**Building a Bankruptcy Model to Classify Global Public Companies Using Various Financial Ratios**

**Authors: 1. Rohit Kumar Gupta, 2. Sai Siri Sunkara, 3. Shradda Pai Kasargod**

**Executive Summary**

Bankruptcy is a legal state when an entity is unable to repay its debt. With access to global audiences, there is an increase in research to find a global pattern to predict bankruptcy for firms. This paper describes the impact of different financial variables and ratios on the predictability of bankruptcy. The financial health of a company is highly affected by three major factors and these factors are Profitability, Debt, and Current assets. Though various economic and legal factors play a major role in deciding the bankruptcy trend of a firm, our paper supports the fact that financial factors are the most crucial ones which help in predicting bankruptcy with precision. For this research, we have analyzed data of 468 companies from 20 countries of 3 major continents- Asia, Europe, and America. The data was collected from S&P’s Compustat dataset which provided an overview of various financial ratios that affected bankruptcy on a higher scale. We developed two models, logistic regression and decision tree to test our analysis of predicting bankruptcy. Logistic regression was chosen as the dependent variable was dichotomous. While doing this analysis, we also tested the relationship between various to understand the magnitude of the correlation between them. We also tested the fitness of our models and compared their accuracy. We took a step further to suggest improvement for the models by tuning the hyperparameters of the logistics regression and decision tree algorithm. Thus, our research and analysis concluded that the 10 financial ratios are the decisive factors of predicting bankruptcy.

# **1. Introduction**

Why companies go bankrupt? How effective are the financial variables in predicting the bankruptcy? These are some of the questions that our team is trying to answer in this paper. A company can go bankrupt due to various factors, for example, external factors such as a rise in competition, economic policies, internal factors such as weak management, monetary problems, tax problems etc.[[1]](#footnote-1) The scope of this project is limited to evaluating the effect of financial factors such as Earnings, Assets, Liabilities, Working Capital, Sales, Debt, Net worth on bankruptcy.

The dynamic nature of the market and the great financial crisis in 2008 has resulted in the increase in research activities in developing predictive models for bankruptcy. This approach to the concept of bankruptcy from the perspective of financial variables may produce very different results to the ones drawn from other reasons.

Though many researchers have been done which focuses on country-specific or region-specific prediction of bankruptcy. With the rise of globalization, companies are not only competing against in their home country but also with companies in other countries. We wanted to see if we could build a model that could be used for global prediction of bankruptcy. [[2]](#footnote-2)

**2. Finding the right data**

Data plays a crucial part in building an effective model. We focused on getting the data from a reliable source so that we can be assured of its quality. We chose COMPUSTAT dataset from S&P for building our model for evaluating the effect of financial factors on bankruptcy.

The dataset had global sample of 468 companies, in which 234 companies are in legal situation of Bankruptcy and 234 non-bankrupt companies belonging to three regions: America (United States, Canada and Bermuda), Europe (Denmark, Austria, France, Ireland, Germany, Italy, Norway, Holland, Portugal, Sweden, Spain, United Kingdom and Switzerland) and Asia ( South Korea, Japan, Taiwan and Singapore). The annual data for these companies were obtained for the period 1990-2013.

**3. Variables**

In this independent research study, 10 variables pertaining to the financial condition are considered. These variables are independent and are the most effective ones considering they cover major aspects of financial measures of a company. Variables cover aspects of profitability (Earnings/total assets, Retained Earnings/ Total assets, EBIT/Total Assets, Earning), Debt, Liquidity (Current assets/ Current Liabilities, Working Capital). [[3]](#footnote-3)

|  |  |
| --- | --- |
| **Code of Variable** | **Variable** |
| V1 | Earnings/Total Assets |
| V2 | Current Assets/Current Liabilities |
| V3 | Working Capital/Total assets |
| V4 | Retained Earnings/Total Assets |
| V5 | EBIT/Total Assets |
| V6 | Sales/Total Assets |
| V7 | (Current Assets + Cash Flow)/Current Liabilities |
| V8 | Total Debt/ Total Assets |
| V9 | Current Assets/Total Assets |
| V10 | Earnings/ Net Worth |
| V11 | GICS |

Table: Variables and their description

**3.1 Variables Distribution and Outliers**

Once we had data we decided to run exploratory data analysis. We visualized distributions of companies with respect to countries. Since the data was collected by Compustat, there was no specific pattern in its distribution. Then we analyzed each variable to understand its distributional characteristics (see Appendix A).

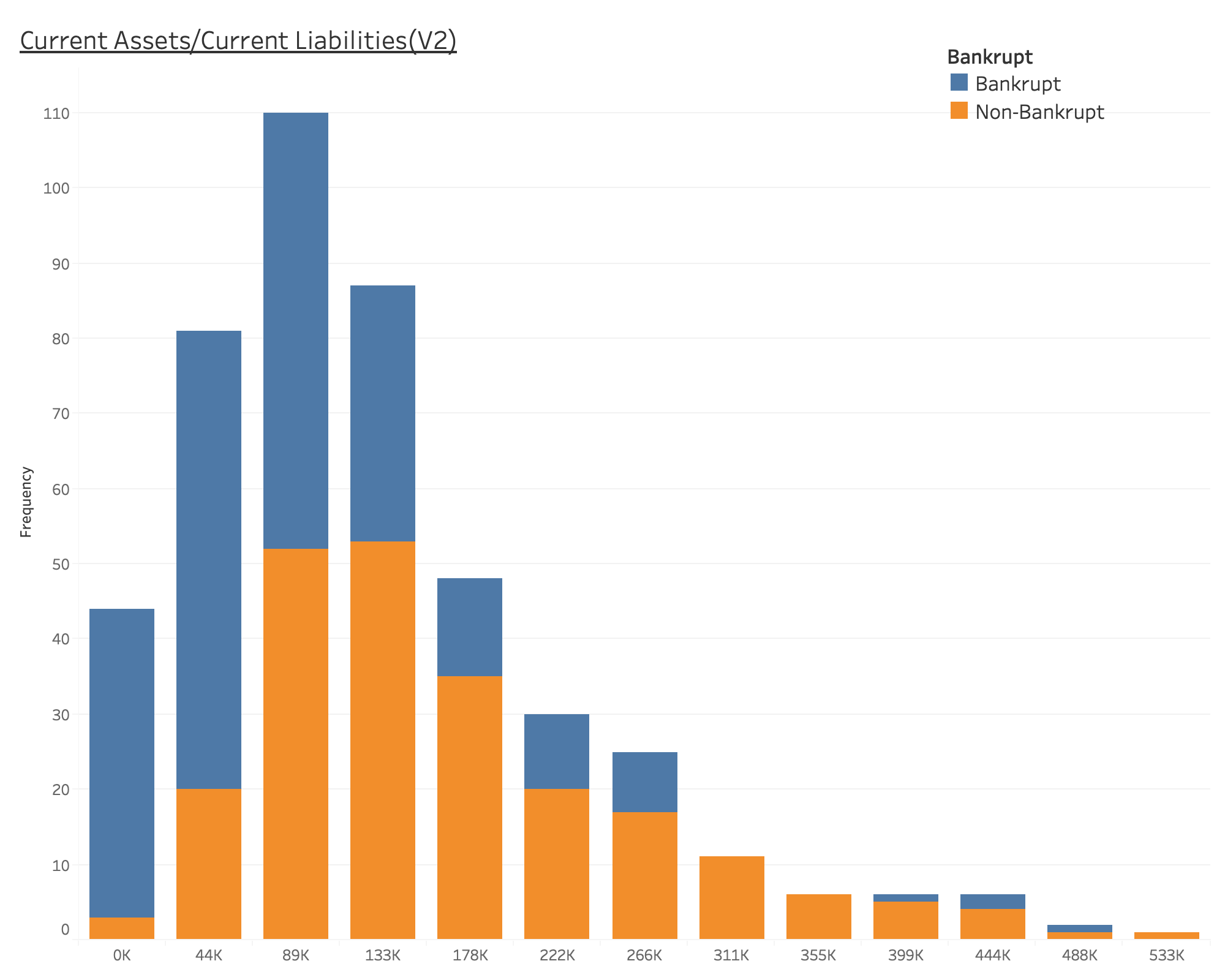
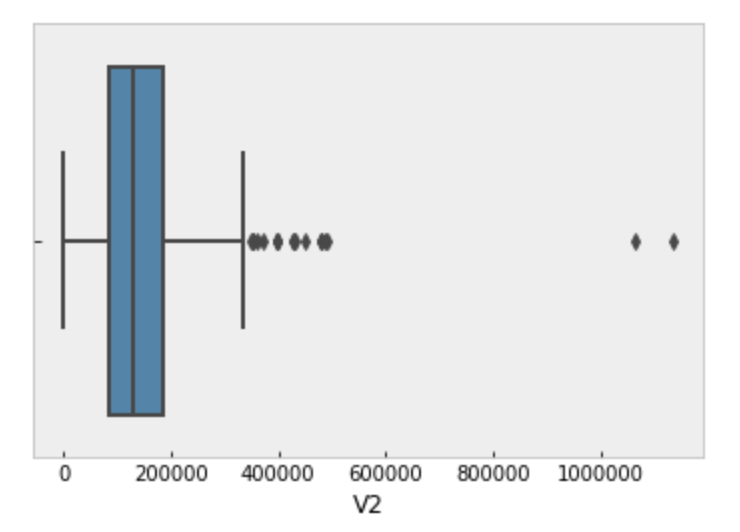


Figure: Box plot and trend chart of V2 (Current assets/current liabilities)

For example, for variable V2 which represents Current Ratio as shown in the graph. Current Ratio is equal to the ratio of current assets to current liabilities. We observed that the variable distribution is little skewed which might influence the model designed. Further, we tried analyzing the outliers found in the box plot. We observed a trend which behaved in accordance with the financial aspect. There are higher chances for a firm to go bankrupt if their current ratio was less. As this ratio increases, Firms are more financially stable and hence they would be less probable to go bankrupt. But from our data, we found that there were few bankrupt companies even when the ratios were high. We were not able to find any strong reason for these outliers, so we concluded that there might be other factors apart from financial factors such as political situations, competitions from other firms etc., which might lead to firm going bankrupt.

