COSC757 Assignment 1

Matthew Quander
Dept. of Computer and Information Sciences
Towson University
Towson, MD, USA
mquand1@students.towson.edu

***Abstract*—This paper provides an analysis of the Istanbul Stock Exchange data set from the UCI Machine Learning Repository. The data set includes the percentage returns of the Istanbul Stock Exchange, along with seven other international indexes, from January 5, 2009 to February 22, 2011. The data set is characterized as multivariate, univariate, and time-series, and is classified for the tasks of regression and classification. Given the Istanbul Stock Exchange data, I will focus on the returns themselves along with their relation to other indexes.**

***Keywords*–-exploratory data analysis, data preprocessing, regression, stock market**

I. INTRODUCTION

This paper will focus primarily on the Istanbul Stock Exchange's (ISE) daily returns in Turkish Lira currency (TL), between January 5, 2009 and February 22, 2011[1]. First I will explore the data set using statistical tools in R. This will include visualizations and descriptive statistics of the data. Although the source data was likely gathered as percentages from open and close index values, the ISE data set was preprocessed into decimal values. Notwithstanding, other normalization techniques were applied, providing more insight into the data. In addition, binning methods were used to observe a frequency distribution. Lastly, I proposed a prediction question and developed a regression model in an attempt to answer it.

II. DATA SET DESCRIPTION

The data set is made up of 8 attributes that include the percentage returns of the Istanbul Stock Exchange (ISE), S&P500 Index, DAX (German stock index), Financial Times Stock Exchange Index (FTSE, UK stock index), Tokyo Stock Exchange Index (NIKKEI), Brazilian Stock Exchange Index (BOVESPA), MSCI European Union Index, and MSCI Emerging Markets Index. The data was collected by the Borsa Istanbul, which the Istanbul Stock Exchange (and many other Turkish exchanges) is now a part of, and Yahoo Finance[1]. The data set contains a total of 536 instances for each of the 8 attributes. Although the US based returns of the ISE are included, the analysis in this paper will primarily focus on the returns in Turkish Lira. The 536 instances consist of the percentage returns of trading days for the above indexes, and does not contain any missing values. A value of 0 is used for weekdays in which markets were closed, since no loss or gain was realized on such days[1].

III. EXPLORATORY DATA ANALYSIS

Loading and preparing the data set into R Studio allowed for the following summary statistics to be derived. Using the `sum` function for attribute values greater than 0 and dividing by the number of instances, I obtained the probability of positive returns. The TL Based ISE had a 0.5708955 probability (57.08955% chance) of yielding positive returns for any given trading day during the observed time period. In exploring the data visually, a distribution of the attributes is displayed through histograms using the `hist` function with the appropriate parameters. The indexes' percentage returns appear to be correlated since most of their data points range between -2% and 2%. However, the ISE's mean appears to be less correlated to the other indexes' means. The ISE's mean for the observed time period is .1629%, while the other indexes' mean range from 0.03077% to 0.09359%. This suggests that the ISE had a higher percentage return over the given time period. An additional observation is that the median value is greater than the mean, which indicates that this attribute in the data set is negatively skewed. These results were derived using the `summary` function on each attribute.

TABLE I. ISE(TL BASSED)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.062208	-0.006669	0.002189	0.001629	0.010584	0.068952

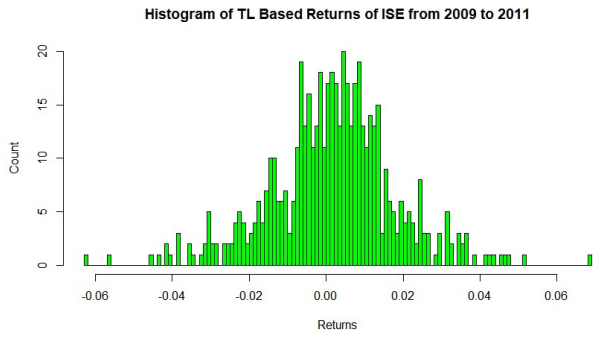

Fig. 1. Histogram of ISE returns.

