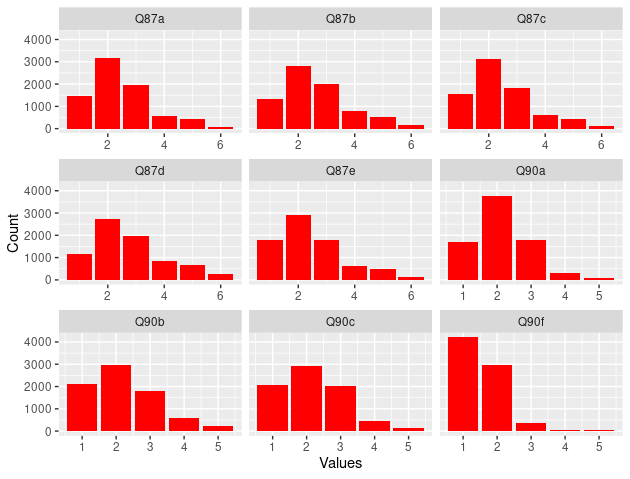
# Task 1 ( European Working Conditions Survey)

I am using a dataset that contains the response of people regarding their work satisfaction rate.

## Data Visualization

The dataset contains multiple answers to the questions to see the satisfaction rate of the employee at their jobs.

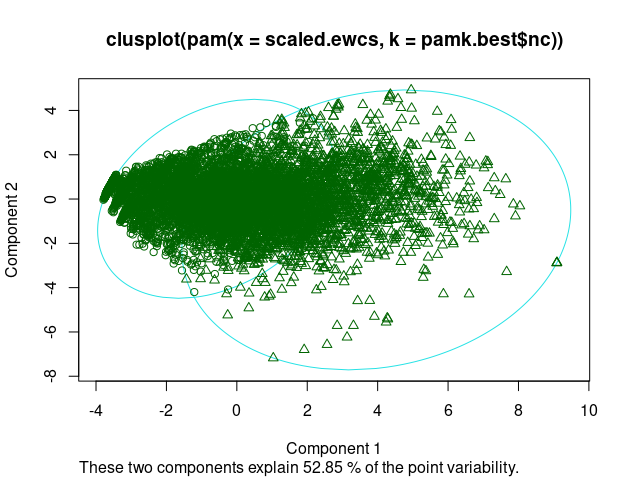


Above plot depicts that most people have chosen option 2 as their answer to most of the question. However, for last question we find more answer for options 1 and both of the options are perfectly postive but there is some ratio of people with negative or neutral responses. But, overall satisfaction rate can be considered postive.

## Cluster Analysis

In this analysis I have used different techniques to analyze my data. Cluster analysis is done to map the data and extract meaningful information like patters. In this method, we usually try to assign the observations to different clusters to better understand people with similar attributes.

To find optimal number of clusters or groups I have used PAM method.

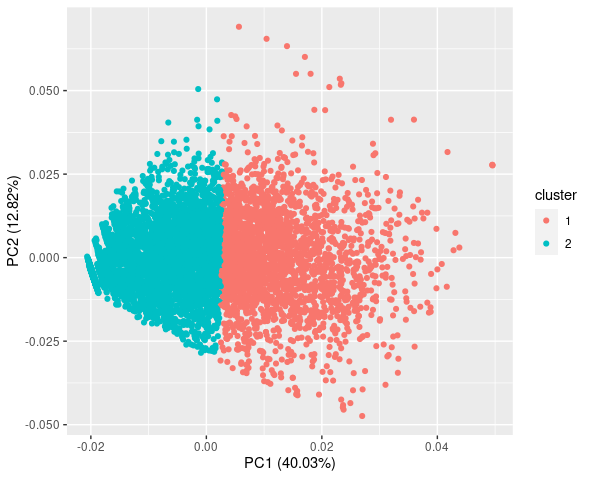


Optimal number of clusters according to PAM method is 2.

Observation grouping per cluster is given below:

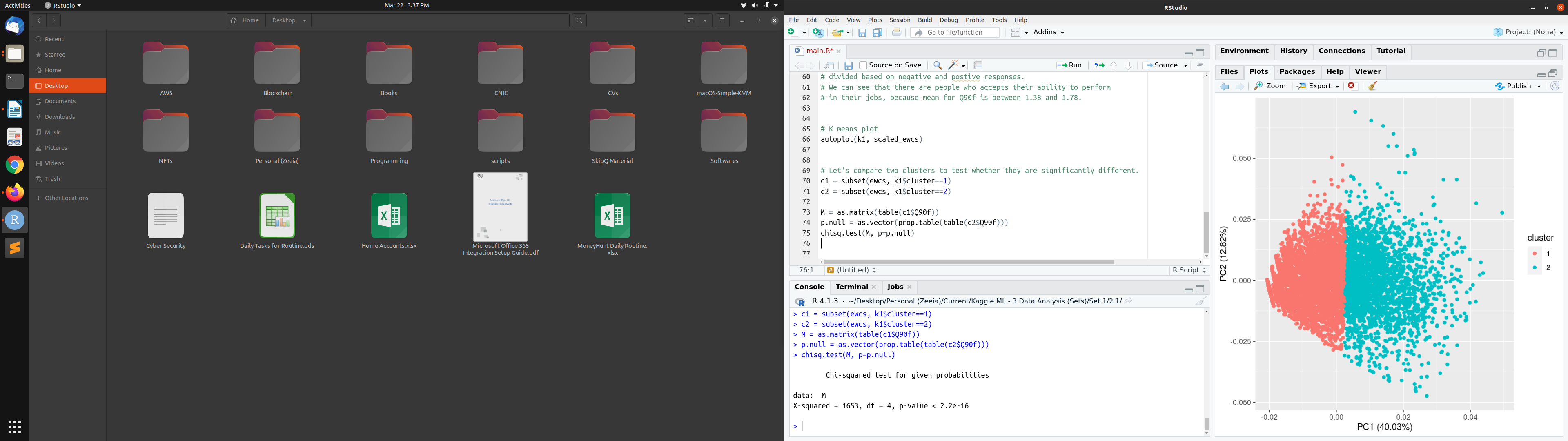
* Cluster 1 Observations: 4869
* Cluster 2 Observations: 2778

Let’s plot observations of cluster 1 and cluster 2 to see it visually.



The above plot shows data distribution for cluster 1 and cluster 2. We see that data is beautifully separated between two clusters and no significant overlap among points can be observed.

The variable Q90f was very similar between in both the clusters which was about one being good at one’s job. However, let’s make a hypothesis test to see if our intuition is correct.



We can see that p-value is << 0.05 which is our significant level. Hence, Null hypothesis is rejected and our intuition that both groups despite their responses to other answers they agreed to being good at their job.

## Summary

This section was all about cluster analysis the data and understand what information it gives us. Cluster analysis is done by grouping the datasets into multiple groups where an individual group represents similar type of employees who answered the question in the similar way. The analysis can be used to further understand the behavior of current or upcoming employee. Here we have two groups. One of them is postive about the job and workplace satisfaction while other group was either neutral or negative at some point.

# Task 2

In this task, we have two school grading datasets. Both were merge to single datasets. The final datasets has 32 features and a target variable. Here, I will using multiple machine learning techniques to train our model on our data and better predict the table without using highly correlated features.

I used Lasso Regression method to find best features and used them to train and test our models.

Details of the models are given below:

## Linear Regression

**Best Features**

Call:

lm(formula = G3 ~ ., data = train)

Coefficients:

(Intercept) sex age higher

11.08240 0.04061 -0.34073 3.07396

**All Features**

Call:

lm(formula = G3 ~ ., data = train)

Coefficients:

(Intercept) school sex age address famsize Pstatus Medu Fedu

11.563583 -0.419307 -0.021308 -0.100449 0.428099 0.151424 -0.055078 0.227011 0.068218

Mjob Fjob reason guardian traveltime studytime failures schoolsup famsup

-0.039088 0.261772 0.063964 -0.009086 -0.090311 0.302305 -1.612483 -1.459076 0.094238

paid activities nursery higher internet romantic famrel freetime goout

-1.079390 0.225797 -0.289931 1.555411 0.396624 -0.544465 0.205881 -0.020497 -0.206257

Dalc Walc health absences

-0.060089 -0.042841 -0.127025 0.014006

The above table shows the summary of our linear regression model with best features that we found using lasso regression and second analysis is done using all the features.

## Random Forest

**Best Features**

Call:

randomForest(formula = G3 ~ ., data = train)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 1

Mean of squared residuals: 14.32706

% Var explained: 6.05

**All Features**

Call:

randomForest(formula = G3 ~ ., data = train)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 10

Mean of squared residuals: 11.30038

% Var explained: 25.89

Random Forest model is using 500 subtrees with 1 variable split at each node using best feature analysis and 10 variable split at each node using all features analysis. We are having difference of 4% in the mean residual which is not too great but as far as variance is concerned, model with all the features is able to explain 26% of the total variance which is great improvement over model 1.

## Support Vector Machine

**Best Features**

Call:

svm(formula = G3 ~ ., data = train, method = "anova")

Parameters:

SVM-Type: eps-regression

SVM-Kernel: radial

cost: 1

gamma: 0.3333333

epsilon: 0.1

Number of Support Vectors: 690

**All Features**

Call:

svm(formula = G3 ~ ., data = train, method = "anova")

Parameters:

SVM-Type: eps-regression

SVM-Kernel: radial

cost: 1

gamma: 0.03333333

epsilon: 0.1

Number of Support Vectors: 642

Above SVM is using radial kernel, 0.03/0.3 as gamma, 1 as cost for both models and 0.1 as epsilon for both models. Also, we see minor change of approx 30 more support vectors.

