Logo

Description automatically generatedA picture containing shape

Description automatically generated

Rohan Sudhir

Military Equipment Analysis

Table of Contents

[Executive Summary 4](#_Toc88940190)

[Data Understanding 5](#_Toc88940191)

[Data Cleaning 6](#_Toc88940192)

[Results and Explanation 7](#_Toc88940193)

[Summary and Expert Judgment 8](#_Toc88940194)

# Executive Summary

In this report I explore data about the transfer of Military equipment to local law enforcement departments. I explored the data and found answers to these questions:

* State:
  + How much each state spent totally (all time)?
  + How much the top 5 states spend per year?
  + Which state spent the most per year?
* How much was spent each year totally?
* Most profitable item? How much was sold per year?
* Items:
  + How many shipments/transactions were made each year?
  + Which items resulted in the most total cost (all-time)?
  + Which items resulted in the most total cost (yearly)?
  + Which were the most common items of all time?
  + Which were the most common items per year?
* Is there a trend between DEMIL Code and total cost sum?

I also designed a Neural Network model to predict the DEMIL Code of a given row of attributes. I created a model using a Sequential neural network that has a predictive accuracy of 98% and a loss rate of 8%.

My recommendation was to increase shipment of items with DEMIL Code C if generating revenue is the main focus; Mine Resistant Vehicles particularly generate a disproportionate amount of revenue when compared to other items. If generating volume is the main focus, then focus on shipments with DEMIL Code D as they are far more frequently shipped historically. A special case is the Unmanned Vehicle which falls into DEMIL code Q but also generates significant revenue.

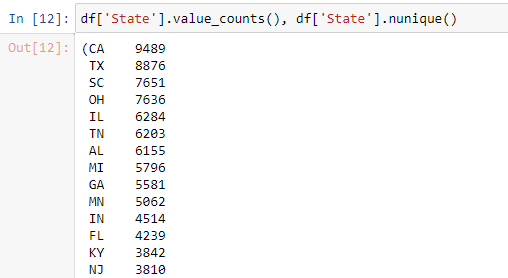
# Data Understanding

The dataset concerns the transfer of excess equipment of the DoD (Department of Defense) by the DLA (Defense Logistics Agency) to the general public, which entails law enforcement agencies around the US on a state level.

The columns are as follows: State, Agency Name, NSN, Item Name, Quantity, UI, Acquisition Value, DEMIL Code, DEMIL IC, Ship Date and Station Type.

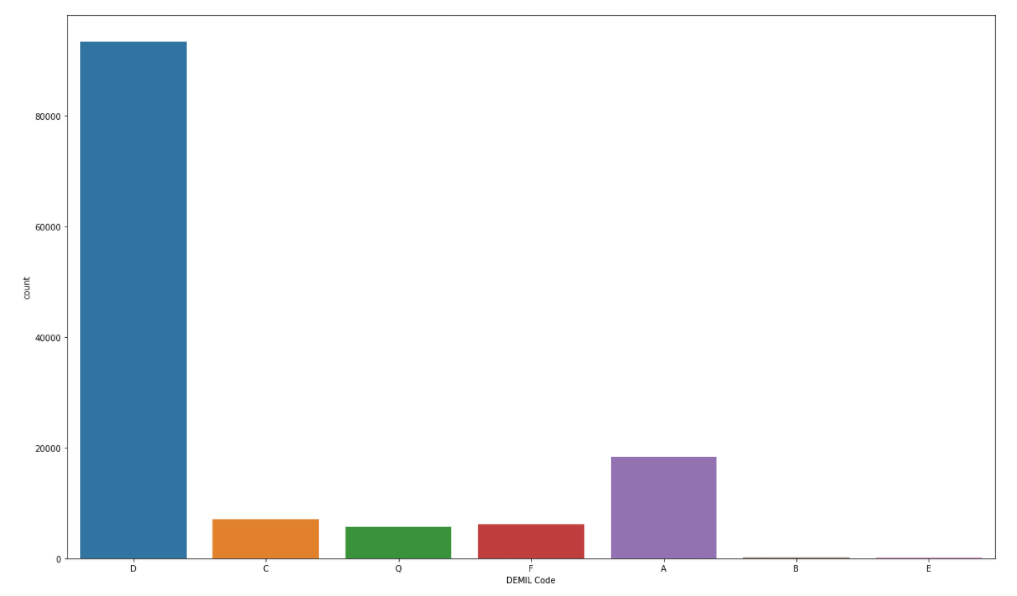
I began by concatenating the multiple sheets into one data frame so that all states were accessible at once.

State Value Count



I found that California, Texas, South Caroline, Ohio and Illinois were the states with the most entries and hence the most orders.

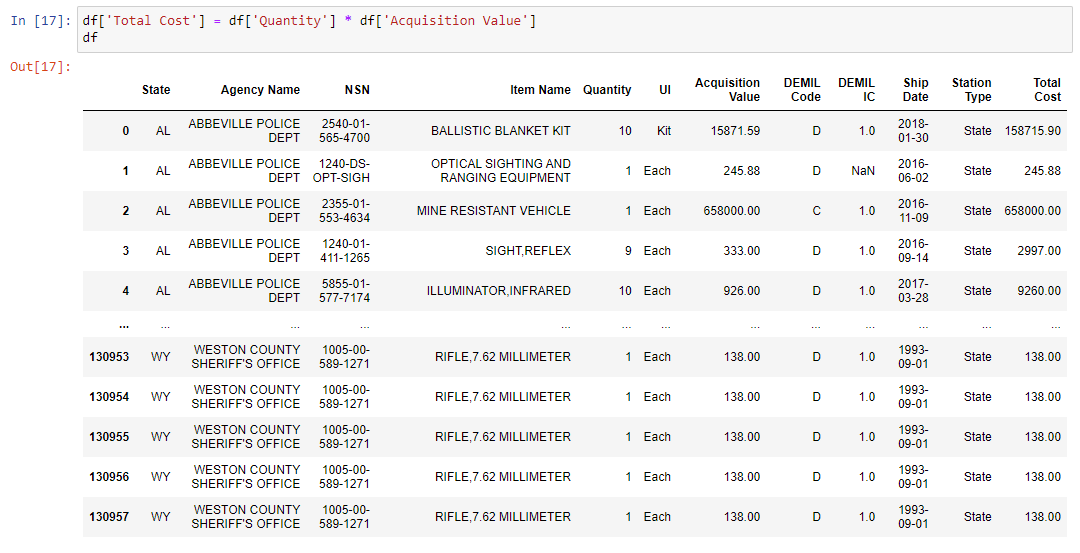
DEMIL Count



I looked at the DEMIL count and found that the most frequently appearing codes were D then A and then C.

Total cost

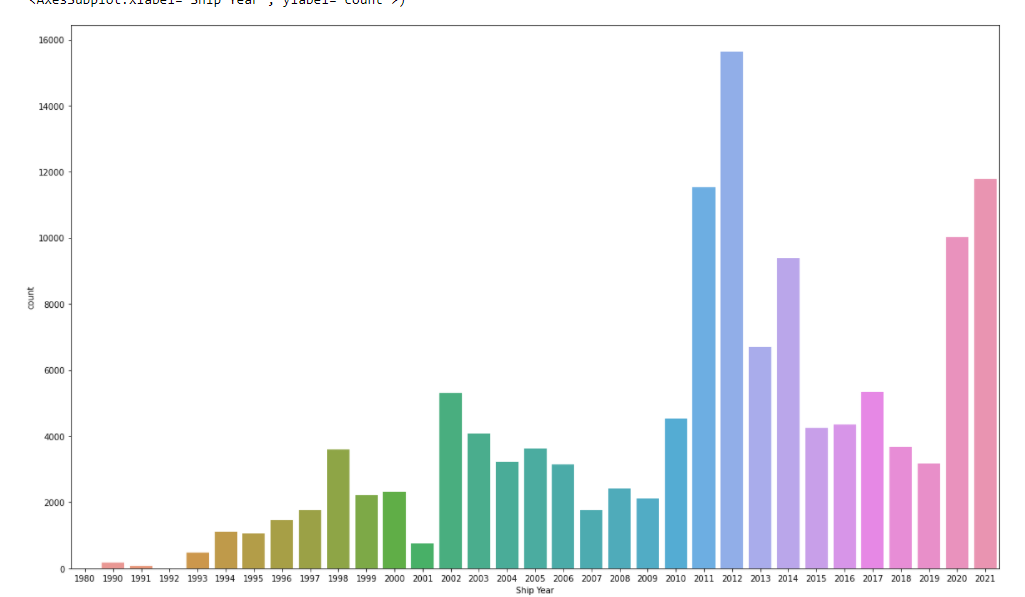
I realized that acquisition cost was the price per unit of an item and so to better understand the total cost per item entry, I created a new column ‘Total Cost’ which was quantity\* acquisition value for each row.



