**Medical Device Proposal (Final Report)**

**Problem Statement**

TG Medical (USA) Inc. has been a worldwide leader in manufacturing and specialization in disposable gloves. Using the Medicare Durable Medical Equipment, Devices & Supplies dataset (via Data.CMS.gov - https://data.cms.gov/resources/medicare-durable-medical-equipment-devices-supplies-by-geography-and-service-data-dictionary), I will be proposing to management a new medical device to add on to our product list and predicting whether it will generate profit for the company.

**Data Wrangling**

The raw dataset contains 40267 rows and 18 columns, there were 2 columns with missing values and half of the columns contain strings which will need to be converted into dummy values later.

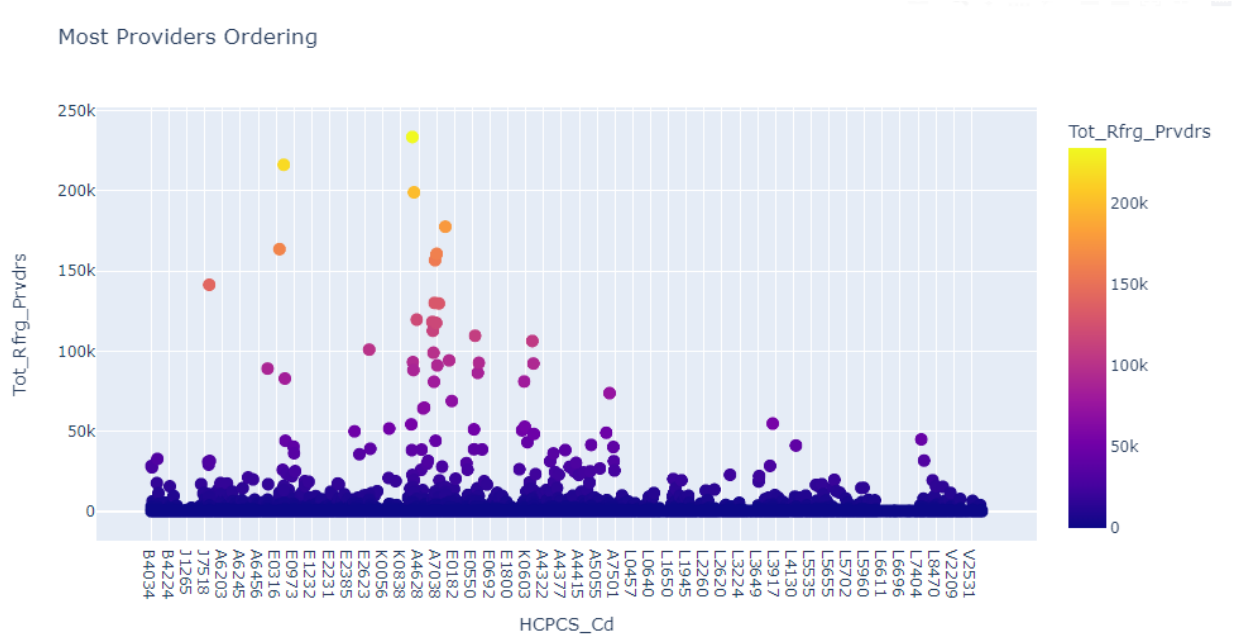
The various columns are:

- Rfrg_Prvdr_Geo_Lvl – Geography Level ('State' or 'National')
- Rfrg_Prvdr_Geo_Cd – Referring Provider Geography Code
- Rfrg_Prvdr_Geo_Desc – Referring Provider Geography Description
- BETOS_Lvl – BETOS Level (3 classifications – Durable Medical Equipment, Prosthetic and Orthotic Devices, and Drugs and Nutritional Products)
- BETOS_Cd – BETOS Code (BETOS code that is assigned to the HCPCS code)
- BETOS_Desc – BETOS Description
- HCPCS_Cd – HCPCS Code (for the specific product/service)
- HCPCS_Desc – HCPCS Description
- Suplr_Rentl_Ind – Supplier Rental Indicator (if the product/service is a rental)
- Tot_Rfrg_Prvdrs – Number of Referring Providers
- Tot_Suplrs – Number of Suppliers
- Tot_Suplr_Benes – Number of Supplier Beneficiaries
- Tot_Suplr_Clms – Number of Supplier Claims
- Tot_Suplr_Srvcs – Number of Supplier Services
- Avg_Suplr_Sbmtd_Chrg – Average Supplier Submitted Charges
- Avg_Suplr_Mdcr_Alowd_Amt – Average Supplier Medicare Allowed Amount
- Avg_Suplr_Mdcr_Pymt_Amt – Average Supplier Medicare Payment Amount
- Avg_Suplr_Mdcr_Stdzd_Amt – Average Supplier Medicare Standard Payment Amount