TABLE II. SP

Min.	1st Qu.	Median	Mean
-0.0542620	-0.0046748	0.0008764	0.0006433
3rd Qu.	Max.		
0.0067056	0.0683664		

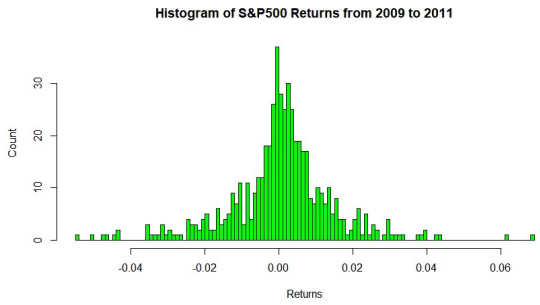

Fig. 2. Histogram of S&P500 returns.

TABLE III. DAX

Min.	1st Qu.	Median	Mean
-0.0523312	-0.0062121	0.0008875	0.0007208
3rd Qu.	Max.		
0.0082235	0.0589505		

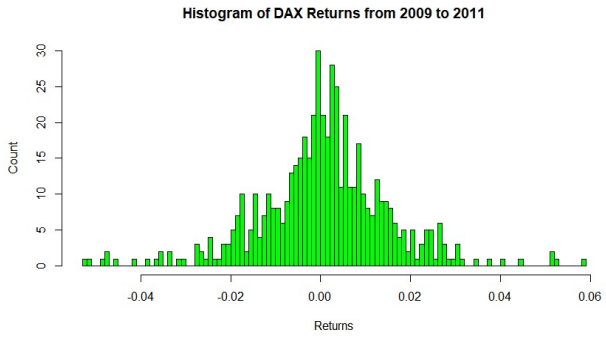

Fig. 3. Histogram of DAX returns.

TABLE IV. FTSE

Min.	1st Qu.	Median	Mean
-0.0548160	-0.0058084	0.0004086	0.0005103
3rd Qu.	Max.		
0.0074282	0.0503227		

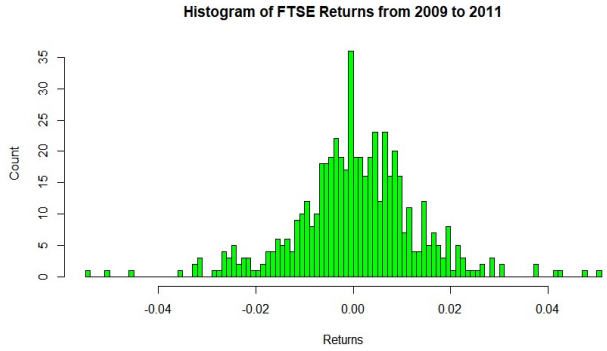

Fig. 4. Histogram of FTSE returns.

TABLE V. NIKKEI

Min.	1st Qu.	Median	Mean
-0.0504476	-0.0074072	0.0000000	0.0003077
3rd Qu.	Max.		
0.0078821	0.0612293		

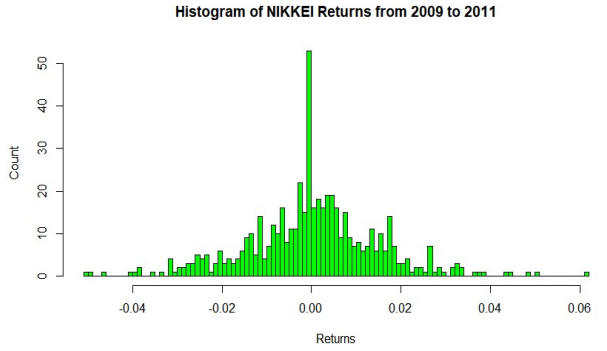

Fig. 5. Histogram of NIKKEI returns.