Shipments per year

I then wanted to see how many shipments were made per year. I did this by creating a new column called “Ship Year” by formatting the “Ship Date” columns and then plotted a count plot of the new “Ship Year” column.

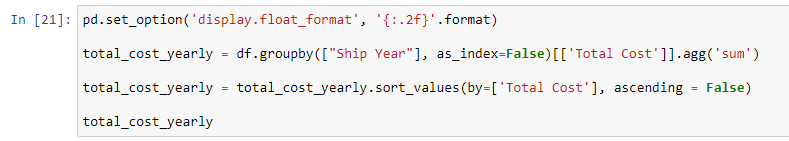


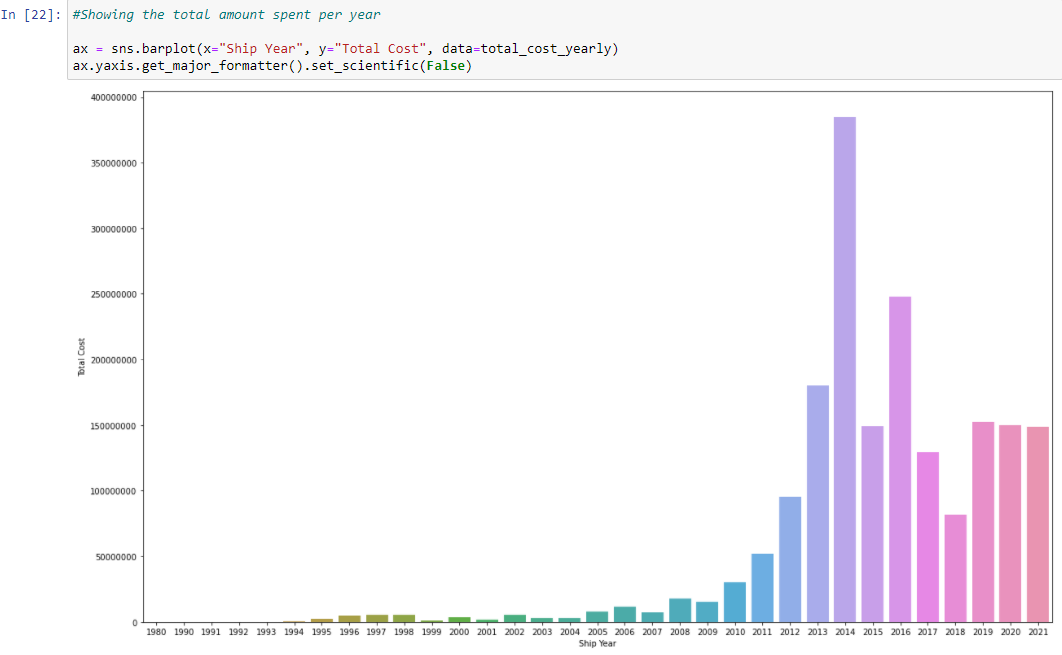


2012 had the most shipments out of any year.

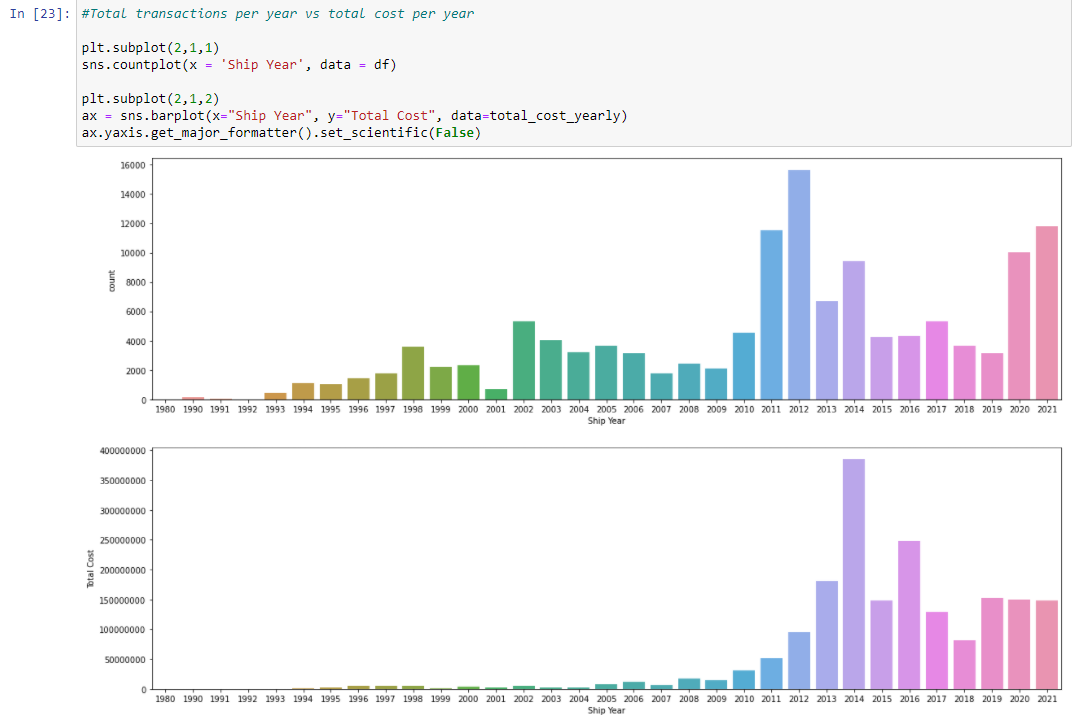
Total cost generated per year

I then wanted to see whether the total cost correlated with the number of shipments per year. I did this by creating a dataframe that grouped columns by Ship Year and then summed the total cost per ship year. I then plotted this using a barplot.





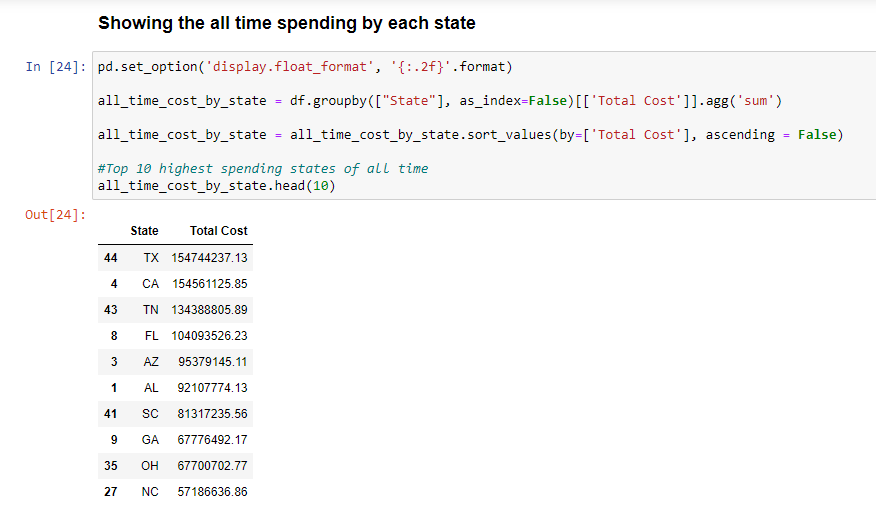
Despite 2012 being the year with the most shipments, 2014 was the year most revenue was generated. A comparison of shipments per year vs total cost per year can be seen below:



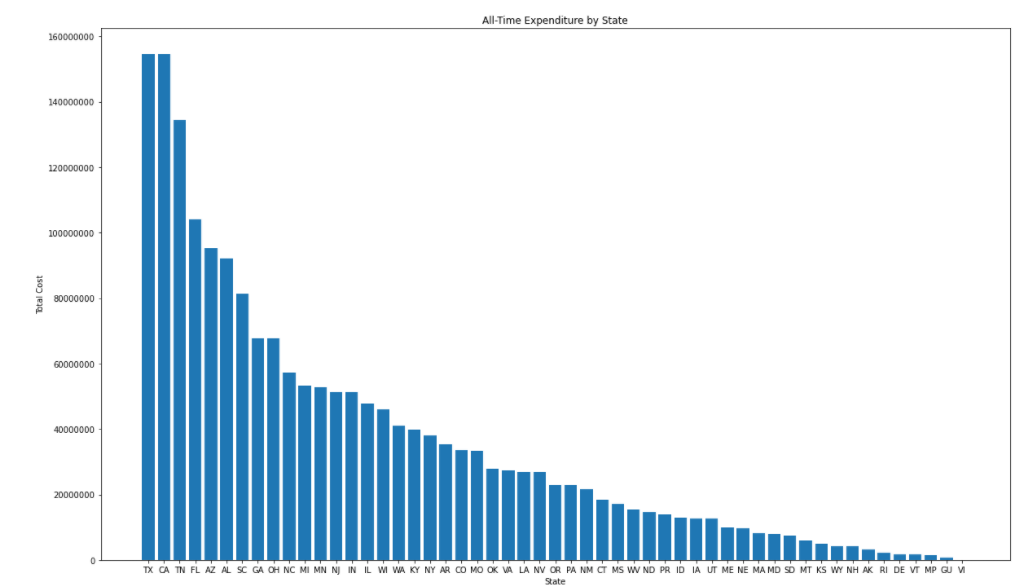
This hinted at the fact that there are perhaps disproportionately priced items that generate more total cost when sold in lower quantities because despite the fewer shipments in 2014, most total cost was generated that year.

Highest spending states

I then took a look at which states spent the most by creating another dataframe that grouped the State column and summed the Total Cost.

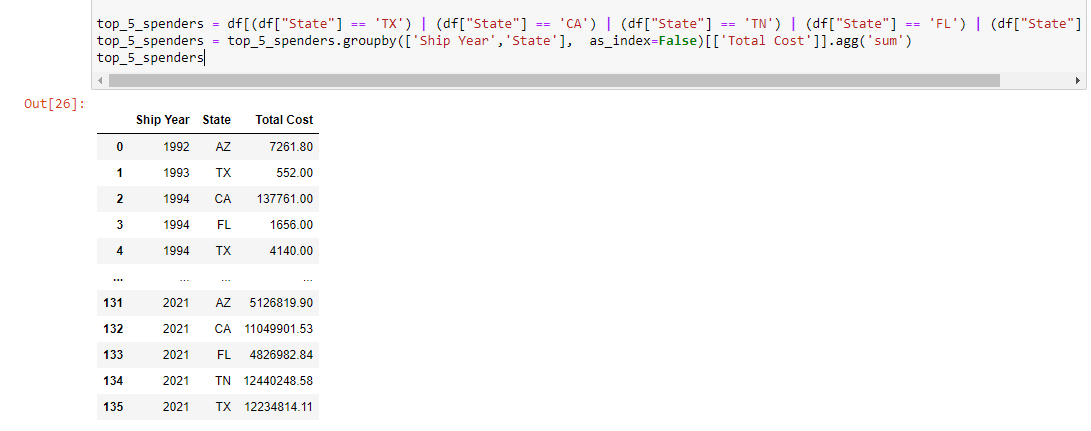


I then plotted this dataframe:

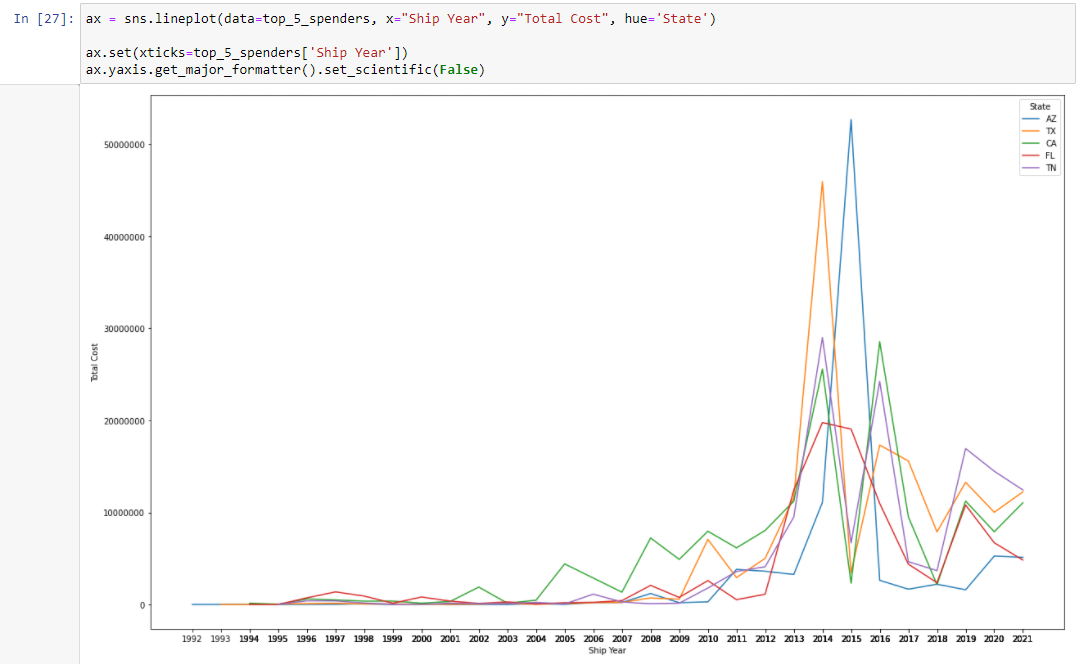


The top 5 highest spending states are: Texas, California, Tennessee, Florida and Arizona.

I then looked at the spending patterns of these 5 states. I created a dataframe that grouped by the top 5 spending States and Ship year and then summed Total Cost to see how much each State was generating in total cost per year.

