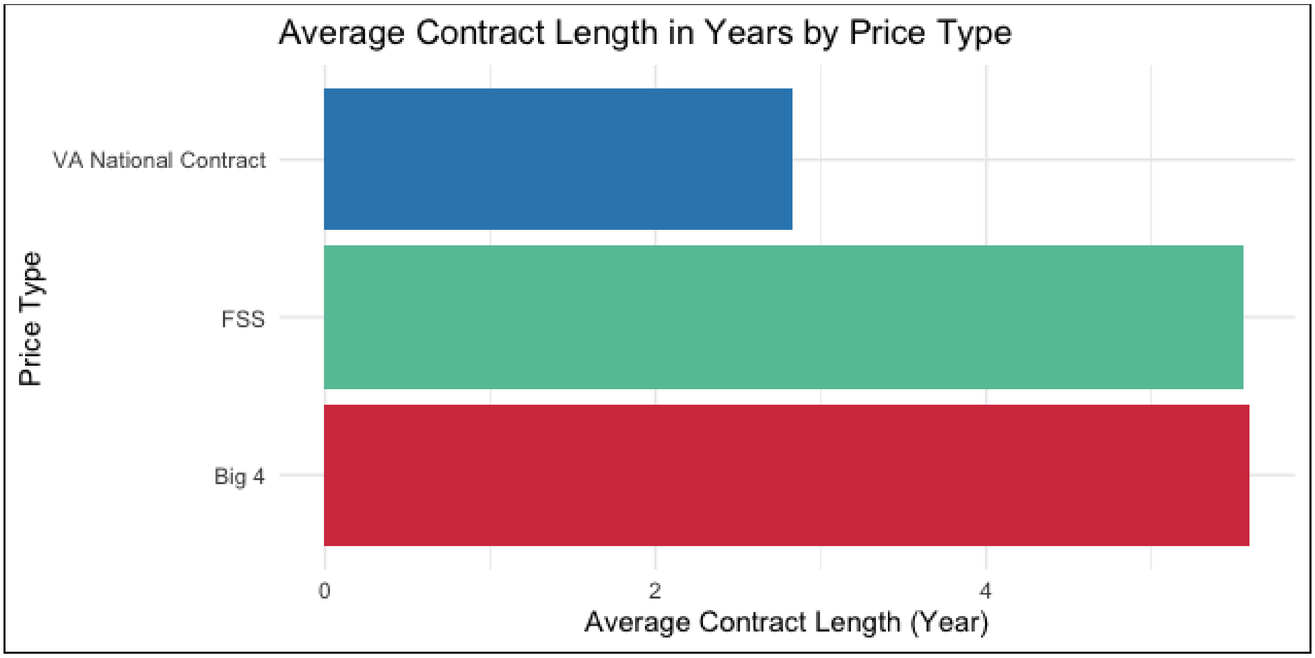
Big 4: Subset of the FSS. Products may not be sold over the Federal Ceiling Price (FCP) to the Department of Defense (DoD), Public Health Services (PHS), the Coast Guard, and the VA. "Big 4" as a price type denotes that the typical FSS price was above the FCP, and the VA thus purchased the given product at the reduced (FCP) Big 4 price. Since our data is focused on the VA, we can conclude that any FSS purchase was, in fact, priced at the FCP or below and there was no special Big 4 reduced cost. The FCP is a minimum of 24% lower than the average manufacturer price [2].

VA National Contract: Reduced pricing specifically negotiated between the VA and the vendor. In exchange for prices lower than even the FCP, the VA agrees to commitment to the vendor for the given product, less than full and open competition [6].

Figure 3.1.1: Average Contract Length by Price Type

One of the first and most prominent interesting relationships we were able to derive from contracts of different price types was that of contract length. As shown in the below figure, VA National Contract average contract length varied greatly from the averages of the FSS and Big 4, the two of which being nearly identical in contract length. The overall average contract length across all price types was 5.42 years, with over half of the data (12,617 observations) having a contract length of precisely 1,825 days (exactly 5 years). This corresponds to the fact that government contracts can't be more than 5 years unless specifically approved in accordance with the agency's procedures [7].

Figure 3.1.1



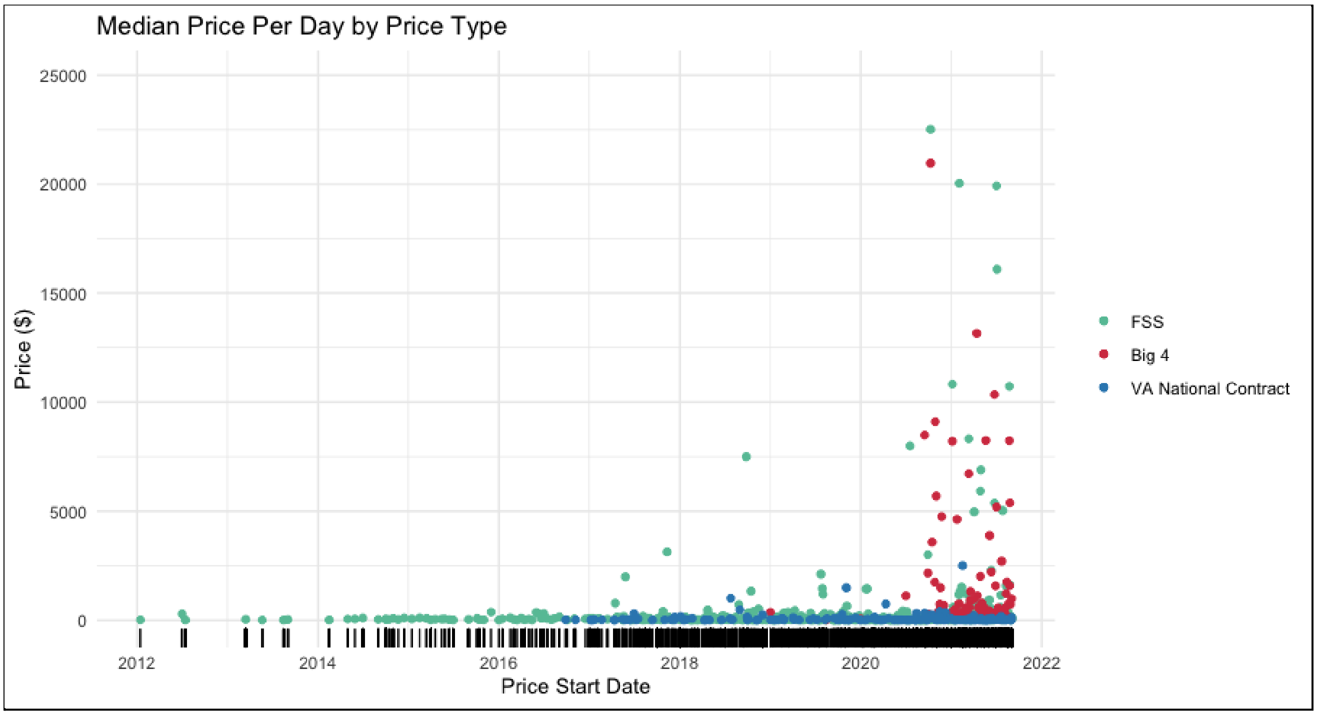
We believe that the VA National Contract, being a price type reserved for those vendors with which the VA has a special relationship, is strongly associated with shorter contract lengths due to vendors' reluctance to agree to exceptionally low prices and the VA's reluctance to agree to sole source provision of a product for extended periods of time. Given the VA National Contract price type typically entails an acquiescence of full and open competition in exchange for reduced prices, it logically follows that, while both the VA and the vendor benefit from this arrangement, shorter contract periods would be preferred by both groups. The VA would only be inclined to limit competition for a brief amount of time, and the vendor would only be inclined to offer extremely low prices for a brief amount of time. This could explain the significantly shorter average contract length we observe with contracts of the VA National Contract price type.

Figure 3.1.2: Median Price Per Day by Price Type

One question we specifically sought to answer when beginning our analysis of this dataset was how the price changed on a day-by-day basis in relation to price type. Since there were so many data points, traditional scatter plots proved ineffective at representing actual price change, as the visual clutter made it hard to derive any meaning. We chose to instead plot median prices per day.

Two key points that arise from this graph, which mostly affirms relationships we had already suspected exist, are how contract lengths and product prices are affected by their price type. We can see from the below figure that VA National Contract products only began to appear in the second half of our price start range—something we would expect to see given our previously established finding that VA National Contract products tend to have shorter contract periods. We wouldn't expect a VA National Contract product with a price start date in 2013 to still be active in 2021, as this would imply a contract length of at least 8 years. Another clear relationship we can observe from this graph is how VA National Contract prices are consistently lower than that of the FSS or Big 4 contracts. Given the very definition of the VA National Contract implies that prices should be even further reduced from the FCP, we would question if this relationship was *not* supported in the data. The VA National Contract median prices never reach the same degree of vertical spread as do the FSS and Big 4 prices towards the later dates.

Figure 3.1.2



This figure mostly served to affirm relationships we had already suspected were true, though it did also provide further insight into how the number of observations during a given time period might skew the image of the data. While this graph certainly seems to paint a picture of prices increasing over time (and a later graph of ours also re-affirms this point), it's important to recognize that, given we are using the median, and the number of observations drastically increases for the latest dates (shown by the rug at the bottom of the graph), it is possible, if not likely, that much of the "relationship" we see in this figure between price and date is not actually about the date at all, but because there are more observations in the later dates. We would naturally expect a wider range of prices on a day with more observations, which we would subsequently expect to affect the median. We are thus hesitant to conclude from this graph alone that date has a significant impact on price.

