Data Mining- Lab Exam

Time: 24 hours Marks:100

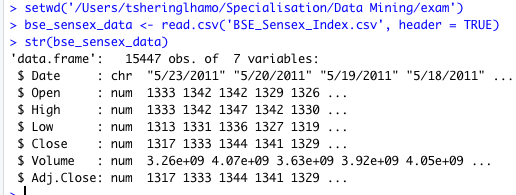
Open a document and update document with your answers for each question and submit it.

1. a) For the dataset BSE\_Sensex\_Index.csv, create an extra column of successive differences for each column of numeric values in this data file. Extract two simple random samples with replacement of 1000 and 3000 observations (rows). Show your R commands for doing this.

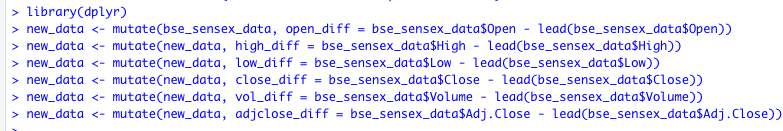
Do the same thing by using Excel. Show your Excel commands.

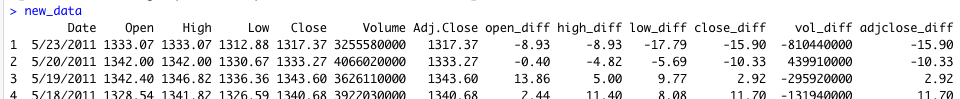
**Note:** successive difference for date d1=(date d1 value - immediate available previous date of d1 value)/immediate available previous date of d1 value. For the last row fill up values with mean of its immediate three previous row values.

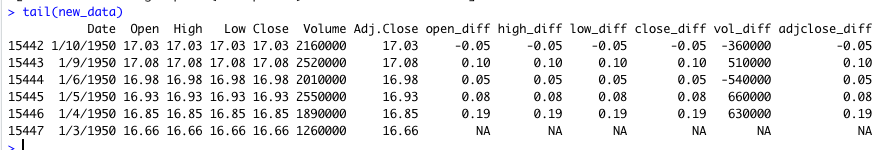
Ans. read data and see structure.



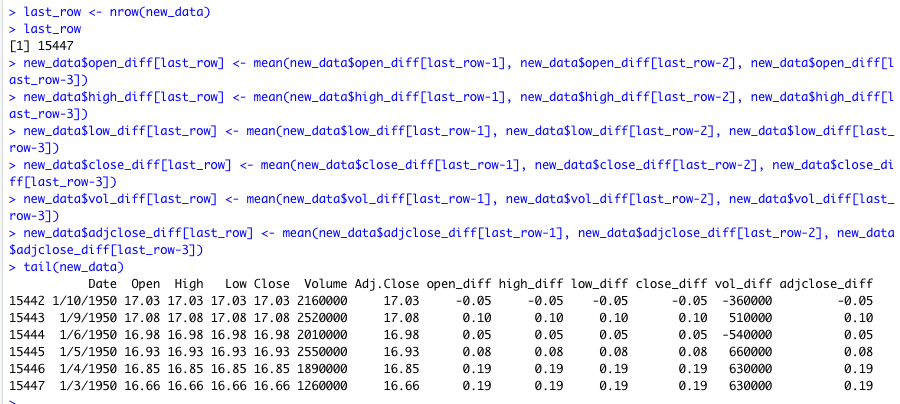
Add six new numerix columns.



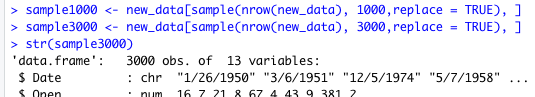




Last rows of the six new columns have NA values.



Generate sample length 1000 and 3000 with replacement.



Excel commands:

=B1-B2 #successive difference

=AVERAGE(H2999:H2997) #find mean of last row

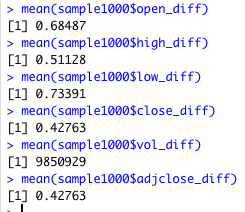
=ARRAY\_CONSTRAIN(A1:H3000, 1000, 8) #generate sample

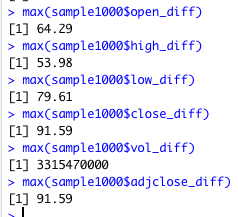
b) For your samples, use the functions mean(), max(), var() and quartile(,.25) to compute the mean, maximum, variance and 1st quartile respectively for each column which has successive differences. Show your R code and the resulting values.