The key findings from our exploratory data analysis were as follows:

1. Most of the variables were prone to outliers. This is evident from the fact that our data was a global dataset and the financial variables vary from country to country.
2. The data was very representative as it has companies from different countries and the United States has more companies compared to another country
3. Only 10% of the total bankrupt companies had a significant value of V1 (Earning/Total assets) which shows that 10% of the companies went bankrupt partly because of non-financial reasons too.
4. Based on V2 (Current Assets/Current Liabilities) distribution, as the value of V2 increase, fewer companies are on the right side of the distribution which shows an obvious fact that bankrupt companies have more liabilities compared to nonbankrupt companies. A similar observation was made for variable V6 and V8.

**3.2 Missing Data Imputation**

We explored various options for missing data imputation. The first option was to drop the rows having missing values. The second option was to replace missing values with median or mean. The third option was to use machine learning algorithms such as Linear Regression, KNN, MICE to fill in the missing values. We chose simplicity over complexity and decided to use median or mean imputation for the missing values. Our decision to impute missing values was based on outliers and variable distribution. If the variable was continuous and had significant outliers, then we decided to replace the missing values with a median. If the variable was continuous and was not affected by outliers, we replaced missing values with mean. As you can see in Appendix A.1 of missing value imputation using python, the variables V1 to V8 and V10 are continuous variables with a significant number of outliers, therefore we replaced missing values in those variables with a median. Variable V9 was a continuous variable but was not having outliers, so we used mean to replace missing values. Variable V11 was a categorical variable, and we created a category called 40 for the missing values.

**3.3 Relationship between variables**

There may be a complex relationship between various variables in the dataset. Hence, it is important to discover and quantify the degree to which these are affecting the output parameters and are dependent on each other. Consider, the variables V7 ((Current Assets + Cash Flow)/Current Liabilities and V2, V3 respectively. The correlation between these variables shows positive growth and direction. This is evident as the V7 variable ratio is highly dependent on V2 and V3 ratios i.e when the operating liquid income is high the financial ratio V7 would also be high. Likewise, consider the financial ratios V4 and V5. These effects are the same for most of the variables except V8, V7, and V6 which show a negative correlation in this scenario. When inspected in detail, it was found that these ratios showed a leverage ratio imbalance.

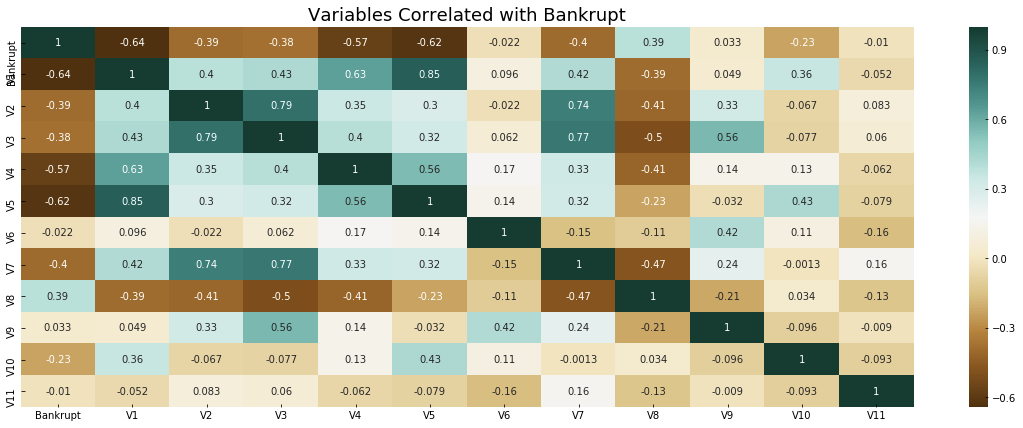


Figure: Correlation Heatmap of all the variables

**4. Selecting models and improving their accuracy**

We have split our data into two subsets: training data and testing data and fit our model on the train data, to make predictions on the test data. From Random selection, we reserved 80% of the data to build training samples and the remaining 20% to obtain testing samples. Since we had only 468 datasets, we divided it using Pareto Rule, training data has 375 datasets and 93 in the test file.

**4.1 Models**

In this section, we are going to discuss the models we used to classify the companies based on their financial variables.

**4.1.1 Model Selection**

We used Logistics regression and decision tree model for classifying the bankruptcy of the firms based on their financial variables. We selected these two models because both are very different from each other. The difference arises due to the decision boundary. In the case of the decision tree, the decision boundary is decided by dividing the space into smaller regions while logistics regression fits a single line to divide the region into two halves. Another advantage of using two models was that we can compare the result of one model with the other (i.e comparing results with a non-linear decision boundary with results with linear decision boundary).

**4.1.2 Fitness of the model**

As discussed earlier, we divided the complete dataset into training and test dataset. We trained our models (i.e Logistics regression and decision tree) using the training dataset and calculated the accuracy of the trained model on the test dataset. Before training the data, we converted our categorical variable (i.e V11) into a format that machine can recognize using one hot encoding instead of label encoding. We used one hot encoding because label encoding can be misunderstood as having order or hierarchy.[[4]](#footnote-4) Therefore, after converting all of our data into machine recognizable format, we trained our models and assessed the fitness of the model using test dataset accuracy and AUC.

**4.2 Model Improvement**

To improve the model, we decided to tune the hyperparameters of the logistics regression and decision tree algorithm. Before going into details of the hyperparameters we decided to tune, we would like you to explain the concept of hyperparameters. The hyperparameters can be thought of as knobs that help in optimizing the model. Hyper-parameters are parameters that are not directly learned within estimators. In scikit-learn, they are passed as arguments to the constructor of the estimator classes.[[5]](#footnote-5) The hyperparameters that were optimized for logistics regression were penalty and Mallow’s Cp. The hyperparameters that were optimized for decision tree were maximum features to be considered before splitting the tree, minimum samples to consider before splitting up the node and minimum samples that are required to be at leaf nodes. We used GridSearchCV algorithm with 10-fold cross-validation to find the optimum values of the hyperparameters. The values of logistics regression hyperparameters after the optimization was l1 for penalty and 1.0 for Mallow’s Cp. The values of decision tree hyperparameters after the optimization was auto for max\_features, 3 for minimum samples split and 2 for minimum samples leaf.

**5. Results**

In this section, we will be discussing the results that we got before and after model optimization. We will be starting with results from the models using default values of hyperparameters.

**5.1 Models with default hyperparameters**

Using the default parameters, our results for accuracy on training and test datasets for logistics and decision tree models are as follows:

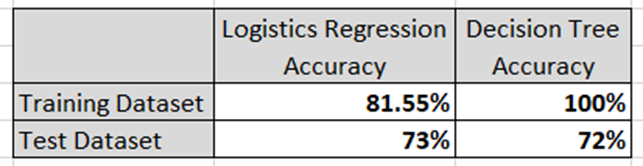


Figure: Accuracy of the models with default hyperparameters

As we can see, the decision tree model is performing perfectly on the training set with an accuracy of 100%, while the logistics regression has an accuracy of 81.55% on the training set. However, on a test dataset, logistics regression is performing better than the decision tree. Model accuracy on test dataset is extremely important as it shows us how a model performs on a data that it has not seen. Now, we will be comparing receiver operating characteristics (ROC) curves for both the models with default hyperparameters. The ROC curves for both the models are as follows:

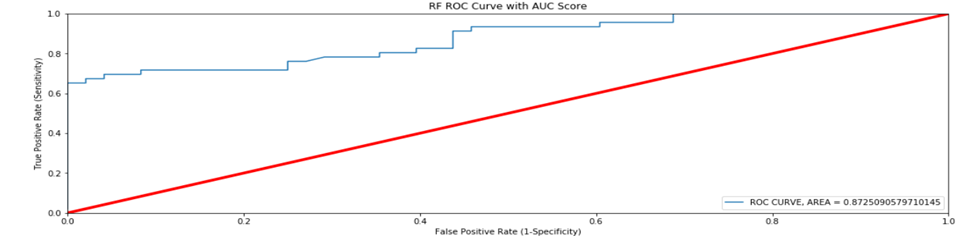


Figure: ROC curve for Logistics Regression Model

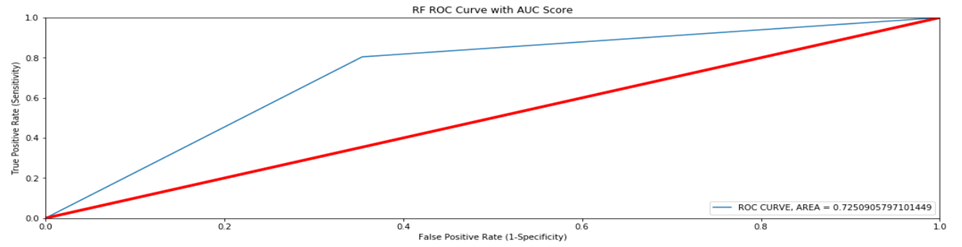


Figure: ROC curve for Decision Tree Model

As we can see that AUC for logistic regression (i.e AUC =0.87) is far better than AUC for decision tree (0.72), which shows that irrespective of the classification thresholds, logistics regression is performing better than the decision tree. Therefore, with default hyperparameters, after comparing test dataset accuracy and AUC values, logistics regression is better compared to decision tree in classifying firm as bankrupt based on financial variables.

**5.2 Models after hyperparameters tuning**

After tuning the hyperparameters for logistics regression and decision tree using the GridSearchCV algorithm with 10-fold cross-validation, our results for accuracy on training and test datasets are as follows:

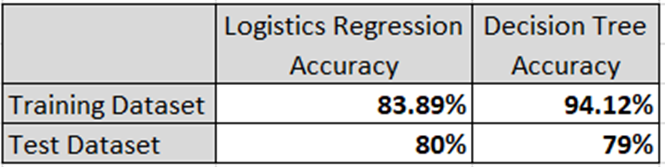


Figure: Accuracy of the models with optimized hyperparameters

As we can see, the decision tree model is performing better on the training set with accuracy of 94.12% which is less than its training accuracy on training dataset with default hyperparameters, while the logistics regression has an accuracy of 83.89% on training set which is higher than its accuracy on training dataset with default hyperparameters. However, on a test dataset, logistics regression is again performing better than the decision tree. Model accuracy on test dataset is extremely important as it shows us how a model performs on a data that it has not seen. Now, we will be comparing receiver operating characteristics (ROC) curves for both the models with optimized hyperparameters. The ROC curves for both the models are as follows:

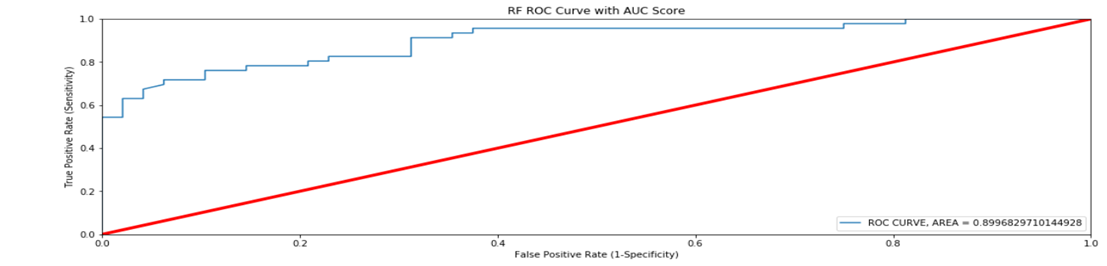


Figure: ROC curve for Logistics Regression Model

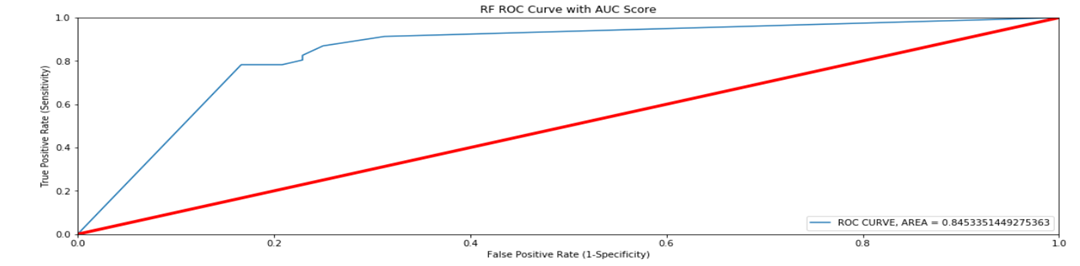


Figure: ROC curve for Decision Tree Model

As we can see that AUC for logistic regression (i.e AUC =0.89) is far better than AUC for decision tree (0.84), which shows that irrespective of the classification thresholds, logistics regression is performing better than the decision tree. Therefore, with optimizing hyperparameters, after comparing test dataset accuracy and AUC values, logistics regression is better compared to the decision tree in classifying firm as bankrupt based on financial variables. Moreover, we also observed an increase in accuracy and AUC values for both the models after optimizing the hyperparameters.

**5.3 Conclusion**

On comparing both the models before and after hyperparameters tuning, we found that logistics regression with optimized hyperparameters is a preferred choice for classifying a firm as bankrupt or non-bankrupt based on the financial variables. This conclusion is drawn after carefully looking at the results of both the models as described in section 5.1 and 5.2.

**6. Future Scope**

We believe that no model is perfect, but we can constantly work to make a model more useful and accurate. As the scope of our project was limited to considering financial variables, we believe that the accuracy of the model can be improved in future by considering other non-financial variables or attributes such as external factors which are the rise in competition, economic factors, internal factors which include weak management, ineffective leadership. Moreover, since our model has data for 468 companies, we believe that the model accuracy can be increased by having data for more number of companies from a reliable source such as Compustat.

**A. Appendix**

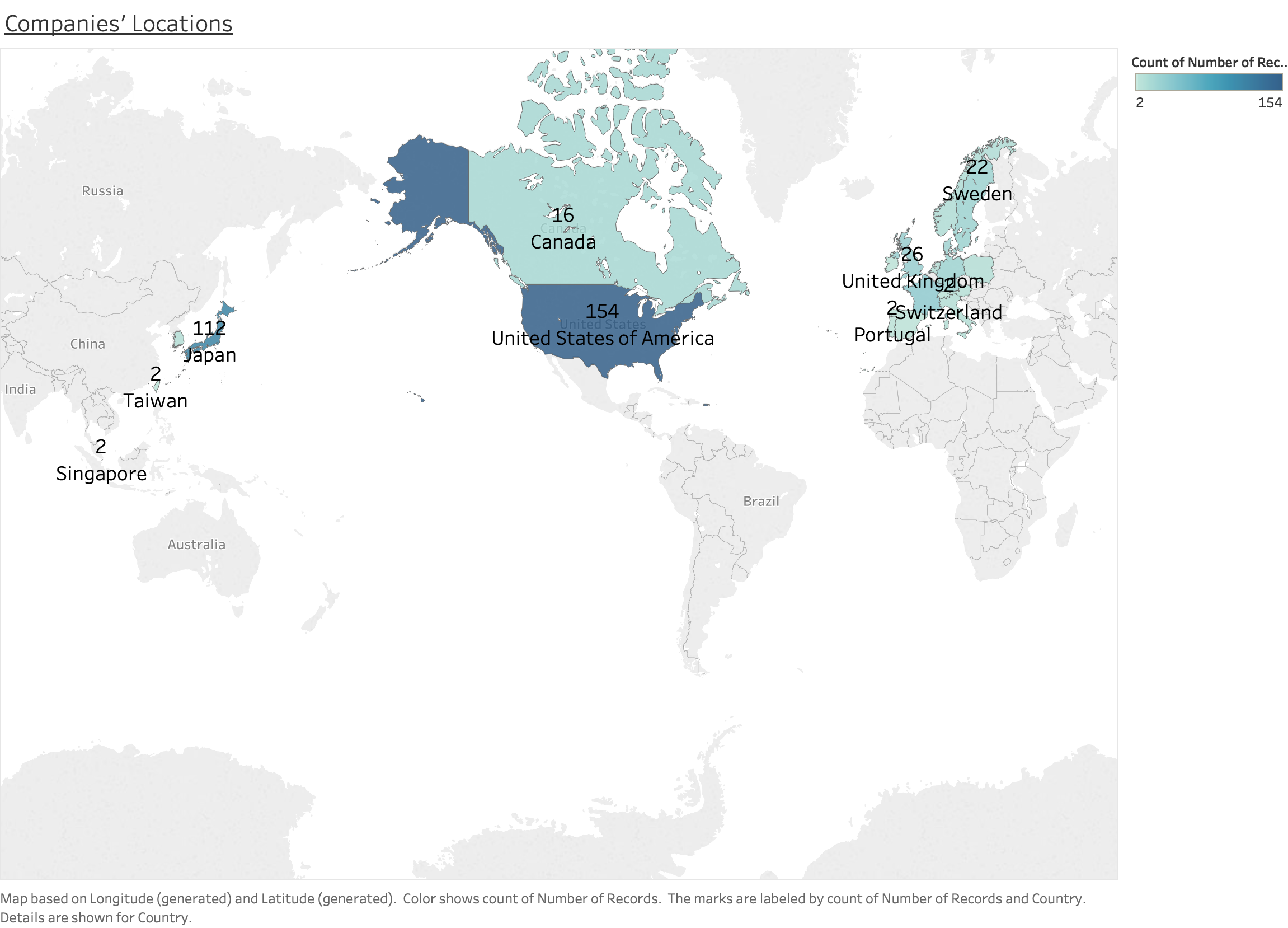
****

Fig 1: Distribution of complete data with respect to countries.