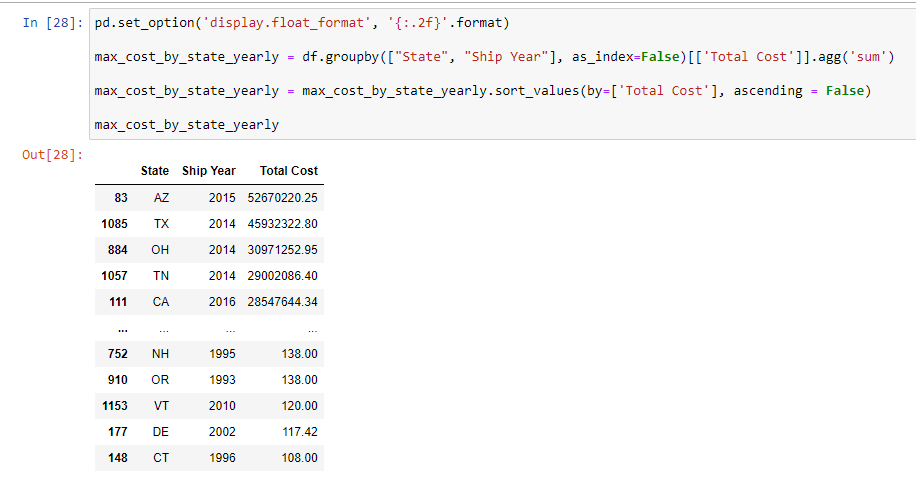


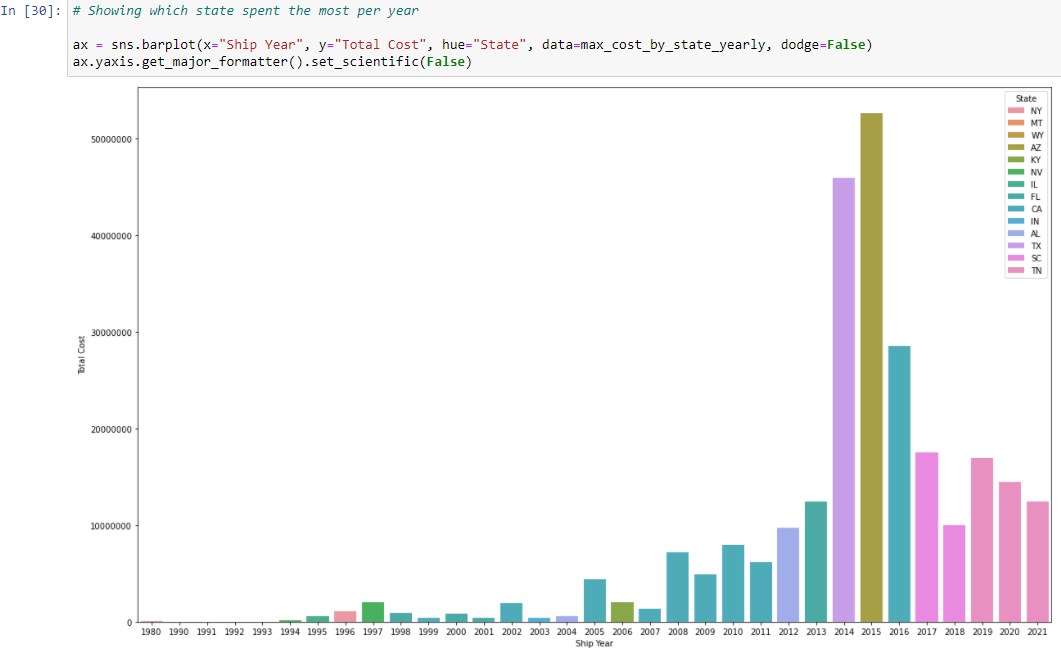
Line plot of this dataframe:



This showed that 4 out of the 5 highest spending states spent a lot of money in the year 2014, the only exception being Arizona which spent much more money in 2015.

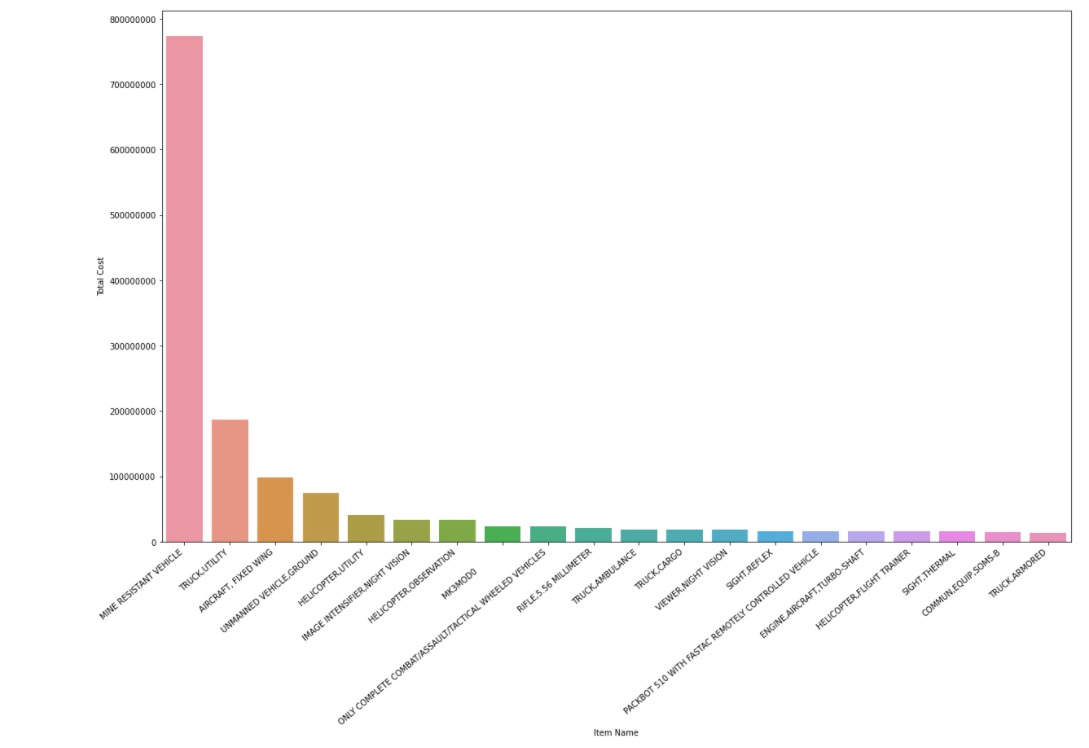
Lastly, I wanted to see which state was spending the most per year. I created a dataframe that grouped by all states, ship year and summed total cost. Then plotted a barplot to visualise the dataframe.





Total cost generated per item

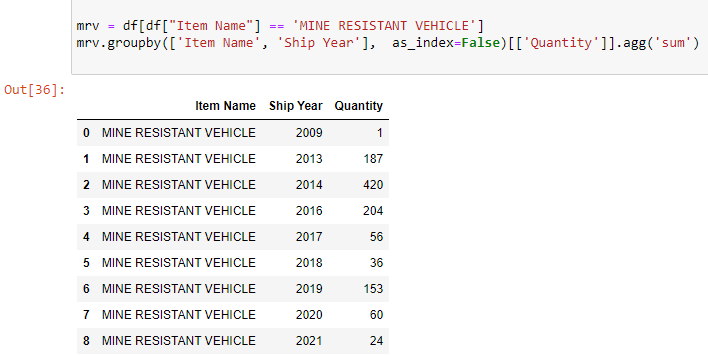
I wanted to see which items were generating the most money. I created a dataframe that grouped by Item name and summed total cost. The plot of this data frame showed:



Mine Resistant Vehicles (MRVs) generated the most money by far.

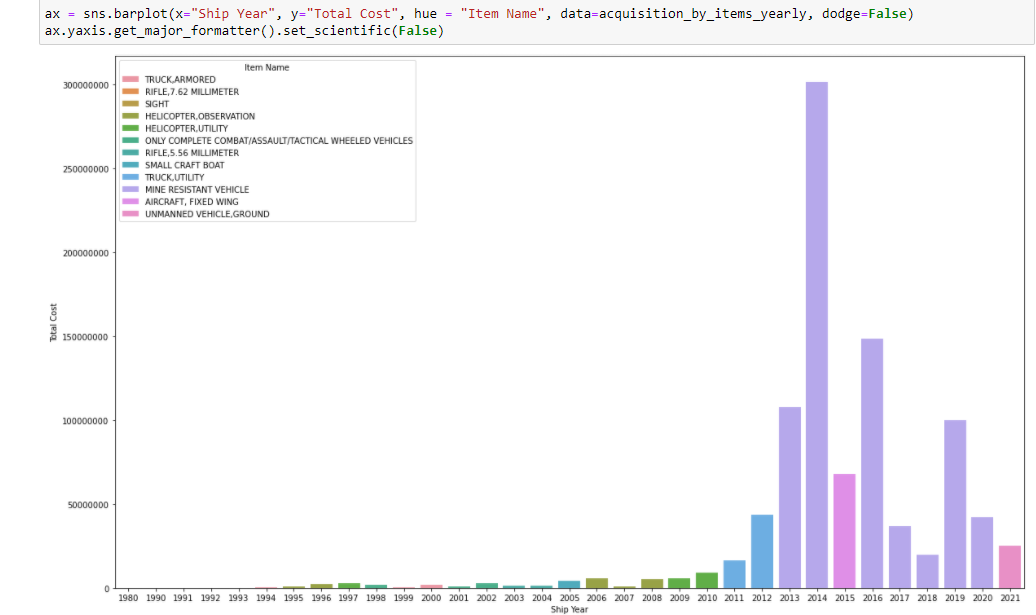


After looking into the quantities of MRVs being sold per year, I found a correlation.



I see here that the highest years of distribution of MRVs were in 2014, then 2016 and finally 2013. The years with the highest total cost overall were also 2014, then 2016 and then 2013. MRVs are a significant contributor to those years being the highest generators.