Figure 3.1.3: Renegotiation of Price Type with a Common Vendor

We've established how price type is associated with contract lengths and how it may or may not relate to price change over time. Another question we specifically wanted to answer as soon as we got our data was how a vendor that has an established history with the Department of Veteran Affairs may or may not renegotiate price typing. To determine this, we found the 170 vendors common to both the 2016 and 2021 data and found their contract price type distributions in each dataset. The below table shows the price type distributions of the top 5 vendors common to the two datasets.

*VNC for VA National Contract

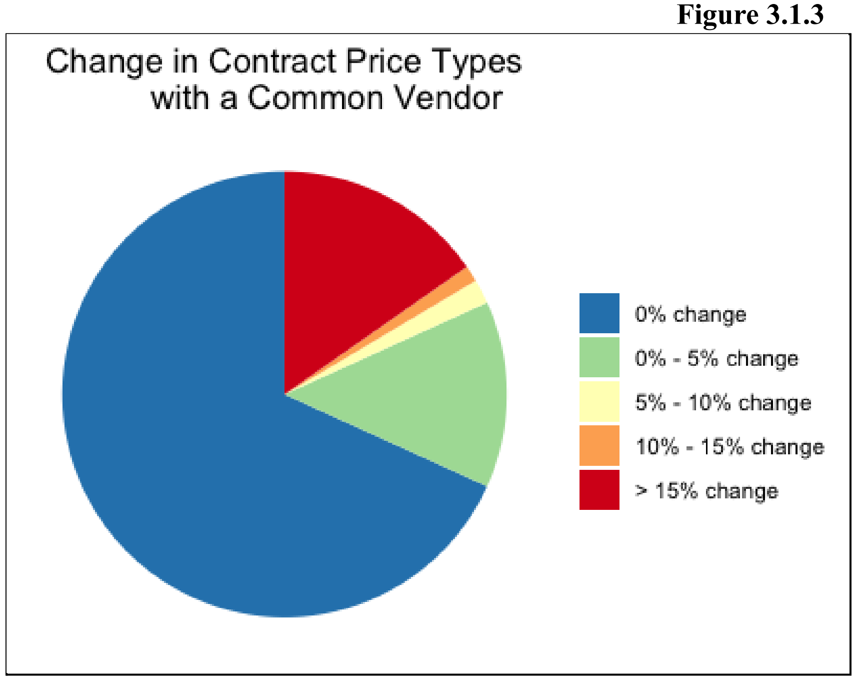
	'16 FSS	'21 FSS	'16 Big 4	'21 Big 4	'16 VNC	'21 VNC
Golden State Medical Supply Inc.	81.69%	86.04%	0.00%	0.00%	18.31%	13.96%
Pfizer Inc.	51.98%	51.61%	48.02%	47.35%	0.00%	1.04%
Major Pharmaceuticals	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
Merck Sharp & Dohme Corp.	50.00%	50.11%	50.00%	49.89%	0.00%	0.00%
Novartis Pharmaceuticals Corp.	50.49%	50.00%	49.51%	50.00%	0.00%	0.00%

Clearly, at least for the top five vendors, the distribution of contracts in each price type doesn't change too greatly for most, and not at all for some. The graph below shows the overall percent change in price type distribution for all 170 common vendors.

68% of the 170 common vendors maintained the exact same proportion of price typing contracts, while another 13.5% saw at most a 5% variation in one or more of the price types. Across all 170 vendors, only 14.7% saw a >15% change.

This trend suggests that price type renegotiations are uncommon. When we first found this data we had expected to potentially see an increase in the more reduced price typing categories, but we didn't observe this in the data. Still, why we see such constancy in price typing is explainable.

Figure 3.1.3



First, it is possible that vendors and the VA negotiate price-typed contracts that are desirable to both parties the first time they work together. We wouldn't expect to see a great change in price typing if both parties had already established price typing that was not only workable but preferable.

A secondary reason for the overall lack of change could be that if we assume that vendors do not change the products they are selling all too frequently and that price typing has at least some relationship with the types of products being sold (see Figure 3.3.3), we wouldn't expect a vast difference in price typing with a constant vendor, given their products are likely relatively constant. This explanation is additionally supported by the breakdown of specific vendors that saw variations in price typing. Small vendors—those vendors who only sell a couple types of products—very frequently had 100% of their contracts be from one price type in both the 2016 and 2021 data. In essence, if the products themselves are constant, we could expect that the price typing of contracts for those products may also be near-constant.

3.2 Product Classification and Vendors' Breadth of Products

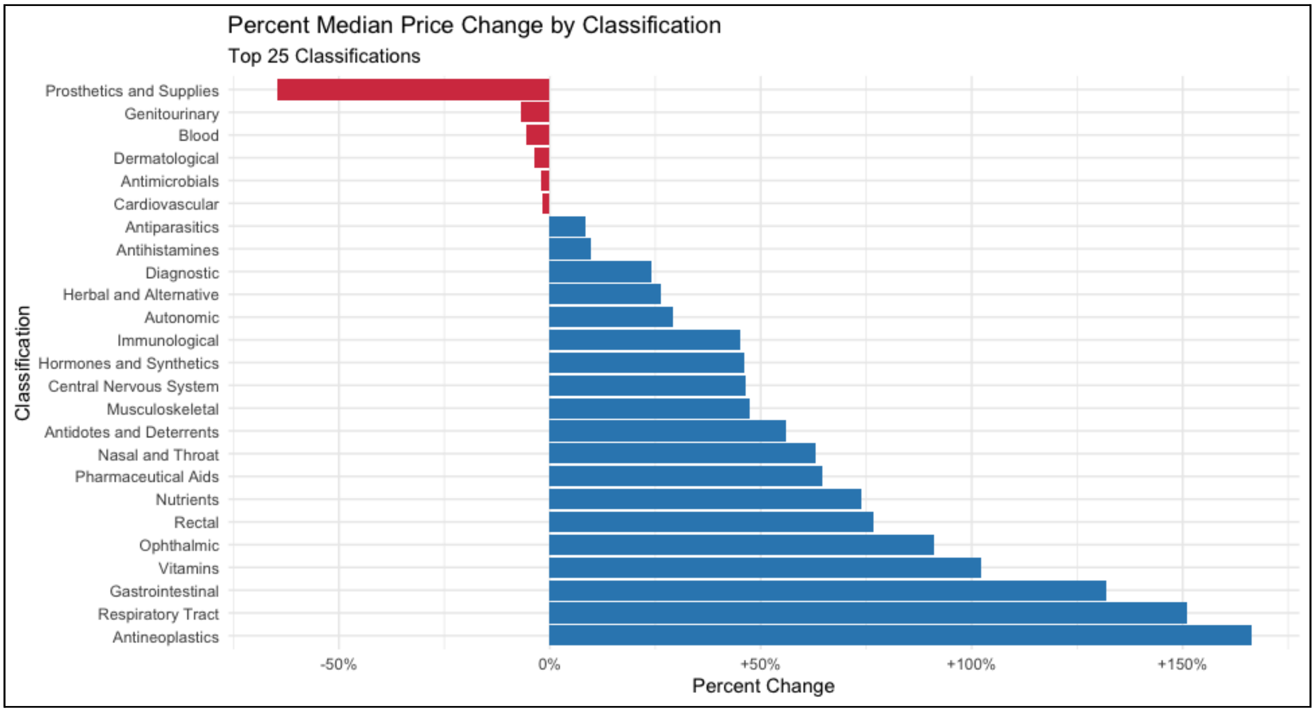
While price type unexpectedly came to the forefront of our investigation, classification was an immediate choice in our initial approach of the data. The opportunity to more precisely see the behavior of each type of medication within this defining bubble of the pharmaceutical market was exactly what we needed to be able to confidently reach significant conclusions regarding the shifts in market pricing.

An observation that we did not initially think of making was that of the interactions between specific vendors and the observed classifications. Though not necessarily related to our primary topic of prices, seeing the scope of vendors involved in the market opened a window of opportunity to observe if any vendors had a stronger control of certain classifications or the market as a whole.

Figure 3.2.1: Percent Price Change by Classification

Naturally, the primary intrigue for most in a dataset of pharmaceutical contracts is that of price, and we too were particularly interested in seeing long-term price changes, especially in relation to the types of products being sold. Using the 2016 data in conjunction with the 2021 data, we found the overall percent median price change from all of the contracts in the 2016 data to all of the contracts in the 2021 data by classification. Excluding the 6 least represented classifications (each of which had <50 observations and some of which skewed the image of the data as a result), we plotted the percent median price changes we found.

Figure 3.2.1



The very first thing we can see from this graph is that the vast majority of products did, as I think most would guess, increase in price. In fact, of the top 25 classifications, only 6 types of products decreased in price, and only one, Prosthetics and Supplies, decreased by a significant margin. What's notable about this increase in price for almost all types of products is that the percent change is not proportionate with inflation alone [8]. 10 of the top 25 classifications show a percent increase of over 50%, and 16 show a percent increase of over 25%. Astonishingly, 4 of the 25 classifications show a median price increase of over 100%, meaning the median price has more than doubled from 2016 to 2021 for these types of products. This graph helps to explain the public perception of drug prices bubbling, and examining some of the specific classifications that saw the most drastic increases (or decreases) in price further illuminates how we (as the public) tend to perceive the pharmaceutical market, and how COVID-19 may have impacted the VA's pharmaceutical negotiations.