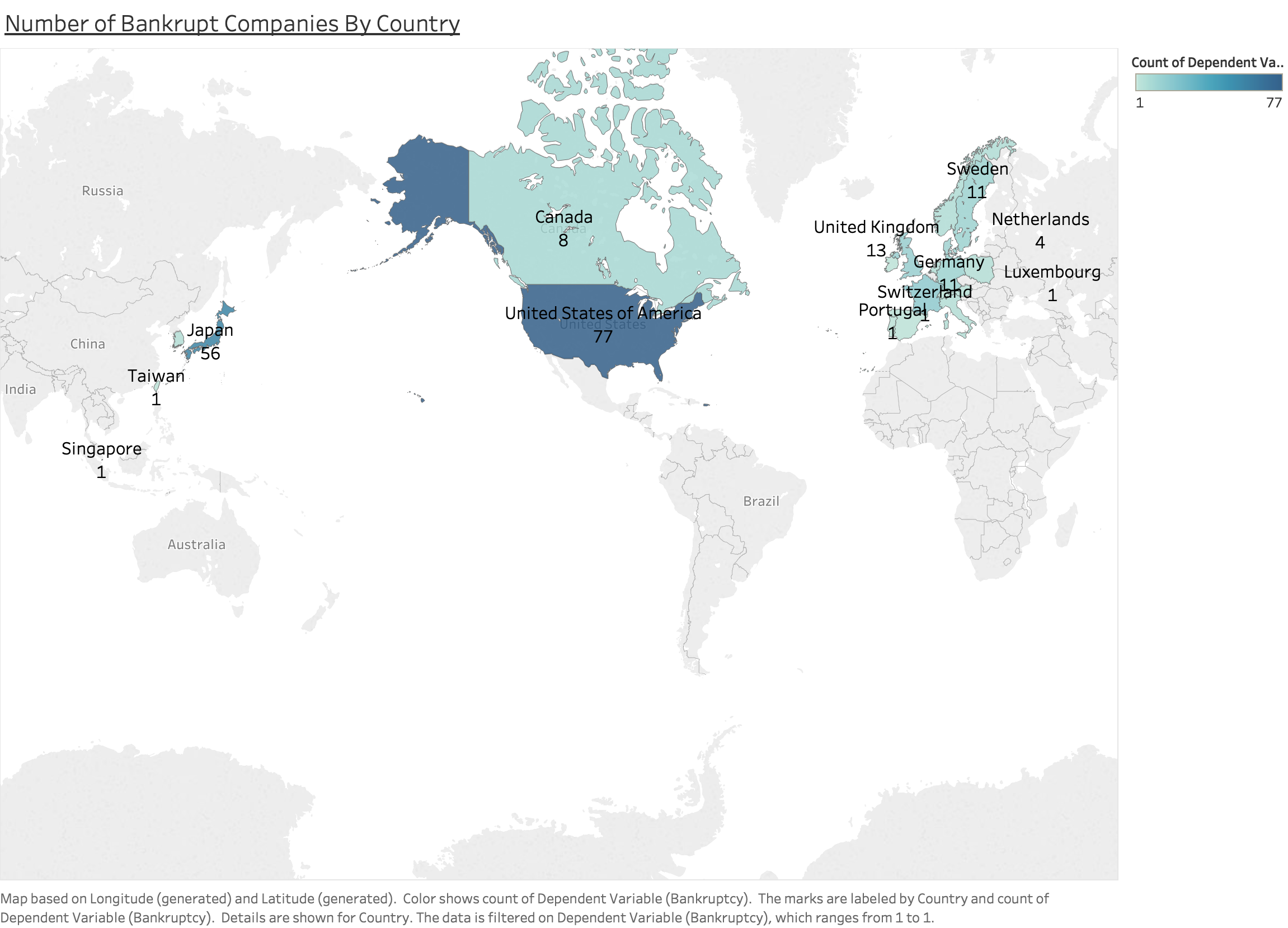
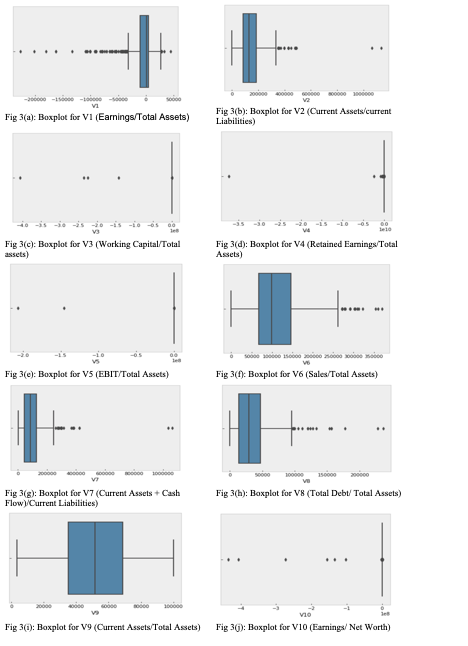
****

Fig 2: Distribution of Bankrupt Firm from dataset with respect to countries.