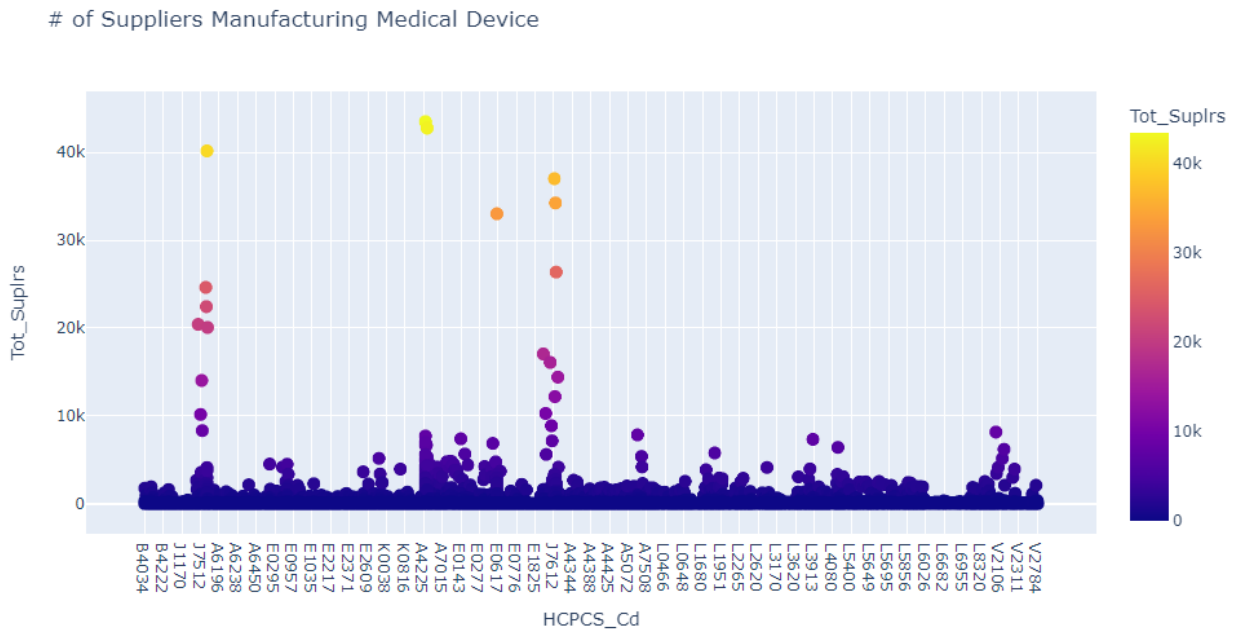
Using the describe() method, we were able to conclude that there are an average of 447 referring providers for an item, an average of 89 suppliers for an item, and an average charge of $644.40 to a provider. Rfrg_Prvdr_Geo_Cd contains categorical values, the mode was used to fill in the missing values. Tot_Suplr_Benes contains numerical values, the mean was used to fill in the missing values. There were no duplicated rows and the transformed dataset was saved to a new file 'Wrangled.csv'.

**Explanatory Data Analysis**

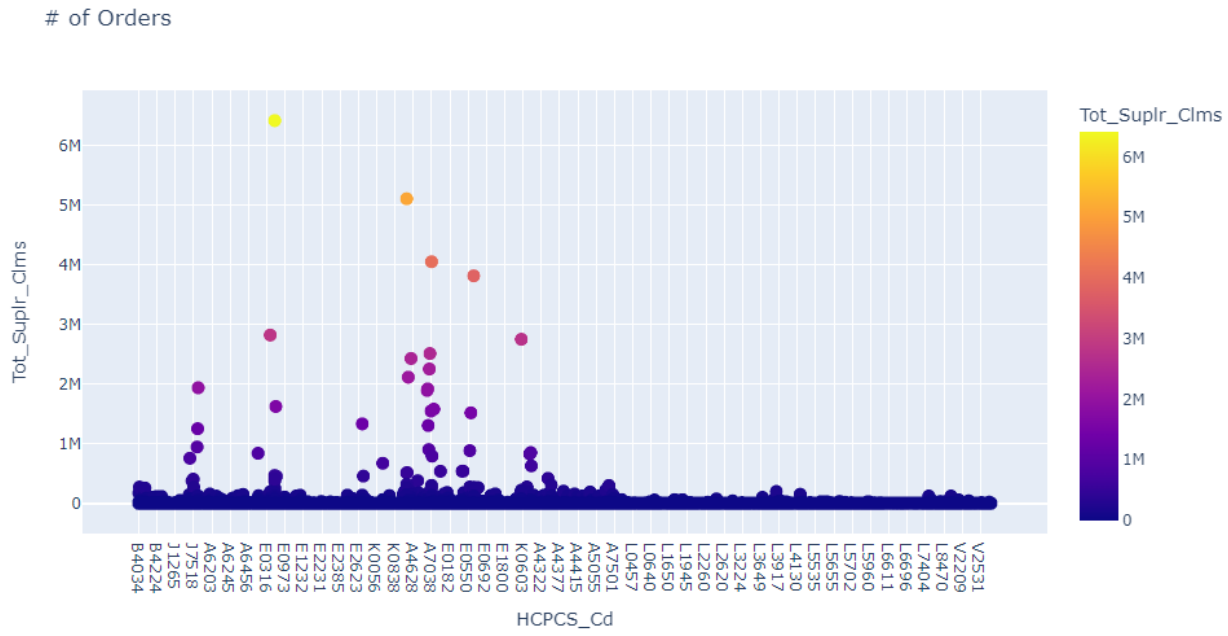
Created several scatter plots to explore the relationship between different columns:



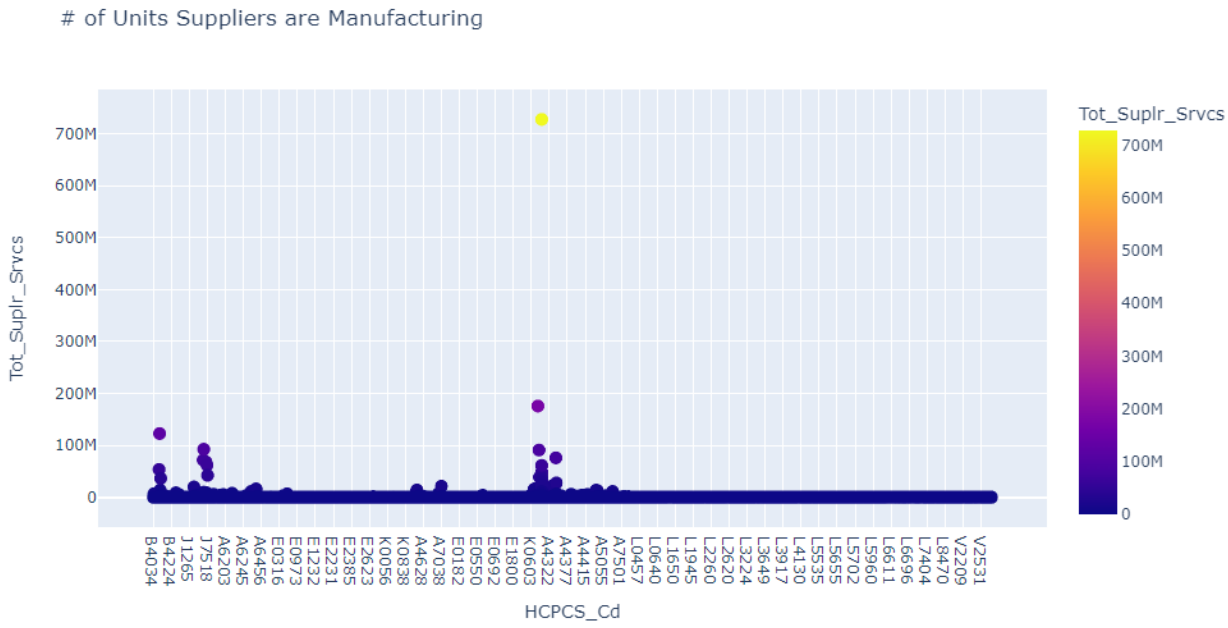

The blood glucose test/reagent strips has the most referring providers (233762 providers) meaning the most providers are ordering this item.



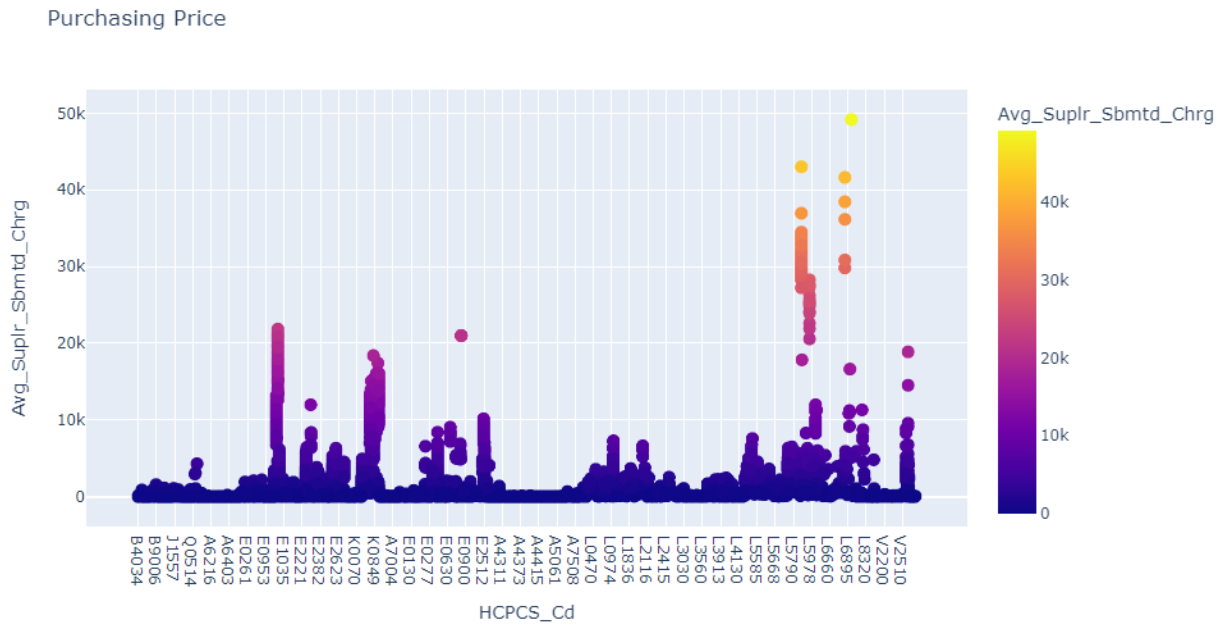

The blood glucose test/reagent strips also have the most suppliers manufacturing it.



