CSC-272
Yutian Nan, Qichen Liu, Champ Clay
Final Project
Dr. Kevin Treu

# Drawing Up a Drug Selling Plan for Cocaine Distribution

## Introduction

After attending Dr.Treu's final exam in Furman, MLGB and Snow White decided to go to the underground bar in Greenville for a drink. After several rounds, MLGB and Snow White met Choncho and had a pleasurable conversation. Unintentionally, these three guys found out that they have the same ultimate goal in life - earn money, and earn a fortune. These three guys decided to form a drug distribution known as Good Girl together, and sell cocaine in the Dream Market. In order to get the huge amount of money in the shortest time, they decided to come up with a selling schedule to earn the highest profit.

The Dream Market is a dark web for online drug distribution. The "Dream Market Cocaine Listings" contains about 4600 cocaine products information in the Dream Market, which was one of the predominant online darknet markets founded in 2013. Although the Dream Market shut down on Apr.30$^{th}$, 2019, we would still like to predict the trend of selling price of the cocaine product for the purpose of this project. The data comes from previous purchase histories and we would like to build a regression model on this dataset in order to develop a better understanding of selling products within the market.

After several experimentations, we have found that grams in our dataset is the most predictive attribute. We also found out that as we increase grams and purity of the cocaine result in high profit. We will expand more and explain in detail about how we obtain this conclusion in

the data analysis and preliminary result section, and we will also come up with a selling plan later in this report.

## Dataset Description

The dataset used in this article is downloaded from Kaggle (site url: https://www.kaggle.com/everling/cocaine-listings). The original data was obtained by David Skipper Everling, a data scientist, in July 2017 by scraping the Dream Market web. It includes the information of the product title, shipping information, quality of the product, weight, price, rating of the product, escrow, vendor name, etc. The shipping information provides the region where the drug was shipped and the place where the drug was received. Drug quality records the percentage of the cocaine's purity.

| Attribute | Descriptions | Data Type |
|---|---|---|
| grams (g) | Grams of cocaine sold | Numeric |
| quality (%) | Purity of the cocaine sold | Numeric |
| escrow | Conditional money transfer | Binary |
| successful_transaction | Times of cocaine within a specific store in the market successfully sold | Numeric |
| Ship_to (24 regions in total) | The region where the cocaine is shipped to | Binary |
| Ship_from (24 regions in total) | The region where the cocaine is shipped from | Binary |
| rating | How customers rate our store on a scale of 0 to 5 | Numeric |
| usd_price ($) (class attribute) | The price of the cocaine sold | Numeric |

## Data Preparation

In order to get the most predictive regression equation, we deleted some non-predictive attributes that are used in the raw dataset such as site link and vendor link. In the meanwhile, we deleted the redundant attribute, ship_from_to, to eliminate the possibility of adding to much weight on that attribute. We also deleted some variables that are dependant with other attributes,

such as cost_per_gram, and cost_per_gram_pure, to eliminate the possibility of disturbing the regression model. And for ship_from and ship_to attributes, we changed their values to 0 or 1 in association with their dummy representations of TRUE (1) or FALSE (0). When using visualization in Weka, we detected an outlier, which is in the case of grams equal 5000, and deleted the outlier instance to ensure the accuracy of the result. We included the new attribute usd_price in the dataset as our class attribute instead of the original class attribute btc_price since the value for btc_price is too small comparing with other values for other attributes in the dataset.

## Data analysis and Preliminary Result

Since the project is a supervised learning called numeric estimation, where we would build a linear regression model. We also incorporated the tree model, SVR model, and a neural network model. Percentage split is a test option that would split the original dataset into two part that 66% of the data is used as the training data, and 34% of the data is used as test data. We will use 66% split to test all the models built for this project.

### *Model I - Linear Regression Algorithm*

When using Weka to run the linear regression algorithm, there are two methods, which are M5 and no attribute selection. We will analyze the results generated by no attribute selection method in this section, and the regression model generated by M5 method will be analyzed under the improvement section later in this report. Linear regression belongs to numeric estimation, which is a supervised learning, that it has a class attribute, and the class attribute has to be numeric. For this algorithm, a linear regression model will be generated with class attribute usd_price as the dependent variable, and other attributes in the dataset as independent variables.