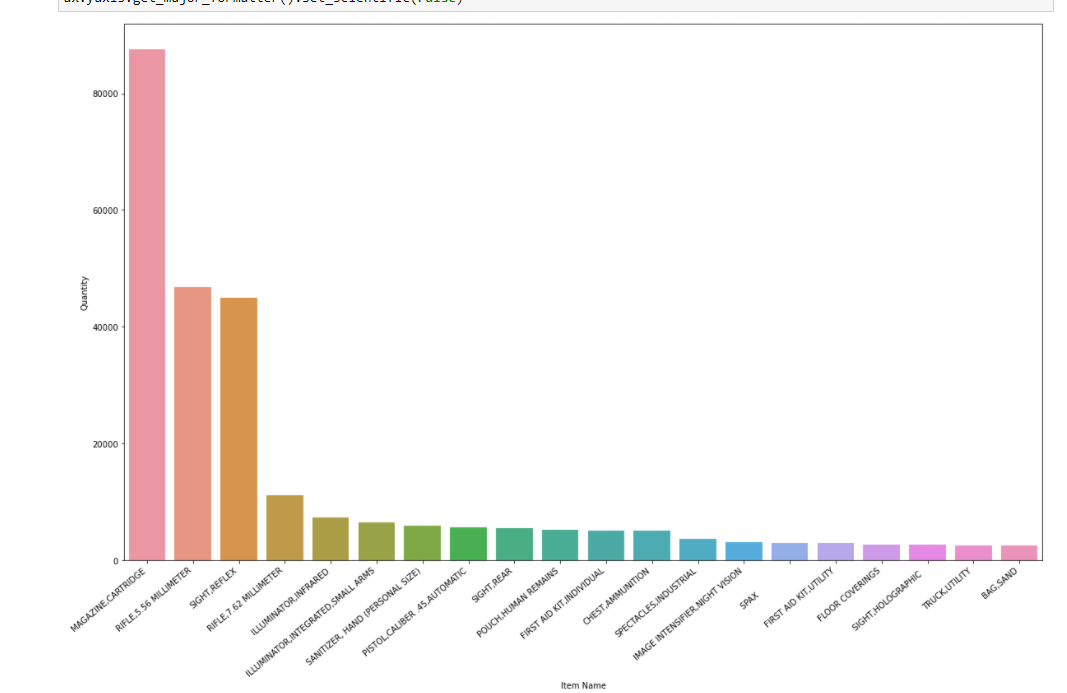
I then looked at the most profitable items per year. I created a data frame that grouped by Item Name and Ship Year and summed the Total Cost. After plotting this data frame I can see:



MRVs shipments generated the most total cost for the most amount of years (7).

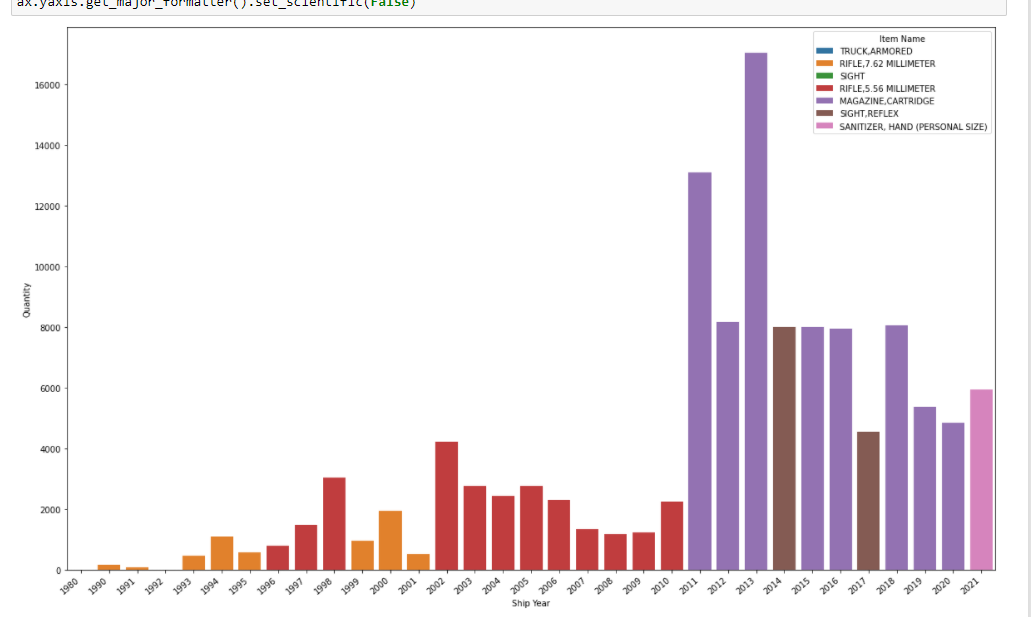
Most common items sold

I created a dataframe that grouped by Item name and summed the quantity. Plotting this showed:



Magazine cartridges were the most shipped item in terms of quantity. Rifle 5.56, millimeter and then sight,reflex were the next most shipped.

By year:

5.56 Millimeter rifles were the most shipped item for 12 years. Magazine cartridges were the most shipped for 8 years however they were shipped in much higher quantities than the rifles.

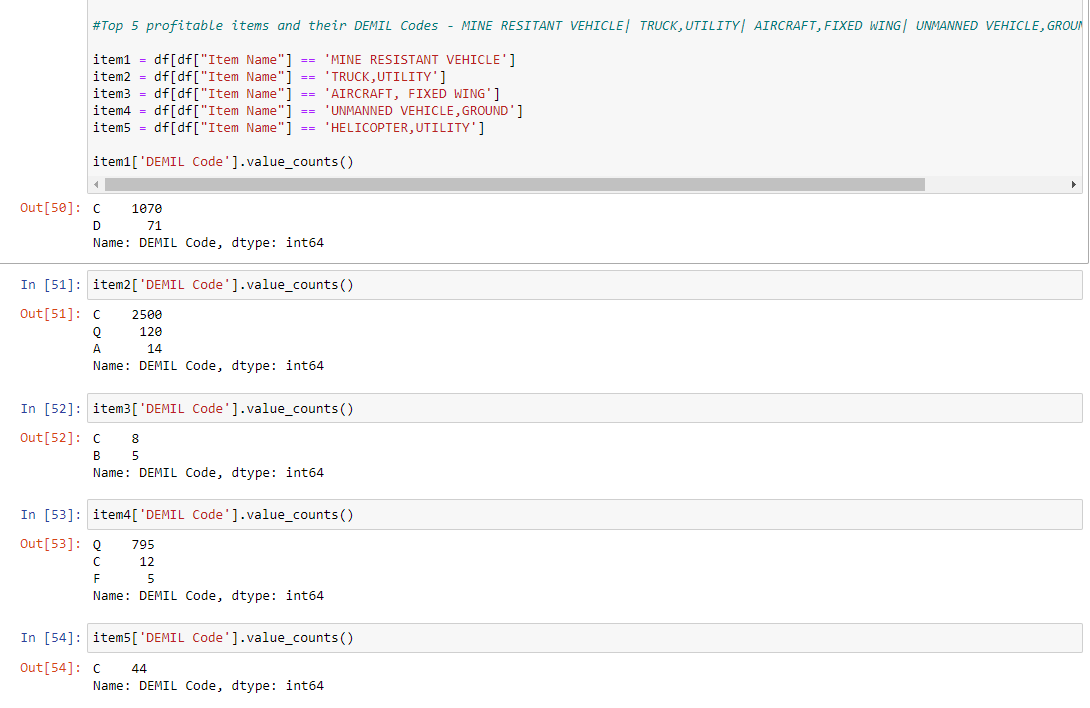
Trend between DEMIL Code and Total Cost

I plotted DEMIL Codes against the total cost that they generated.