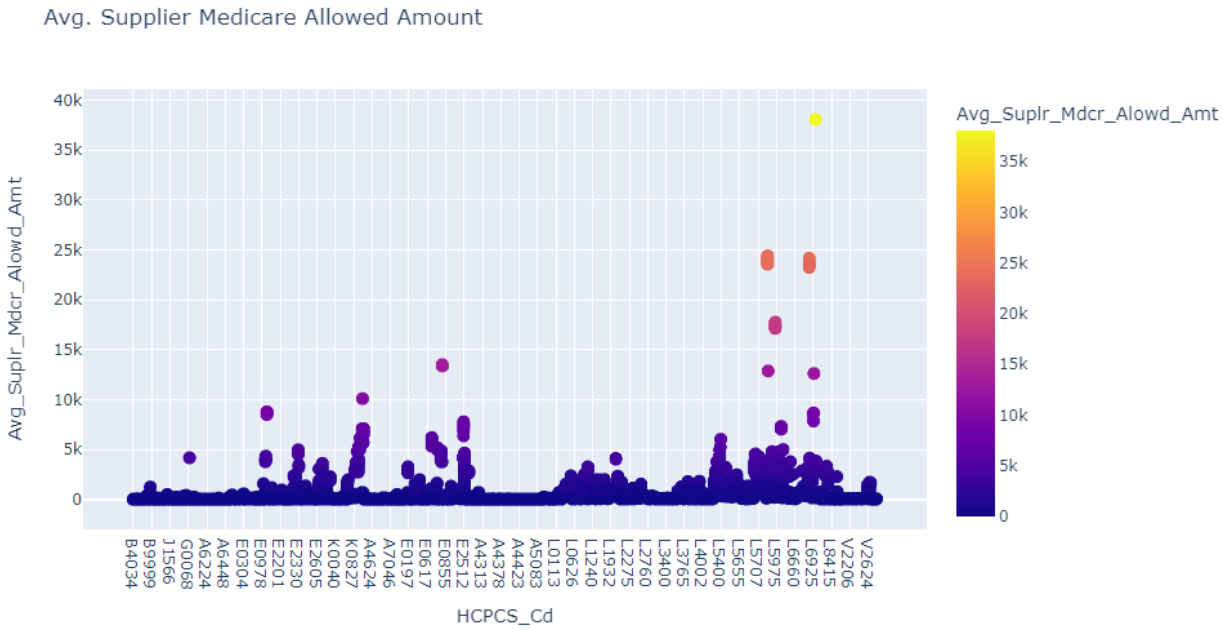

The blood glucose test/reagent strip is the 2nd most ordered item, the most ordered item is the oxygen concentrator.



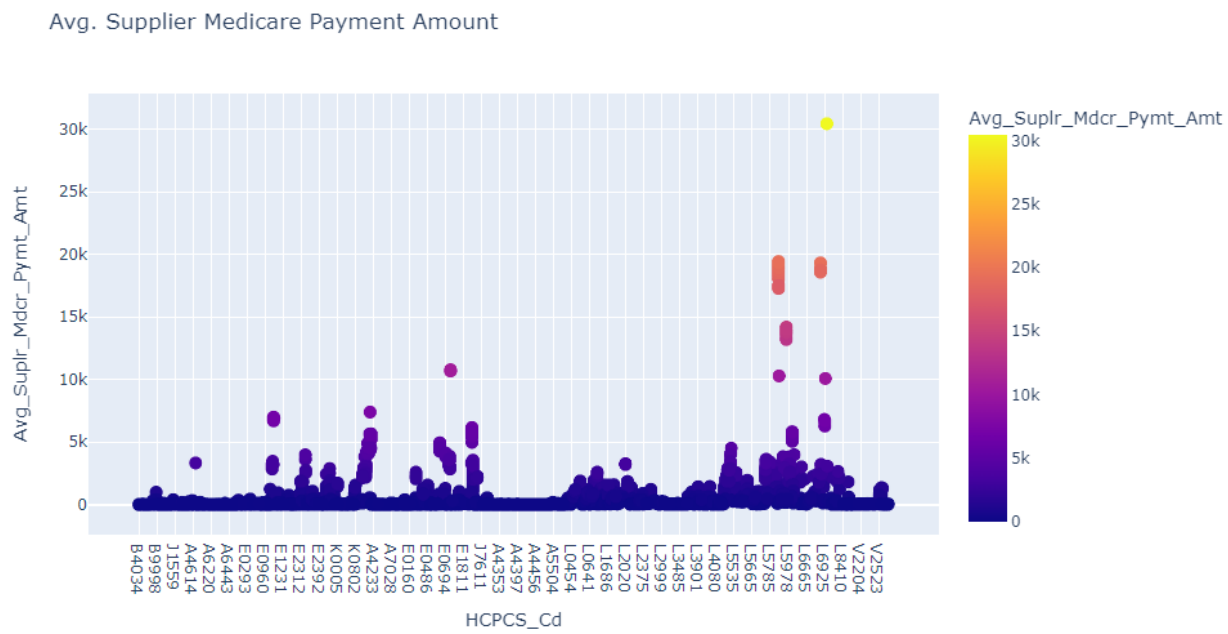

Only under 200M units of the blood glucose test/reagent strips were manufactured, the most manufactured item is the revefenancin inhalation solution.



