STAT 542 / CS 598: Project 2

Fall 2019, by Prathamesh(satpute3), Vivek(vivekg3) and Athul(as81)

Due: Monday, Dec 16 by 11:59 PM Pacific Time

[10 Points, half a page] Project description and summary. This part should summerise your goal, your approach, and your results.

half page description goes here

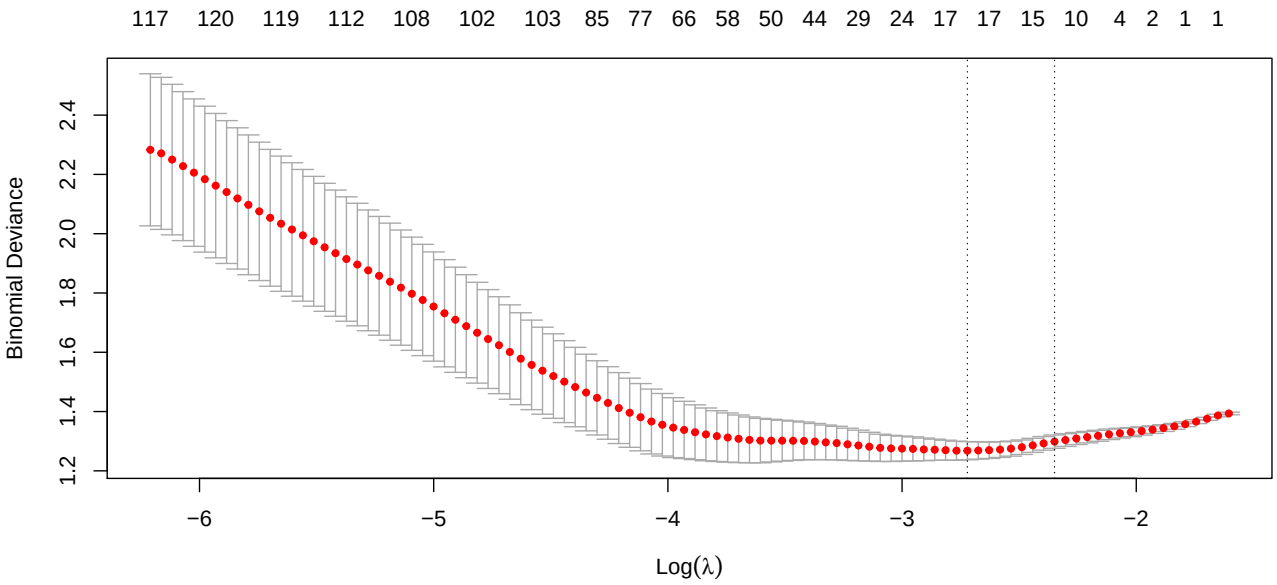
[5 Points, half a page] Data processing for Question 1. Describe how you process the data so that it can be analyzed to answer question 1.

half page description goes here

[30 Points, within 5 pages] Classification models based on pixels.

```
  # kable(results, caption = "KNN")
```

```
# use parallel for performace
  registerDoMC(cores = 4)
  cv_glmnet_model <- cv.glmnet(train.data[, -1], train.data[, 1], parallel = TRUE, alpha=1, family="binomi
```




```
  best_lambda = cv_glmnet_model$lambda.1se
  # training with best lambda selected from the cv
  train_glmnet_model <- glmnet(train.data[, -1], train.data[, 1], lambda = best_lambda, alpha=1, family="b

  summary(train_glmnet_model)
```