For no attribute selection setting in linear regression algorithm in Weka, all the attributes will contribute to building in the regression model. After running this algorithm, the regression model built looks as follows.

*Equation*:

$$usd\ price = 41.9831 * grams + 2.38 * quality + (-78.3255) * escrow + 0.001 * successfulTransactions + (-80.409) * rating + -(845.6827) * shipstoUS + 451.6418 * shipsfromUS + (-749.5152) * shipstoNL + 208.5243 * shipsfromNL + 517.4167 * shipstoFR + 0.8212 * shipsfromFR + (-230.4865) * shipstoGB + 30.8554 * shipsfromGB + (-270.7281) * shipstoCA + (-10.0942) * shipsfromCA + (-72.175) * shipstoDE + 278.295 * shipsfromDE + 3952.6749 * shipstoAU + (-2923.4486) * shipsfromAU + 9.2043 * shipstoEU + (-674.2615) * shipsfromEU + (-225.3559) * shipsfromES + (-481.5656) * shipstoN.America + -181.768 * shipsfromBE + (-84.2749) * shipstoWW + (-109.1408) * shipsfromWW + (-367.5327) * shipstoSI + 442.0835 * shipsfromIT + (-160.2648) * shipstoCH + (-160.1554) * shipsfromCH + (-364.2121) * shipsfromBR + 71.0949 * shipsfromCZ + (-350.8405) * shipsfromSE + (-33927.8233) * shipsfromCN + 679.5981$$

```
=== Summary ===

Correlation coefficient                  0.9837
Mean absolute error                    739.8467
Root mean squared error               2038.3253
Relative absolute error                 19.0372 %
Root relative squared error             22.0584 %
Total Number of Instances              511
```

From the regression model above, we can see that the correlation coefficient is 98.37%, which means there's a strong positive linear relationship between the class attribute and attributes that are considered as independent variables.

To interpret the model, we can see from the positive coefficients of grams that as grams increase by 1, usd_price of the product increase by $41.9831. And as the percentage of the purity of cocaine increase by 1%, usd_price of the products increases by $2.38. These two coefficients in front grams and quality make intuitive sense since the price of a product will increase as the quantity and quality of that product increase in real life. We can also see from the coefficients in

front of our dummy variables in this regression model that it is most profitable to ship the product to AU as we will gain $3952.6749 in profit. We would also like to avoid shipping from CN as we will have a $33927.8233 loss in profit. This gain and loss in profit have a hidden knowledge that we can infer, which we can interpret it as regions like CN have a strict rules and examinations on packages leaving the country to avoid cocaine dealings. If we want to have our product shipped from CN, we need to pay more in the process of hiding and transferring the drug. And it is vice versa for AU that we would like to sell more of our products to AU customers in order to gain a higher profit.

