1. **A hospital is interested in learning about the risk of readmission for people who were treated for a specific chronic ailment. They want to know what contributes to an individual’s risk of readmission. Risk of readmission is based on the likelihood that the person will come back to the hospital within a short period of time. There are several data mining tasks that are relevant in this situation. Discuss the relevant ones and how the analysis may be of value to the hospital.**

**Solution:** The data mining tasks that are relevant in learning about the risk of readmission are:

* **Develop an Understanding of the purpose of Data Mining Project:** Business understanding is the most important preliminary task of the data mining process. It helps the hospital understand the business challenge better.
* **Obtain the Dataset to be used in analysis:** The data relevant for the analysis may be lying in different departments of the organization. Collection of relevant data files is very important for further analysis.
* **Explore, Clean and Preprocess the Data:** The obtained dataset may be in different formats and not all the data obtained will be required for the analysis. Hence, data cleansing, blending and formatting is necessary in the data preparation phase. Also, not all patients’ data will be required for analysis, such as one with null values can be eliminated.
* **Reduce the data dimension, if necessary:** Performing descriptive analytics to measure understand the data attributes, the outliers, errors and reduce overfitting will help in reducing data dimension.
* **Determine the Data Mining Task:** This is a transition which the business problem is translated into a data problem and hence choosing the right data mining method.
* **Partition the Data (for supervised tasks):** Data must be divided into training, validation and testing datasets.
* **Choose the data mining techniques and algorithms to be used:** There are lot of data mining techniques and algorithms, such as regression and clustering. Choosing the one that fits the dataset and the business problem is necessary to find the correlation.
* **Evaluate and deploy the Model:** Evaluating the model and making necessary before deployment will help the organization in achieving better results.

1. **A. Describe a situation where you want to keep outliers. Describe one where you don’t. Give an example of each.**

**Solution:**

**When to drop an outlier:** If the outlier is the reason for false results in the process, we should drop the outlier. For instance, a survey of High School Kids had an outlier where the weight of a person was entered as 15lbs. It may have been 105lbs, 151lbs, 115lbs or 215lbs. Since we are unable to rectify it, we should drop the entry.

**When to keep an outlier:** If the mistake is common enough for us to rectify by ourselves, we can keep the outlier. For example, In the same survey the city of a student was spelt as Tacoma. Since we know it is Tacoma, we should just rectify and keep it.

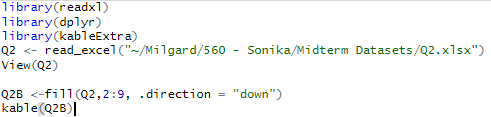
**B. Missing data can be problematic. There are several ways for handling missing data.**

**Given the following table define two *data imputation* methods which we have covered for filling in the missing values.**