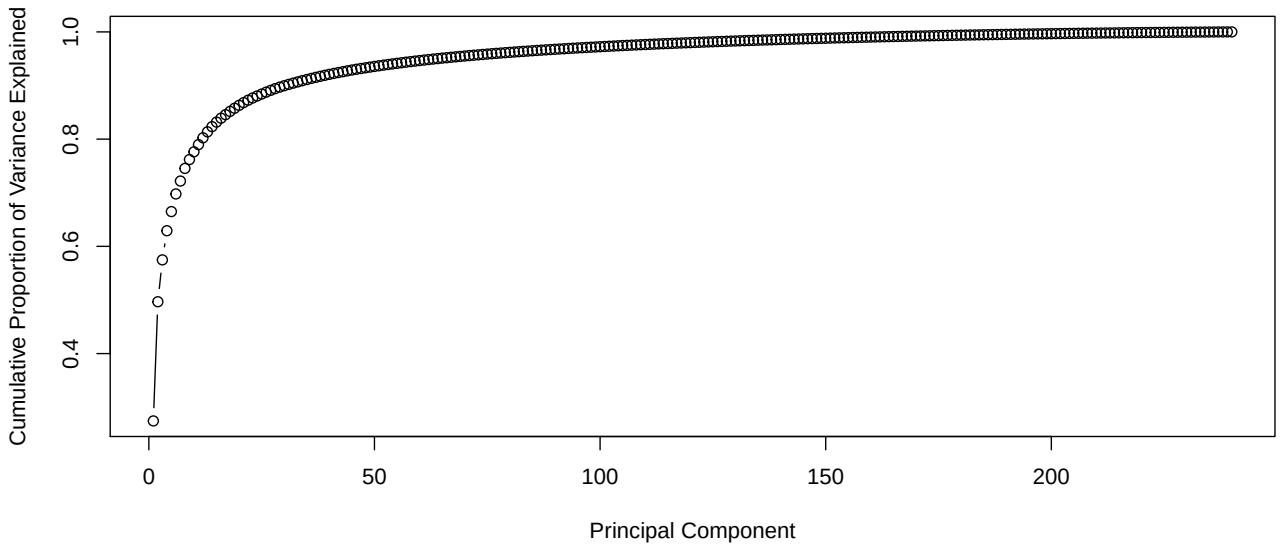
Table 1: Lasso Regression Results

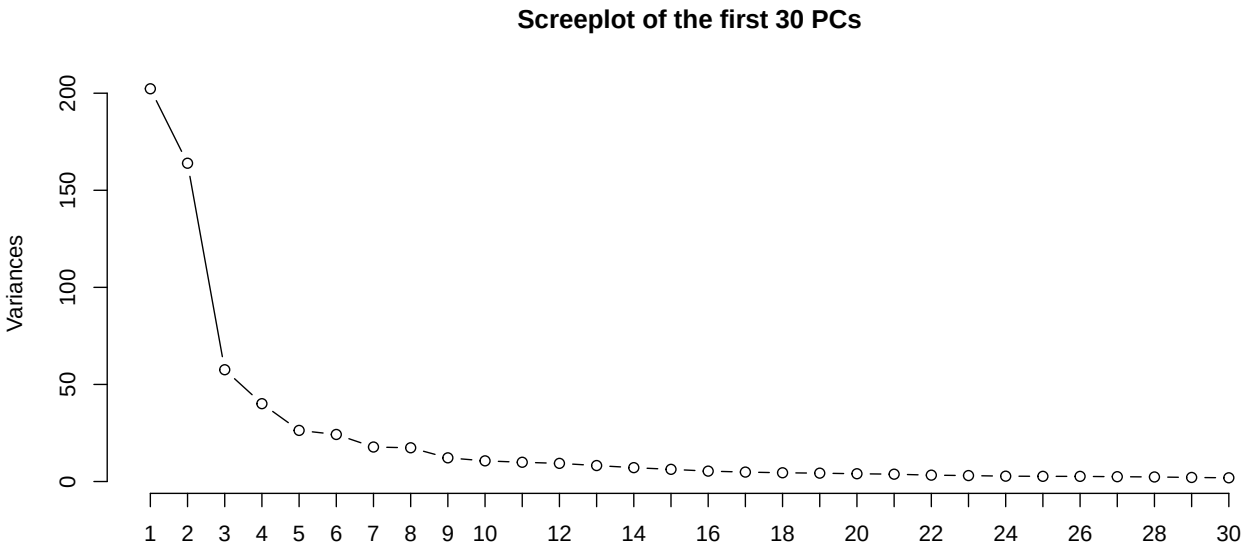
Best.lambda	Accuracy
0.0954646	0.6333333