Looking at those types of products that have decreased in price, the first thing we noticed was that the majority of decreases were minor. Most of the classifications that saw a decrease in price—cardiovascular, dermatological, antimicrobial, and blood—were also some of the most represented classifications in our dataset (see Figure 3.3.2 to compare the number of observations per classification). Given how minor these decreases were for these classifications, we are inclined to believe that the majority of the decreases in price that we see are insignificant: a consequence of using the median price rather than all price data. Likewise, those classifications that saw minor price increases may not indicate a trend, but expected variation in such large datasets.

Prosthetics and Supplies, however, was one of the lesser represented classifications in our dataset; it is also the only classification that saw a significantly decreased median price. We believe this could be a result of COVID-19, and of the limitations of the VA's classification system [5]. Given prosthetics and supplies are grouped together under the VA's classification system, and prosthetics and supplies vary widely in their prices, we find it likely that we saw such a sharp decrease in median price in the 2021 data for this classification not because the prices of these products actually decreased, but because the 2021 data skewed towards over-representing the "supplies" part of Prosthetics and Supplies. Since our main dataset is as recently updated as September of 2021, we believe that COVID-19 could have resulted in the VA purchasing a greater ratio of supplies to prosthetics in the 2021 data than in the 2016 data. Since supplies are always going to cost less than prosthetics, if the 2021 data truly did contain more supplies than it did prosthetics in comparison to 2016, we would expect a lowered median price for that particular classification—not because prices for prosthetics or supplies *actually* decreased, but because the grouping of these two types of products together skewed the image of price in the 2021 dataset. We are unfortunately limited in our analysis by the classifications these products are grouped into.

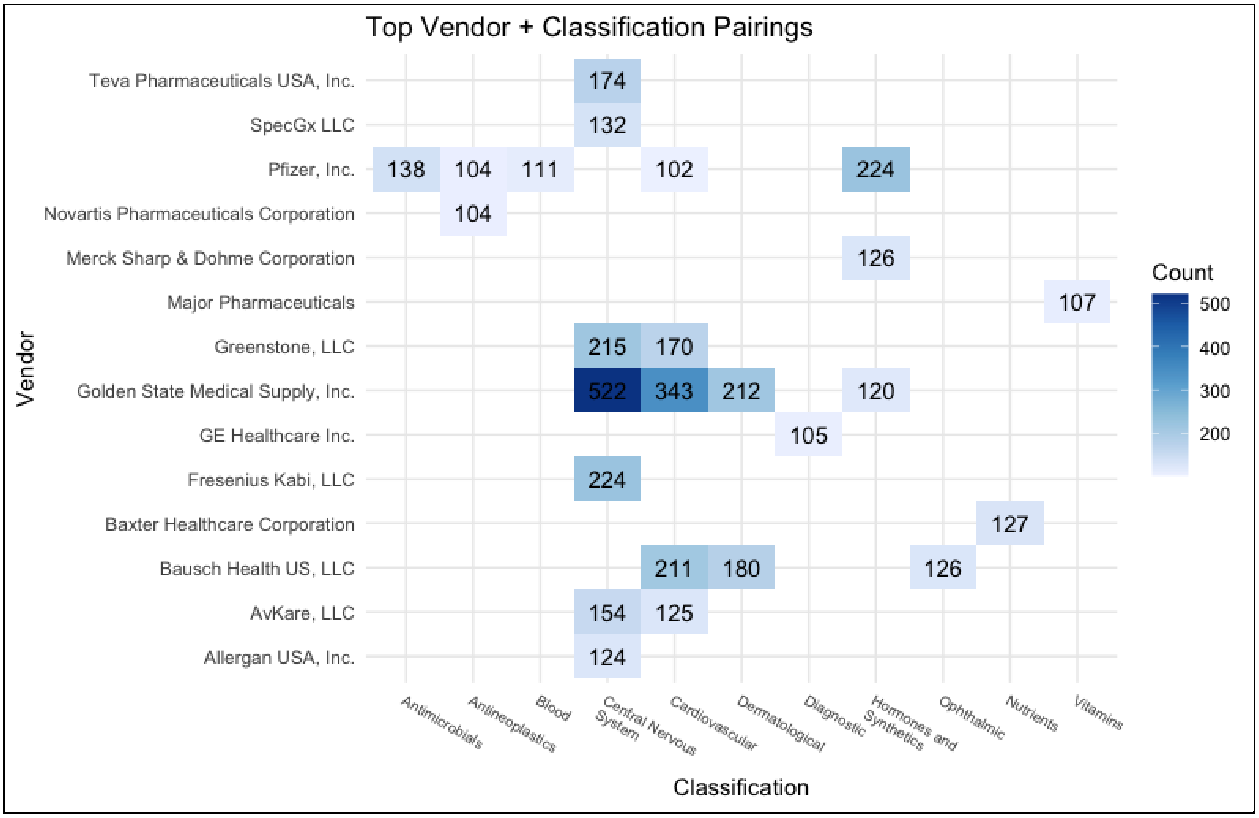
On the other hand, 19 of the top 25 classifications saw increases in price. The two most increased, Respiratory Tract and Antineoplastics, are also two of the most interesting in terms of why their prices might have increased so dramatically and how their increases have affected the public perception of pharmaceutical prices. Respiratory tract, having to do with breathing, likely saw such an extreme increase in price again due to COVID-19 [9].

Antineoplastics, the most increased classification (over 150%), is an especially interesting classification to examine. Though, as we can see from this graph, most types of products have seen an increase in median price from 2016 to 2021, none have seen it to the degree that antineoplastics have. Defined by the CDC as medications to treat cancer, antineoplastics are chemotherapy drugs [10]. It's especially interesting that, of all classifications, cancer medications saw such an extreme price increase, as it begins to crystallize a potential reason why drug prices are discussed in the manner they are. When most people talk about medical care costs becoming exorbitant and unattainable, it's human nature to focus on the most extreme cases—life-saving surgery, transplants, and cancer treatment. One antineoplastic drug, lomustine, reportedly increased in price by nearly 1,400% percent from 2013 to 2018 [11]. When, as Figure 3.2.1 shows us, cancer drugs in particular have increased by such a large margin, it illuminates how the discussion of drug prices increasing has become one of injustice and righteous anger. It's not just that most types of drugs have seen an increase in price that cannot be explained by inflation alone, it's that, in the most desperate of times, the drugs needed are bubbling in cost faster than any other.

Figure 3.2.2: A Closer Look at Some of the Most Represented Vendors

By this point, we had already found some interesting relationships using the classification variable, but we also found it worthy of pursuing a more general picture of precisely which classifications and which vendors were most represented in the data. The heat map below shows the vendor + classification pairings with the most observations.

Figure 3.2.2

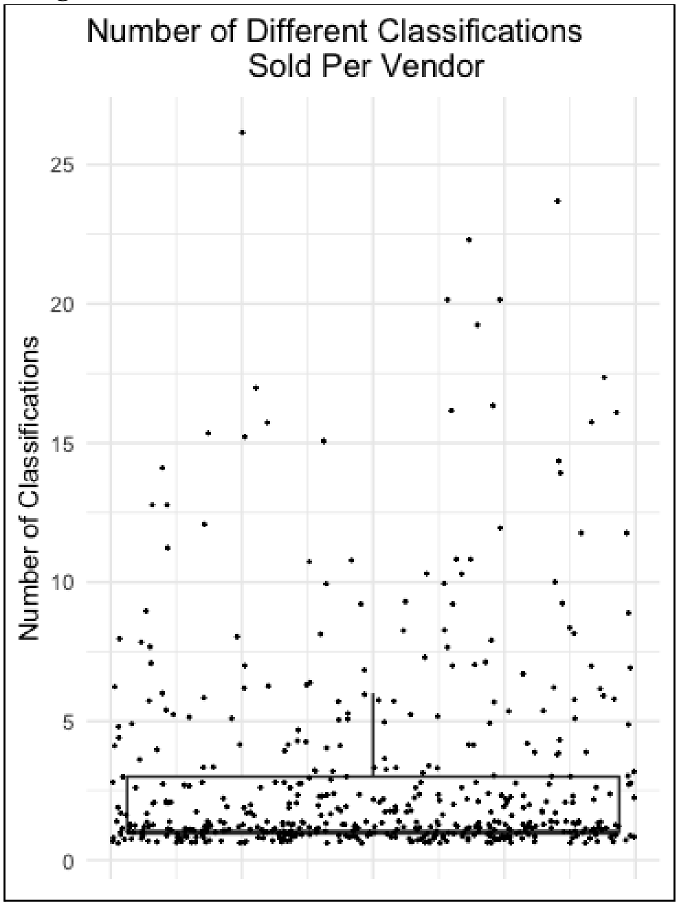


While we initially created this heat map mostly so that we, in doing our research, could have a better understanding of some of the most common classifications and vendors in the dataset, we found this heatmap actually revealed something about our data we had not yet extrapolated. Summing the number of observations of these top 25 pairings, we realized that 4,217 observations (out of a total of 20,226) were represented in this heat map alone, meaning that over a fifth of all of the data in our dataset came from these 14 vendors and 11 classifications. If so few vendors could represent over 20% of the data, we began to question how large the other vendors could possibly be, leading us into our next graph.

Figures 3.2.3 & 3.2.4: Number of Classifications Sold Per Individual Vendor

As we saw in our earlier heat map, only 14 vendors comprise a minimum of 20% of our data. This is remarkable, as there are 500 vendors in our dataset. When we saw this, another question we had not yet thought to ask naturally arose: if a few vendors are so overwhelmingly present in our data, how many classifications of drugs do the vast majority of vendors produce? If so much of our data truly was isolated to just the top 25 vendors or so, we realized the remaining 400+ could not possibly offer more than a few types of products. We first created a boxplot to check this hypothesis.