TABLE VI. BOVESPA

Min.	1st Qu.	Median	Mean
-0.0538495	-0.0072146	0.0002790	0.0009353
3rd Qu.	Max.		
0.0088808	0.0637915		

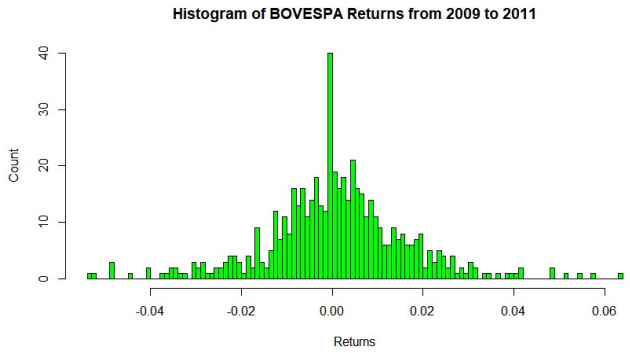

Fig. 6. Histogram of BOVESPA returns.

TABLE VII. EU

Min.	1st Qu.	Median	Mean
-0.0488168	-0.0059518	0.0001958	0.0004706
3rd Qu.	Max.		
0.0077915	0.0670425		

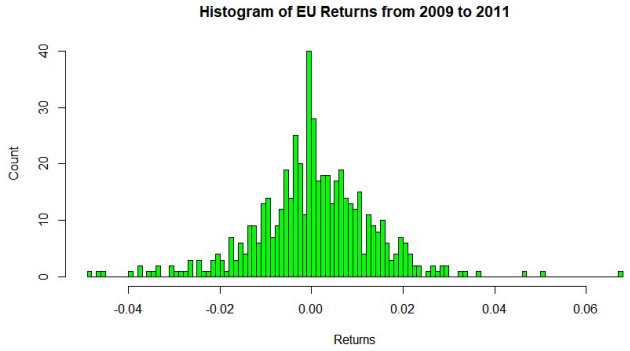

Fig. 7. Histogram of EU returns.

TABLE VIII. EM

Min.	1st Qu.	Median	Mean
-0.0385645	-0.0049114	0.0010772	0.0009359
3rd Qu.	Max.		
0.0064226	0.0478045		

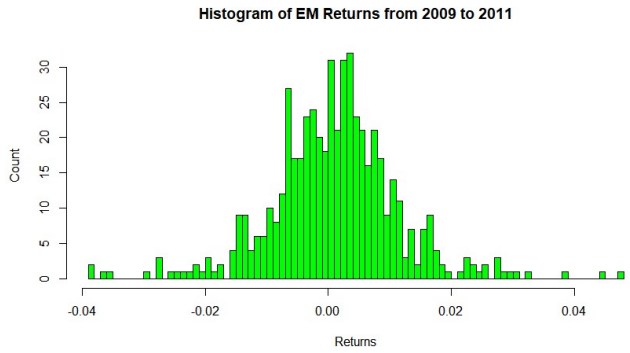

Fig. 8. Histogram of EM returns.

IV. DATA PREPROCESSING

A. *Normalization*

Upon downloading the data set, the data was already converted into decimal values representing the percentage gain or loss. After applying the min-max normalization formula using R, the below descriptive statistics were obtained.

TABLE IX. ISE(TL) MIN-MAX NORMALIZATION

Min.	1st Qu.	Median	Mean	3rd Qu.
0.0000	0.4234	0.4910	0.4867	0.5550
Max.				
1.0000				

Similarly, using the z-score standardization formula in R gave some insightful information about the data. The minimum and maximum z-scores range from -3.92498 and 4.13931. Since these values are outside of the -3 to 3 range, there exists outliers in the ISE(TL) subset.

TABLE X. Z-SCORE STANDARDIZATION

Min.	1st Qu.	Median	Mean	3rd Qu.
-3.92498	-0.51017	0.03442	0.00000	0.55059
Max.				
4.13931				

In further processing of the data, a skewness of -0.1032719 was calculated for the ISE TL Based returns using the skewness formula. A negative skew could have also been inferred by the fact that this subset's mean is less than its median, as shown in the descriptive statistics above.