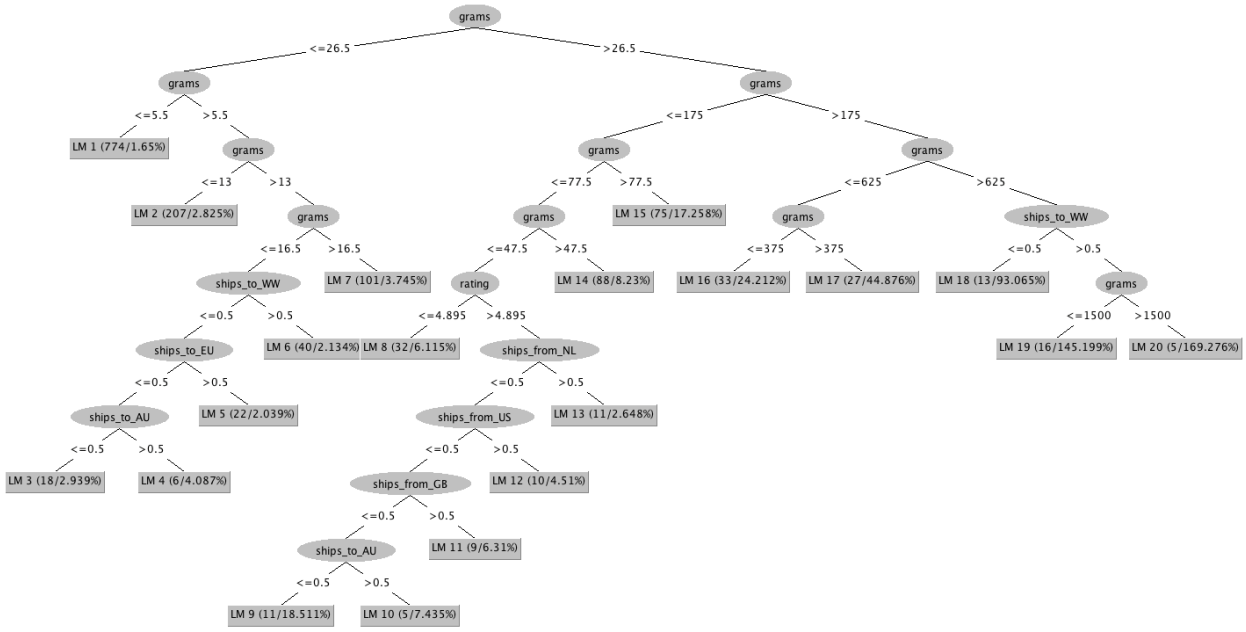
There's also an improvement that we can make by examining this linear regression model above since rating has a negative coefficient of -80.409 in front of it. It means as rating increase by 1, the price of cocaine will decrease by $80.409. This seems counterintuitive at the first glance since the rating of a product should be positively correlated with its price. This particular attribute rating is an improvement that we can make in future studies to make the overall result better. We will explain in detail about the reason behind this negative coefficient in the improvement section later in this report.

***Model II - Regression Tree and Model Tree***

We also used the M5P tree algorithm in Weka to build a regression tree and a model tree. Building trees is a "divide and conquer" process that each node involves testing a specific attribute. Numeric values, like our dataset, may appear more than one time in a tree. The difference between these two trees is that a regression tree has numeric quantities as results on its leaves, but model tree gives us regression models, where we need to manually plug into values to get the final result, on its leaves. After running the algorithm, the trees generated look as follows.

```
=== Summary ===

Correlation coefficient                  0.9103
Mean absolute error                    972.5282
Root mean squared error               4554.6856
Relative absolute error                 25.0244 %
Root relative squared error             49.29   %
Total Number of Instances              511
```

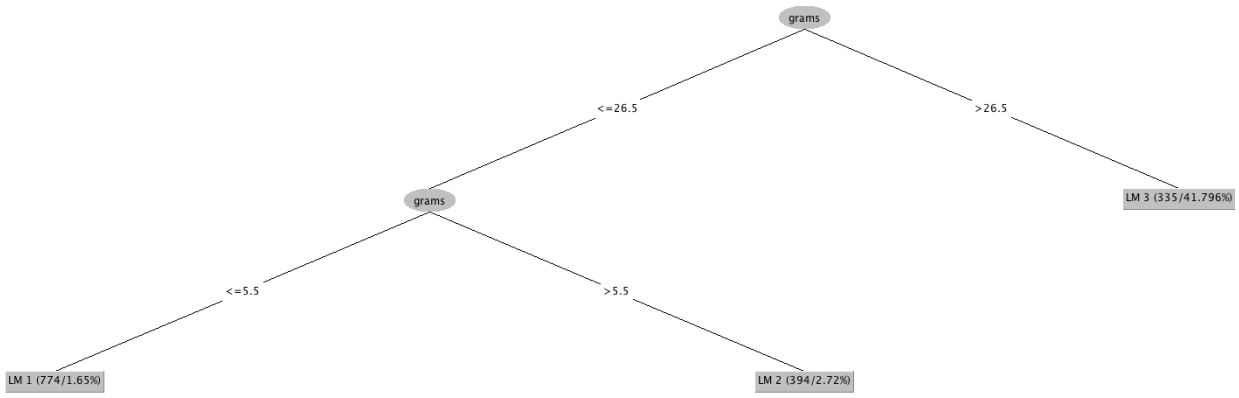



From this regression tree, we can see the correlation coefficient decreases from 98.37% to 91.03% comparing to the result of the linear regression algorithm, but it still shows a relatively high positive linear relationship between the class attribute and other attributes. We can see from the root node of the tree that this node is the attribute grams, which indicate that grams is the most predictive attribute in predicting the class attribute usd_price. We have held out some instances from the dataset to manually plug into the tree to test the validity, and the results confirms that the tree is valid.