Figure 3.2.3



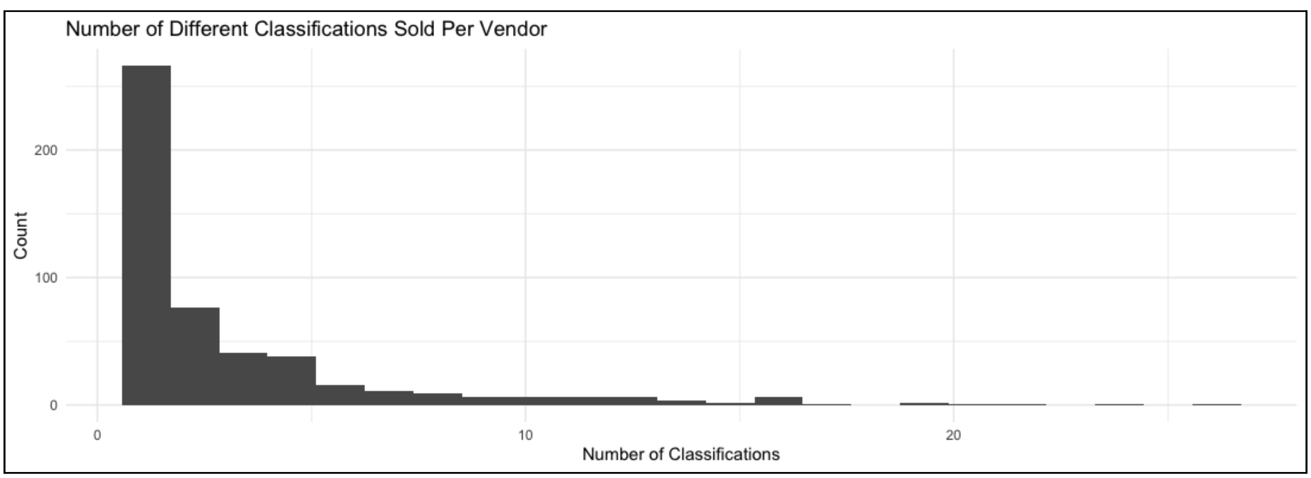
Looking at this boxplot, we can immediately derive a few conclusions about the size, scope, and resources of the typical vendor that works with the Department of Veteran Affairs.

The median, first quartile, and lower bound are all equal at 1. This means that at least 50% of all vendors in the dataset produced only 1 type of drug. The third quartile, at 3, tells us that 75% of all vendors produced 3 or fewer types of drugs, and all the data points above the upper tail are outliers. These outliers are also, as we might at this point have come to expect, those vendors that are over-represented in the data.

This graph helps to solidify our earlier notion that the vast majority of vendors in the dataset *must* be "small," only producing a few types of drugs, contributing relatively few contracts with the VA.

To better visualize the skew, we plotted a histogram of the same data.

Figure 3.2.4



From this graph, we can better see that the vast majority of vendors produce only 1 classification, with a moderate number of vendors producing 2-4 types, and the number of vendors petering out towards the higher numbers of classifications.

3.3 Additional Analysis: The Covered Variable

After a hefty amount of research to find a reputable definition of covered, we quickly realized that it had immense potential as a point of observation. Incidentally, the qualities of what allows a product to be covered were not ones that we could readily observe within the scope of our selected data sets, but this was no barrier to making basic observations of the variable in correlation with our other topics discussed.

To better understand why we wanted to observe covered products, we posed questions of how it could be used to see a more direct relation between previously observed categories and actual consumers. This in turn reflects on the vast medical market as a whole and how it has changed to meet the demands of both large organizations such as the VA as well as the clients under their care. Firstly, a definition of the covered variable is necessary so that we may efficiently move through our analysis [12].

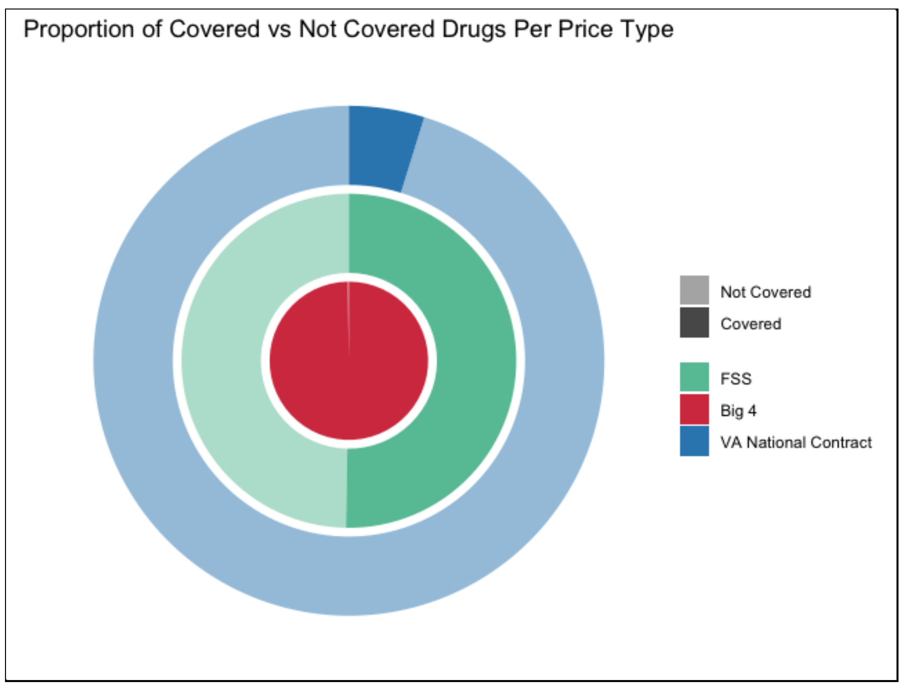
Covered: A complicated legal term with many caveats, the main defining component of coverage is whether or not the product in question is a prescription drug. Generally, a prescription drug is considered covered and other types of products are not, including over-the-counter medications, drugs administered in facilities, and products that are not drugs. While there are more exceptions and special cases, for the purposes of our analysis, it is most important to recognize that covered predominantly means prescription vs non-prescription, and that medications to be used in-house are not considered covered.

Figure 3.3.1: Covered by Price Type

Our first observation of the covered variable was its relation to price type. This provided insight into the activity of the three types as a whole.

The Big 4 relies almost entirely on covered products, with only 0.2% of contracts being for non-covered products. Conversely, only 49.7% of FSS products were uncovered. One could assume that the VA would heavily favor covered products, but quite the contrary appears to be the case. Only 4.74% of VA National Contract products were covered.

Figure 3.3.1



Given the definition of covered, this drastic difference in covered products being used by the VA compared to other types could be explained by the capabilities of the VA as a result of their own organizational activities.

Because covered products are those that can be distributed through prescription, the scrutiny these products face is far higher than other medicinal products. The biggest barrier of qualification for a covered medication is the approval of safety and effectiveness that is provided by the FDA through rigorous examination in accordance with the standards established by sections 505, 505(j) and 507 of the FFDCA [12]. As a corollary, within the definition of covered products is that outpatient drugs, (those used in situations that do not require patients to stay in the care of an institution over the span of a night or full day of operation), do not qualify as covered as they are handled and administered at the discretion of in house operatives. Examples of these sorts of products include insulin or anesthesia to be used in medical procedures.

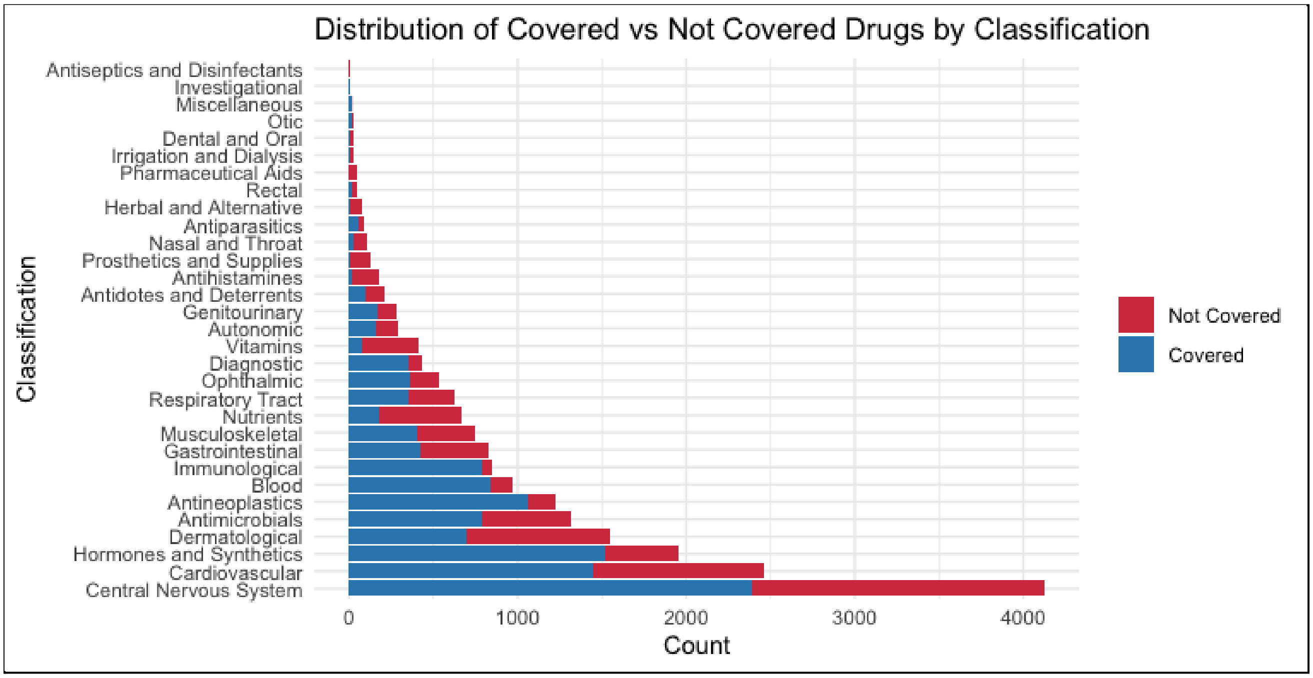
From this corollary, we can now draw conclusions of the VAs use of covered products. Due to operating as their own independent Federal institution, a vast majority of the products being administered through its facilities count as being administered in house. Thus, more direct contracts with vendors can be made for products that the VA may have more justification in using compared to the other types. Hospital beds marketed under the VA National Contract price type to the VA for use within their facilities, expectedly, are non-covered products [13], and are just one example of the types of products that are typically purchased under the VA National Contract price type.