Since the dataset was already preprocessed into decimal format, I applied the inverse (multiplying by 100) to display the descriptive statistics as percentages for the ISE TL Based subset:

TABLE XI. ISE(TL) PERCENTAGES

```
Min.    1st Qu.  Median   Mean    3rd Qu.
-6.2208 -0.6669  0.2189   0.1629  1.0584
Max.
6.8952
```

B. Binning

Binning the variables by k-means clustering into k=3 clusters using R's `kmeans` function resulted in a majority of the data values falling between .998 (inclusive) and 1.67 (non-inclusive).

TABLE XII. BINNING WITH K=3 CLUSTERS

```
(0.998,1.67]   (1.67,2.33]      (2.33,3]
         311           116           109
```

Binning the subset using R's `cut` and `table` functions into 5 arbitrary bins of equal widths of roughly 0.0262, the following levels and frequency distribution were established:

TABLE XIII. BINNING WITH EQUAL WIDTH

```
(-0.0623,-0.036](-0.036,-0.00974](-0.00974,0.0165]
          10               99              354

(0.0165,0.0427]   (0.0427,0.0691]
          66                7
```

This distribution indicates that the vast majority of the percentage returns for the ISE (TL) fall within the range of -0.00974 (non-inclusive) and 0.0165 (inclusive), which is within one standard deviation (0.01626427) from the mean (0.001629).

C. Preprocessing

Since the data was already preprocessed and also contains negative values for the percentage loss days, I was unable to perform a meaningful normalized distribution using the square root, inverse square root, and natural log functions. Furthermore, attempting to use those functions on negative numbers resulted in several undefined numbers in the resulting data frame.

V. REGRESSION ANALYSIS

From the exploratory data analysis, it appears there may be some correlation between the ISE's TL Based stock returns and the returns of the other indexes. This seems possible from the patterns in the histograms and descriptive statistics. Therefore I pose the prediction question:

Is there any correlation between the index returns of the ISE and the S&P500 or the NIKKEI, on the trading days between January 1, 2009 and February 22, 2011?

A. Correlation with the S&P500 and the NIKKEI

Using the regression analysis `lm` function in R, I obtained a low R^2 value of 19.32% (adjusted: 19.16%) for the ISE and S&P500, and a regression line of:

1) $\hat{y} = 0.0013026 + 0.5072017x$

The R^2 value tells us that the linear regression does not fit the data well, since a value of 100% indicates a perfect fit. Using the `plot` and `abline` functions, a scatter plot and regression line were created. In the graph the data points are clustered around the middle rather than along the regression line.

TABLE XIV. ISE(TL) AND S&P500 REGRESSION SUMMARY

```
Call:
lm(formula = tlBasedReturns ~ SP)

Residuals:
 Min        1Q    Median        3Q       Max
-0.065231 -0.008348  0.000653  0.008435  0.045822

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0013026  0.0006323    2.06   0.0399 *
SP          0.5072017  0.0448599   11.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 0.01462 on 534 degrees of
freedom
Multiple R-squared:  0.1932,  Adjusted       R-squared:
0.1916
F-statistic: 127.8 on 1 and 534 DF,  p-value: < 2.2e-16
```

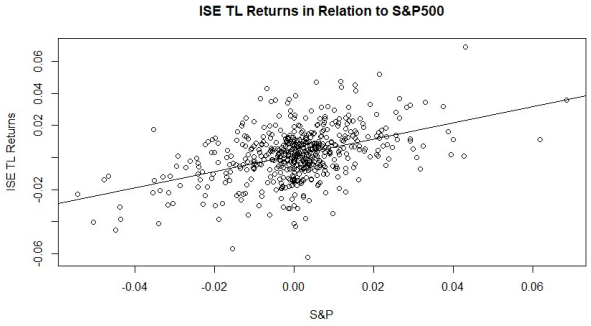

Fig. 9. Scatter plot and regression line of ISE(TL) and S&P500 returns.