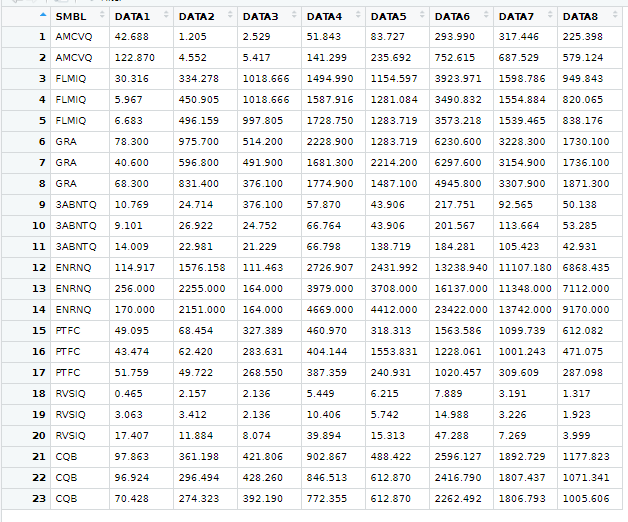
**NOTE: You will want to implement your methods and fill in the missing values in the table.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SMBL** | **DATA1** | **DATA2** | **DATA3** | **DATA4** | **DATA5** | **DATA6** | **DATA7** | **DATA8** |
| **AMCVQ** | **42.688** | **1.205** | **2.529** | **51.843** | **83.727** | **293.99** | **317.446** | **225.398** |
| **AMCVQ** | **122.87** | **4.552** | **5.417** | **141.299** | **235.692** | **752.615** | **687.529** | **579.124** |
| **FLMIQ** | **30.316** | **334.278** | **1018.666** | **1494.99** | **1154.597** | **3923.971** | **1598.786** | **949.843** |
| **FLMIQ** | **5.967** | **450.905** |  | **1587.916** | **1281.084** | **3490.832** | **1554.884** | **820.065** |
| **FLMIQ** | **6.683** | **496.159** | **997.805** | **1728.75** | **1283.719** | **3573.218** | **1539.465** | **838.176** |
| **GRA** | **78.3** | **975.7** | **514.2** | **2228.9** |  | **6230.6** | **3228.3** | **1730.1** |
| **GRA** | **40.6** | **596.8** | **491.9** | **1681.3** | **2214.2** | **6297.6** | **3154.9** | **1736.1** |
| **GRA** | **68.3** | **831.4** | **376.1** | **1774.9** | **1487.1** | **4945.8** | **3307.9** | **1871.3** |
| **3ABNTQ** | **10.769** | **24.714** |  | **57.87** | **43.906** | **217.751** | **92.565** | **50.138** |
| **3ABNTQ** | **9.101** | **26.922** | **24.752** | **66.764** |  | **201.567** | **113.664** | **53.285** |
| **3ABNTQ** | **14.009** | **22.981** | **21.229** | **66.798** | **138.719** | **184.281** | **105.423** | **42.931** |
| **ENRNQ** | **114.917** | **1576.158** | **111.463** | **2726.907** | **2431.992** | **13238.94** | **11107.18** | **6868.435** |
| **ENRNQ** | **256** | **2255** | **164** | **3979** | **3708** | **16137** | **11348** | **7112** |
| **ENRNQ** | **170** | **2151** |  | **4669** | **4412** | **23422** | **13742** | **9170** |
| **PTFC** | **49.095** | **68.454** | **327.389** | **460.97** | **318.313** | **1563.586** | **1099.739** | **612.082** |
| **PTFC** | **43.474** | **62.42** | **283.631** | **404.144** | **1553.831** | **1228.061** | **1001.243** | **471.075** |
| **PTFC** | **51.759** | **49.722** | **268.55** | **387.359** | **240.931** | **1020.457** | **309.609** | **287.098** |
| **RVSIQ** | **0.465** | **2.157** | **2.136** | **5.449** | **6.215** | **7.889** | **3.191** | **1.317** |
| **RVSIQ** | **3.063** | **3.412** |  | **10.406** | **5.742** | **14.988** | **3.226** | **1.923** |
| **RVSIQ** | **17.407** | **11.884** | **8.074** | **39.894** | **15.313** | **47.288** | **7.269** | **3.999** |
| **CQB** | **97.863** | **361.198** | **421.806** | **902.867** | **488.422** | **2596.127** | **1892.729** | **1177.823** |
| **CQB** | **96.924** | **296.494** | **428.26** | **846.513** | **612.87** | **2416.79** | **1807.437** | **1071.341** |
| **CQB** | **70.428** | **274.323** | **392.19** | **772.355** |  | **2262.492** | **1806.793** | **1005.606** |

**Solution:** The two common methods of data imputation are Imputation of Data using Mean-Median Values or by filling in the missing values by copying the value of variable adjacent to it. Adjacent Values can be obtained from the use of r code as follows:

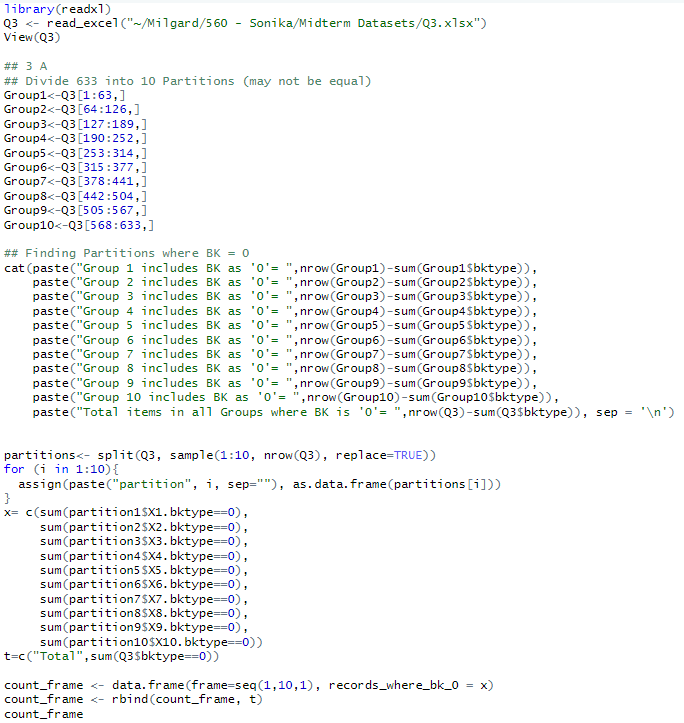


**Output:**

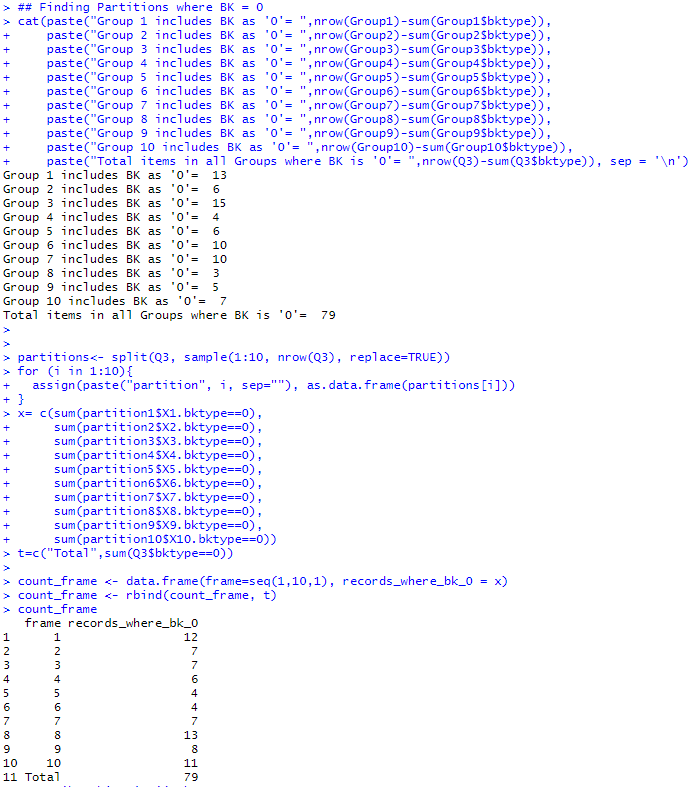


1. **Using the data set “dataforquestion3” included in the Exam module –**
2. **Split the data into 10 equal (roughly) partitions. That means that you will have 633/10 records in each partition. Determine the number of records with a value of BK = 0 in each partition – how many a value of BK = 0 in each partition.**

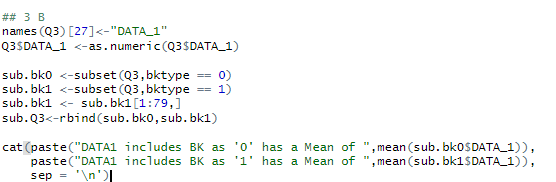
**Solution:**



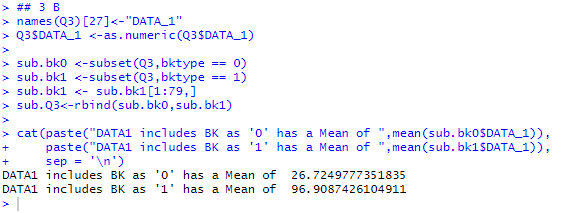
**Output:**



1. **Generate a subset of the data set with all the records that have a value of BK = 0 (should be 79) and an equal number of BK=1 (randomly selected from the remaining records, BK=1). Calculate the mean for DATA1 for each type BK=0, BK=1.**

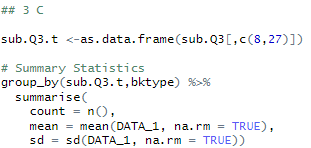
**Solution:** 

**Output:**

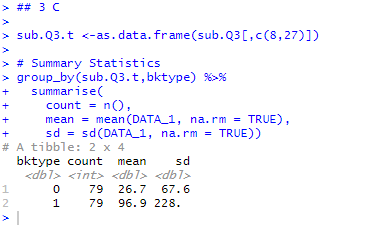


1. **Conduct a t test for difference in the means of the two groups in the subset. Put your results here.**