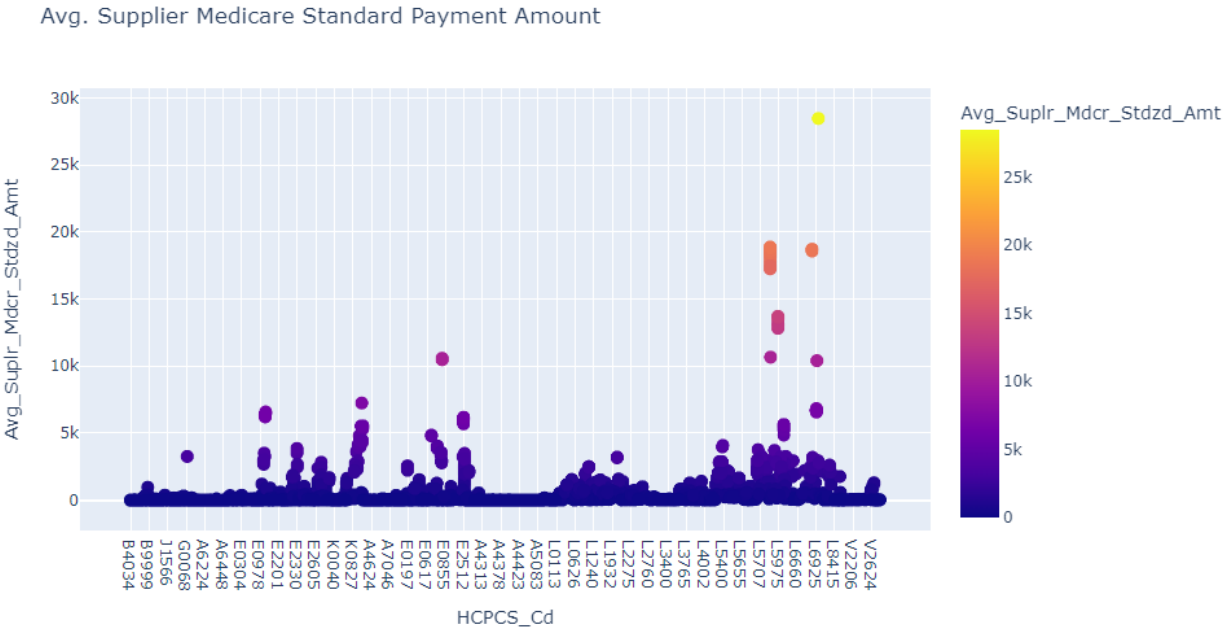

The blood glucose test/reagent strips were being charged on the lower end compared to the rest of the medical devices. The most expensive medical device is the electronic elbow. We also found that the average purchasing price for blood glucose test/reagent strip is $61.44,