****

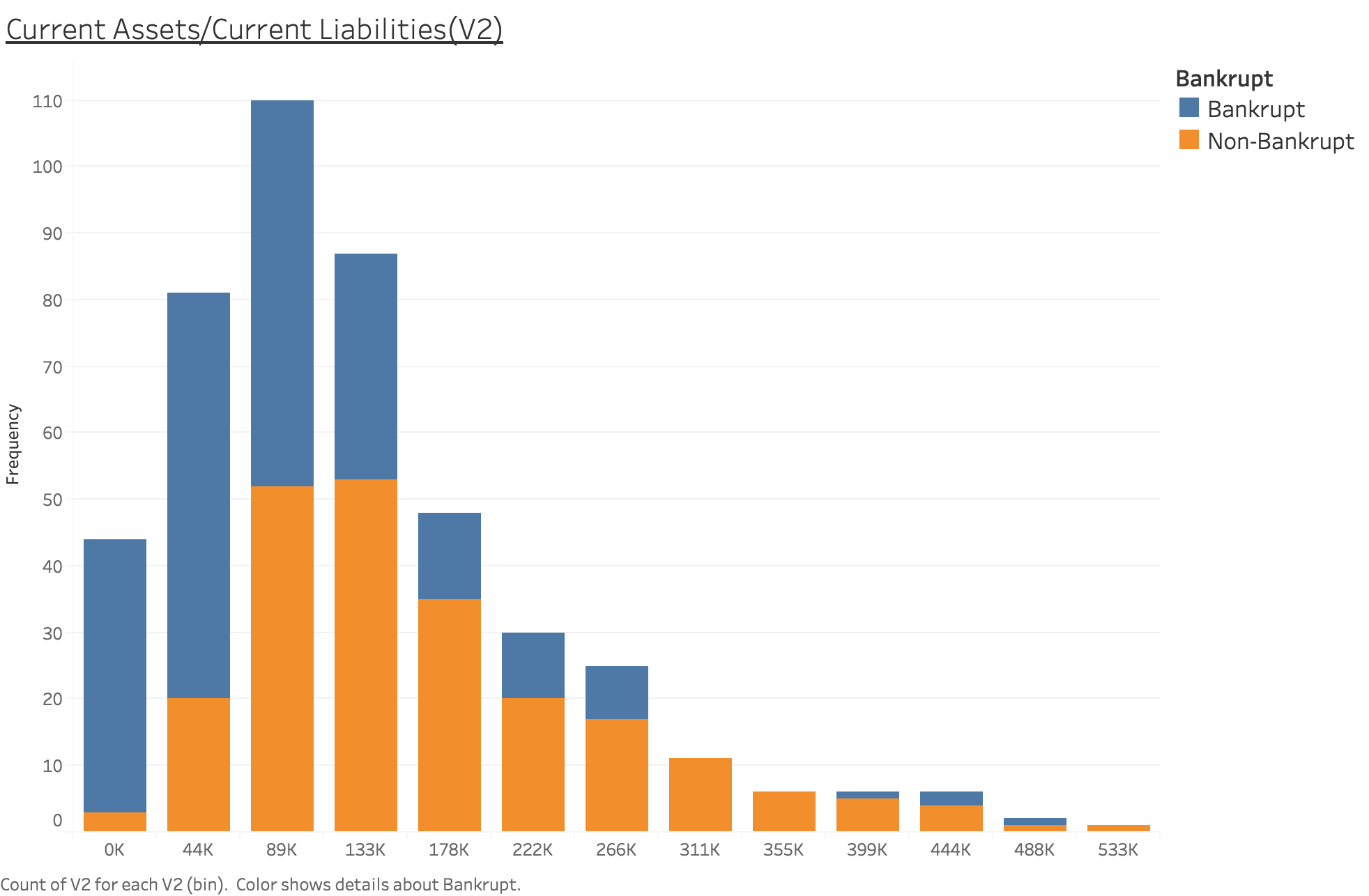
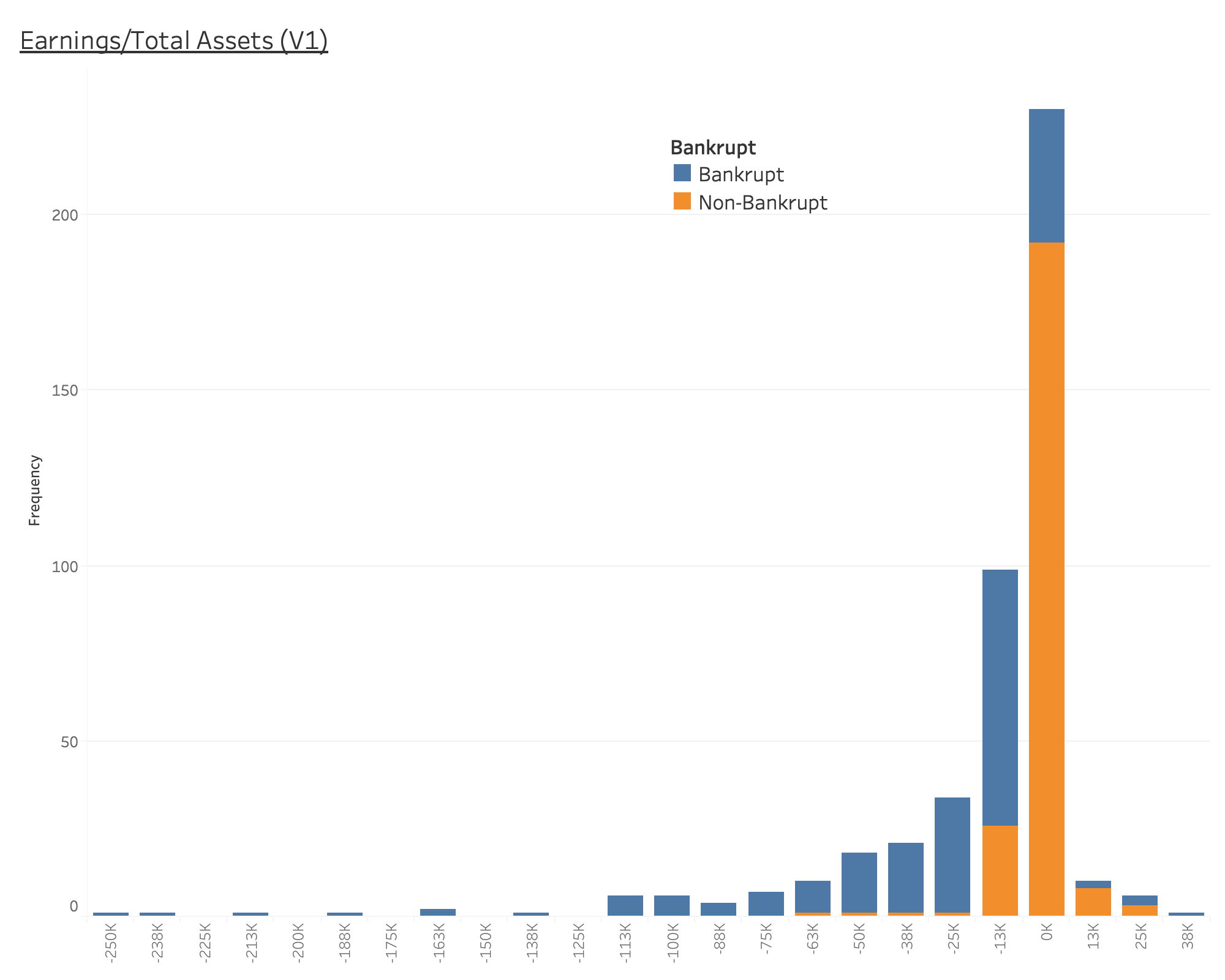
****

Fig 4(a): Trend of V1 (Earnings/Total Assets) Fig 4(b): Trend of V2 (Current Assets/current Liabilities)

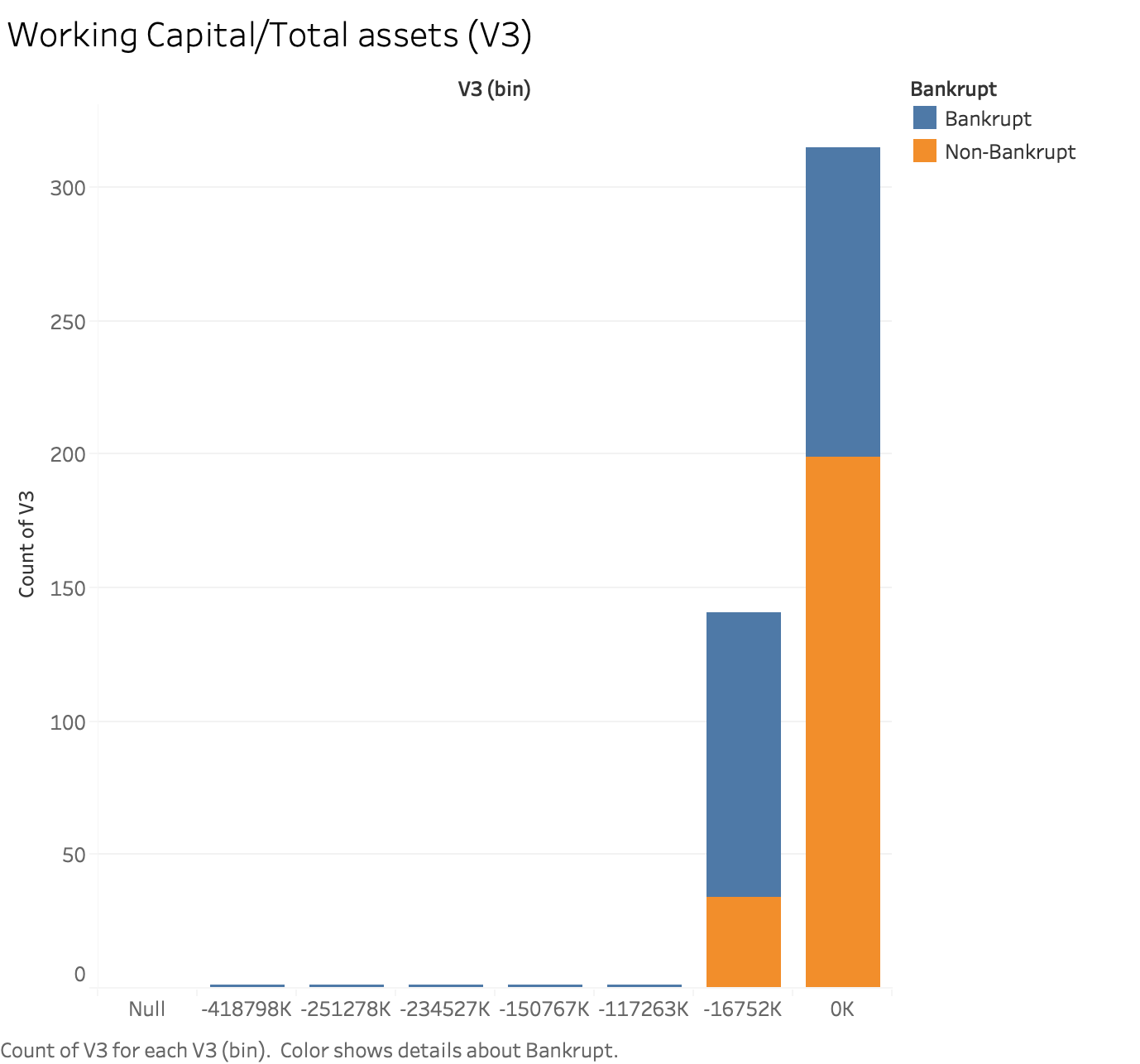
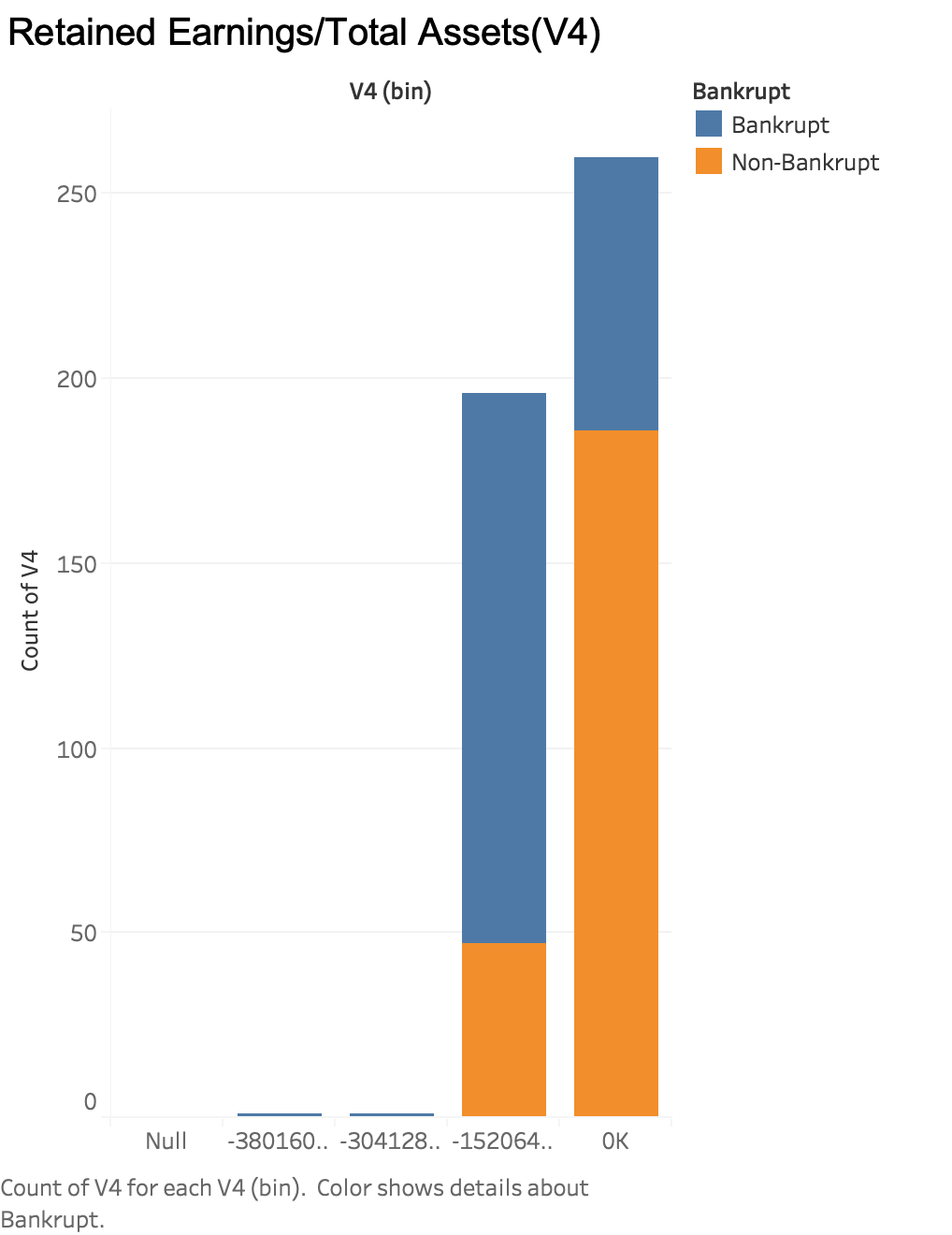
** **