**Solution:**



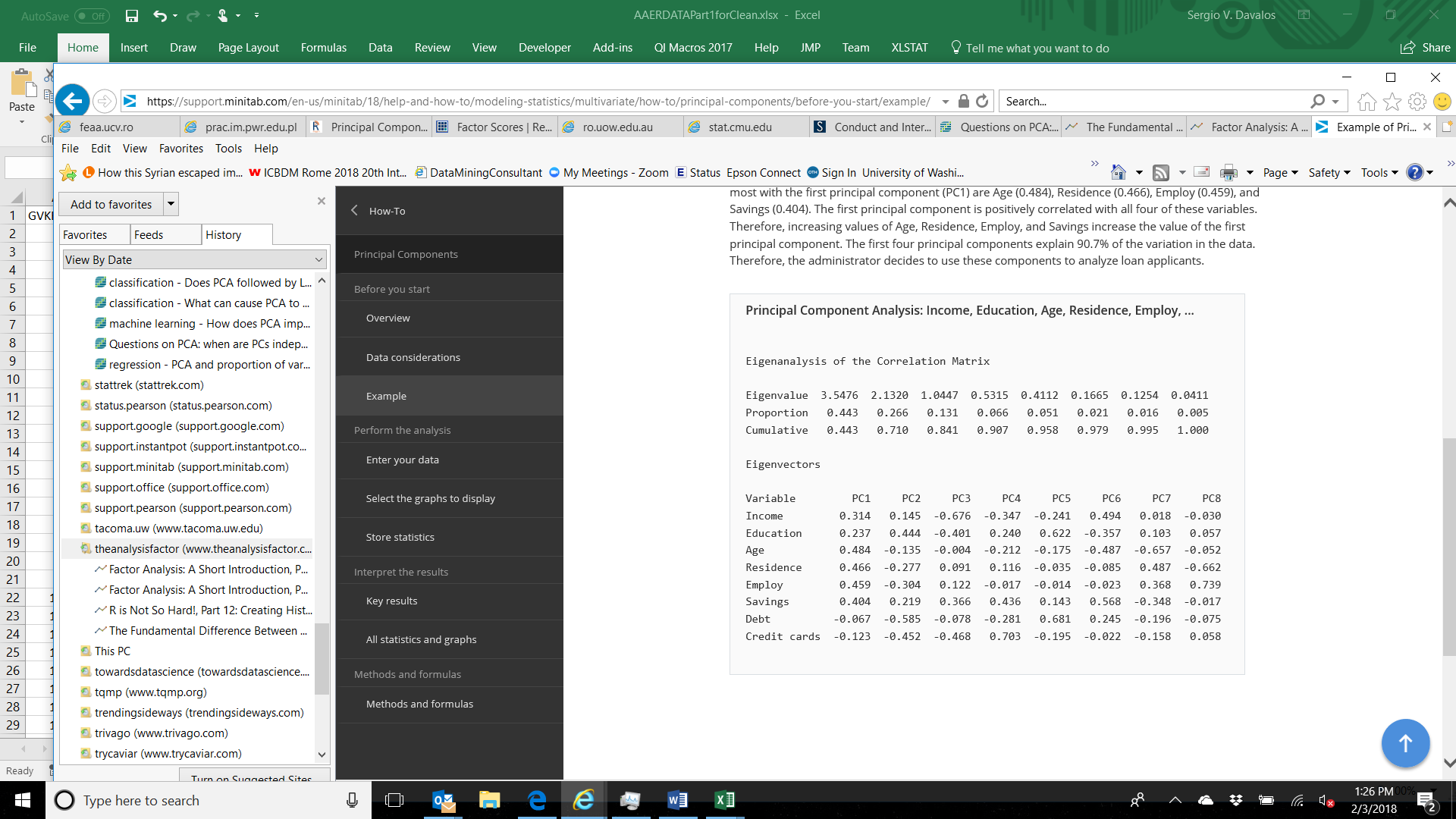
**Output:**



1. **Below is the output from a PCA.**
2. **How many principal components should we use? Explain.**

**Solution:** Eigen values that are greater than 1 are highly significant. Hence, I would choose the first three principal components. Using 3 components, a model can be built without overfitting. I will consider choosing another one if my model turn out to be less accurate.