We also used M5P algorithm with 66% percentage split to generate a model tree, and the result for model tree looks as follows.

```
=== Summary ===

Correlation coefficient                  0.9865
Mean absolute error                    548.102
Root mean squared error               1751.785
Relative absolute error                 14.1034 %
Root relative squared error             18.9575 %
Total Number of Instances              511
```




From this model tree, we can see that the correlation coefficient increases to 98.65% comparing to the correlation coefficient of the regression tree. This confirms the strong positive linear relationship between the class attribute and the attribute grams that appear in the model tree. Again, the result confirms that grams is the most important determining factor for predicting price in the model tree. For the model tree, we also used the hold-out portion of the dataset to test the validity of the tree, and it seems that the results confirm with the validity.

***Model III - Support Vector Regression Algorithm***

In this case, we used Support Vector Regression, which is developed for support vector machine used in regression. Support vector regression helps us to find the line of best fit while minimizing the error from the cost function, where only instances in the training data that are closest to the line are considered in building the model. We used the default setting, polynomial kernel, which has a default setting of 1 and is almost linear, but as the polynomial degree

increases, the line becomes more wiggly. We experimented with setting the exponent to 3, but it takes too long to build the model. We think the correlation coefficient might slightly increase with exponent increased to 3. However, the current result is good enough to predict price in the market, therefore, we will use the result acquired from the default setting to do the analysis.

```
=== Summary ===

Correlation coefficient                  0.9832
Mean absolute error                    641.6153
Root mean squared error               2389.1657
Relative absolute error                 16.5096 %
Root relative squared error             25.8551 %
Total Number of Instances              511
```

```
SMOreg

weights (not support vectors):
 +       1.2813 * (normalized) grams
 +       0.0021 * (normalized) quality
 -       0.0002 * (normalized) escrow
 +       0.0004 * (normalized) successful_transactions
 -       0.0002 * (normalized) rating
 -       0.0003 * (normalized) ships_to_US
 +       0.0229 * (normalized) ships_from_US
 -       0.0015 * (normalized) ships_to_NL
 +       0.0225 * (normalized) ships_from_NL
 -       0.0004 * (normalized) ships_to_FR
 +       0.0232 * (normalized) ships_from_FR
 -       0.0001 * (normalized) ships_to_GB
 +       0.0224 * (normalized) ships_from_GB
 -       0.0019 * (normalized) ships_to_CA
 +       0.0235 * (normalized) ships_from_CA
 -       0.0005 * (normalized) ships_to_DE
 +       0.023  * (normalized) ships_from_DE
 +       0.0211 * (normalized) ships_to_AU
 +       0.0067 * (normalized) ships_from_AU
 +       0.0001 * (normalized) ships_to_EU
 +       0.0216 * (normalized) ships_from_EU
 +       0      * (normalized) ships_to_ES
 +       0.0222 * (normalized) ships_from_ES
 -       0.0017 * (normalized) ships_to_N. America
 +       0      * (normalized) ships_from_N. America
 +       0      * (normalized) ships_to_BE
 +       0.0219 * (normalized) ships_from_BE
 -       0.0002 * (normalized) ships_to_WW
 +       0.0221 * (normalized) ships_from_WW
 -       0.0003 * (normalized) ships_to_SI
 +       0      * (normalized) ships_from_SI
 +       0      * (normalized) ships_to_IT
 +       0.0236 * (normalized) ships_from_IT
 +       0      * (normalized) ships_to_DK
 +       0      * (normalized) ships_from_DK
 +       0      * (normalized) ships_to_S. America
 +       0      * (normalized) ships_from_S. America
 +       0.0102 * (normalized) ships_to_CH
 +       0.0102 * (normalized) ships_from_CH
 +       0      * (normalized) ships_to_BR
 +       0.0201 * (normalized) ships_from_BR
 +       0      * (normalized) ships_to_CZ
 +       0.0227 * (normalized) ships_from_CZ
 +       0      * (normalized) ships_to_SE
 +       0.0201 * (normalized) ships_from_SE
 +       0      * (normalized) ships_to_CO
 +       0      * (normalized) ships_from_CO
 +       0      * (normalized) ships_to_CN
 -       0.3284 * (normalized) ships_from_CN
 +       0      * (normalized) ships_to_PL
 +       0      * (normalized) ships_from_PL
 +       0      * (normalized) ships_to_GR
 +       0      * (normalized) ships_from_GR
 -       0.022
```