Recalling the drastic increases in prices for almost all classifications that were observed in Figure 3.2.1, this yet again solidifies the VAs preference towards non-covered products. Given these increases in prices, it is easy to assert that the common client would not have the means to pay for these products. Consequently, the VAs ability to more freely select the products needed to care for their clients is supported by a backbone of non-covered products, allowing for more refined care of clients with the additional benefit of lower costs for all involved due to more direct contracting with vendors.

Figure 3.3.2: Covered by Classification

Moving forward with our observation of the covered variable, we were interested in the interactions of covered products and the specific classifications of these products. The intention of this was to gain a better understanding regarding classification of medications primarily provided via prescription compared to ones easily administered at the discretion of a care-provider as well as the volumes of these products being acquired. This in turn reinforces the inferences made regarding the three price types and their use of covered medications.

Figure 3.3.2



Antineoplastics, the classification we identified as experiencing the greatest price increase from 2016 to 2021, are almost entirely covered: 86.3% are covered, to be precise. This observation, and our conception of covered as (usually) meaning prescription versus non-prescription, aligns with our understanding of what antineoplastics are: chemotherapy drugs. Looking at antineoplastics' price increase as well as it's overwhelming covered representation, it may be easy to conclude that covered products in particular saw greater price increases where non-covered products saw increases of a lesser extent. If this conclusion were supported, it would again further bolster much of the public perception around drug prices increasing: that prescription medication in particular was rising in price to an extreme degree. However, upon closer inspection of Figures 3.2.1 and Figure 3.3.2 together, we see that this assumption is not based in reality. Respiratory tract medications were nearly split 50/50 in terms of covered vs non-covered, and as we saw in Figure 3.2.1, Respiratory Tract saw the second greatest price increase. Additionally, we can clearly see from Figure 3.3.2 that the gastrointestinal and vitamin classifications, the next two largest price increases, are not majority covered products, and in fact, vitamins are largely not covered. Looking at these two graphs together, we are inclined to conclude that there is no discernible relationship between price changes and coverage.

Figure 3.3.3: Price Type by Classification

One of the very first lines of inquiry we considered, the relationship between price type and classification, proved an especially interesting avenue of pursuit after we had begun to look into the covered variable. While one of our initial questions upon finding this data had always been which types of products were most associated with which price types, when we examined the covered variable and saw how strong the relationship between covered and price type truly was, the answer to this question became all the more important to answer.

Figure 3.3.3

The first interesting observation we can make from this graph, and indeed the primary question we sought to answer from this graph, is that the Prosthetics and Supplies classification and Nutrients classification stand out in their high proportion of VA National Contract price typing. As we saw in Figure 3.2.1, Prosthetics and Supplies was one of the only classifications to see a decrease in price. Especially considering Prosthetics and Supplies not only decreased in price but also decreased by the largest margin, when we made this plot, we naturally referred back to Figure 3.2.1 to see if Nutrients too had any particularly interesting relationship with price change. Figure 3.2.1 reveals that the Nutrients classification saw a pretty typical price increase, making us hesitant to make any conclusions about Prosthetics and Supplies' price decrease being in any way related to the VA National Contract.

The other thing this graph does is confirms some of the relationships we began to see in Figures 3.3.1 and 3.3.2. Those classifications that are largely not prescription drugs and therefore largely not covered, such as Prosthetics and Supplies, Nutrients, Pharmaceutical Aids, Herbal and Alternative, et cetera, do show somewhat higher proportions of VA National Contract than other types of products, which aligns with the trend we saw in Figure 3.3.1—that VA National Contract products are largely not covered. These types of products also certainly fit our conceptual understanding that much of the products that are purchased under the VA National Contract price type are to be used in VA facilities, and are thus not covered.

Interpreting this graph from the perspective of the covered variable also helps to illuminate trends with the Big 4 proportionality that would otherwise be hard to see. We saw in Figure 3.3.1 that almost 100% of all Big 4 products are covered. This observation from Figure 3.3.1 in addition to the information we can glean from Figure 3.3.2 explains some of the classifications we see in Figure 3.3.3 that have no Big 4 presence. Pharmaceutical Aids and Herbal and

Alternative, being majority uncovered products, do not have *any* Big 4 contracts. Prosthetics and Supplies and Antihistamines also being majority uncovered have very little Big 4 presence.

Though our initial intention in creating this graph was only to see which classifications were most associated with the VA National Contract price type, when we began to include the covered variable in our analysis, it helped to explain even more so not just which classifications were associated with which price types, but why these classifications were associated with these price types.

4 Conclusion

Following the discovery of unexpectedly compelling variables when cleaning our data, we quickly began to see a picture of the medical market begin to form. After realizing the potential of price type, our approach faced some adjustment to account for this somewhat broad variable so that we could establish a strong foundation of information to help build up our other findings. Covering these fundamental variables, we established the basic relations between the price types, primarily the VA National Contract. We found that the VA National Contract notably favored shorter contracts that faced little change over time.

Using these foundational observations, moving into 3.2 we sought to find evidence of prices changing. Figure 3.2.1 was one of our most important figures, as it showed us that prices had in fact been increasing even for the federal government, and it highlighted those classifications that had seen even greater price increases than average. From this observation, we began to understand why the discussion of drug prices had become so impassioned in recent years. We also looked at the vendors themselves, seeing the level of influence that a few vendors had on the entire dataset.From this, the picture of the market gained a new clarity after seeing the depth of vendor involvement. We concluded that not only are vendor relationships fairly constant, but that a few vendors have great presence in the pharmaceutical market, in some ways drowning out the impact of those smaller vendors. From the covered variable, we were able to conclude that the VA National Contract price type is likely used predominantly for products to be used in VA facilities, rather than more general federal usage. While this might have been an expected conclusion, it was not immediately clear from the data until we expanded the variables we were analyzing, namely the covered variable. We believe, thus, that there are certainly more interesting conclusions and findings to be obtained from our dataset in the future. If we were to focus on different variables, including the `prime_vendor` variable, as an example, we believe future analyses of this data could yield even more interesting results. Additional comparison against even older data than 2016, or against consumer price data, could also reveal compelling trends.

Resoundingly, we feel that we have found significant results within our data, and have reason to believe that with further research we may be able to establish clear results and reasons as to why the price of medical care has changed so drastically. Seeing such drastic increases in median price alone opened the door into a vast world calling to be observed, with many reasons to affirm the suspicions of medical care that are held by many. We believe additional analysis could provide even further enlightenment.

5 References

[1] Crawford, Krysten. "America's Medical Debt Is Much Worse than We Think." *Stanford Institute for Economic Policy Research*, Stanford University, 20 July 2021, siepr.stanford.edu/news/americas-medical-debt-much-worse-we-think. Accessed 18 Nov. 2021.

[2] United States, Congressional Budget Office. *Prices for Brand-Name Drugs Under Selected Federal Programs*. Government Publishing Office, June 2005. *Congressional Budget Office*, www.cbo.gov/sites/default/files/109th-congress-2005-2006/reports/06-16-prescriptdrug.pdf. Accessed 7 Nov. 2021.

[3] "Historical VA Pharmaceutical Prices." 15 Sept. 2021. *U.S. Department of Veteran Affairs*, www.va.gov/opal/docs/foia/VApharmaceuticalPrices20210915.xlsx. Accessed 16 Nov. 2021. Raw data.

[4] "Historical VA Pharmaceutical Prices." 31 Dec. 2016. *U.S. Department of Veteran Affairs*, www.va.gov/opal/docs/foia/VApharmaceuticalPrices2016.zip. Accessed 16 Nov. 2021. Raw data.

Links to a zip folder of all 2016 datasets. We used the last in the folder, representing the data from December of 2016.

[5] "VA Drug Classification System." *National Drug File*. *Indian Health Service*, Sept. 1998, www.ihs.gov/RPMS/PackageDocs/PSN/psn318u2.pdf. Accessed 10 Nov. 2021.

[6] *Hospital Beds for Home Use | Application of FAR 52. 217-8 Option to Extend Services from: November 23, 2019 to January 22, 2020. GovTribe*, 2 Dec. 2019, govtribe.com/file/government-file/36c10g19r0008-va797n-14-c-0015-p00016025-dot-pdf. Accessed 18 Nov. 2021.

[7] United States, Federal Acquisition Regulation. *Code of Federal Regulations*. Vol. 48, 1996, section 17.104. *Legal Information Institute*, Cornell, www.law.cornell.edu/cfr/text/48/17.104. Accessed 16 Nov. 2021.