Fig 4(c): Trend of V3 (Working Capital/Total assets) Fig 4(d): Trend of V4 (Retained Earnings/Total Assets)

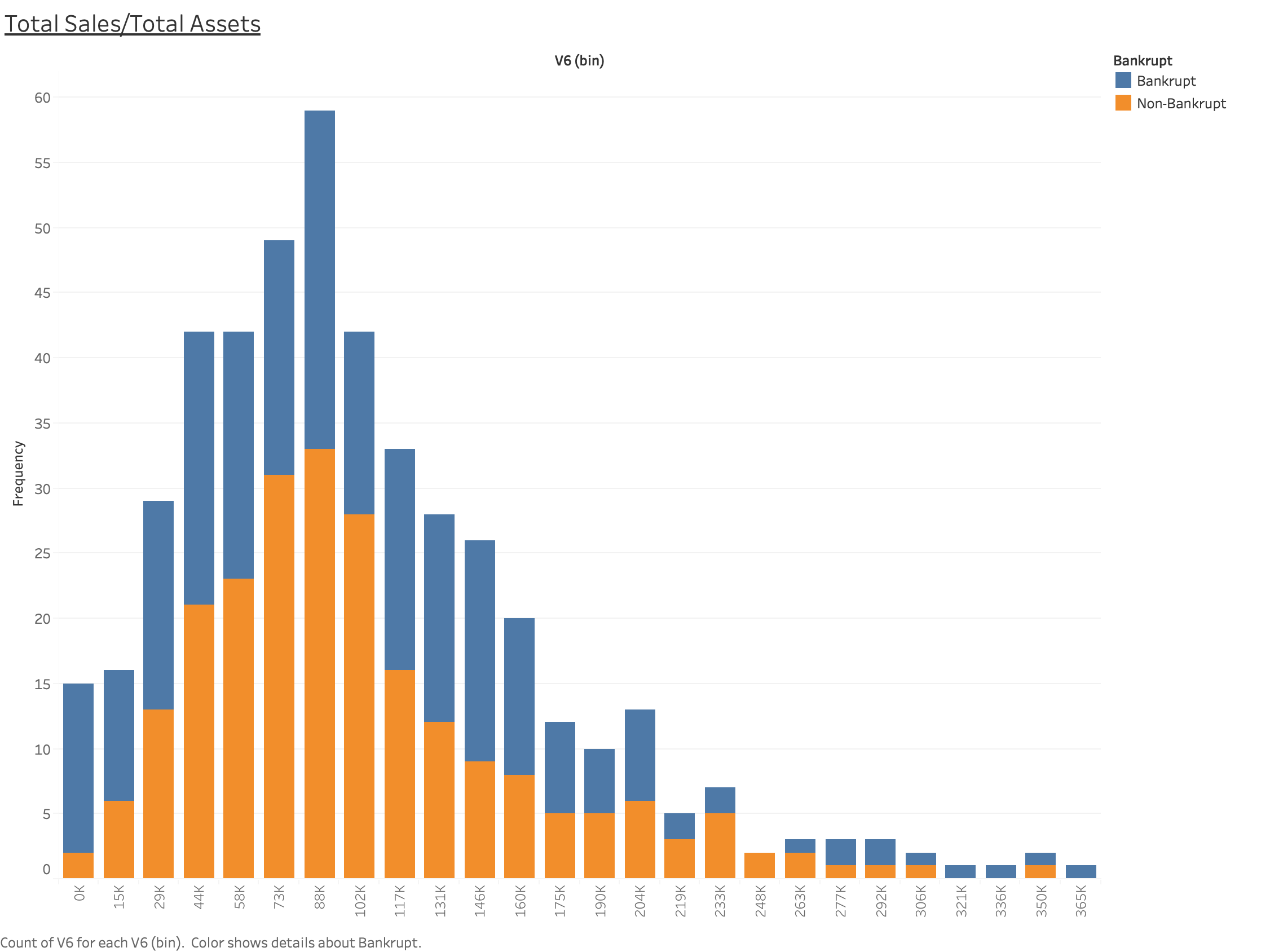
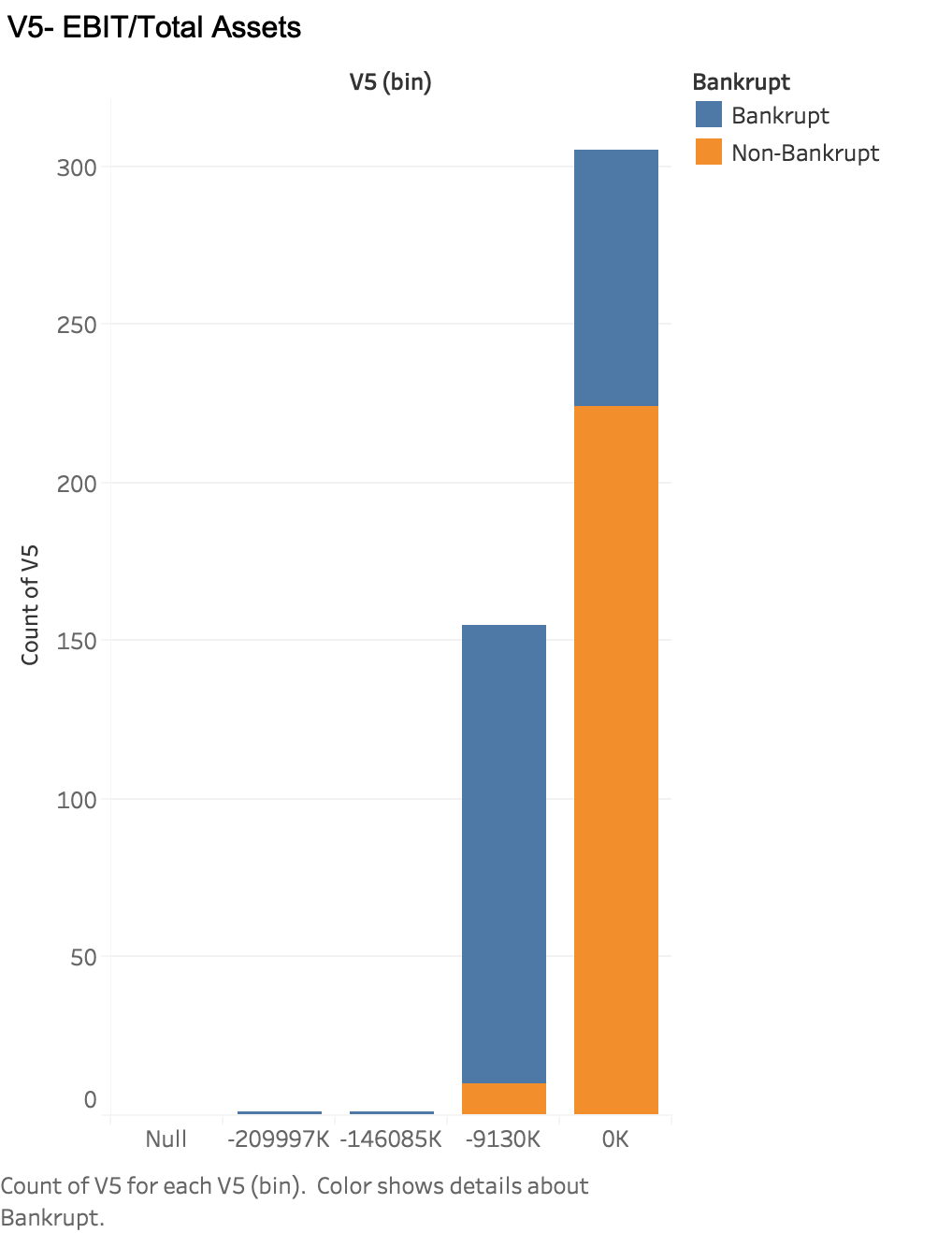
****