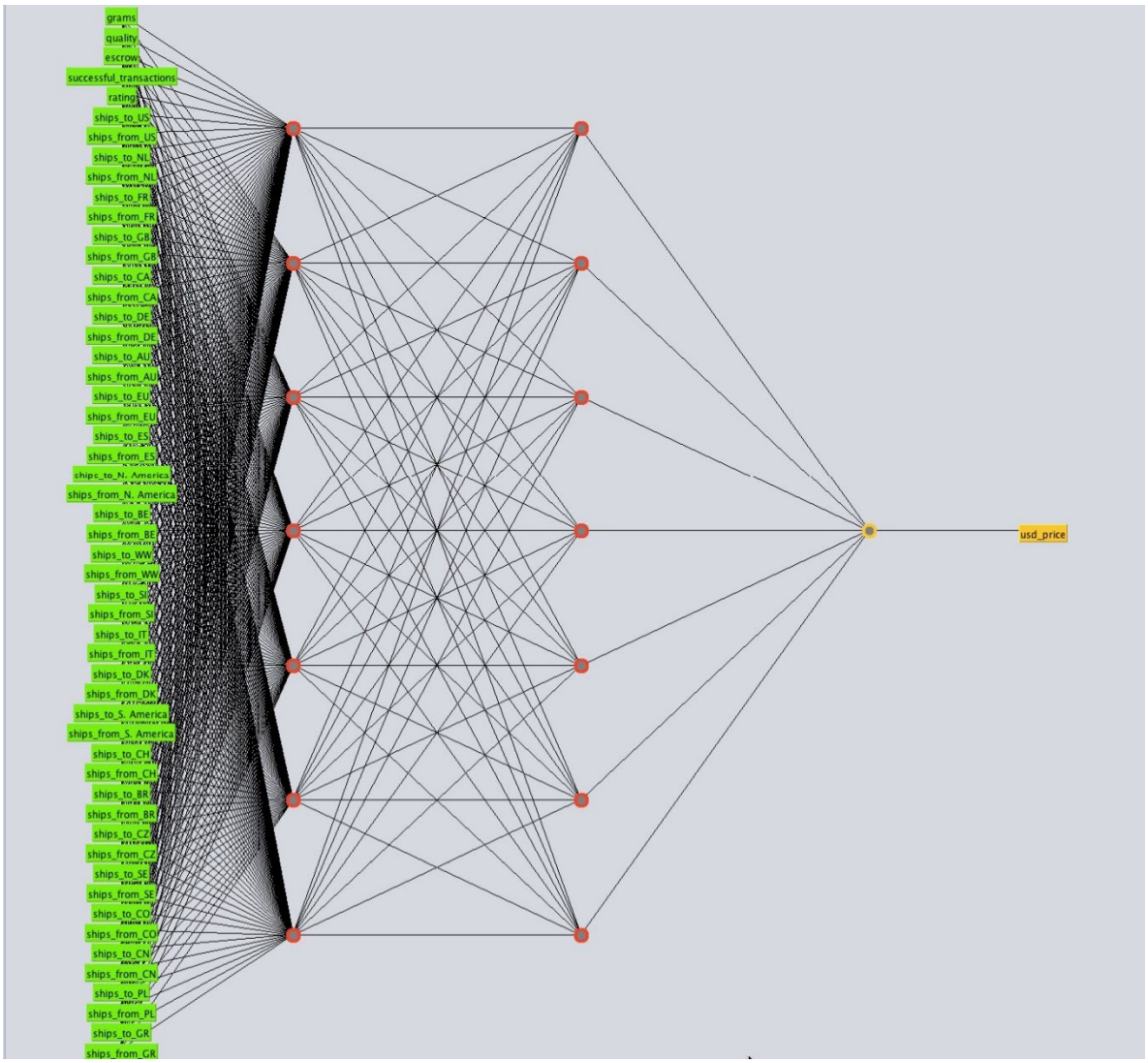
From the graphs above, we found that grams is still the most predominant attribute in predicting price. Comparing this model with Model I, the difference is that using SVR, it is most