[8] Avery, Taylor. "Prescription drug prices increased twice the inflation rate of US economy in 2020, AARP report finds." *USA Today*, 8 June 2021, www.usatoday.com/story/news/health/2021/06/08/aarp-prescription-drug-price-us-inflation-2020/7601608002/. Accessed 18 Nov. 2021.

[9] Owermohle, Sarah. "Drug Prices Steadily Rise amid Pandemic, Data Shows." *Politico*, 7 July 2020, www.politico.com/news/2020/07/07/drug-prices-coronavirus-351729. Accessed 16 Nov. 2021.

[10]"Antineoplastic (Chemotherapy) Drugs – Reproductive Health." *Centers for Disease Control and Prevention*, www.cdc.gov/niosh/topics/repro/antineoplastic.html. Accessed 17 Nov. 2021.

[11] Berr, Jonathan. "Price of 40-year-old cancer drug hiked 1,400% by new owners." *CBS News*, 2 Jan. 2018, www.cbsnews.com/news/cancer-drug-lomustine-price-hiked-1400-percent-by-new-owners/. Accessed 18 Nov. 2021.

[12] United States, Centers for Medicare and Medicaid Services, Department of Health and Human Services. *Code of Federal Regulations*. Vol. 42, 2016, section 447.502. *Legal Information Institute*, Cornell, www.law.cornell.edu/cfr/text/42/447.502. Accessed 10 Nov. 2021.

[13] *Invacare G-Series Bed*. Invacare, 2016. *Invacare*, dealer.invacare.com/file/general/16-037VA_G-Series_Bed.pdf. Accessed 15 Nov. 2021.

