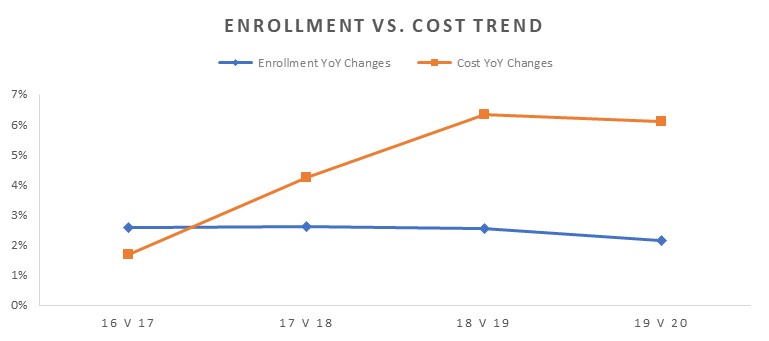
|  |  |  |
| --- | --- | --- |
| SNELL FAMILY FARMGROUP 4: Final Report | | |
| cdp group4 (@CdpGroup4) / Twitter |  | |
| Due: 6/19/2022**Team Members**: **Catherine Zhang**  **Elias Joseph**  **Joshua McCleary**  **Lan Tran**  **Nicholas Lichtsinn**  **Valerie Jones** | |  |

# Introduction (Title Page):

# Specification:

According to the Center for Medicare and Medicaid Services (CMS), total Medicare enrollment has increased 2%-3% on a year over year basis between 2016 through 2020. However, when comparing pharmacy costs year over year, the increase is quite drastic prior to the pandemic. The scope of our analysis is to determine whether new drugs introduced to the market has been impacted by the pandemic and to project drug costs for new drugs.





**Research Question**:

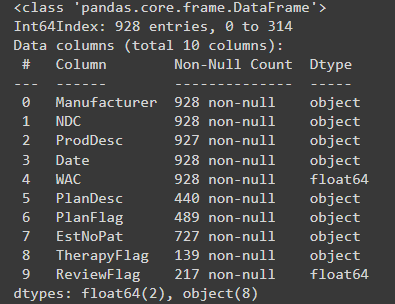
* 1. What is the pricing trend for new drugs introduced to market increased over the past couple of years?
  2. Has the pandemic impacted cost of new drugs introduced to market?
  3. Are some models more accurate at predicting drug cost than others?

**Hypotheses**:

Assumption is that pricing trend is not impacted by the pandemic and that there are certain models that can help predicting drug costs than others.

**Data for Analysis**:

(Datasets: refer to Colab location [here](https://colab.research.google.com/drive/1TWX8JzWv5MHcFKUh4KhxX-s4qHBo0ik9?usp=chrome_ntp#scrollTo=6fYB2g-E0WxI))

* *Prescription Drugs introduced to Market*: The data represents new prescription drugs introduced to market in California with a Wholesale Acquisition Cost (WAC) that exceeds Medicare Part D specialty drug cost threshold
  + 2019 through 2022 with date stamp of when the drug was introduced into the market
  + File format obtained as an xlsx using panda to read in excel file

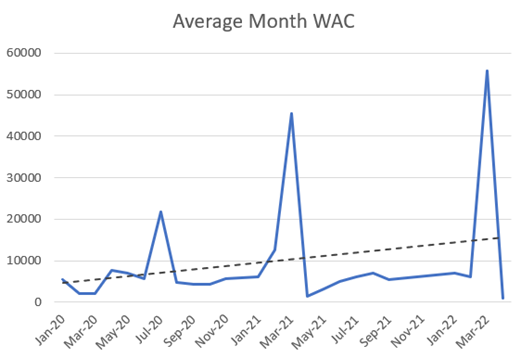
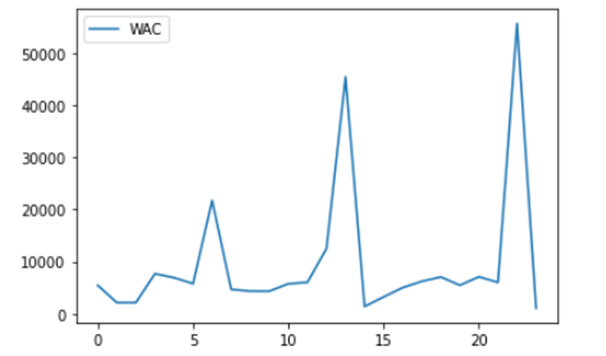
# Observation:

***Data Exploration:***

**Trend**:

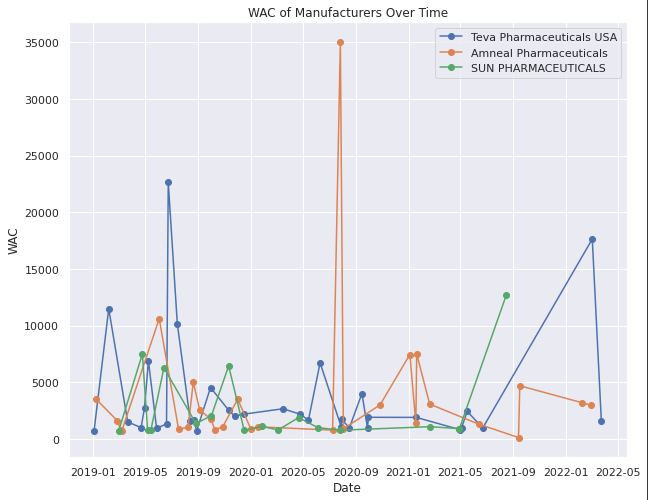
From 2020 to 2022 the drug WAC price has been steadily increasing across all manufacturers.

There are annual spikes around June which are growing larger each year. Visually, these spikes are causing the increase in WAC over time, as the cost seems to hover around 7,500 in the areas outside these spikes.



If we look at the WAC price by manufacturer, there are unfortunately a lot of missing values, but looking at the three largest manufacturers, it is interesting to note that the manufacturers have spikes around the same time of year, but one will usually be much larger than the others. For example, in the plot below, Amneal spiked massively around mid 2020, and while Teva had an increase in WAC price around that time, it was a much smaller increase.

**Top Manufacturer – Year over Year comparisons**

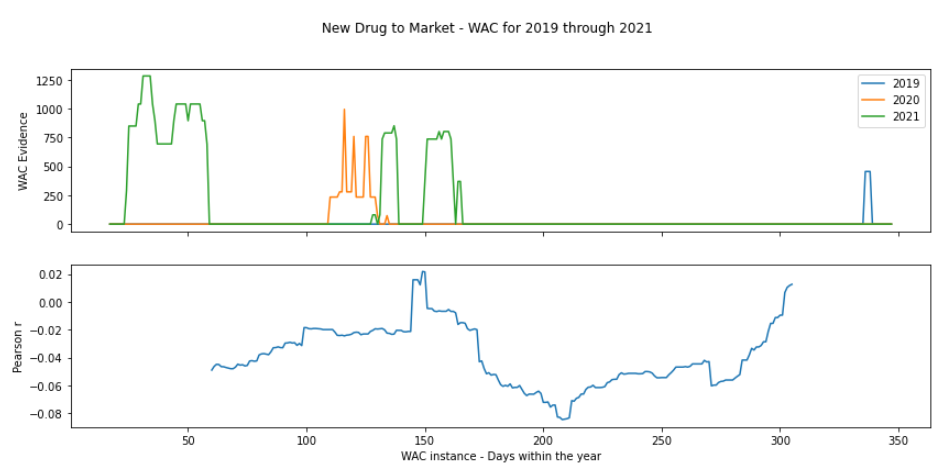


**Time Series Correlation**

When dealing with time series, there will be instances where data are not captured for set dates between years. Initial cleaning of data involves populating all days within the year with zero for instances where new drugs were not captured for the day.