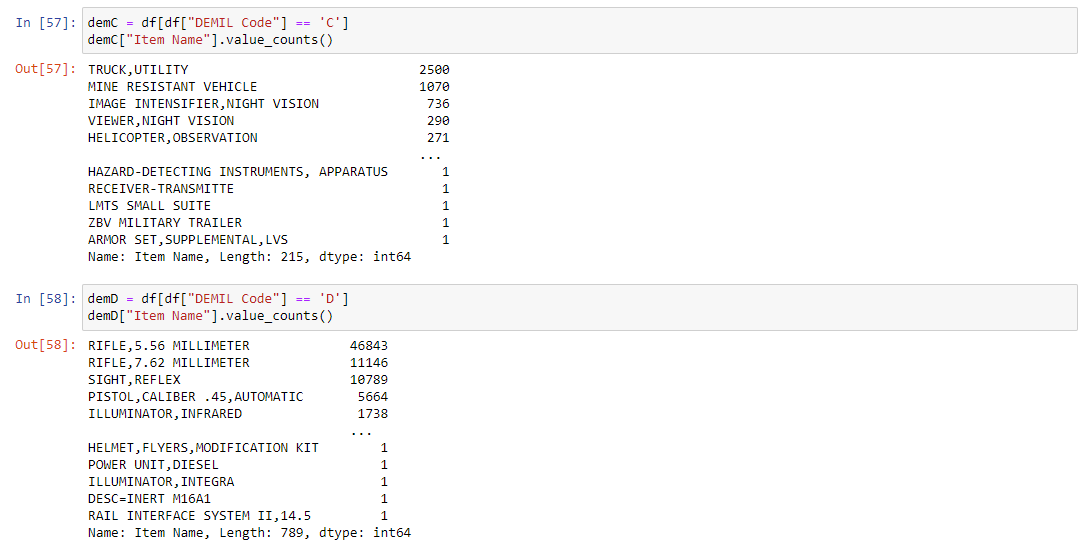
Items with DEMIL code C generated the most total cost.

I then checked which code the 5 most profitable items were associated with.



Out of the top 5 most profitable items, 4/5 have DEMIL Code C as the majority.

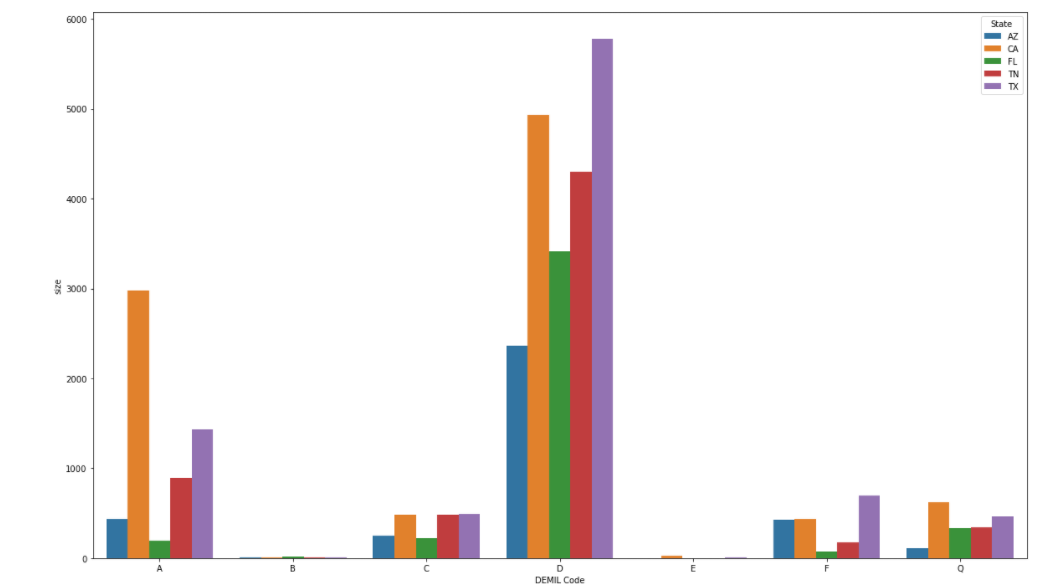
I then checked the highest frequency of items in each DEMIL Code. Of interest were code C and D:



DEMIL Code C has a high frequency of Truck, Utility and Mine Resistant Vehicles which were determined to be the two most profitable items.

Code D has a very high frequency of rifles and rifle related gear, however the acquisition cost of these items is not as a high as the items found in class C.

I looked at the DEMIL Codes that the items bought by the top 5 spending states fell into:



The top 5 spenders buy item types that mostly fit into DEMIL Code D.

# Data Cleaning

\*Note Total cost and ship year are custom made columns from earlier.

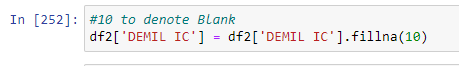
1. Dropped 'Agency Name', 'NSN', 'Quantity', 'Acquisition Value', 'Ship Date', 'Station Type', 'Ship Year' from as feature columns. (Dropped acquisition and quantity because the newly made total cost column was a multiplication of those two rows)



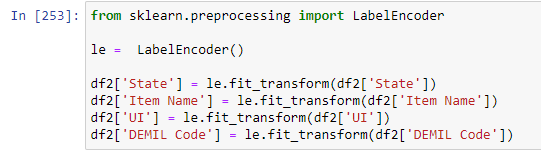
1. Removed rows with null values in item name and UI



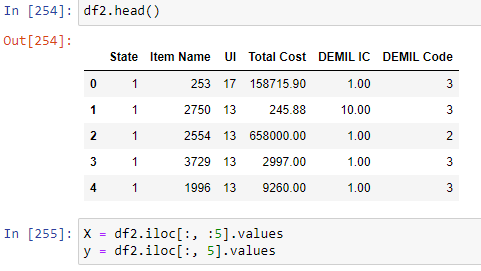
1. Filled in blank values in DEMIL IC with placeholder value 10.



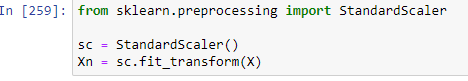
1. Label encoded Categorical columns (State, Item Name, UI and DEMIL Code)



1. Set X to feature columns (State, Item Name, UI, Total Cost, DEMIL IC)



1. Standardized all feature columns using StandardScaler from sklearn



1. Created a train test split of 70/30.



# Results and Explanation

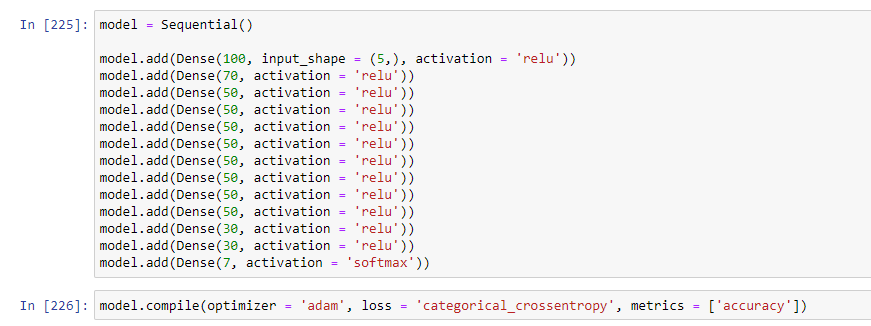
Predicting DEMIL Code:

With every subsequent model in this section, the parameters of the previous model were maintained unless indicated otherwise.

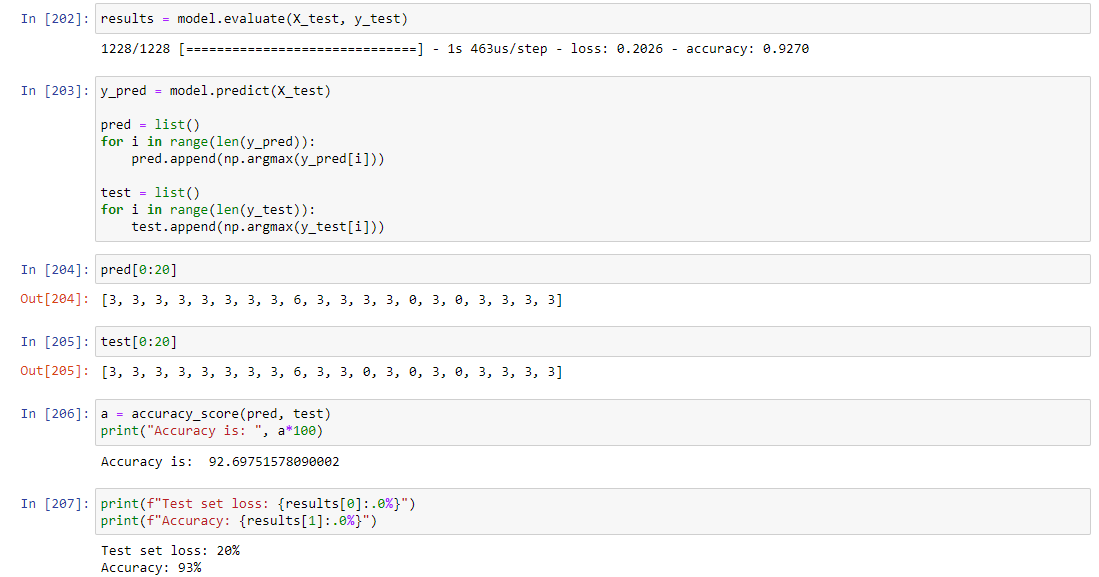
Testing model (random nodes and layers)

I used a Sequential neural network from the keras library.