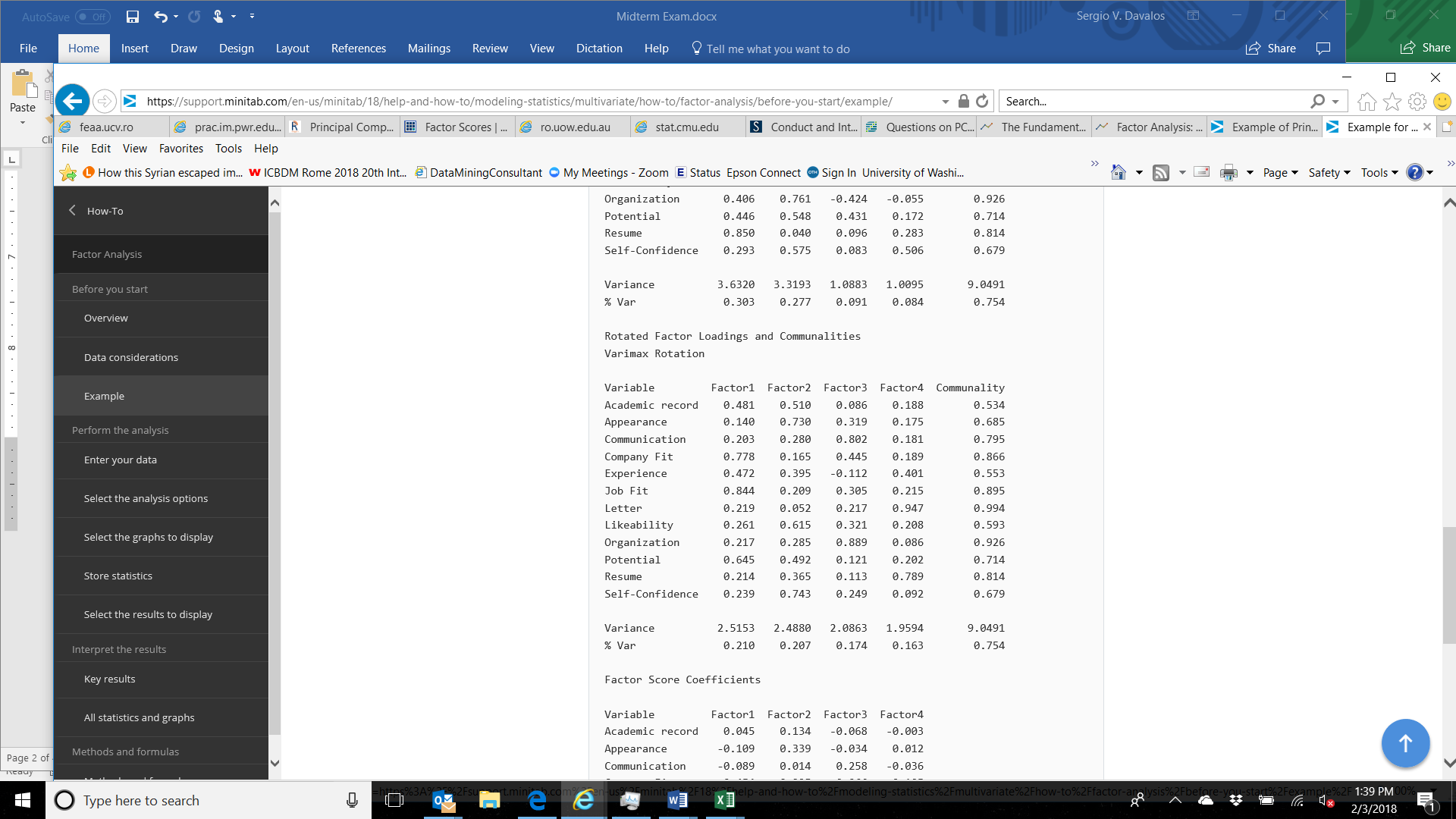
1. **Discuss the weights for the first four PCs – PC1 to PC4. How do we interpret them?**

****

**Solution:**

* **PC1:** This represents the group of people who have positive correlation with income, education, age, savings, employment and residence which indicates that they are financially stable.
* **PC2:** This group of people show negative correlation with age, employment, debt, credit cards and residence. They are slightly positively correlated with income, education and savings, which could be teenagers still studying.
* **PC3:** Although this set of people are positively correlated with residence, employment and savings, they are negatively correlated with income, education, age, debt and credit cards which could mean that these are high school students with part-time jobs who earn but do not have a channel for credit.
* **PC4:** This group of people are negatively correlated with income, age, debt and employment, which could mean that they are retired. Adding to this is the positive correlation with savings and credit cards.

1. **Using the Factor loading table below, identify the latent factors and give them a label. Justify your labels.**

****

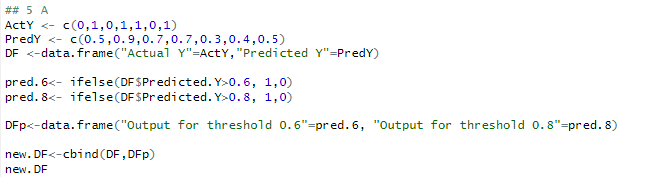
**Solution:**

* **High Potential factor:** High scores in potential, job fit, and company fit. Good scores in experience and academic record. Low score in personality.
* **Confidence Factor:** High scores in appearance, likeability and self-confidence. Good scores in academic record. Low score in technical skills.
* **Communication Factor:** High scores in communication, being organized. Some improvements required. Low scores in job experience.
* **Resume Factor:** High scores in resume and recommendation. Good score in job experience. Low scores on technical skills and personality.