6 Appendix

```
library(ggplot2)
library(dplyr)
library(RColorBrewer)
library(reshape2)
library(stringr)
library(tibble)
library(forcats)

# DATA CLEANING

# 2021 Dataset

pharm_21 <- read.csv("main_data.csv")

# 2016 Dataset

pharm_16 <- read.csv("secondary_data.csv")

# Rename columns.
names(pharm_21) <- c("contract_number", "vendor", "contract_start",
                "contract_end", "NDC", "sub_item_identifier", "package",
                "generic_drug", "trade_name_drug", "class", "covered",
                "prime_vendor", "price", "price_start", "price_end",
                "price_type", "compliant")

names(pharm_16) <- c("contract_number", "vendor", "contract_start",
                "contract_end", "NDC", "sub_item_identifier", "package",
                "generic_drug", "trade_name_drug", "class", "covered",
                "prime_vendor", "price", "price_start", "price_end",
                "price_type")

# Remove the sub item identifier column
pharm_21 <- select(pharm_21, -sub_item_identifier)
pharm_16 <- select(pharm_16, -sub_item_identifier)

# Convert variables to logical variables.
pharm_21$covered <- as.logical(pharm_21$covered)
pharm_16$covered <- as.logical(pharm_16$covered)
```

```
pharm_21$prime_vendor <- as.logical(pharm_21$prime_vendor)
pharm_16$prime_vendor <- as.logical(pharm_16$prime_vendor)
pharm_21$compliant <- as.logical(factor(pharm_21$compliant))

# Simplify class variable -- the first two letters of a drug's class represent
# the family of drugs they belong to, which is more than enough information for
# our analysis. There is a loss of specific information as a result, but our
# drug classes are now generalized down to groupings like "Immunological Agents"
# rather than "Immunoglobulins".
pharm_21$class <- substring(pharm_21$class, 1, 2)
pharm_16$class <- substring(pharm_16$class, 1, 2)

# Clean up fringe errors in the class data. If the fringe case is clearly
# identifiable with what class it should fall into (Using a 0 instead of a O,
# lowercase instead of uppercase, the same drugs in the file are all classified
# under a valid code) and comparison and outside verification affirms that the
# nearest class is correct for the particular drug, the drug is reclassified
# appropriately. Else, the drug's class is reassigned to NA.
pharm_21$class <- toupper(pharm_21$class)
pharm_16$class <- toupper(pharm_16$class)

pharm_21$class[pharm_21$class == "0T"] <- "OT"
pharm_16$class[pharm_16$class == "0T"] <- "OT"

pharm_21$class[pharm_21$class == "1M"] <- "IM"
pharm_16$class[pharm_16$class == "1M" | pharm_16$class == "I,"] <- "IM"

pharm_21$class[pharm_21$class == "CA" | pharm_21$class == "CH"] <- "CN"
pharm_16$class[pharm_16$class == "CA"] <- "CN"

pharm_16$class[pharm_16$class == " C"] <- "CV"

pharm_21$class[pharm_21$class == "HF"] <- "HS"

pharm_21$class[pharm_21$class == "10" |
        pharm_21$class == "44" |
        pharm_21$class == "C0" |
        pharm_21$class == "C1" |
        pharm_21$class == "CA" |
        pharm_21$class == "CH" |
        pharm_21$class == "GO" |
```

```
            pharm_21$class == "IO" |
            pharm_21$class == "OY" |
            pharm_21$class == "XS"] <- NA

pharm_16$class[pharm_16$class == "CT" |
            pharm_16$class == "GO" |
            pharm_16$class == "IO" |
            pharm_16$class == "NO" |
            pharm_16$class == "TX"] <- NA


# Store class abbreviations and names in their respective character vectors
class_abbreviations <- c("AD", "AH", "AM", "AN", "AP", "AS", "AU", "BL", "CN",
                    "CV", "DE", "DX", "GA", "GU", "HA", "HS", "IM", "IN",
                    "IR", "MS", "NT", "OP", "OR", "OT", "PH", "RE", "RS",
                    "TN", "VT", "XA", "XX")

class_names <- c("Antidotes and Deterrents", "Antihistamines", "Antimicrobials",
            "Antineoplastics", "Antiparasitics",
            "Antiseptics and Disinfectants", "Autonomic", "Blood",
            "Central Nervous System", "Cardiovascular", "Dermatological",
            "Diagnostic", "Gastrointestinal", "Genitourinary",
            "Herbal and Alternative", "Hormones and Synthetics",
            "Immunological", "Investigational", "Irrigation and Dialysis",
            "Musculoskeletal", "Nasal and Throat", "Ophthalmic",
            "Dental and Oral", "Otic", "Pharmaceutical Aids",
            "Respiratory Tract", "Rectal", "Nutrients", "Vitamins",
            "Prosthetics and Supplies", "Miscellaneous")

# Factor the class variable and rename each factor to correspond with each class
# code's meaning.
pharm_21$class <- factor(pharm_21$class,
                    levels = class_abbreviations,
                    labels = class_names)

pharm_16$class <- factor(pharm_16$class,
                    levels = class_abbreviations,
                    labels = class_names)

# Factor the price_type variable
pharm_21$price_type <- factor(pharm_21$price_type,
```

```
                  labels = c("Big 4",
                         "FSS",
                         "VA National Contract"))
pharm_16$price_type <- factor(pharm_16$price_type,
                  labels = c("Big 4",
                         "FSS",
                         "VA National Contract"))

# Convert the price variable from character to numeric
pharm_21$price <- as.numeric(pharm_21$price)
pharm_16$price <- as.numeric(pharm_16$price)

# Create a factored_price variable that categorizes drugs into log(10)
# price ranges.
price_groups <- c("< $1",  "$1 - $9", "$10 - $99", "$100 - $999",
           "$1,000 - $9,999", "$10,000 - $99,999", "$100,000 - $999,999",
           "> $1,000,000")

pharm_21$factored_price <- cut(log10(pharm_21$price),
                   breaks = -1:7,
                   labels = price_groups)
pharm_16$factored_price <- cut(log10(pharm_16$price),
                   breaks = -1:7,
                   labels = price_groups)

pharm_21$factored_price <- factor(pharm_21$factored_price, ordered = TRUE)
pharm_16$factored_price <- factor(pharm_16$factored_price, ordered = TRUE)

# Coerce date variables into Date objects

# 2021 price start
pharm_21$price_start <- paste(substring(pharm_21$price_start, 1, 6),
                  substring(pharm_21$price_start, 9, 10),
                  sep = "")
pharm_21$price_start <- as.Date(pharm_21$price_start, "%m/%d/%y")

# 2021 price end
pharm_21$price_end <- paste(substring(pharm_21$price_end, 1, 6),
                 substring(pharm_21$price_end, 9, 10),
                 sep = "")
pharm_21$price_end <- as.Date(pharm_21$price_end, "%m/%d/%y")
```

```
# 2016 price start
pharm_16$price_start <- paste(substring(pharm_16$price_start, 1, 6),
                              substring(pharm_16$price_start, 9, 10),
                              sep = "")
pharm_16$price_start <- as.Date(pharm_16$price_start, "%m/%d/%y")

# 2016 price end
pharm_16$price_end <- paste(substring(pharm_16$price_end, 1, 6),
                            substring(pharm_16$price_end, 9, 10),
                            sep = "")
pharm_16$price_end <- as.Date(pharm_16$price_end, "%m/%d/%y")

# 2021 contract start
pharm_21$contract_start <- paste(substring(pharm_21$contract_start, 1, 6),
                                 substring(pharm_21$contract_start, 9, 10),
                                 sep = "")
pharm_21$contract_start <- as.Date(pharm_21$contract_start, "%m/%d/%y")

# 2021 contract end
pharm_21$contract_end <- paste(substring(pharm_21$contract_end, 1, 6),
                               substring(pharm_21$contract_end, 9, 10),
                               sep = "")
pharm_21$contract_end <- as.Date(pharm_21$contract_end, "%m/%d/%y")

# 2016 contract start
pharm_16$contract_start <- paste(substring(pharm_16$contract_start, 1, 6),
                                 substring(pharm_16$contract_start, 9, 10),
                                 sep = "")
pharm_16$contract_start <- as.Date(pharm_16$contract_start, "%m/%d/%y")

# 2016 contract end
pharm_16$contract_end <- paste(substring(pharm_16$contract_end, 1, 6),
                               substring(pharm_16$contract_end, 9, 10),
                               sep = "")
pharm_16$contract_end <- as.Date(pharm_16$contract_end, "%m/%d/%y")

# Clean up fringe errors in the date data.

pharm_16$price_end[substring(pharm_16$price_end, 1, 4) == "2007"] <- NA
```

```
pharm_21$price_start[substring(pharm_21$price_start, 1, 4) == "2002" |
                substring(pharm_21$price_start, 1, 4) == "2000"] <- NA

pharm_21$price_end[substring(pharm_21$price_end, 1, 4) == "2029"] <- NA

# Store price start dates as characters to be used later
price_starts <- levels(factor(pharm_21$price_start))

# Create contract_length variable
pharm_21$contract_length <- pharm_21$contract_end - pharm_21$contract_start
pharm_16$contract_length <- pharm_16$contract_end - pharm_16$contract_start

# Create vector of common vendors and subset x and x16 into only those drugs
# with common vendors
common_vendors <- unique((inner_join(pharm_21, pharm_16, by = "vendor"))$vendor)

# Store subsetted versions of the data containing only the common vendors.
common_21 <- filter(pharm_21, vendor %in% common_vendors)
common_16 <- filter(pharm_16, vendor %in% common_vendors)

# Store all vendors as characters to be used later
vendor_names <- levels(factor(pharm_21$vendor))

# Store main color palette in character vector to more easily access specific
# colors
spectral_palette <- brewer.pal(11, "Spectral")


# GRAPHS

# Figure 3.1.1

# Store the contract length means by price type
mean_contract_length_by_price_type <- by(pharm_21$contract_length / 365.0,
                              pharm_21$price_type,
                              mean)

# Mean Contract Length by Price Type
ggplot(mapping = aes(x = mean_contract_length_by_price_type,
                y = levels(pharm_21$price_type),
                fill = levels(pharm_21$price_type))) +
```

```
  geom_col(show.legend = FALSE) +
  theme_minimal() +
  labs(x = "Average Contract Length (Year)", y = "Price Type",
       title = "Average Contract Length in Years by Price Type") +
  scale_fill_manual(values = c(spectral_palette[2],
                               spectral_palette[9],
                               spectral_palette[10]))

# Figure 3.1.2

# Store the summary data of price by price_start dates by price types
summary_price_FSS <- by(pharm_21$price[pharm_21$price_type == "FSS"],
                        pharm_21$price_start[pharm_21$price_type == "FSS"],
                        summary)
summary_price_Big4 <- by(pharm_21$price[pharm_21$price_type == "Big 4"],
                         pharm_21$price_start[pharm_21$price_type == "Big 4"],
                         summary)
summary_price_VNC <- by(pharm_21$price[pharm_21$price_type ==
                                         "VA National Contract"],
                        pharm_21$price_start[pharm_21$price_type ==
                                               "VA National Contract"],
                        summary)

# Create vectors of the appropriate size.
median_price_FSS <- 1:length(price_starts)
median_price_Big4 <- 1:length(price_starts)
median_price_VNC <- 1:length(price_starts)

# Iterate through each date
for (i in 1:length(price_starts)) {

  if (as.Date(price_starts[i]) %in%
      pharm_21$price_start[pharm_21$price_type == "FSS"])
    median_price_FSS[i] <- summary_price_FSS[[price_starts[i]]][["Median"]]
  else
    median_price_FSS[i] <- NA

  if (as.Date(price_starts[i]) %in%
      pharm_21$price_start[pharm_21$price_type == "Big 4"])
    median_price_Big4[i] <- summary_price_Big4[[price_starts[i]]][["Median"]]
  else
```

```
    median_price_Big4[i] <- NA

  if (as.Date(price_starts[i]) %in%
      pharm_21$price_start[pharm_21$price_type == "VA National Contract"])
    median_price_VNC[i] <- summary_price_VNC[[price_starts[i]]][["Median"]]
  else
    median_price_VNC[i] <- NA
}

# Create a data frame to hold the medians and rename the columns.
med_price_date <- data.frame(cbind(price_starts,
                                   median_price_FSS,
                                   median_price_Big4,
                                   median_price_VNC))
names(med_price_date) <- c("price_start", "median_price_FSS",
                           "median_price_Big4", "median_price_VNC")

# Coerce price_start back to Date type and medians into numeric type
med_price_date$price_start <- as.Date(med_price_date$price_start)
med_price_date$median_price_FSS <- as.numeric(med_price_date$median_price_FSS)
med_price_date$median_price_Big4 <- as.numeric(med_price_date$median_price_Big4)
med_price_date$median_price_VNC <- as.numeric(med_price_date$median_price_VNC)

# Plot the data
ggplot(mapping = aes(x = med_price_date$price_start)) +
  scale_color_manual(name = "Price Type",
                     values = c(FSS_color = spectral_palette[9],
                                Big4_color = spectral_palette[2],
                                VNC_color = spectral_palette[10]),
                     labels = c("FSS", "Big 4", "VA National Contract")) +
  theme_minimal() +
  theme(legend.title = element_blank()) +
  geom_point(mapping = aes(y = med_price_date$median_price_FSS,
                           color = "FSS_color")) +
  geom_point(mapping = aes(y = med_price_date$median_price_Big4,
                           color = "Big4_color")) +
  geom_point(mapping = aes(y = med_price_date$median_price_VNC,
                           color = "VNC_color")) +
  geom_rug(data = subset(pharm_21, price < 25000),
           mapping = aes(x = price_start, y = price),
           sides = "b",
```

```
          inherit.aes = FALSE) +
  labs(title = "Median Price Per Day by Price Type",
       x = "Price Start Date",
       y = "Price ($)")