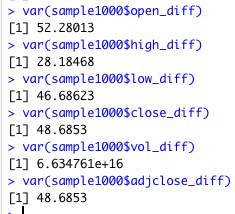
Do the same thing by using Excel. Show your Excel commands.

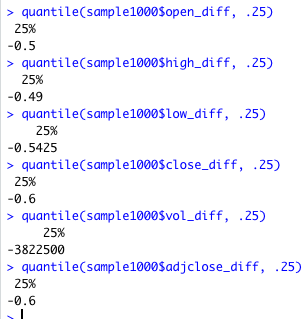
Ans.

For sample 1000

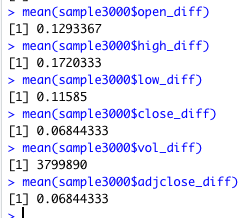


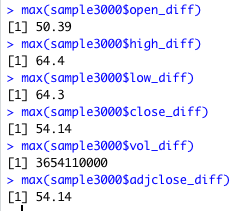


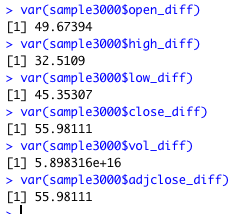


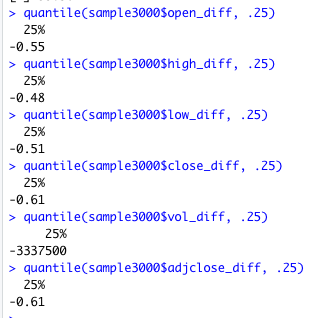


For sample 3000









Excel commands:

=AVERAGE(B2:B1000)

=VAR(B2:B1000)

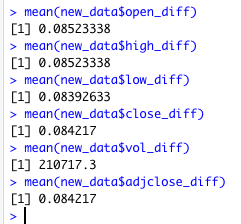
=MAX(B2:B1000)

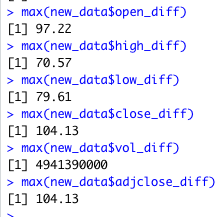
=QUARTILE(B2:B1000, 0.25)

c) Compute the same quantities in part b on the entire data set and show your answers. How much do they differ from your answers in part b? Do you find any significant difference between two sample values like mean in comparison with entire data? If so what explanation you can give for that?

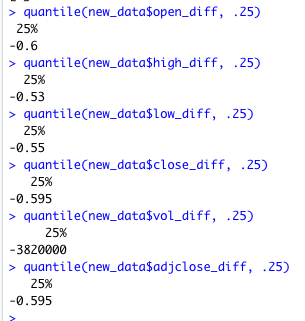
Do the same thing by using Excel. Show your Excel commands.

Ans.









Excel commands:

=AVERAGE(J2:J3000)

=VAR(J2:J3000)

=MAX(J2:J3000)

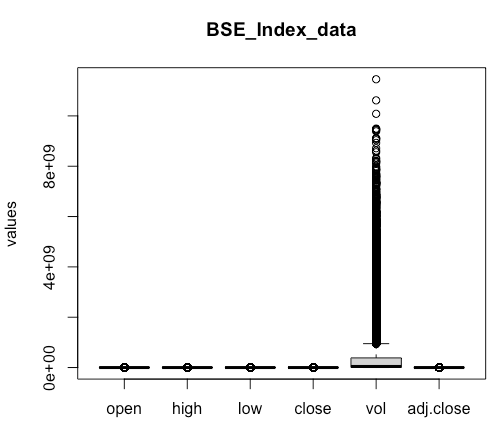
=QUARTILE(J2:J3000, 0.25)

d) Use R to produce a single graph displaying a boxplot for open, close, high and low. Include the R commands and the plot.

Do the same thing by using Excel. Show your Excel commands

Ans.



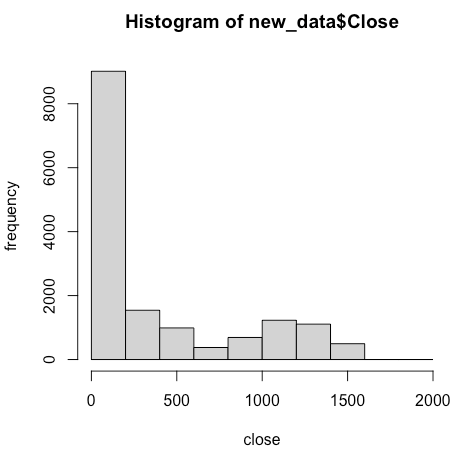


e) Use R to produce a frequency histogram for Close values. Use intervals of width 2000 beginning at 0. Include the R commands and the plot.