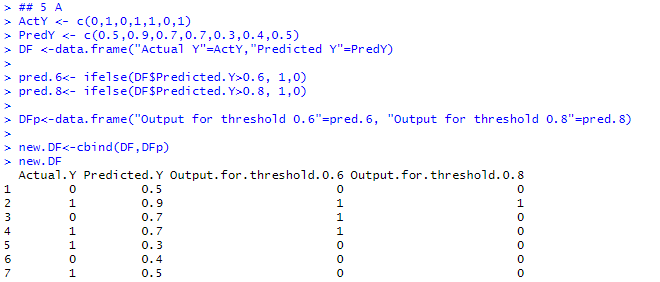
1. **A. Given the table below calculate the confusion matrix for cutoff of 0.6 and 0.8.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Actual Y** | **Predicted y** | **Output for threshold 0.6** | **Output for threshold 0.8** |
| **0** | **0.5** |  |  |
| **1** | **0.9** |  |  |
| **0** | **0.7** |  |  |
| **1** | **0.7** |  |  |
| **1** | **0.3** |  |  |
| **0** | **0.4** |  |  |
| **1** | **0.5** |  |  |

**Solution:**

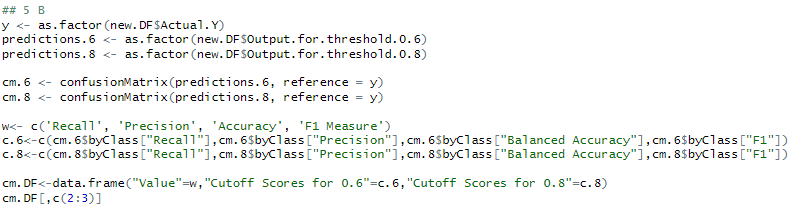


**Output:**

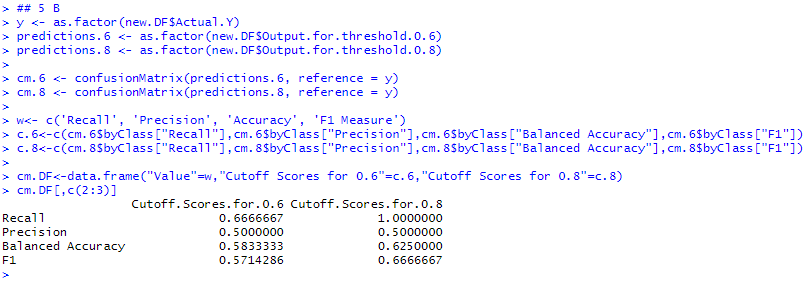


**B. Provide the recall, precision, accuracy, and F1 Measure values for both 0.6 and 0.8.**

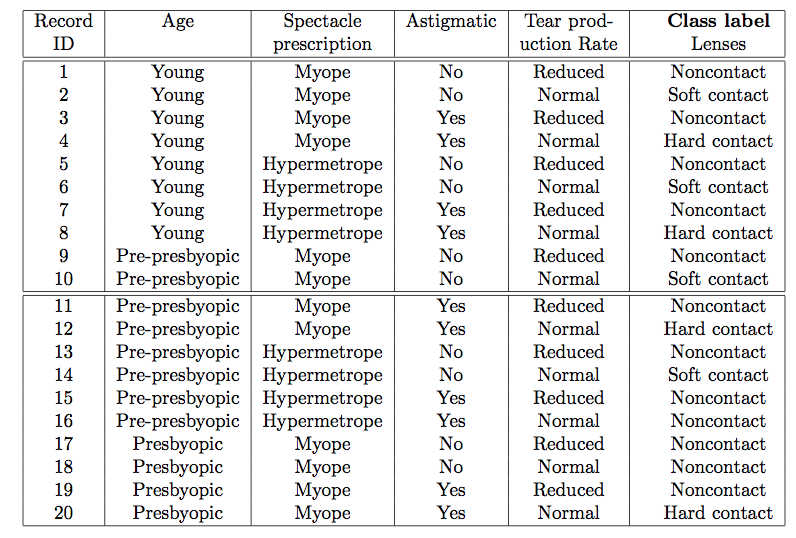
**Solution:**



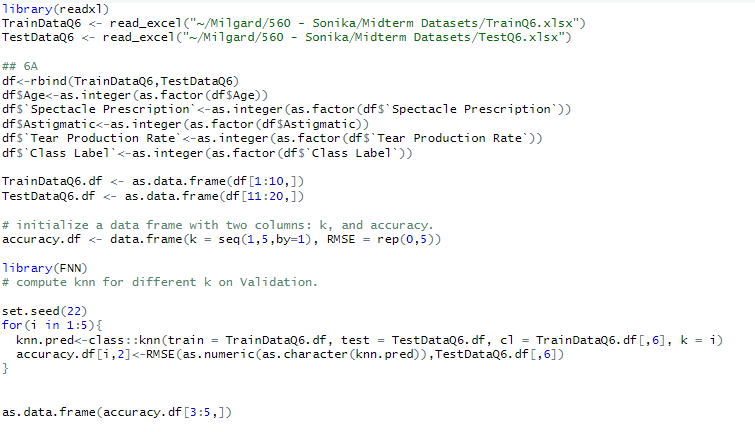
**Output:**



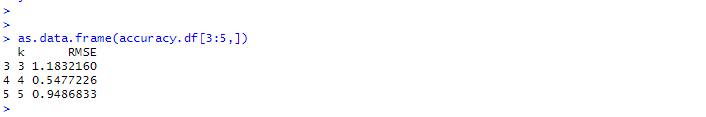
1. **A. The following table below contains 10 records for the training data and 10 for the test. Using KNN identify the test data classes. Determine