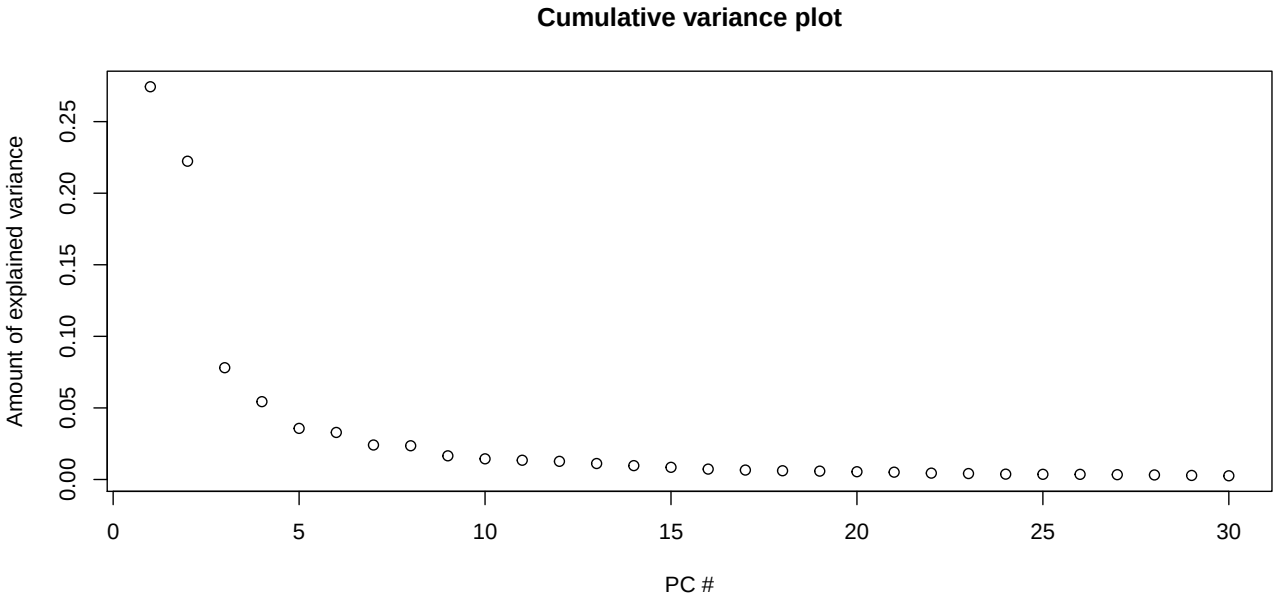
```
##            Length Class     Mode
## a0              1 -none-    numeric
## beta        30000 dgCMatrix S4
## df              1 -none-    numeric
## dim             2 -none-    numeric
## lambda          1 -none-    numeric
## dev.ratio       1 -none-    numeric
## nulldev         1 -none-    numeric
## npasses         1 -none-    numeric
## jerr            1 -none-    numeric
## offset          1 -none-    logical
## classnames      2 -none-    character
## call            6 -none-    call
## nobs            1 -none-    numeric
```

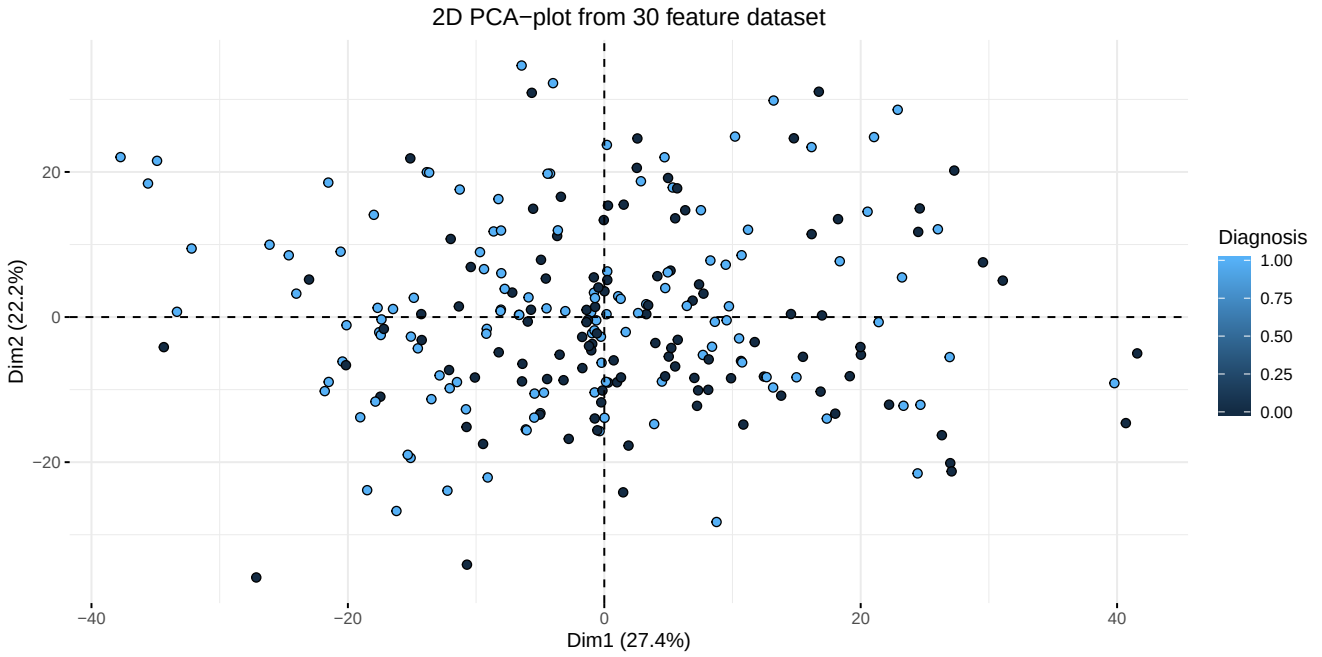
```
pred <- predict(train_glmnet_model, s = best_lambda, newx = test.data[, -1], type = "class")
accuracy = mean(test.data[, 1] == pred)

results = data.frame("Best lambda" = best_lambda, "Accuracy" = accuracy)
```

```
kable(results, caption = "Lasso Regression Results")
```




Screeplot of the first 30 PCs
Variances
0
50
100
150
200
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30


Cumulative variance plot
Amount of explained variance
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
20
25
30
PC #



[10 Points, 1 page] Literature review. You should search and read existing literature and summarize clinically relevant characteristics that could be used for skin cancer image diagnosis. There is no limitation on what type of literature you could use. However, the goal should be motivating your feature engineering approaches from a clinical and analytic point of view. Please give appropriate citations to the literature you read.