# Figure 3.1.3

# Store the total number of contracts in each price type by vendor
# (of the vendors that are common between 2016 and 2021 data).
price_types_21 <- by(common_21$price_type, common_21$vendor, summary)
price_types_16 <- by(common_16$price_type, common_16$vendor, summary)

# Create vectors of the appropriate length for storage of information about
# vendor contracts for future usage.
VNC <- 1:length(common_vendors)
FSS <- 1:length(common_vendors)
Big4 <- 1:length(common_vendors)

VNC16 <- 1:length(common_vendors)
FSS16 <- 1:length(common_vendors)
Big416 <- 1:length(common_vendors)

# Iterate through each common vendor and store the total number of contracts
# in each price type in the respective vectors.
for (i in 1:length(common_vendors)) {
  VNC[i] <- price_types_21[[common_vendors[i]]][["VA National Contract"]]
  FSS[i] <- price_types_21[[common_vendors[i]]][["FSS"]]
  Big4[i] <- price_types_21[[common_vendors[i]]][["Big 4"]]
}

for (i in 1:length(common_vendors)) {
  VNC16[i] <- price_types_16[[common_vendors[i]]][["VA National Contract"]]
  FSS16[i] <- price_types_16[[common_vendors[i]]][["FSS"]]
  Big416[i] <- price_types_16[[common_vendors[i]]][["Big 4"]]
}

# Create a data frame of the common vendors and the respective totals
# of each price type.
total_price_types_21 <- data.frame(cbind(common_vendors,
                                   VNC,
                                   FSS,
```

```
                                  Big4))
total_price_types_16 <- data.frame(cbind(common_vendors,
                                  VNC16,
                                  FSS16,
                                  Big416))

# Coerce into numeric types.
total_price_types_21$VNC <- as.numeric(total_price_types_21$VNC)
total_price_types_21$FSS <- as.numeric(total_price_types_21$FSS)
total_price_types_21$Big4 <- as.numeric(total_price_types_21$Big4)

total_price_types_16$VNC16 <- as.numeric(total_price_types_16$VNC16)
total_price_types_16$FSS16 <- as.numeric(total_price_types_16$FSS16)
total_price_types_16$Big416 <- as.numeric(total_price_types_16$Big416)

# Create a variable to keep track of the total number of contracts
# by a given vendor.
total_price_types_21$total_contracts <- total_price_types_21$VNC +
                                total_price_types_21$FSS +
                                total_price_types_21$Big4
total_price_types_16$total_contracts16 <- total_price_types_16$VNC16 +
                                 total_price_types_16$FSS16 +
                                 total_price_types_16$Big416

# Reassign the price type variables to the proportion of the total number
# of contracts they represent, rather than the raw count.
total_price_types_21$VNC <- total_price_types_21$VNC /
                      total_price_types_21$total_contracts
total_price_types_21$FSS <- total_price_types_21$FSS /
                      total_price_types_21$total_contracts
total_price_types_21$Big4 <- total_price_types_21$Big4 /
                       total_price_types_21$total_contracts

total_price_types_16$VNC16 <- total_price_types_16$VNC16 /
                        total_price_types_16$total_contracts16
total_price_types_16$FSS16 <- total_price_types_16$FSS16 /
                        total_price_types_16$total_contracts16
total_price_types_16$Big416 <- total_price_types_16$Big416 /
                         total_price_types_16$total_contracts16

# Combine the two data frames.
```

```
total_price_types <- inner_join(total_price_types_16,
                                total_price_types_21,
                                by = "common_vendors")

# Create a vector of the appropriate size and initialize counter
# variables to zero.
total_price_types$variations <- 1:length(total_price_types$FSS)
price_type_0range <- 0
price_type_5range <- 0
price_type_10range <- 0
price_type_15range <- 0
price_type_over15range <- 0

# Iterate through each vendor
for (i in 1:length(total_price_types$FSS)) {

  if (((total_price_types$FSS16[i] == total_price_types$FSS[i]) &
      (total_price_types$Big416[i] == total_price_types$Big4[i]) &
      (total_price_types$VNC16[i] == total_price_types$VNC[i]))) {

    total_price_types$variations[i] <- "0% change"
    price_type_0range <- price_type_0range + 1

  } else if (between(total_price_types$FSS16[i],
                     (total_price_types$FSS[i] - 0.05),
                     (total_price_types$FSS[i] + 0.05)) &
             between(total_price_types$Big416[i],
                     (total_price_types$Big4[i] - 0.05),
                     (total_price_types$Big4[i] + 0.05)) &
             between(total_price_types$VNC16[i],
                     (total_price_types$VNC[i] - 0.05),
                     (total_price_types$VNC[i] + 0.05))) {

    total_price_types$variations[i] <- "0% - 5% change"
    price_type_5range <- price_type_5range + 1

  } else if (between(total_price_types$FSS16[i],
                     (total_price_types$FSS[i] - 0.10),
                     (total_price_types$FSS[i] + 0.10)) &
             between(total_price_types$Big416[i],
                     (total_price_types$Big4[i] - 0.10),
```

```
                (total_price_types$Big4[i] + 0.10)) &
          between(total_price_types$VNC16[i],
                (total_price_types$VNC[i] - 0.10),
                (total_price_types$VNC[i] + 0.10))) {

    total_price_types$variations[i] <- "5% - 10% change"
    price_type_10range <- price_type_10range + 1

  } else if (between(total_price_types$FSS16[i],
                (total_price_types$FSS[i] - 0.15),
                (total_price_types$FSS[i] + 0.15)) &
          between(total_price_types$Big416[i],
                (total_price_types$Big4[i] - 0.15),
                (total_price_types$Big4[i] + 0.15)) &
          between(total_price_types$VNC16[i],
                (total_price_types$VNC[i] - 0.15),
                (total_price_types$VNC[i] + 0.15))) {

    total_price_types$variations[i] <- "10% - 15% change"
    price_type_15range <- price_type_15range + 1

  } else {

    total_price_types$variations[i] <- "> 15% change"
    price_type_over15range <- price_type_over15range + 1

  }
}

total_price_types$variations <- factor(total_price_types$variations,
                                       ordered = TRUE,
                                       levels = c(1, 2, 3, 4, 5))

variation_distribution <- c(price_type_0range / 170,
                            price_type_5range / 170,
                            price_type_10range / 170,
                            price_type_15range / 170,
                            price_type_over15range / 170)

# Plot the pie chart.
ggplot(mapping = aes(x = "", y = variation_distribution)) +
```

```
  geom_bar(stat = "identity",
           width = 1,
           mapping = aes(fill = levels(total_price_types$variations))) +
  coord_polar("y", start = 0) +
  theme_void() +
  scale_fill_brewer(palette = "Spectral",
                    direction = -1,
                    labels = c("0% change", "0% - 5% change",
                               "5% - 10% change", "10% - 15% change",
                               "> 15% change")) +
  theme(legend.title = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  labs(title = "Change in Contract Price Types
       with a Common Vendor")

# Figure 3.2.1

# Store the summary data of price by class
summary_price_21 <- by(pharm_21$price, pharm_21$class, summary)
summary_price_16 <- by(pharm_16$price, pharm_16$class, summary)

# Create a vector of the appropriate size
median_price_21 <- 1:length(class_names)
median_price_16 <- 1:length(class_names)

# Iterate through each class and store the median prices in the
# previously created vector.
for (i in 1:length(class_names)) {

  # Store the median prices only in the vector
  median_price_21[i] <- summary_price_21[[class_names[i]]][["Median"]]
  median_price_16[i] <- summary_price_16[[class_names[i]]][["Median"]]

}

# Create a data frame out of the useful information and rename the columns
median_price <- data.frame(cbind(class_names,
                                 median_price_21,
                                 median_price_16))
names(median_price) <- c("class", "median21", "median16")
```

```
# Coerce median variables into numeric types
median_price$median16 <- as.numeric(median_price$median16)
median_price$median21 <- as.numeric(median_price$median21)

# Create a variable to hold the difference in median prices from 2016 to 2021
median_price$price_change <- median_price$median21 - median_price$median16

# Create a variable to hold the percent price change from 2016 to 2021
median_price$percent_change <- median_price$price_change / median_price$median16

# Plot the percent change in median price for each class
ggplot(data = subset(median_price, class != "Antiseptics and Disinfectants" &
                       class != "Otic" &
                       class != "Investigational" &
                       class != "Miscellaneous" &
                       class != "Dental and Oral" &
                       class != "Irrigation and Dialysis")) +
  geom_col(mapping = aes(x = percent_change,
                         y = fct_reorder(class,
                                         percent_change,
                                         sum,
                                         .desc = TRUE),
                         fill = percent_change > 0),
           show.legend = FALSE) +
  scale_x_continuous(breaks = c(-1, -0.5, 0, 0.5, 1, 1.5),
                     labels = c("-100%", "-50%", "0%",
                                "+50%", "+100%", "+150%")) +
  theme_minimal() +
  labs(title = "Percent Median Price Change by Classification",
       x = "Percent Change",
       subtitle = "Top 25 Classifications",
       y = "Classification") +
  scale_fill_manual(values = c(spectral_palette[2], spectral_palette[10]))

# Figure 3.2.2

t <- pharm_21[, c(2, 9)]

vendor_class_map <- t %>%
  group_by(vendor, class) %>%
  summarise(Count = n())
```

```
vendor_class_map <- vendor_class_map[order(vendor_class_map$Count,
                                    decreasing = T), ]

top_25 <- vendor_class_map[c(1:25), ]

ggplot(top_25, aes(x = class, y = vendor)) +
  geom_tile(aes(fill = Count))+
  geom_text(aes(label = Count), col = "black") +
  labs(title = "Top Vendor + Classification Pairings",
       y = "Vendor",
       x = "Classification") +
  theme_minimal() +
  scale_fill_distiller(palette = "Blues", direction = 1) +
  scale_x_discrete(labels = function(pharm_21) str_wrap(pharm_21, width = 20)) +
  theme(axis.text.x.bottom = element_text(angle = 330, hjust = 0, size = 7))

# Figure 3.2.3

# Find summary data about the number of classifications each vendor
# typically sells
num_vendor_appearances <- data.frame(table(vendor_class_map$vendor))
names(num_vendor_appearances) <- c("vendors", "num_classes")

# Plot a boxplot of the number of classifications per vendor
ggplot(data = num_vendor_appearances, mapping = aes(y = num_classes, x = 1)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(size = 0.3) +
  theme_minimal() +
  scale_y_continuous(breaks = c(0, 5, 10, 15, 20, 25, 30)) +
  theme(axis.text.x.bottom = element_blank(),
        axis.title.x.bottom = element_blank(),
        axis.ticks.x.bottom = element_blank()) +
  labs(title = "Number of Different Classifications
                Sold Per Vendor",
       y = "Number of Classifications")

# Figure 3.2.4

# Plot a histogram of the number of classifications per vendor
ggplot(data = num_vendor_appearances, mapping = aes(x = num_classes)) +
```

```
  geom_histogram(bins = ceiling(sqrt(length(num_vendor_appearances$vendors)))) +
  theme_minimal() +
  labs(title = "Number of Different Classifications Sold Per Vendor",
       y = "Count",
       x = "Number of Classifications")

# Figure 3.3.1

ggplot(data = pharm_21, mapping = aes(y = price_type,
                                      x = ..count..,
                                      fill = price_type,
                                      alpha = covered)) +
  scale_alpha_manual(values = c(0.5, 1), labels = c("Not Covered", "Covered")) +
  scale_fill_manual(values = c("FSS" = spectral_palette[9],
                               "Big 4" = spectral_palette[2],
                               "VA National Contract" = spectral_palette[10])) +
  theme_void() +
  theme(legend.title = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  geom_bar(position = "fill") +
  coord_polar() +
  labs(title = "Proportion of Covered vs Not Covered Drugs Per Price Type")

# Figure 3.3.2

ggplot(data = subset(pharm_21, !is.na(pharm_21$class))) +
  geom_bar(mapping = aes(y = fct_infreq(factor(class)),
                         fill = covered)) +
  theme_minimal() +
  theme(legend.title = element_blank()) +
  scale_fill_manual(values = c(spectral_palette[2], spectral_palette[10]),
                    labels = c("Not Covered", "Covered")) +
  labs(title = "Distribution of Covered vs Not Covered Drugs by Classification",
       x = "Count",
       y = "Classification")

# Figure 3.3.3

ggplot(data = subset(pharm_21, !is.na(class) &
                       class != "Antiseptics and Disinfectants" &
                       class != "Otic" &
```

```
                class != "Investigational" &
                class != "Miscellaneous" &
                class != "Dental and Oral" &
                class != "Irrigation and Dialysis"),
    mapping = aes(y = fct_infreq(factor(class)), fill = price_type)) +
geom_bar(position = "fill") +
theme_minimal() +
scale_fill_manual(values = c("FSS" = spectral_palette[9],
                      "Big 4" = spectral_palette[2],
                      "VA National Contract" = spectral_palette[10])) +
theme(legend.title = element_blank()) +
labs(title = "Distribution of Different Price Types Per Classification",
    x = "Proportion",
    subtitle = "Top 25 Classifications",
    y = "Classification")
```