profitable to ship products from IT as we will gain the highest profit, but in the linear regression algorithm, it indicates that it is most profitable to ship products from U.S.

***Model IV - Neural Network***

In this model, we used the multilayer perceptron algorithm to build a neural network on our dataset. Multilayer perceptrons are network of perceptrons, and the purpose for the neural network is to reconstruct the input through a condensed dimensional representation. A node in the network combines input from the data with a set of coefficients or weights. Each node is organized in layer, and a node is a place where computation happens, like loosely patterned on a neuron in the human brain that fires when it encounters efficient stimuli.

Since the neural networks are a complex algorithm to use for predictive modeling because there are so many configuration parameters that can only be tuned effectively through intuition and a lot of trial and error, the result we have currently does not contain a satisfied correlation coefficient, as it is 16.03%. We may find the more appropriate hidden layers setting to use in future studies. From the network, it has an input layer, hidden layer, and an output layer. The green nodes showing on the picture are input nodes, the yellow node is the output node, and red nodes represent hidden layers. The yellow node performs a weighted sum, and each of the connections have a weight. There are two hidden layers with 7 numbers of neurons in each of the hidden layers. Each node performs a weighted sum of its inputs, and thresholds the result.



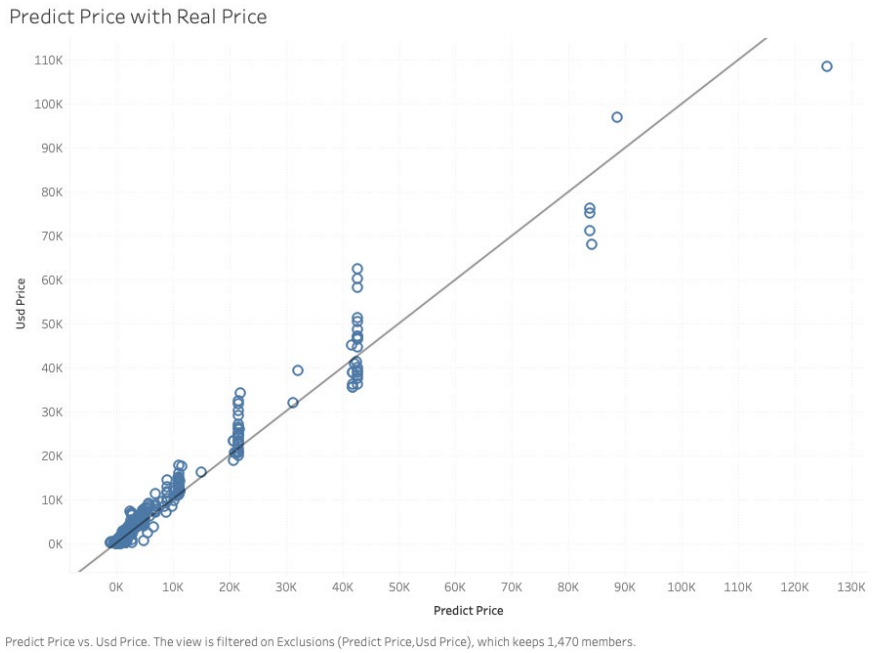

```
=== Summary ===

Correlation coefficient                  0.1603
Mean absolute error                  45657.6943
Root mean squared error              46168.2672
Relative absolute error               1174.8316 %
Root relative squared error            499.6245 %
Total Number of Instances              511
```