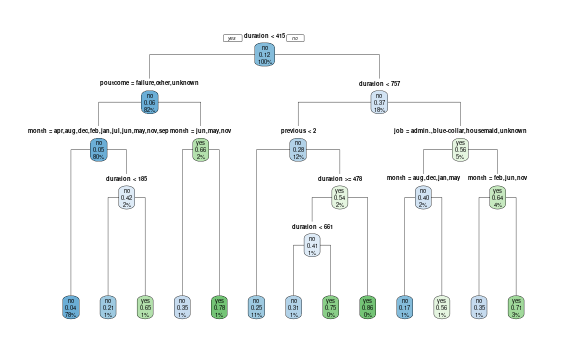
## Summary

The models are performing worst on the best features extracted using Lasso Regression. All three models have giving R2 around 0.02 which is very bad performance. Hence, we say that using best feature sex, age and higher does not make a difference in the performance of the model. As far as the all feature models are concerned, svm performed with 0.3 R2, random forest performed with 0.36 R2 and logistic regression performed with 0.24 R2. From this analysis we can say that using all the features, random forest model outperforms all the other models.

# Task 3

In this problem, I am training a model to predict term deposit for customers. For this classification problem I will be using Decision Tree, Random Forest, and Logistic Regression.

## Decision Tree

The decision tree clearly shows each split variable, the decisions are on edges and leaf nodes represents final output.

## Random Forest

I trained a random forest and its configuration is as follows:

* It uses mtry as 3, 500 number of trees.
* 3 variables at each split.
* OOB error rate is 10.0% which is good.
* Our error rate is approx 0.

## Logistic Regression

Logistic Regression is an extension of linear regression except that we use a threshold to make it binary classifier. The model is also performing well on all the features however, the model did specify some features as important and some of them are joblue-collar, maritalmarried, contactunknown and more. We can prove this from the models’ summary given below:

Call:

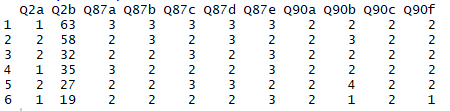
* glm(formula = y ~ ., family = "binomial", data = train\_dataset)
* Deviance Residuals:
* Min 1Q Median 3Q Max
* -2.9943 -0.3717 -0.2382 -0.1499 3.0721
* Coefficients:
* Estimate Std. Error z value Pr(>|z|)
* (Intercept) -1.804e+00 7.299e-01 -2.472 0.013428 \*
* age -4.764e-03 8.739e-03 -0.545 0.585606
* jobblue-collar -6.865e-01 3.018e-01 -2.275 0.022905 \*
* jobentrepreneur -1.821e-01 4.363e-01 -0.417 0.676402
* jobhousemaid -3.449e-01 4.887e-01 -0.706 0.480380
* jobmanagement -1.529e-01 2.917e-01 -0.524 0.600268
* jobretired 7.269e-01 3.767e-01 1.930 0.053634 .
* jobself-employed 1.350e-01 4.024e-01 0.336 0.737195
* jobservices -2.594e-01 3.409e-01 -0.761 0.446753
* jobstudent 3.273e-01 4.890e-01 0.669 0.503291
* jobtechnician -1.669e-01 2.774e-01 -0.602 0.547388
* jobunemployed -9.854e-01 5.459e-01 -1.805 0.071070 .
* jobunknown -5.945e-01 9.158e-01 -0.649 0.516247
* maritalmarried -4.550e-01 2.108e-01 -2.158 0.030915 \*
* maritalsingle -3.094e-01 2.457e-01 -1.259 0.207995
* educationsecondary -2.362e-01 2.442e-01 -0.967 0.333335
* educationtertiary -1.340e-02 2.801e-01 -0.048 0.961861
* educationunknown -1.048e+00 4.901e-01 -2.138 0.032550 \*
* defaultyes 5.319e-01 5.239e-01 1.015 0.309983
* balance -1.701e-05 2.003e-05 -0.849 0.395841
* housingyes -3.336e-01 1.699e-01 -1.963 0.049623 \*
* loanyes -6.111e-01 2.428e-01 -2.517 0.011852 \*
* contacttelephone -2.511e-01 3.092e-01 -0.812 0.416737
* contactunknown -1.252e+00 2.726e-01 -4.593 4.36e-06 \*\*\*
* day 1.782e-02 9.963e-03 1.788 0.073731 .
* monthaug -2.340e-01 2.961e-01 -0.790 0.429249
* monthdec -4.041e-01 8.987e-01 -0.450 0.652937
* monthfeb -2.046e-02 3.669e-01 -0.056 0.955537
* monthjan -1.431e+00 5.011e-01 -2.856 0.004291 \*\*
* monthjul -1.048e+00 3.090e-01 -3.393 0.000692 \*\*\*
* monthjun 2.743e-01 3.706e-01 0.740 0.459089
* monthmar 1.703e+00 4.618e-01 3.687 0.000227 \*\*\*
* monthmay -6.084e-01 2.850e-01 -2.135 0.032743 \*
* monthnov -1.174e+00 3.365e-01 -3.490 0.000484 \*\*\*
* monthoct 1.669e+00 3.984e-01 4.188 2.81e-05 \*\*\*
* monthsep 6.182e-01 4.973e-01 1.243 0.213819
* duration 4.456e-03 2.478e-04 17.980 < 2e-16 \*\*\*
* campaign -6.490e-02 3.390e-02 -1.915 0.055549 .
* pdays -5.803e-04 1.193e-03 -0.486 0.626665
* previous -3.353e-02 4.882e-02 -0.687 0.492191
* poutcomeother 5.222e-01 3.352e-01 1.558 0.119260
* poutcomesuccess 2.528e+00 3.533e-01 7.156 8.34e-13 \*\*\*
* poutcomeunknown -4.116e-01 3.851e-01 -1.069 0.285117
* ---
* Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
* (Dispersion parameter for binomial family taken to be 1)
* Null deviance: 2274.9 on 3163 degrees of freedom
* Residual deviance: 1470.1 on 3121 degrees of freedom
* AIC: 1556.1
* Number of Fisher Scoring iterations: 6