Do the same thing by using Excel. Show your Excel commands. (10+10=20M)

Ans.

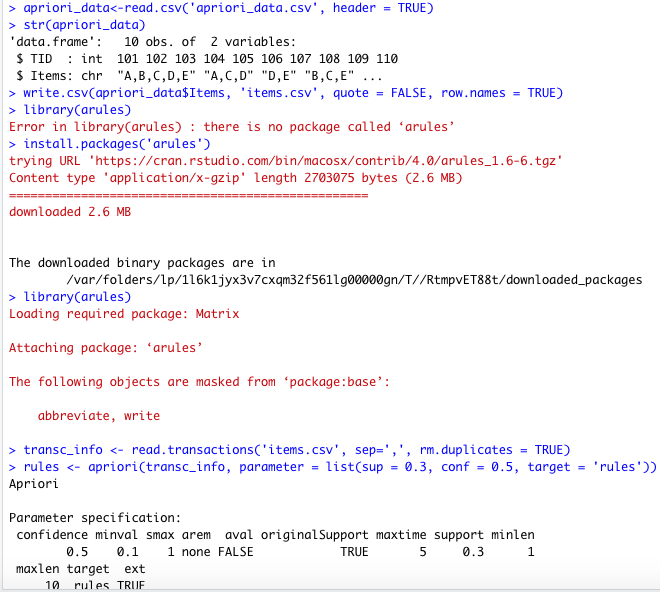




1. Implement Apriori Algorithm or use built in packages to find out the frequent itemsets and generate rules for frequent itemsets. Trace and submit the program output for the following given dataset of transactions with a minimum support of 3. (10M)

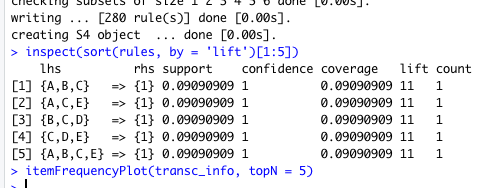
|  |
| --- |
| TID, Items |
| 101, A,B,C,D,E |
| 102, A,C,D |
| 103, D,E |
| 104, B,C,E |
| 105, A,B,D,E |
| 106, A,B |
| 107, B,D,E |
| 108, A,B,D |
| 109, A,D |
| 110, D,E |

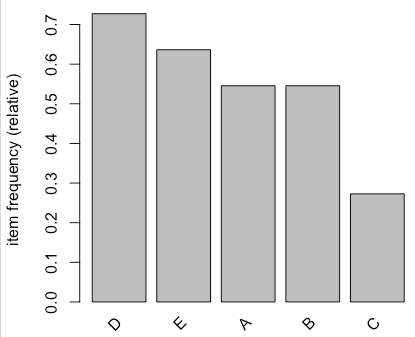
Ans.



Load package arules for apriori algorithm.

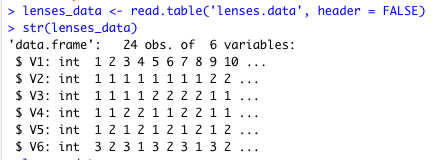






1. Build Decision Trees by using i) information gain and ii) misclassification error rate for Lenses Data Set provided at <http://archive.ics.uci.edu/ml/datasets/Lenses>. In terms of tree size what do you conclude comparing these two? (10M)

Ans.



Attribute Information:

-- 3 Classes

1 : the patient should be fitted with hard contact lenses,

2 : the patient should be fitted with soft contact lenses,

3 : the patient should not be fitted with contact lenses.

1. age of the patient: (1) young, (2) pre-presbyopic, (3) presbyopic

2. spectacle prescription: (1) myope, (2) hypermetrope

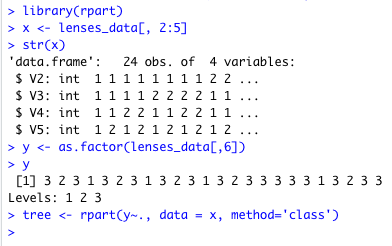
3. astigmatic: (1) no, (2) yes

4. tear production rate: (1) reduced, (2) normal

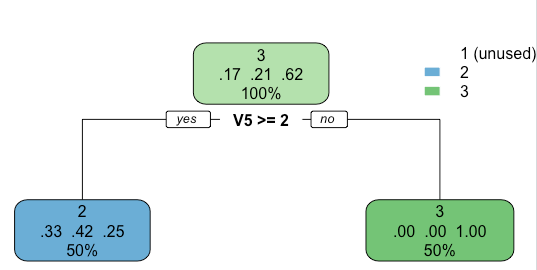
As per the attribute information given:

Independent variables are V2 to V5

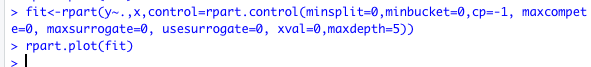
dependent/target variable is V6 with three classes.