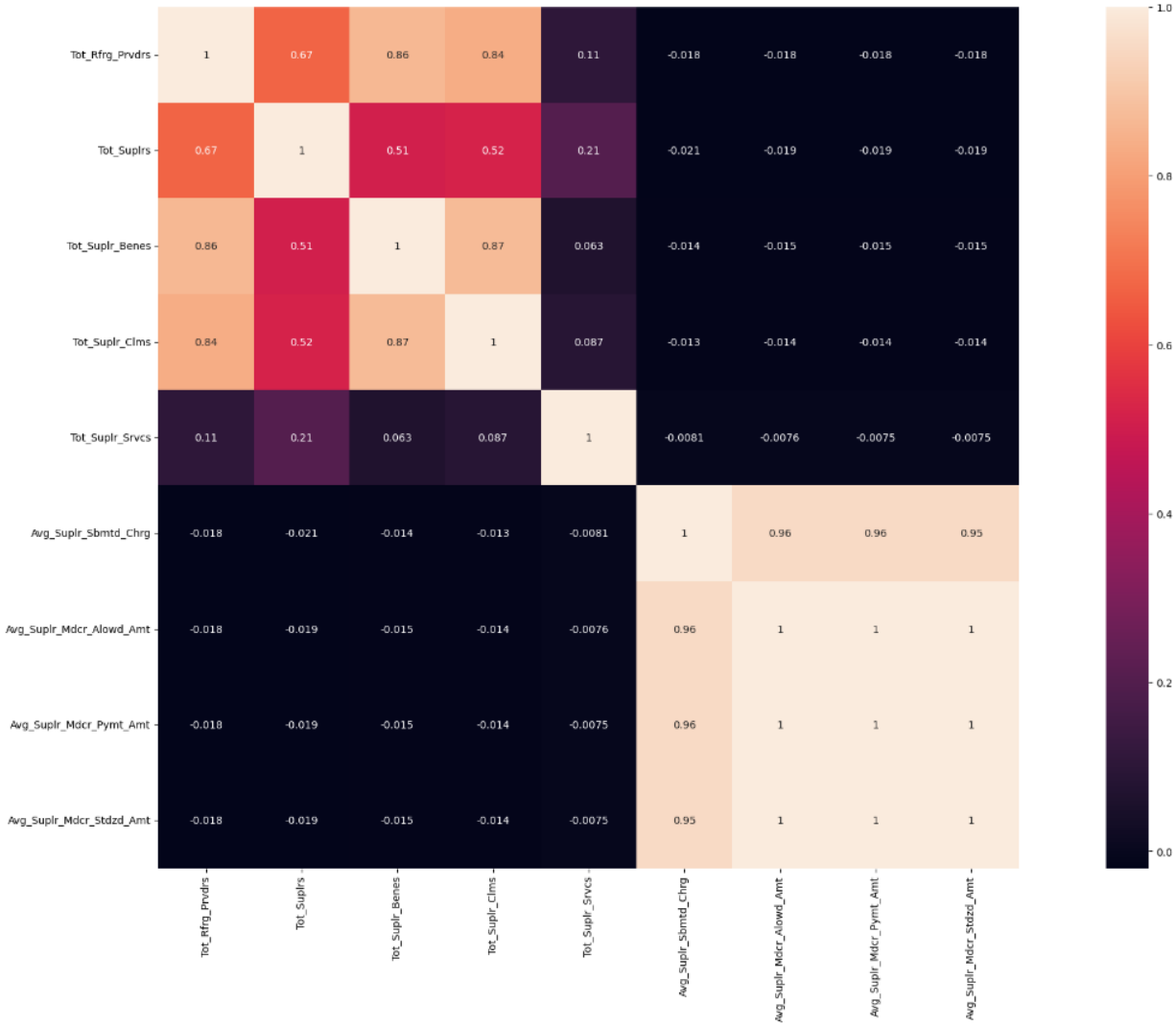

The blood glucose test/reagent strips were being covered by Medicare on the lower end compared to the rest of the medical devices, this makes sense as the blood glucose test/reagent strips do not cost as much as other items. The most covered by Medicare is the electronic elbow which is also the most expensive item.