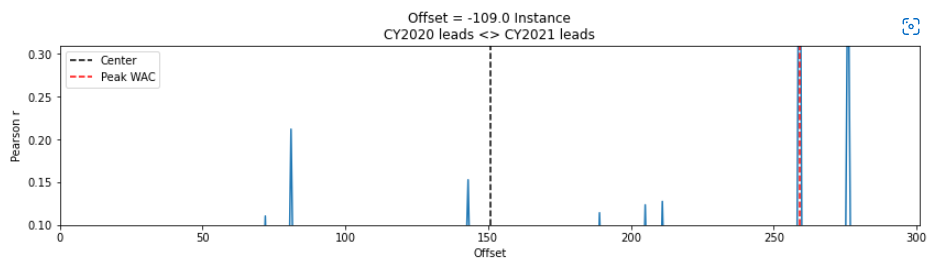
***Pearson Correlation***

*This is a measurement using Pearson correlation to identify signals along a rolling timeframe. The top image shows the instance where introduction of new drugs (WAC price is presented) to market. The bottom graph shows the comparison of 2020 to 2021 using the Pearson correlation method. Early summer is where the two years show the most correlation as far as WAC prices and introduction of new drug to market.*



***Time Lagged Cross Correlation (TLCC)***

*TLCC measures the incremental shifts of two time series by measuring the correlation between the two signals. The peak correlation here shows the center at 150 days (about 5 months) into the year. This shows that the two-time series are most aligned at that instance. The inference here is that, correlation is maximized when 2021 is pulled backward by 109 days (about 3 and a half months).*



**Heatmap correleation**

A correlation heatmap was used to identify a possible correlation between the national drug cost and the initial cost (WAC) data.



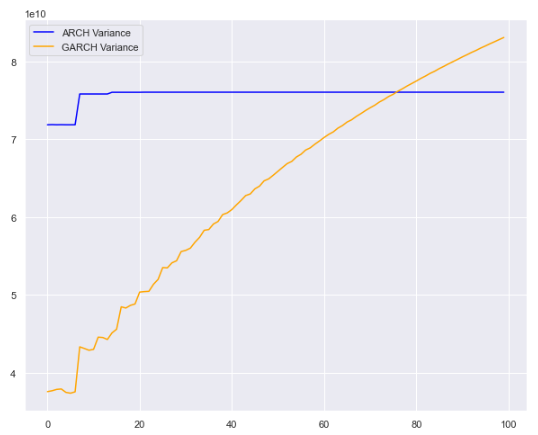
There was a slight negative correlation between the two points of data but nothing statistically significant.

# Analysis:

***MODELS:***

***ARCH & GARCH***

*The ARCH or Autoregressive Conditional Heteroskedasticity method provides a way to model a change in variance in a time series that is time dependent, such as increasing or decreasing volatility. An extension of this approach named GARCH or Generalized Autoregressive Conditional Heteroskedasticity allows the method to support changes in the time dependent volatility, such as increasing and decreasing volatility in the same series.*

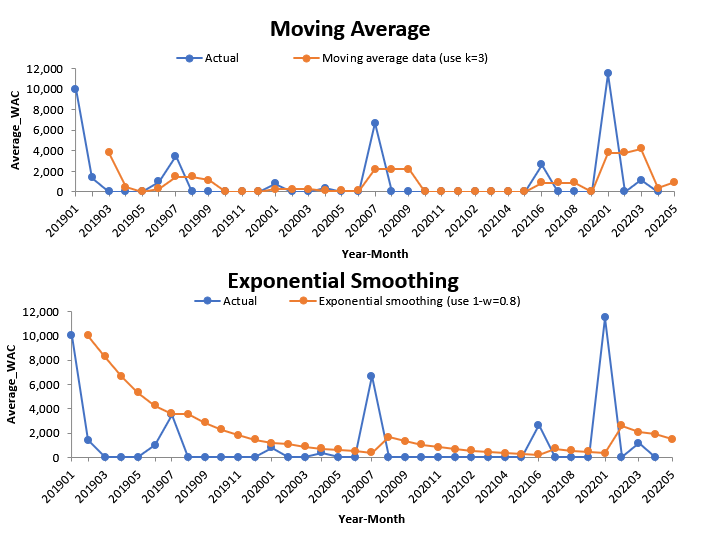


**Moving Average & Exponential Smoothing:**

Due to sporadic peaks of when new drugs are introduced to the market, moving average and exponential smoothing would not be a great model at predicting future introduction of drugs to market. The root mean square error (RMSE) for the exponential smoothing regression model yields an error of 1,697.