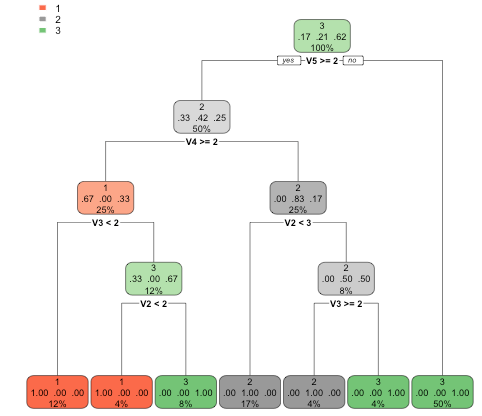


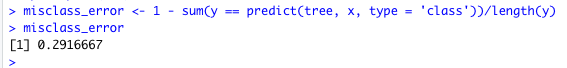


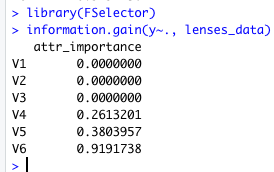


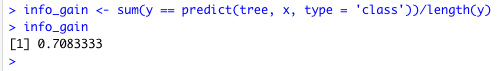
Or detailed view of tree









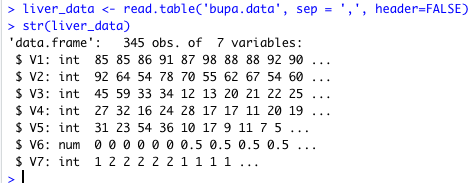


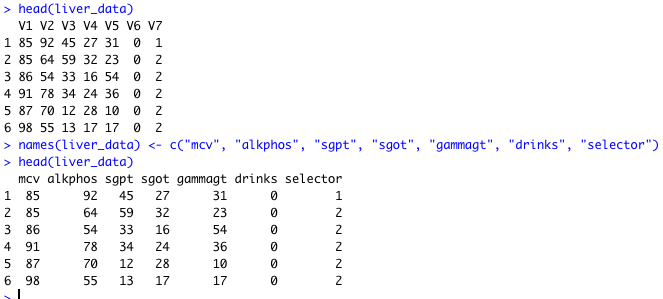
1. Fit 1, 2 and 3-nearest-neighbor classifiers to the Liver Disorders Data Set at

<http://archive.ics.uci.edu/ml/datasets/Liver+Disorders> for measures Euclidean and cosine.

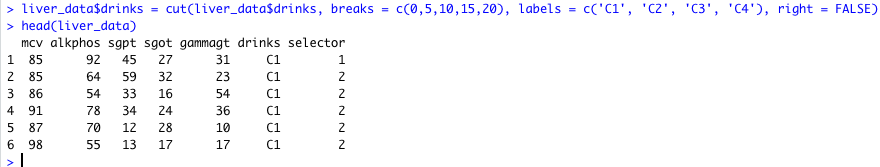
Last but one column is a decision attribute. Replace decision values into 4 classes (0<=c1<5, 5<=c2<10, 10<=c3<15, 15<=c4<=20). Last column is a data split column into training and test sets. 1 means the object is used for training. 2 means the object is used for testing. Explain the input parameters you provided for the classifier. Compute the misclassification error on the training data and also on the test data. Annotate your program. (10M)

Ans.

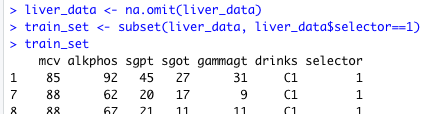


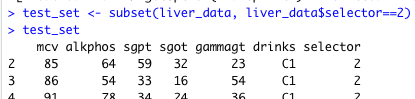
Assign column names according to data description given. 

Convert numeric vector drinks to intervals or factors as it is the decision(target) attribute using cut method.

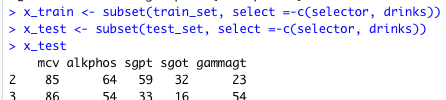


Based on selector values, split train and test set. Train set with selector value = 1 and test set with selector value = 2.