Above table shows the trained configuration of my logistic regression. We can find several features which are statistically significant.

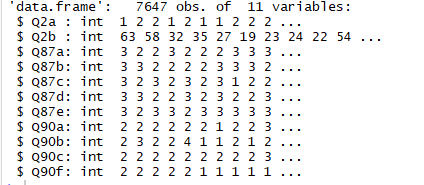
The logistic regression gave us 90% accuracy, decision tree gave us 89% accuracy, random forest gave us 90% accuracy. Here, after testing all the models, we can conclude that random forest and logistic regression best fit the data with approx 90% accuracy.

Appendix

### Data Tables



The above table shows the top 6 data points from the survey responses. Each variable has its own meaning defined in the data description file.



The above table shows each feature, its data type, and some samples. In total, we have 11 features.

**Decision Tree Mathematical Tree**

n= 3164

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 3164 368 no (0.88369153 0.11630847)

2) duration< 414.5 2601 161 no (0.93810073 0.06189927)

4) poutcome=failure,other,unknown 2539 120 no (0.95273730 0.04726270)

8) month=apr,aug,dec,feb,jan,jul,jun,may,nov,sep 2467 90 no (0.96351844 0.03648156) \*

9) month=mar,oct 72 30 no (0.58333333 0.41666667)

18) duration< 185 38 8 no (0.78947368 0.21052632) \*

19) duration>=185 34 12 yes (0.35294118 0.64705882) \*

5) poutcome=success 62 21 yes (0.33870968 0.66129032)

10) month=jun,may,nov 17 6 no (0.64705882 0.35294118) \*

11) month=apr,aug,dec,feb,jan,jul,mar,oct,sep 45 10 yes (0.22222222 0.77777778) \*

3) duration>=414.5 563 207 no (0.63232682 0.36767318)

6) duration< 756.5 393 112 no (0.71501272 0.28498728)

12) previous< 1.5 345 86 no (0.75072464 0.24927536) \*

13) previous>=1.5 48 22 yes (0.45833333 0.54166667)

26) duration>=478 34 14 no (0.58823529 0.41176471)

52) duration< 661 26 8 no (0.69230769 0.30769231) \*

53) duration>=661 8 2 yes (0.25000000 0.75000000) \*

27) duration< 478 14 2 yes (0.14285714 0.85714286) \*

7) duration>=756.5 170 75 yes (0.44117647 0.55882353)

14) job=admin.,blue-collar,housemaid,unknown 57 23 no (0.59649123 0.40350877)

28) month=aug,dec,jan,may 23 4 no (0.82608696 0.17391304) \*

29) month=apr,feb,jul,jun,nov,oct 34 15 yes (0.44117647 0.55882353) \*

15) job=entrepreneur,management,retired,self-employed,services,student,technician,unemployed 113 41 yes (0.36283186 0.63716814)

30) month=feb,jun,nov 23 8 no (0.65217391 0.34782609) \*

31) month=apr,aug,dec,jan,jul,mar,may,oct 90 26 yes (0.28888889 0.71111111) \*