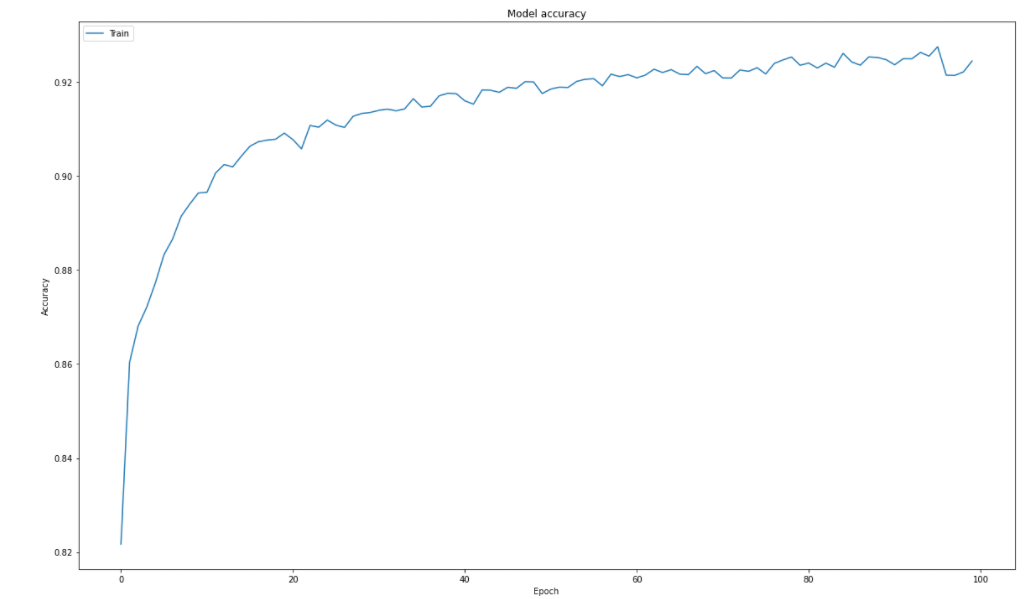
I used a random number of layers and nodes with an epoch of 100 and a batch size of 50. Optimizer used was adam and loss function was categorical\_crossentropy. The metric used was accuracy.



After fitting the model and evaluating I compared the train accuracy to the test accuracy.



The difference between the two was 0.3% which indicates that there was any significant underfitting or overfitting. I then plotted the accuracy against number of epochs:

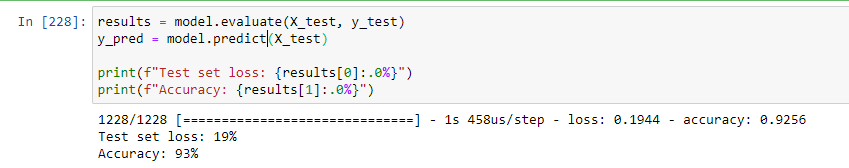


Accuracy:93%  
Loss: 20%

Model 1

The accuracy plateaued at around 80 epochs. So, I set the epochs to 80 and increased the batch size to 128 for faster processing.



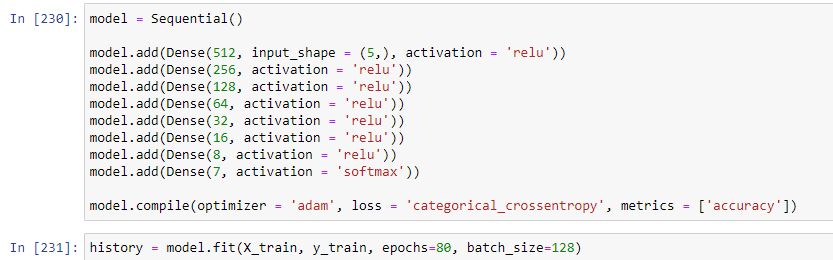


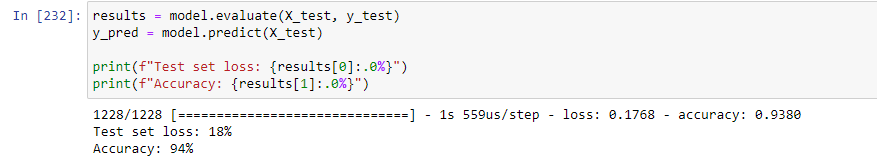
An improvement in test set loss can be seen here after epoch and batch size is increased.

Accuracy:93%  
Loss: 19%

Model 2

Here I changed the number of layers and nodes per layer to find an optimal improvement. I settled on 7 layers (not including output) with nodes in the first layer start at 512 and decreasing by a factor of 2 with every further layer.



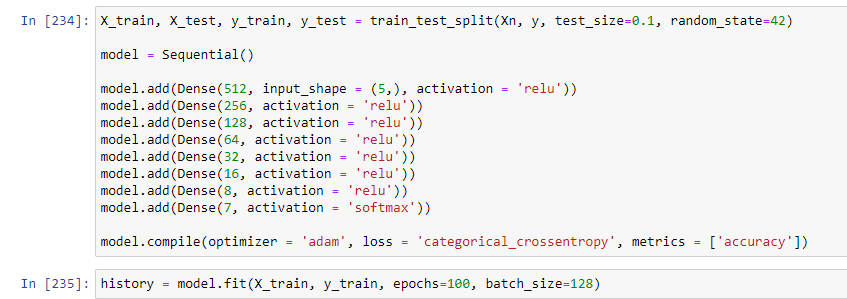


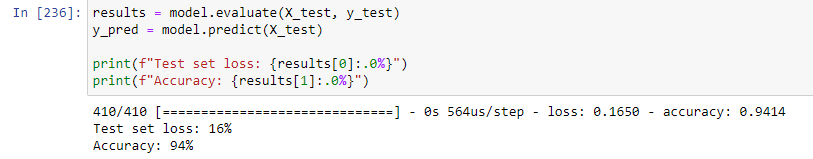
An increase in results were seen up until 7 layers (not including output layer). After that the results plateaued at 94% accuracy around 17/18% test set loss.

Accuracy: 94%  
Loss: 18%

Model 3

With this model, I tried changing the train test split. With bigger datasets it becomes less necessary to have a large test split. I increased the split from 70/30 incrementally and saw the best results at a split of 90/10. I also increased the epoch to 100 as the accuracy was steadily climbing even at 80 epochs.