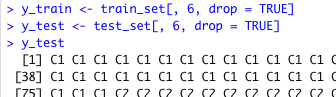




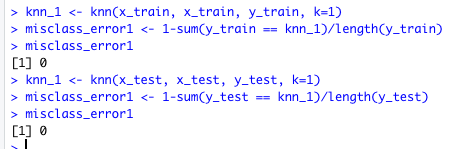
Input variables to the classifier for both train and test sets are mcv, alkphos, sgpt, sgot, and gammagt.



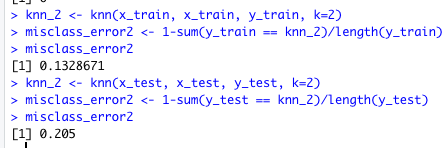
Dependent variable y is drinks, the 6th column.



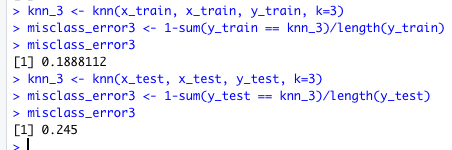
For k = 1, on train and test data



For k = 2



For k = 3



Attribute information is as follows:

1. mcv mean corpuscular volume

2. alkphos alkaline phosphotase

3. sgpt alanine aminotransferase

4. sgot aspartate aminotransferase

5. gammagt gamma-glutamyl transpeptidase

6. drinks number of half-pint equivalents of alcoholic beverages drunk per day

7. selector field created by the BUPA researchers to split the data into train/test sets

1. Use Support Vector machine for above problem. And compare the performance of both. Explain the input parameters you provided for the classifier. (10M)

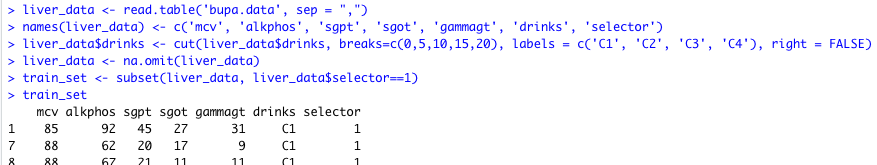
Ans.

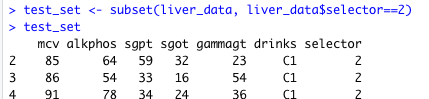
Load input data set.

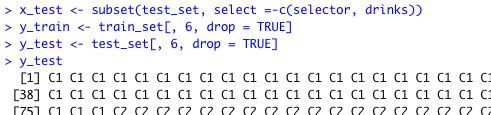
Assign column names as per the data description given.

Convert target attribute drinks to factors or categorical attribute.

Split into train and test set as above.



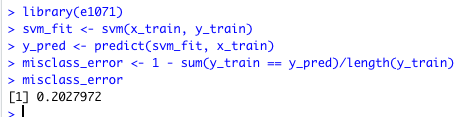


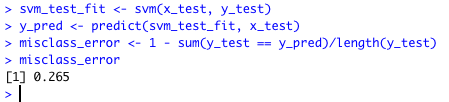


Install package e1071 for svm classifier.



Train set and test set to the classifier is same as the above data.



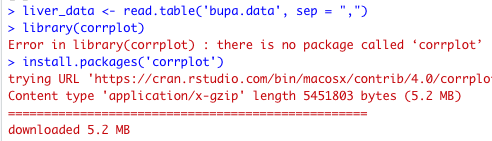


Misclassification error for the knn classifier is lower than the svm classifier. So, Knn is better here.

1. Create k-means clusters for k=4 for the Liver Disorders Data Set at <http://archive.ics.uci.edu/ml/datasets/Liver+Disorders> . Explain the input parameters you provided for the clustering algorithm. Plot the fitted cluster centers using a different color. Finally assign the cluster membership for the points to the nearest cluster center. Color the points according to their cluster membership. (10+10=20M)

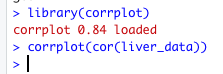
Ans.

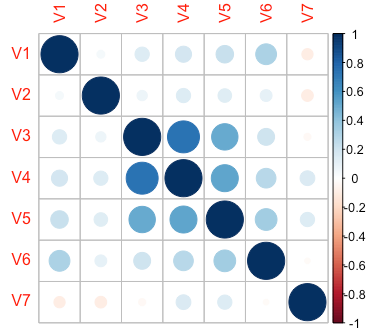
Read the dataset and install the corrplot package to view the correlation matrix.