the K from possible values of 3, 4, and 5 that will give you the best accuracy for the test cases. Note: you will want to convert the variable values to integer to perform KNN.**

****

**Solution:**



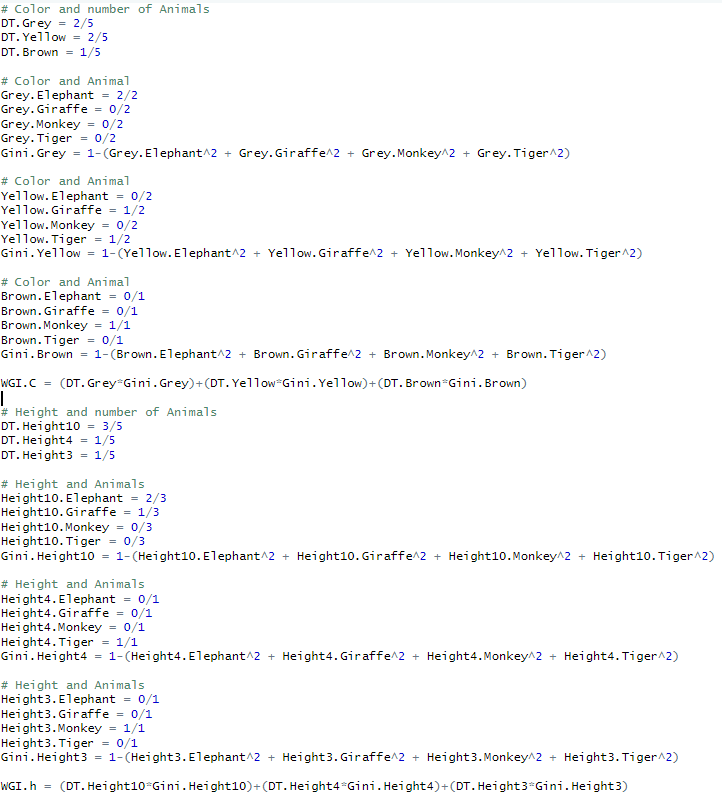
**Output:**

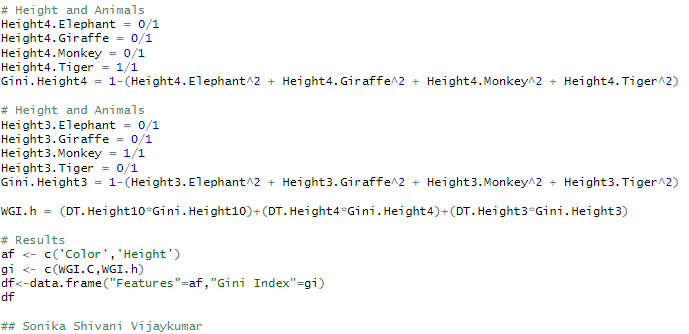


K from 4 will give the best accuracy for the test cases as it has the best fit among the 3 values.

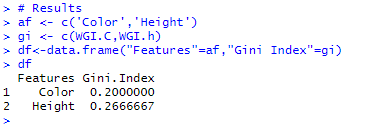
1. **Given the following table construct the decision tree manually using the Gini impurity measure:**

|  |  |  |
| --- | --- | --- |
| **Color** | **Height** | **Label** |
| **Grey** | **10** | **Elephant** |
| **Yellow** | **10** | **Giraffe** |
| **Brown** | **3** | **Monkey** |
| **Grey** | **10** | **Elephant** |
| **Yellow** | **4** | **Tiger** |

**Solution:** 



**Output:**



1. **Say you have 1000 fruits which could be either ‘banana’, ‘orange’ or ‘other’. These are the 3 possible classes of the Y variable.**