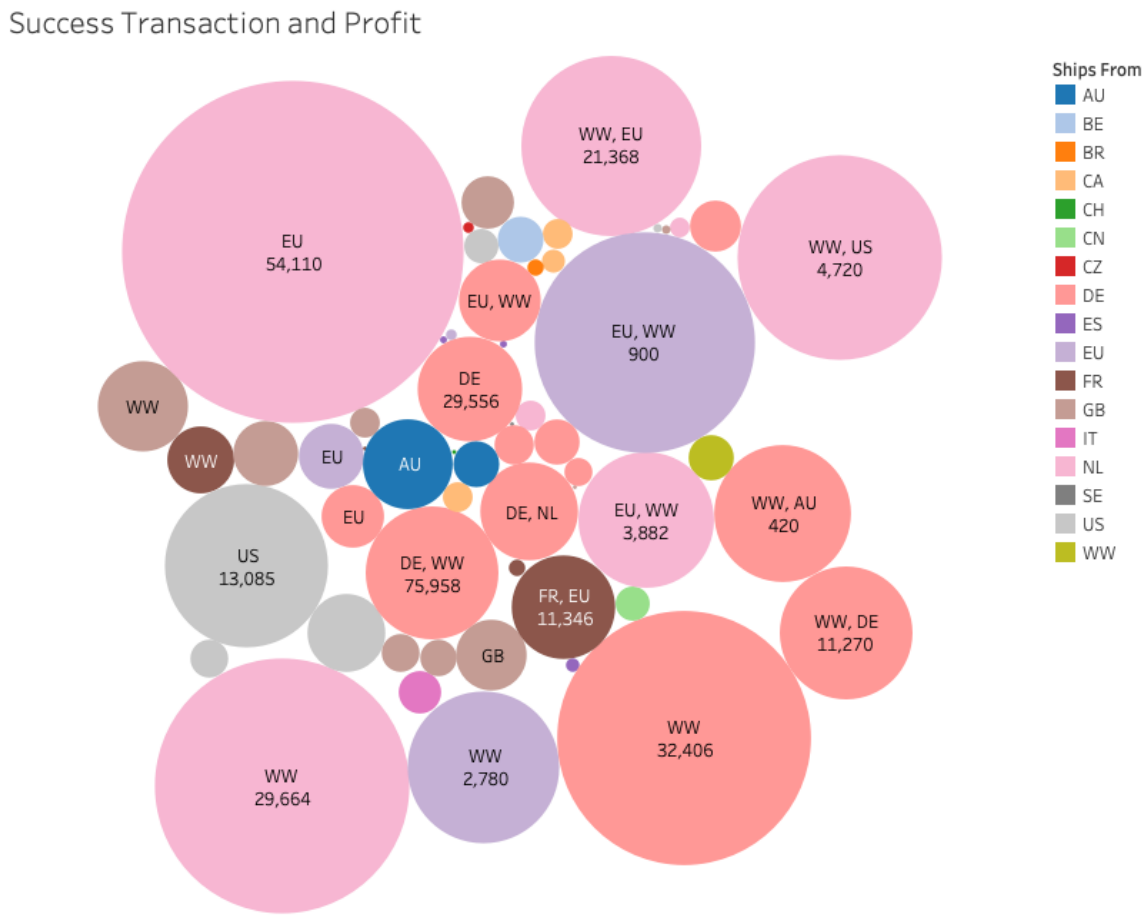
## Data Visualization

The graph below shows the graph of predict price and usd_price. The predic results are presented on the X-axis and the actual prices are presented on the Y-axis. From the trend line(45 degree diagonal line) shown in graph,we could say the predict price of the cocaine product and

by the simple linear regression is close to the actual price of the cocaine. In the meanwhile, it can be clearly seen that when the transaction price is low, the prediction is more close to the actual price. With the increase of the transaction price, the forecast prices deviate a little. The possible reason is that when the transaction price is high, the seller may consider offering a discount to the customer, so that the forecast price is higher than the actual price.




The graph below shows the actual transaction situation. The size of the bubble represents the level of the transaction price, where the higher the total transaction price, the bigger the bubble. The color of the bubble indicates the country where cocaine ship from. On the right side of the image is a detailed color legend. The text on the bubble shows the region where the cocaine was transported to and the volume of the total transactions in that regigon. It can be observed from the whole packed bubble chart that the cocaine trading from Germany and Netherlands accounts for the majority of the trading volume, while a realtively large amount of the cocain product is shipped to the European Union.