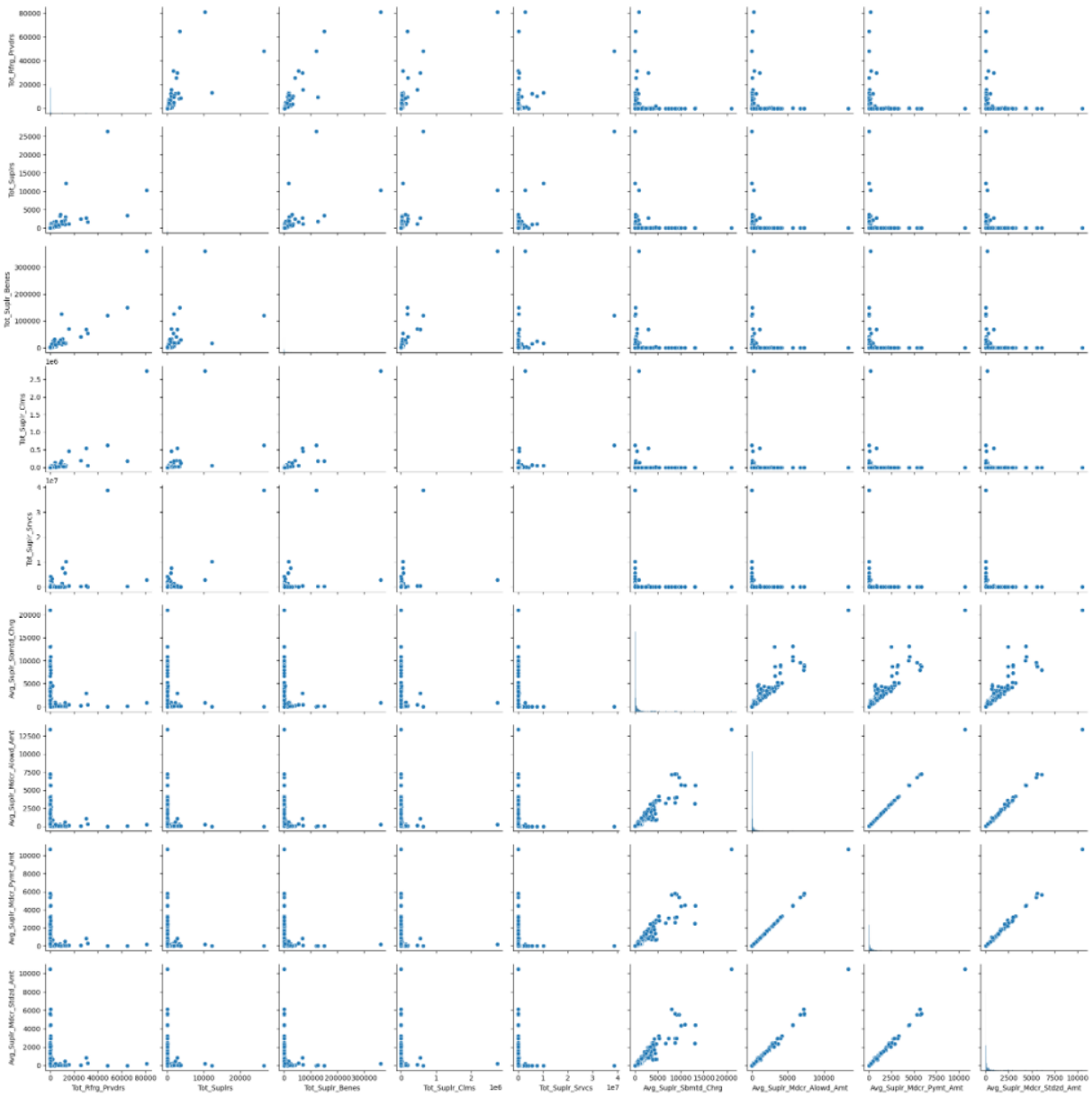



The results for the average supplier Medicare payment amount are the same as the previous distribution.




The results for the average supplier Medicare payment amount are the same as the previous distribution.


Tot_Rfrg_Prvdrs
Tot_Suplrs
Tot_Suplr_Benes
Tot_Suplr_Clms
Tot_Suplr_Srvcs
Avg_Suplr_Sbmtd_Chrg
Avg_Suplr_Mdcr_Alowd_Amt
Avg_Suplr_Mdcr_Pymt_Amt
Avg_Suplr_Mdcr_Stdzd_Amt
1 0.67 0.86 0.84 0.11 -0.018 -0.018 -0.018 -0.018
0.67 1 0.51 0.52 0.21 -0.021 -0.019 -0.019 -0.019
0.86 0.51 1 0.87 0.063 -0.014 -0.015 -0.015 -0.015
0.84 0.52 0.87 1 0.087 -0.013 -0.014 -0.014 -0.014
0.11 0.21 0.063 0.087 1 -0.0081 -0.0076 -0.0075 -0.0075
-0.018 -0.021 -0.014 -0.013 -0.0081 1 0.96 0.96 0.95
-0.018 -0.019 -0.015 -0.014 -0.0076 0.96 1 1 1
-0.018 -0.019 -0.015 -0.014 -0.0075 0.96 1 1 1
-0.018 -0.019 -0.015 -0.014 -0.0075 0.95 1 1 1
1.0
0.8
0.6
0.4
0.2
0.0



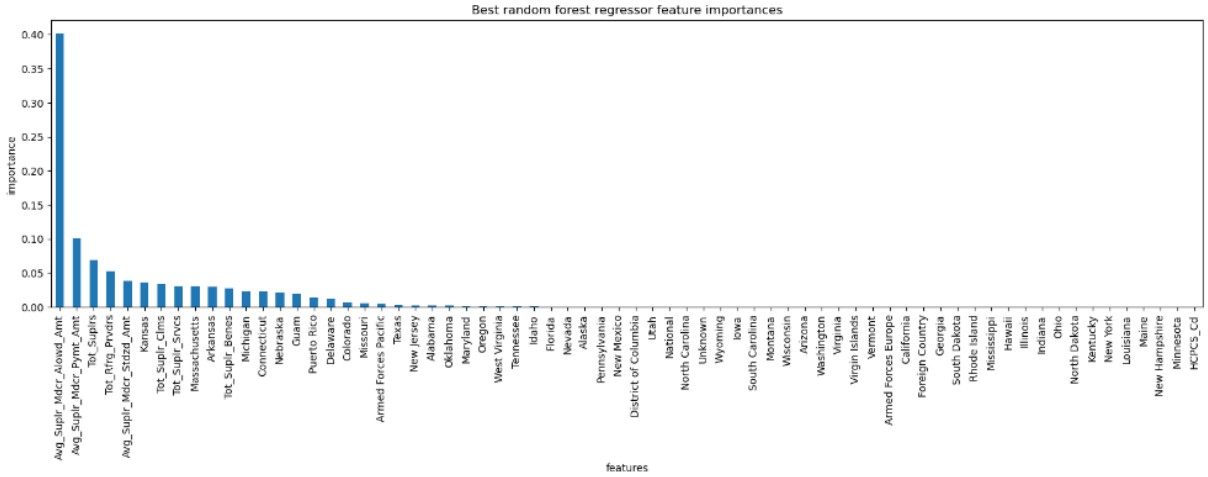

There were a few strong correlations between the Medicare payment columns; however, we did not see any other strong correlations between the other features. The features will need to be further analyzed in the next section.

**Pre-processing and Training Data Development**

Best Estimators:

```
Avg_Suplr_Mdcr_Pymt_Amt     160.274626
Puerto Rico                  -1.384814
Avg_Suplr_Mdcr_Alowd_Amt    -33.559312
Avg_Suplr_Mdcr_Stdzd_Amt   -146.573691
```

The average supplier Medicare payment amount is the biggest positive feature, this makes sense as submitted claims are paid for by Medicare first and may sometimes contribute to the majority of the payment. Average supplier Medicare allowed amount and average supplier Medicare standardized amount is negatively associated with purchasing price, this is suggesting that the higher the Medicare allowed/standardized payment amount then the lower the purchasing price. We can interpret this as some products may be getting paid out by Medicare more than others.