Similar results are found when performing linear regression with the NIKKEI (R^2 value of 6.763%, adjusted: 6.588%) where the data points are clustered in the middle, and with multiple regression of the NIKKEI and the S&P500 together (R^2 value of 23.48%, adjusted: 23.2%). Below are the respective regression lines:

2) $\hat{y} = 0.0015412 + 0.2848160x$

3) $\hat{y} = 0.0012533 + 0.4760111x_1 + 0.2255248x_2$

This leads to a negation of the prediction question – the returns of the ISE(TL) and the S&P500 and NIKKEI indexes are not correlated for the observed time period.

TABLE XV. ISE(TL) AND NIKKEI REGRESSION SUMMARY

```
Call:
lm(formula = tlBasedReturns ~ nikkei)

Residuals:
       Min         1Q     Median         3Q        Max
 -0.052339 -0.008941  0.000627  0.008672  0.062882

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
  Intercept) 0.0015412 0.0006791 2.269   0.0236 *
 nikkei      0.2848160 0.0457645 6.224 9.84e-10***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual standard error: 0.01572 on 534 degrees of
freedom
Multiple R-squared:  0.06763,  Adjusted R-squared:
0.06588
F-statistic: 38.73 on 1 and 534 DF,  p-value: 9.836e-
10
```

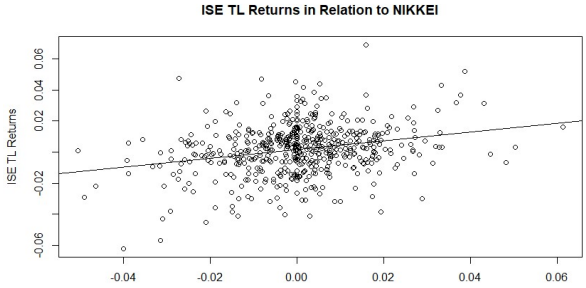

Fig. 10. Scatter plot and regression line of ISE and NIKKEI returns.

TABLE XVI. ISE(TL), AND S&P500 AND NIKKEI REGRESSION SUMMARY

```
Call:
lm(formula = tlBasedReturns ~ SP + nikkei)

Residuals:
Min         1Q     Median         3Q        Max
-0.056041 -0.008616  0.000420  0.007944  0.046565

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0012533  0.0006164   2.033   0.0425 *
SP          0.4760111  0.0441086  10.792  <2e-16***
nikkei      0.2255248  0.0418596  5.388 1.07e-07 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 0.01425 on 533 degrees of
freedom
Multiple R-squared:  0.2348, Adjusted R-squared: 0.232
F-statistic: 81.78 on 2 and 533 DF,  p-value: < 2.2e-16
```

VI. CONCLUSION

From the above processes, only a limited amount of information could be obtained from this data set. The visualizations produced from the exploratory data analysis revealed that the returns of the 8 stock market indexes stayed within a particular range from January 5, 2009 to February 22, 2011. This observation is congruent to the slow economic recovery after the global financial crisis in 2008. In processing, I was able to identify the fact that outliers exist in this data set. This should be expected due to the volatile natural of financial markets, at times. Binning the Istanbul Stock Exchange (TL) subset in multiple ways revealed that the majority of the data points are contained within a certain range, further illustrating the slow economic growth during that time period. Lastly, the linear and multiple regression analyses performed did not provide enough evidence to conclude that the percentage returns of the Istanbul Stock Exchange (TL), and the S&P500 and NIKKEI indexes were correlated. This, however, does not completely rule out the possibility of their correlation which could be uncovered using other methods.

REFERENCES

[1] O.Akbilgoc, *UCI Machine Learning Repository: ISTANBUL STOCK EXCHANGE Data Set*, Center for Machine Learning and Intelligent Systems, University of California, Irvine, May 2013. Accessed on: Mar. 6, 2021. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ISTANBUL+ STOCK+EXCHANGE