## Problem & Improvement

- ***Negative Coefficient of Rating in Linear Regression Model***

As we have mentioned in the data analysis and preliminary result section that the rating attribute has a negative coefficient in the regression model, which is counterintuitive in the common sense. The possible reason behind this result might be that the variation for rating is too little, where the majority of instance are between 4.7 to 5. In future studies, it might be better to use another way for rating the cocaine product to replace the current rating values to achieve making the rating attribute becomes more predictive.

- ***Attribute Quality Absent from Model Built by M5 Setting***

The M5 setting is another option for running linear regression algorithm in Weka, which works by automatically detect and remove correlated attributes in the dataset. When using the M5 setting to build the linear regression model, the attribute quality, which should be a determinant factor in determining price, is not included in the model.

It might be the case that we have too many binary variables in the dataset, which affects the regression model result immensely. Comparing the effect of quality with shipping place attribute, quality's effect may seems less conspicuous. However, from later experiments we explored, we have seen that quality does have a predictive role that it has a positive relationship with price. Therefore, the quality attribute is reasonable in this project.

```
usd_price =

     42.0008 * grams +
   -718.8944 * ships_to_US +
    452.335  * ships_from_US +
   -764.651  * ships_to_NL +
    301.4904 * ships_from_NL +
    699.2686 * ships_to_FR +
    335.1202 * ships_from_DE +
   3931.2304 * ships_to_AU +
  -2705.4832 * ships_from_AU +
   -583.9658 * ships_from_EU +
 -33878.8189 * ships_from_CN +
    283.8902
```

- ***Future Study Direction***

  This project is not relatively complicated since we consider ship from and ship to attributes separately. However, these two attributes need to be considered togehtehr in real life. For example, if country A is the chepest place for supply, and selling the supply we obtain from country A to country B allows us to gain the highest profit. If we examine this carefully, there might be an associated high cost in transportation that will decrease our profit dramatically. As a result, this chain of drug distribution may not be the optimal choice for a merchant to gain profit. Therefore, we need to consider the cost of supply independently or consider ship from and ship to counteires as a pair in future studies. From there, more experimentations need to be carried out in order to acquire the most appropriate product distribution schedule.

## Results

From the linear regression algorithm, Tree, and SVR algorithm that we used, we come up with the conclusion that attribute grams is the most predominant variable in predicting price. Our prediction shows that as grams increases, the price of cocaine also increases. For testing, we used 66% split to test all the model mentioned in this project, and the results show a high correlation coefficient except for neural network, which indicate strong positive linear relationship. For the purpose of this project, we are satisfied with the results we have so far. Although in M5 regression setting that quality is not as predominant as grams, it still plays an important role in predicting the price of cocaine. Besides neural network that may need more trials, the rest models all make intuitive sense in predicting the class attribute and they generalize well. The correlation coefficients of linear regression model, tree, and SVR are approximate with each other, and this offers us confidence about our results.

Through our experiments, the manager of Good Girl would like to avoid have cocaine products shipped from China since the price decreases sharply comparing with price when products are shipped from other regions. A decrease in the selling price leads to a decrease in profit, and Good Girl would like to prevent this from happening. From the results we obtain so far, it is most profitable for Good Girl to acquire their supply from US. In the meantime, the manager decides to mainly ship as large amount of cocaine products with high purity to Australia in order to gain the highest profit. From the overall transaction price, we can conclude that Dream Market's cocaine products have relatively good earnings.

## Conclusion

The initial project objective is to find out the determining factor that most affect the price of cocaine through analysis. Throughout the project, we come up with the conclusion that grams is the most important factor in determining the price of cocaine. When using 66% split to test on the model, the results shows decent validity with high correlation coefficient. Therefore, we can conclude that the models are credible. At the same time, we have provided the cocaine distribution with a good selling plan that the distribution should mainly have its products shipped from US, and mainly sell the cocaine products to Australia to gain the highest profit. In the meantime, Good Girl should avoid have its products shipped from China.We hope that our study will help Good Girl to achieve their optimal goal.