The dominant top four features in common with the linear model are:

- Avg_Suplr_Mdcr_Allowd_Amt
- Avg_Suply_Mdcr_Pymt_Amt
- Tot_Suplrs
- Tot_Rfrg_Prvdrs

For the Linear Regression model:

- The mean cross-validated Mean Absolute Error (MAE) is approximately 3.42, with a standard deviation of approximately 1.06.
- The MAE calculated directly on the test set using the best estimator from grid search is approximately 3.48.

For the Random Forest mode:

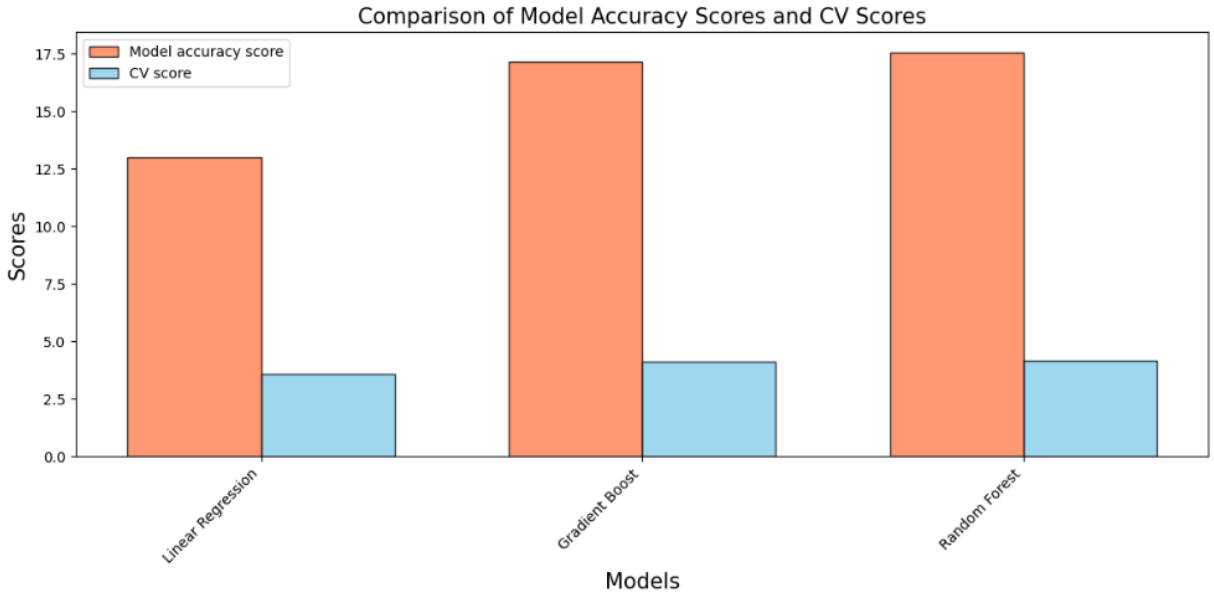
- The mean cross-validated MAE is approximately 4.08, with a standard deviation of approximately 0.51.
- The MAE calculated directly on the test set using the best estimator from grid search is approximately 3.31.

Conclusion: Comparing the mae on the test set, the random forest model appears to perform slightly better than the linear regression model.


There initial rapid improvement in model scores but levels off which indicates that adding more data does not significantly improve model performance beyond a certain point.

**Modeling**

I conducted several models – linear regression, gradient boosting regression, and random forest regression.



```
           Algorithm  Model accuracy score
0  Linear Regression             12.991052
1      Gradient Boost             17.130138
2       Random Forest             17.553441
```

| | Algorithm | CV_Score |
|---|---|---|
| 0 | Linear Regression | 3.604310 |
| 1 | Gradient Boost | 4.138857 |
| 2 | Random Forest | 4.189683 |

Linear Regression appears to perform the best among the three algorithms as it has the lowest score. However, when considering cross-validated scores, Gradient Boost and Random Forest seem to perform better than Linear Regression as they have higher cv_scores but higher cv_scores also indicate slightly worse generalization performance compared to Linear Regression.

```
Train MSE: 3.866538869605473
Validation MSE: 5.111203427959549
Test MSE: 3.7461880052871277
Train R2 Score: 0.8463407132970207
Validation R2 Score: 0.8899563587190521
Test R2 Score: 0.7593648624604183
```

The model appears to perform well on the training set as indicated by the relatively low training MSE and high training R2 score. This suggests that the model explains approximately 81% of the variance in the training data. The similar R2 scores for the validation and test sets indicate that the model's performance generalizes well to unseen data.

**Conclusion**

- Linear Regression model is the best model for predicting whether the blood glucose test/reagent strip will generate profit for the company, it performs slightly better than Gradient Boosting and Random Forest models.
- A lot of features are not important and have no impact on the target variable.

**Future Work**

- Further analysis, such as examining additional evaluation metrics, tuning hyperparameters, or exploring different algorithms to help make a better decision on the best-performing algorithm.
- Integrating data from the company for more accurate predictions.