**We have data for the following X variables, all of which are binary (1 or 0).**

* **Long**
* **Sweet**
* **Yellow**

**The following table is an example of the data:**

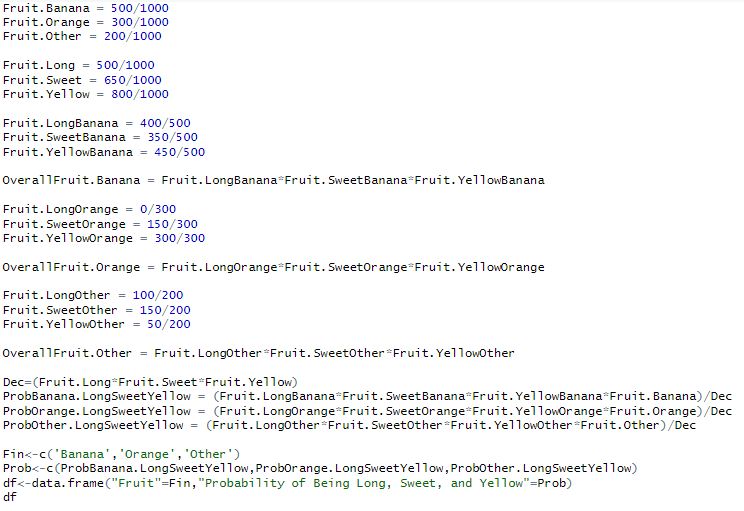
| **Fruit** | **Long (x1)** | **Sweet (x2)** | **Yellow (x3)** |
| --- | --- | --- | --- |
| **Orange** | **0** | **1** | **0** |
| **Banana** | **1** | **0** | **1** |
| **Banana** | **1** | **1** | **1** |
| **Other** | **1** | **1** | **0** |

**For computing the probabilities the training data is aggregated to form a counts table like this.**

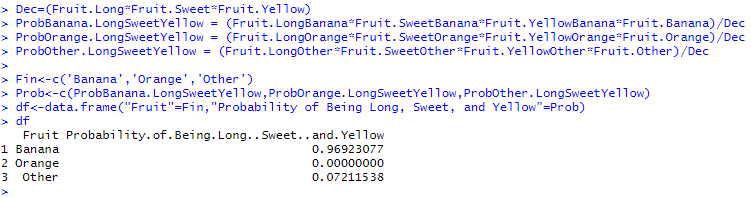
**[](https://www.machinelearningplus.com/wp-content/uploads/2018/11/05_Naive_bayes_example_new.png)**

1. **If a fruit is long, sweet, and yellow what is it? Show your work.**

**Solution:** Banana is the long, sweet and yellow fruit. From the above info, we can derive it using the Naïve Bayes Method



**Output:**



1. **Using a data set that relates stopping distance and speed, the following is the result of conducting linear regression on the results:**

**Provide an interpretation of the results. If I have a speed of 12 miles per hour, what will my stopping distance be?**

**Residuals:**

**Min 1Q Median 3Q Max**

**-29.069 -9.525 -2.272 9.215 43.201**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) -17.5791 6.7584 -2.601 0.0123 \***

**speed 3.9324 0.4155 9.464 1.49e-12 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 15.38 on 48 degrees of freedom**

**Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438**

**F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12**

**Solution:**

* **R-Squared:** It is a measure of how close the data are to the fitted regression line, also known as the coefficient of determination. The above table shows a 65% R-squared, which means that a 65% variation in distance is required to stop.
* **F-Statistic:** F-statistic shows the variation between sample means/ within the sample. A value more than one is highly significant. The above model shows a value of 89.57.

Y = m\*X + C formula can be used to determine the stopping distance.

Y = 3.93\*12 + (-17.57) = 29.6 ft

Note: Y = stopping distance, X = Speed, m = coefficient, C = intercept.

* **Coefficients:** These are the estimates of correlation between the predictor and independent variables. Coefficient have various determining factors.

1. Estimates: For every mile per hour increase in speed, the stopping distance will increase by 3.93 ft provided the car takes -17.57 ft to stop.

* **Residual Standard Error:** This indicates whether the model is suitable for fitting. Values near to 0 represents that the model is a better fit. The current model value is 15.38.
* **Residuals:** Residuals represent the difference between actual and predicted values. The above equation shows significant difference in actual and predicted.

1. **Which variable selection methods are most likely computationally intensive – they will longer run than the other methods due to the amount of computation required or the amount of memory required? Explain.**

**Solution:** Variable selection methods that are computationally intensive are:

1. **RIDGE regression and LASSO regression** – combination of iterative and simple filter methods.
2. **KNN and Decision trees** – KNN finds the nearest neighbor to match and generate an outcome. Decision trees such as classification, boosted, pruned are intensive as they split the data multiple times.
3. **Forward Selection and Backward Elimination** - Generation of series of repetitive outcomes.