Load the package into r session to use.

Find the correlation between variables and plot the result.



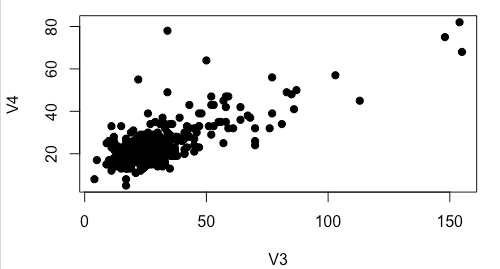


From the correlation plot, V3 and V4 are highly correlated, around 0.7

Select column V3 and V4 and store it in x.

Plot V3 and V4 distribution.



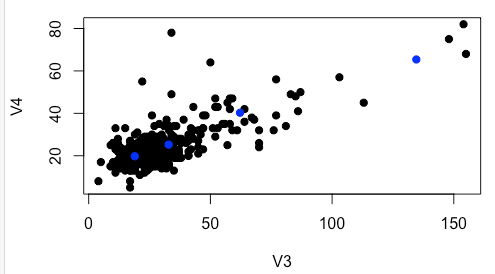


Kmeans algorithm has parameters x and 4. Data x consists of columns 3 and 4 of the liver data since it has the highest correlation as compared to other pairs of variables. 4 is the number of clusters k.

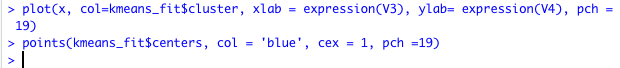


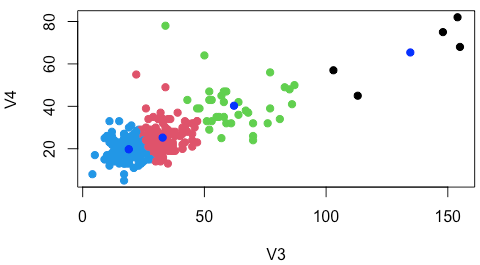
Plot the centroids for the data points.





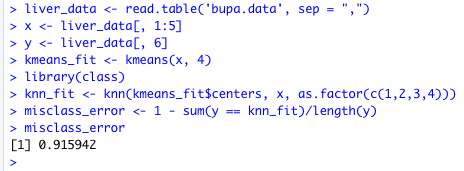
Assign cluster membership for the points to the nearest cluster centroid.





1. Compute the misclassification error that would result if you used your clustering rule to classify the data by assigning the majority class of the cluster. (10M)

Ans.



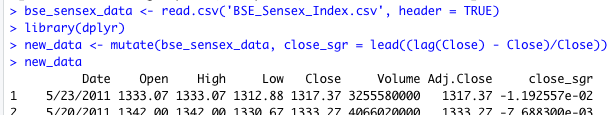
1. Consider the dataset BSE\_Sensex\_Index.csv. Create an extra column of successive growth rate for column close where the successive growth rate is defined as

(value of day x- value of day x-1)/value of day x-1. Use a z score cut off of 3 to identify any outliers. List the respective dates from the csv file on which day these outliers fall. (10M)

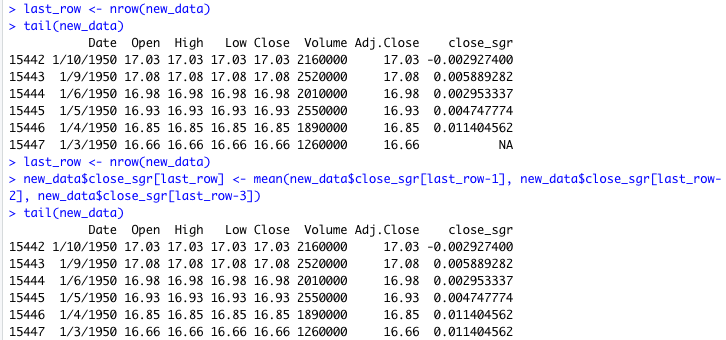
Ans.

Read the dataset.

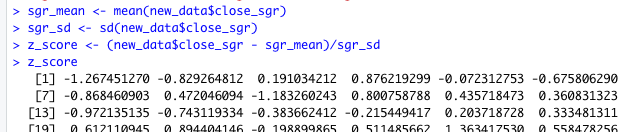
Create a new column named as close\_sgr.



The last row of close\_sgr is na. Replace na by mean of the values of the last three rows of close\_sgr column.



calculate the z-score.



Find the outlier dates which have z-score cutoff greater than or equal to 3.