Fig 4(e): Trend of V5 (EBIT/Total Assets) Fig 4(f): Trend of V6 (Sales/Total Assets)

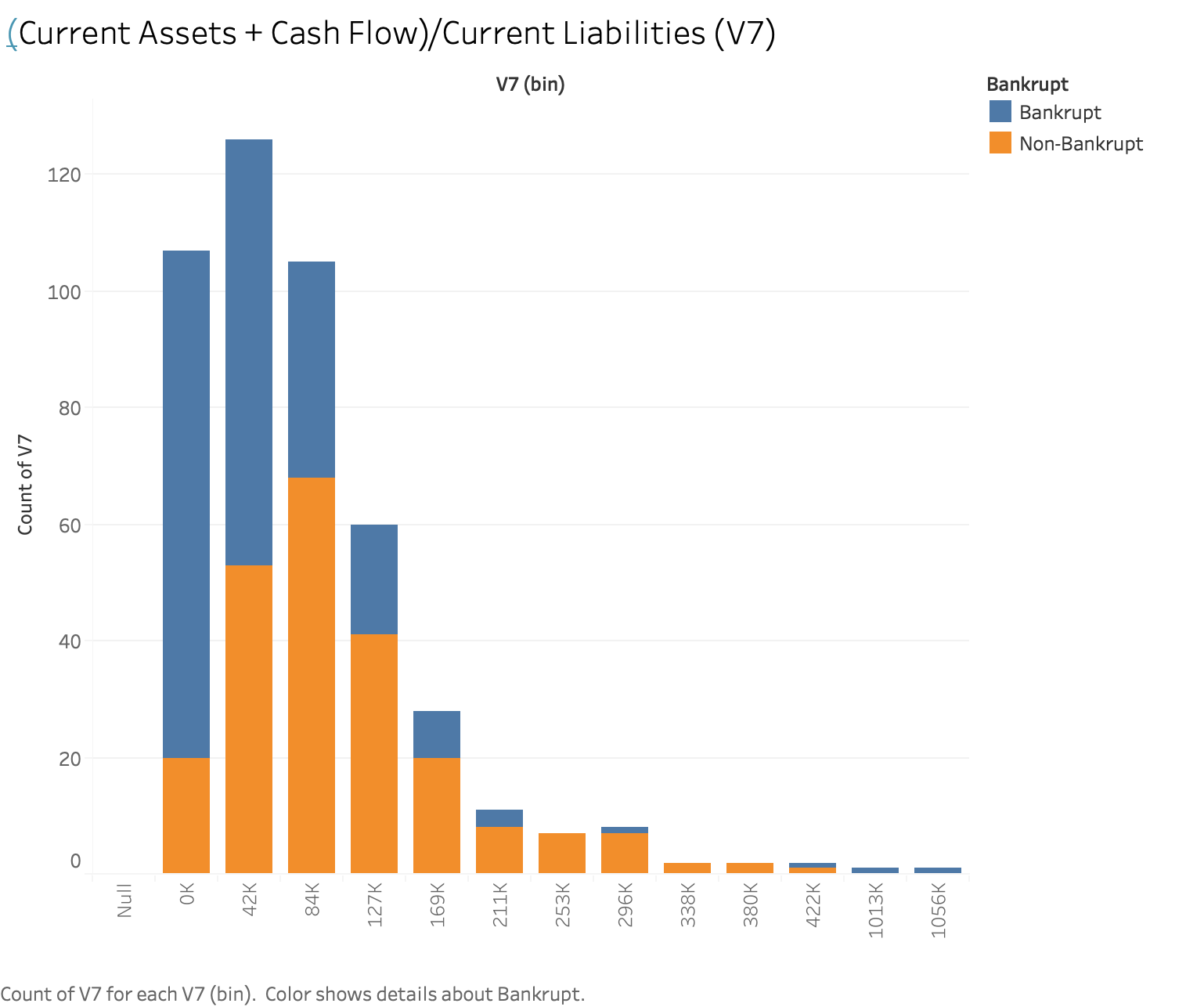
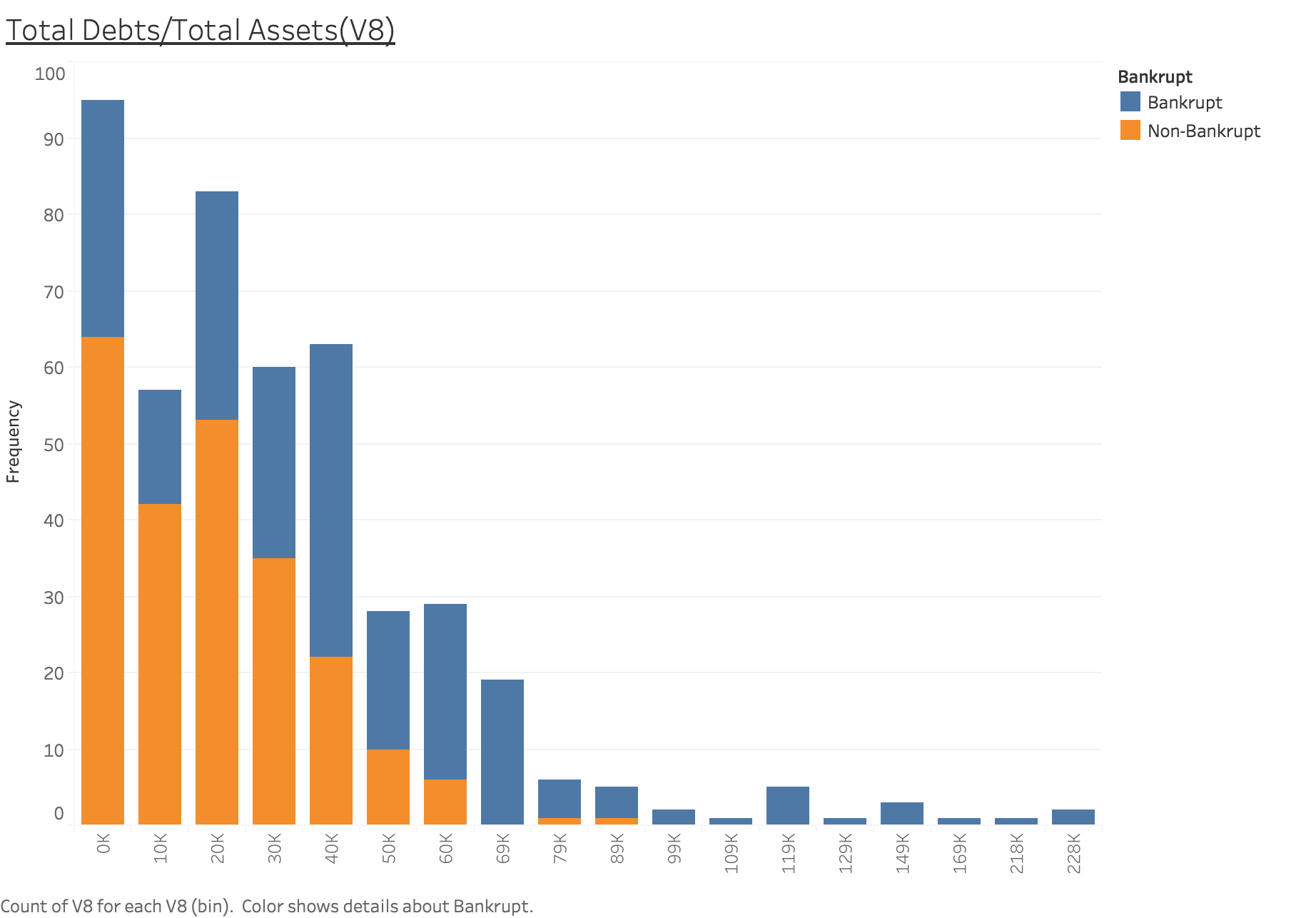
** **