Model Parameters:

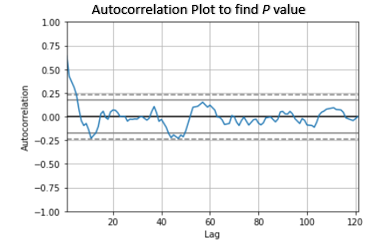
* *Moving Average*: K = 3
* *Exponential smoothing*: uses a Damping factor of 0.8



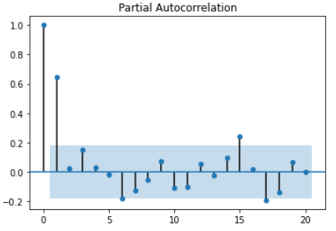
**ARIMA:**

Data was processed, and outliers removed. Residual charts show good results. However, predicted values contain negative numbers. Experimented transformation approaches such as logarithm and Exponential Smoothing, which did not help resolve the issue.

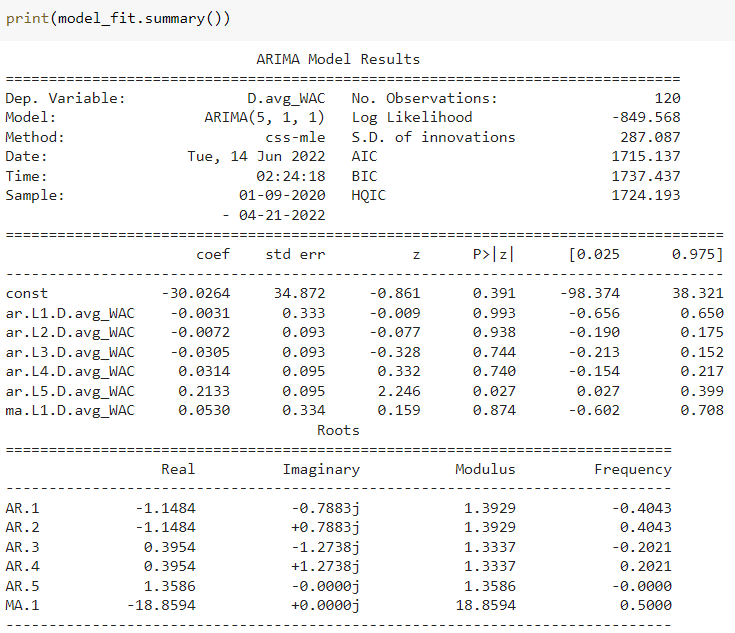
Used Autocorrelation Plot to help find the best p value and PACF to find the best q. Here the best *q was 5, or 6.*



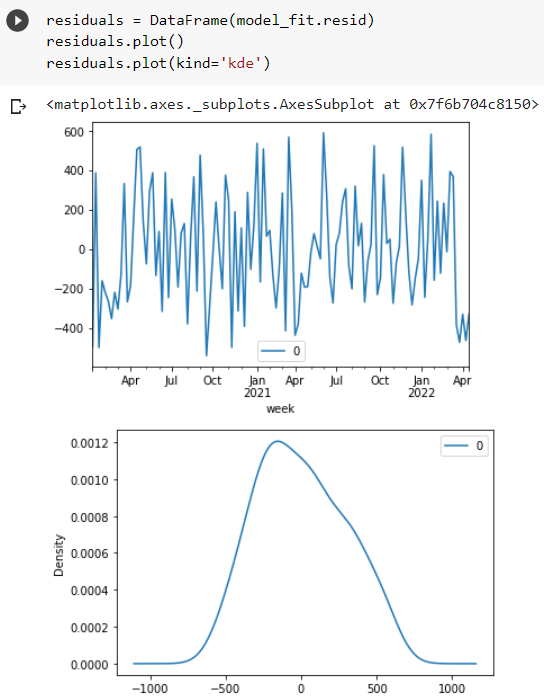
Used adfuller () function, and PACF show a p-value of 0.002903, indicating the time series data is stationary. Partial Autocorrelation chart shows the best *q* should be 1:

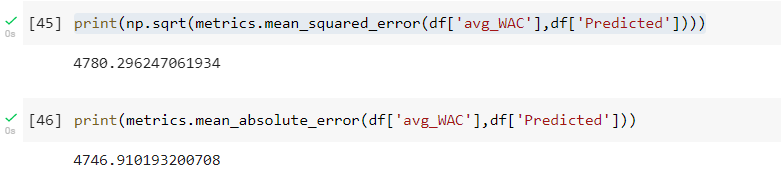


Model results:

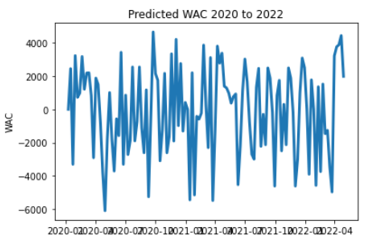


Residual charts appear stationary and normally distributed:





Predicted values, however, include negative numbers. More research needs to be done on how to restrict to positive values and improve overall model performance.

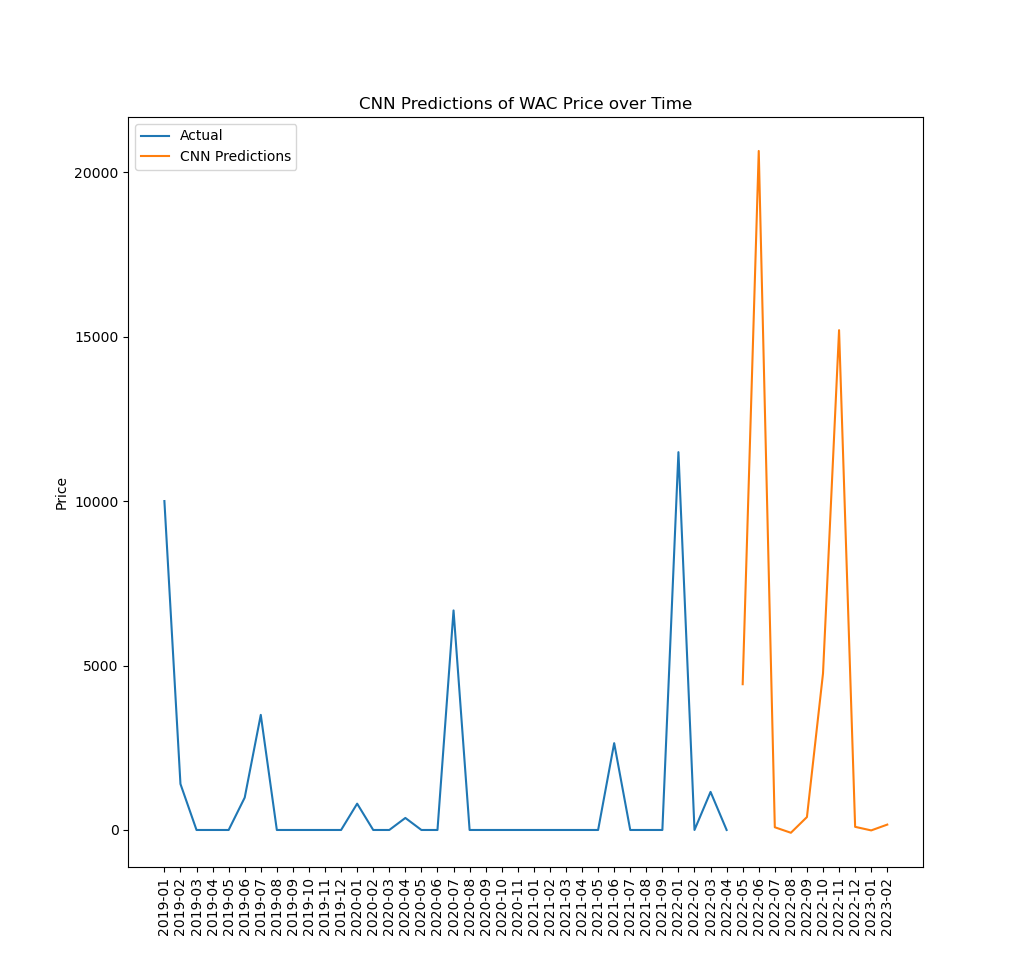


**K-Nearest Neighbors**

The K-nearest neighbor model was used to predict which drug company would have the next spike in drug cost. The previously scrubbed *rpx* dataframe was used to train and test the data. Given the sporadic nature of the data however, only a 17% validation accuracy was reached. This result concludes that this is not an effective method in predicting drug costs.