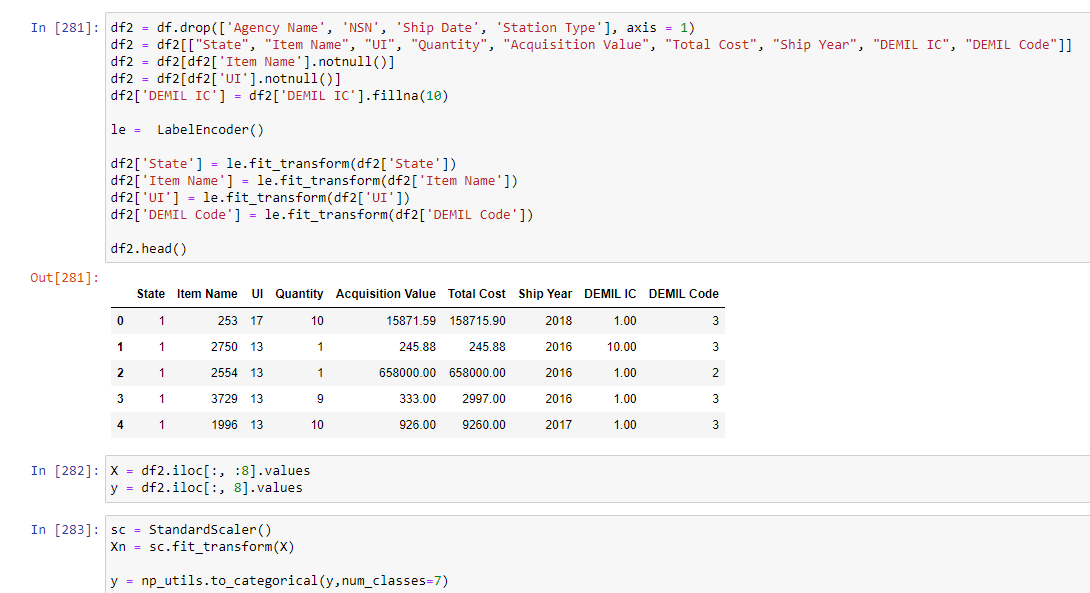
I saw a significant decrease in test set loss with this approach.

Accuracy: 94%  
Loss: 16%

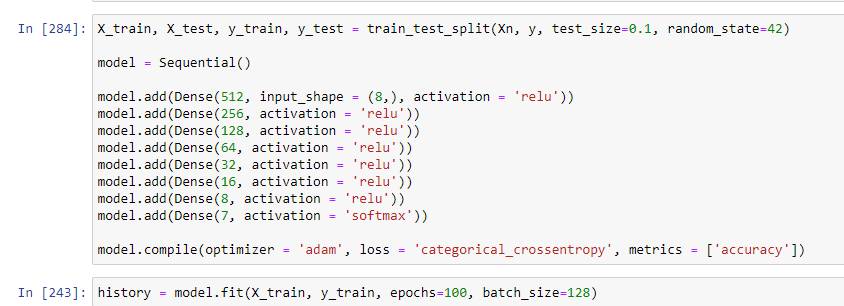
Model 4

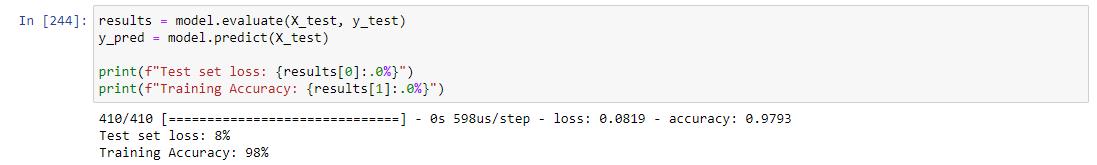
With this model, I tried changing the feature columns to include total cost, acquisition value, quantity and ship year.

Cleaning:



Train test split + model layers and nodes:





With these modifications, I see a drastic improvement in accuracy by 4% and a drop in loss rate by 8%

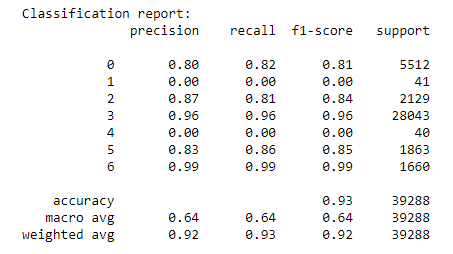
Accuracy: 98%  
Loss: 8%

Classification report comparisons

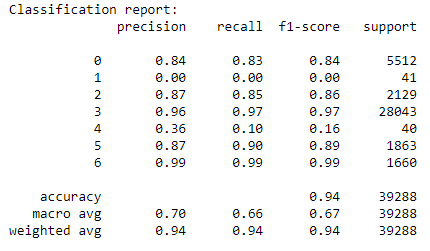
Class Key:

0 – A  
1 – B  
2 – C  
3 – D  
4 – E  
5 – F  
6 – Q

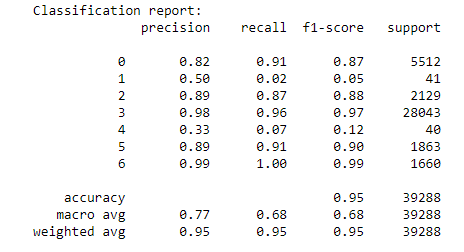
Model 1:



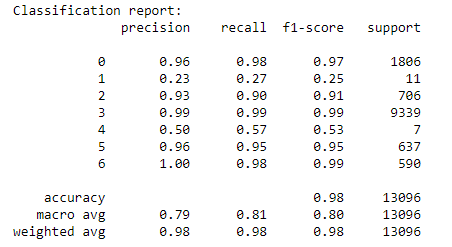
Model 2:



Model 3:



Model 4:

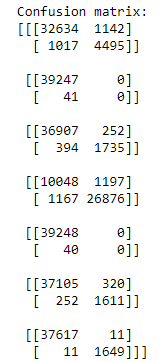


I can see that all the models had difficulty predicting DEMIL Code B and E correctly, likely due to the very low frequency of items that have this code. Model 4 has the highest overall f1 scores for every class prediction accuracy.

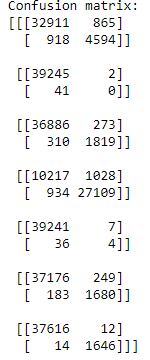
Confusion matrices:

The matrices are in order of the class key above.

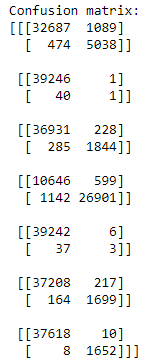
Model 1:



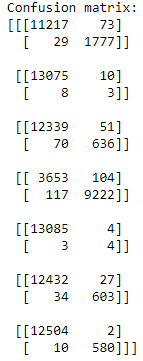
Model 2:



Model 3:



Model 4:



Again model 4 has better ratios of true positives to false positives and true negatives to false negatives overall compared to the other models.

# Summary and Expert Judgment

# Model 4 was the best predictor model for prediction. Here is a summary of the model:

1. X = "State", "Item Name", "UI", "Quantity", "Acquisition Value", "Total Cost", "Ship Year", "DEMIL IC"
2. Label encoded categorical columns and standardized feature columns
3. Sequential Neural Network model with 7 layers (not including output layer), 512 nodes in first layer down to 8 in the last layer
4. Optimizer = adam, loss = categorical\_crossentropy
5. Epochs = 100, batch\_size = 128

This model resulted in a 98% prediction accuracy and a loss rate of 8%.

There were many patterns found within the data. The most notable line of patterns identified revolved around the Mine Resistance Vehicles. These items were responsible for the most amount of money generated historically in the data. The years the sold in high quantities were the years that the most total cost was generated. DEMIL Code C items in general were responsible for generating a lot of total cost and it is partially due to the high acquisition costs of items like the MRVs or the utility trucks.

Despite there being a high frequency of items that had DEMIL code D, these items generated close to the same amount of summed total cost as items with codes that had far fewer occurrences (code A and code Q).

My recommendations are dependent on the motivations of the shipments.

If the priority is generating income, then focus should be put on shipping more items with DEMIL Code C. These items have the largest total cost. In particular, MRVs and utility trucks are disproportionately profitable. One exception to this is the Unmanned vehicle, which is one of the top 5 most profitable items but falls into DEMIL code Q. This item should also be focused on shipping to improve revenue generated.

If the priority is to maintain a high order of shipments, then focus should be kept on repurposing magazine cartridges since they have been ordered in high quantities for 8 out of the past 10 years. In addition, 5.56-millimeter and reflex sights are the next two most common items shipped. All 3 of these items fall into DEMIL Code D.

In conclusion, to generate revenue, focus on DEMIL Code C items. To generate volume, focus on DEMIL Code D items. A special case is the Unmanned vehicle which is a DEMIL Code Q item that also generates a significant amount of revenue.