Fig 4(g): Trend of V7 Fig 4(h): Trend of V8 (Total Debt/ Total Assets)

(Current Assets + Cash Flow)/Current Liabilities)

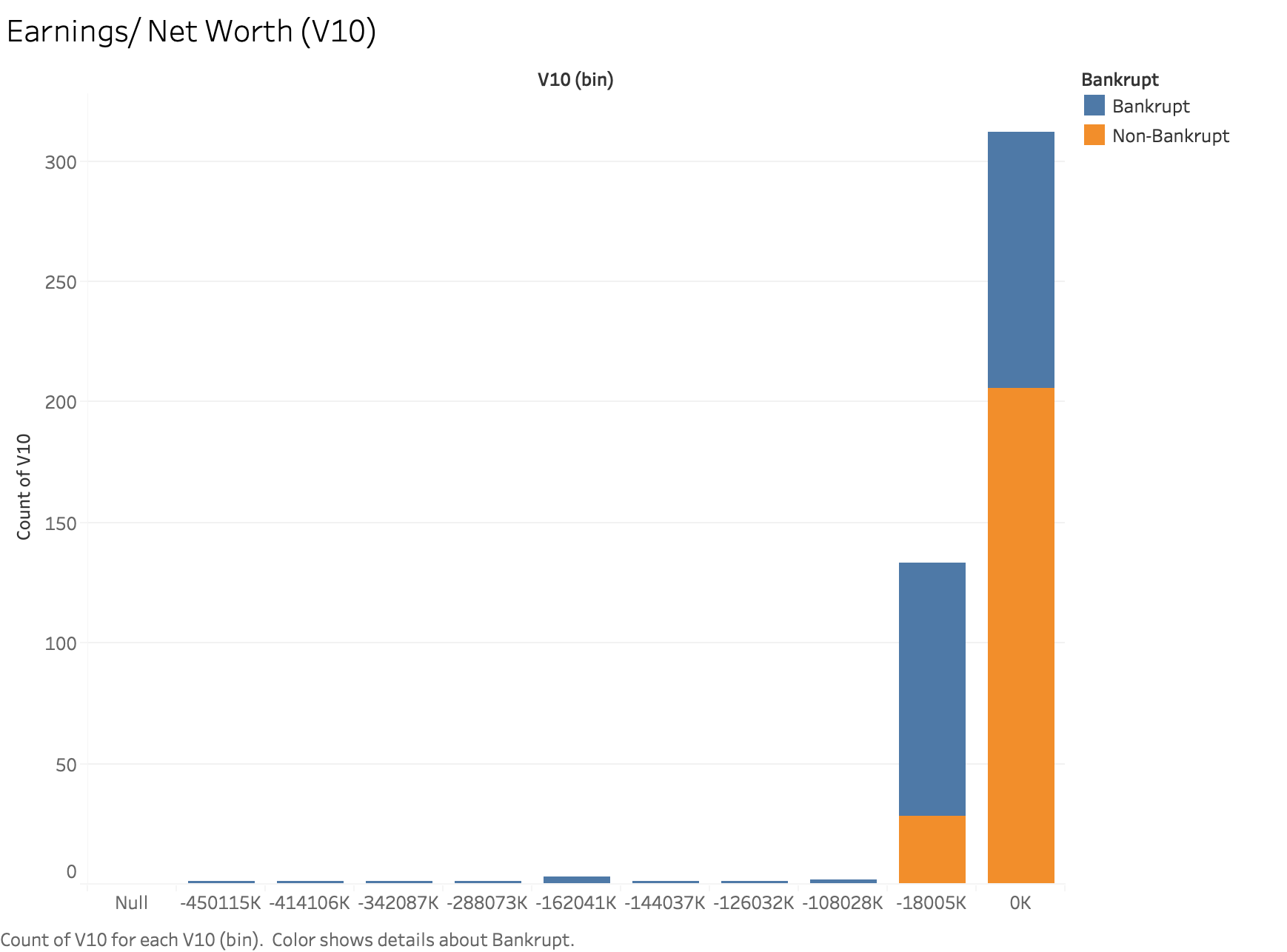
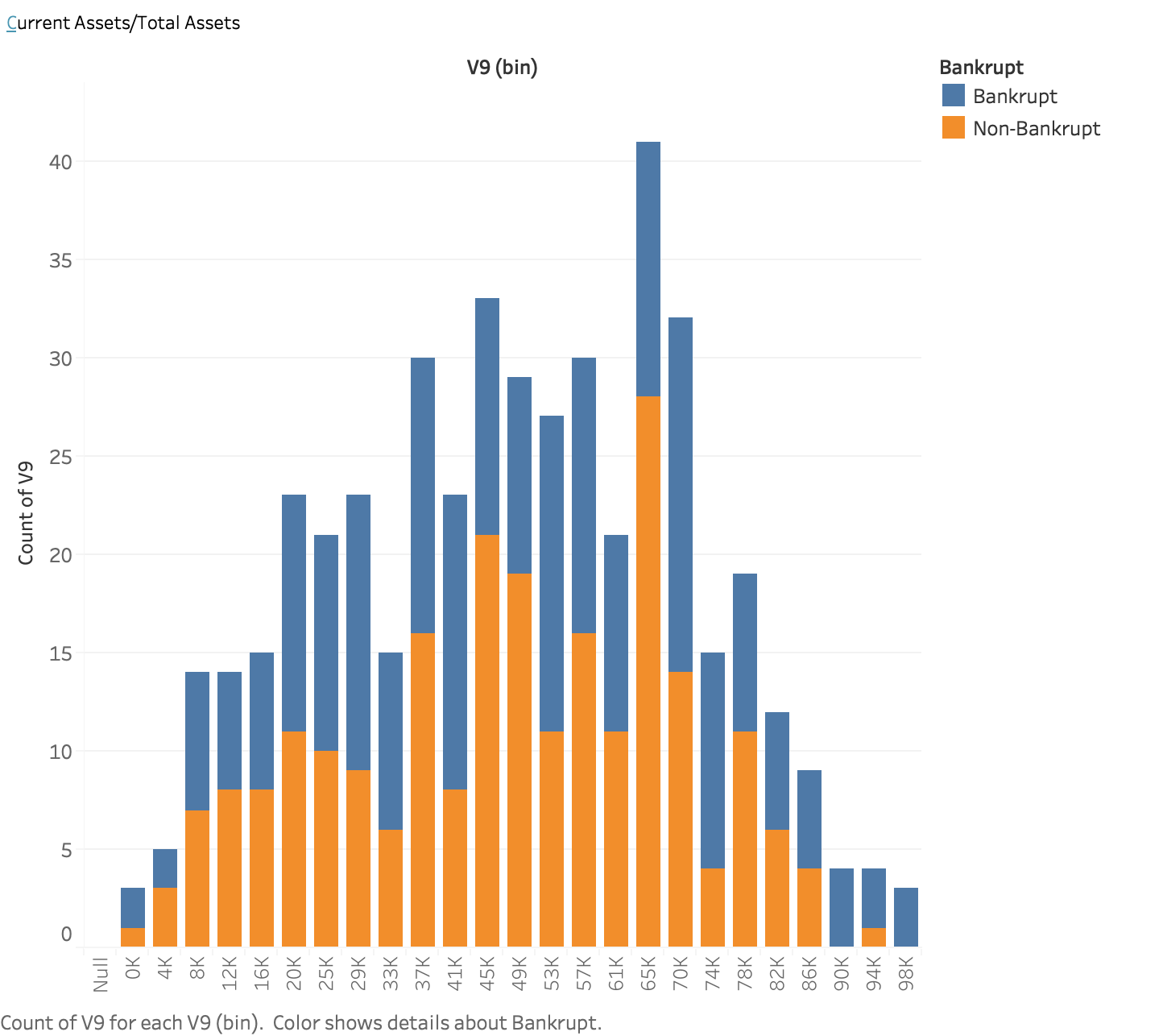
****

Fig 4(i): Trend of V9 (Current Assets/Total Assets) Fig 4(j): Trend of V10 (Earnings/ Net Worth)

1. "Bankruptcy and Business: Causes of Bankruptcy | MoreBusiness.com." <https://www.morebusiness.com/business-bankruptcy/>. [↑](#footnote-ref-1)
2. “A Global Model for Bankruptcy Prediction”

   <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5120822/> [↑](#footnote-ref-2)
3. “A Global Model for Bankruptcy Prediction”

   <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5120822/> [↑](#footnote-ref-3)
4. https://medium.com/@contactsunny/label-encoder-vs-one-hot-encoder-in-machine-learning-3fc273365621 [↑](#footnote-ref-4)
5. https://scikit-learn.org/stable/modules/grid\_search.html [↑](#footnote-ref-5)