**One Dimensional Convolutional Neural Network**

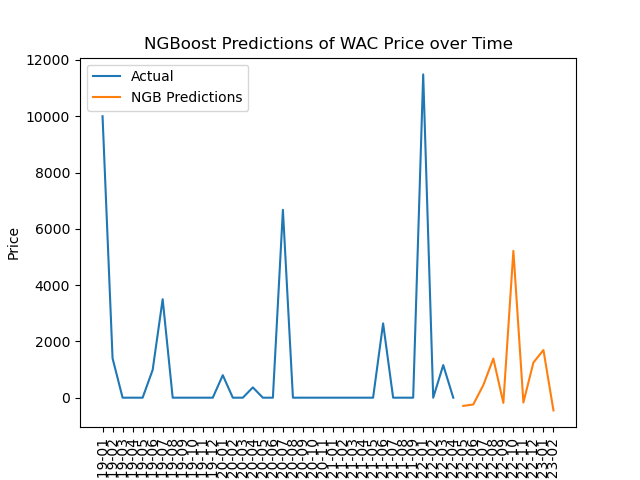
One of the deep learning methods we tried to use was a one dimensional convolutional neural network (1DCNN). This model works by interpreting the sequential time series data as an image with a height of one, and the values in the time series being treated as pixel values.



The 1DCNN produced the above forecast. The spikes it predicts are larger than anything in the actual data. The root mean squared error on the admittedly small test set was 2382.0, which is very high, but makes sense when you look at the fact that the model overestimates spikes in WAC price.

**NGBoost**

NGboost, or Natural Gradient Boosting, is a boosting algorithm developed by Stanford, that has been able to produce good results when used with time series. The data processing didn’t need to be changed from the 1DCNN.



The RMSE from NGBoost was 431.3, which is a lot lower that both ARIMA and 1DCNN. This model, as seen in the forecast, also predicted more reasonable spikes in price, which most likely contributes to the lower RMSE. This model has the issue of making slightly negative predictions, which don’t make sense. However, overall, this model is better than ARIMA and 1DCNN.

# Recommendation:

* Collect more data so results are not so sporadic and less skewed from outliers
* Provided we were to obtain more data, NGBOOST and LSTM models are very promising

# References:

* Medicare Enrollment data: <https://data.cms.gov/browse-data-categories>
* Medicare Data dictionary: <https://data.cms.gov/resources/medicare-part-d-spending-by-drug-data-dictionary>
* Medicare Pharmacy Data (API): <https://data.cms.gov/data-api/v1/dataset/54426646-2108-48e7-a339-730dcfabbe9a/data>
* Prescription Drugs introduced to the Market: <https://data.chhs.ca.gov/dataset/prescription-drugs-introduced-to-market>
* New drug Data 2022: <https://data.chhs.ca.gov/dataset/e54d331c-65d3-4c6e-b4ba-390bd7024248/resource/eded767c-4651-46d6-8b40-d34db9560091/download/newdrugdata-mar2022.xlsx>
* New drug Data 2021: <https://data.chhs.ca.gov/dataset/e54d331c-65d3-4c6e-b4ba-390bd7024248/resource/ba710f43-749d-4c6f-b105-28845cd742b7/download/ctrx-new-drug-report-2021-q1-q2-q3.xlsx>
* New drug Data 2020: <https://data.chhs.ca.gov/dataset/e54d331c-65d3-4c6e-b4ba-390bd7024248/resource/9a37709e-4151-4300-a031-91ea7813da60/download/ctrx-new-drug-report-2020-q1-q2-q3-q4.xlsx>
* New drug Data 2019: <https://data.chhs.ca.gov/dataset/e54d331c-65d3-4c6e-b4ba-390bd7024248/resource/6393cda0-6101-424e-b738-07aee9e6f94f/download/ctrx-new-drug-report-2019-q1-q2-q3-q4.xlsx>
* Time series Correlation - <https://towardsdatascience.com/four-ways-to-quantify-synchrony-between-time-series-data-b99136c4a9c9>