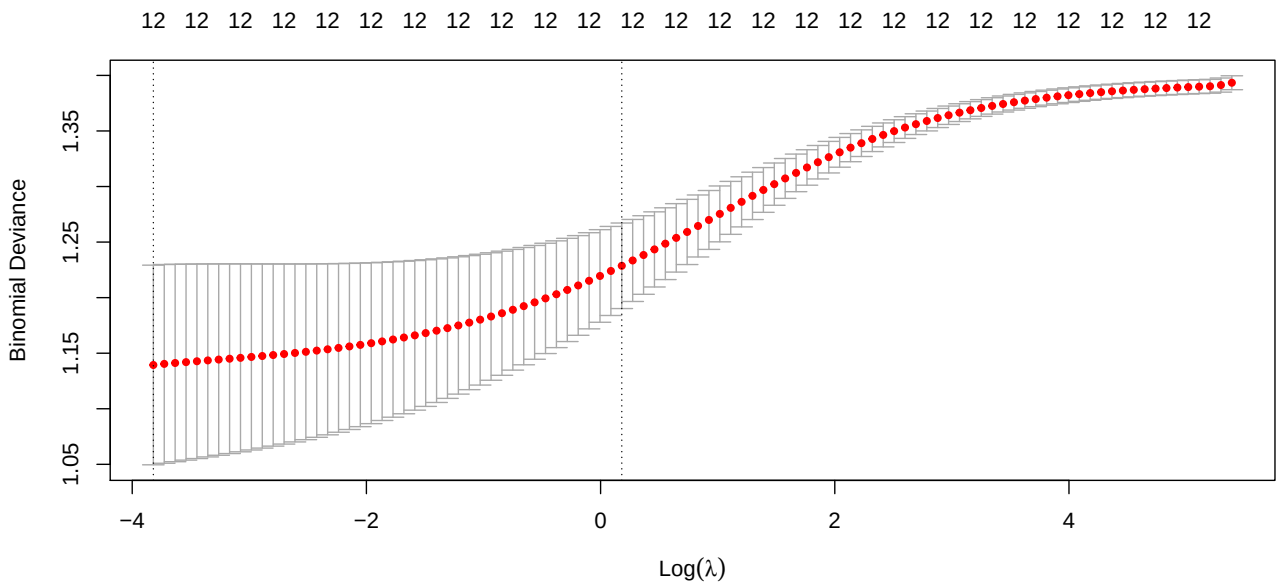
1 page description goes here

[10 Points, 1 page] Feature engineering. Motivated by what you have read (or your understanding), process the data in a reasonable way such that the new variables are more intuitive to your collaborator/clinicians. You need to describe clearly what is your data processing criteria and how your variables are calculated.

1 page description goes here

[20 Points, 2 page] Classification models based on new features. Fit two different classification models to identify malignant moles. You can either use the ones from Question 1 or use some new models if you believe they may perform better on the new features. Same requirements of Question 1 apply to this part. Besides, you should focus more on variable selection and interpretation.

```
# use parallel for performace
  registerDoMC(cores = 4)

  # Ridge Regression
  cv_glmnet_model <- cv.glmnet(train.data[, -1], train.data[, 1], parallel = TRUE, alpha=0, family="binomi
```




```
best_lambda = cv_glmnet_model$lambda.1se
# training with best lambda selected from the cv
train_glmnet_model <- glmnet(train.data[, -1], train.data[, 1], lambda = best_lambda, alpha=0, family="b

summary(train_glmnet_model)
```

```
##            Length Class     Mode
## a0          1     -none-    numeric
## beta       12     dgCMatrix S4
## df          1     -none-    numeric
## dim         2     -none-    numeric
## lambda      1     -none-    numeric
## dev.ratio   1     -none-    numeric
## nulldev     1     -none-    numeric
## npasses     1     -none-    numeric
## jerr        1     -none-    numeric
## offset      1     -none-    logical
## classnames  2     -none-    character
## call        6     -none-    call
## nobs        1     -none-    numeric
```

```
pred <- predict(train_glmnet_model, s = best_lambda, newx = test.data[, -1], type = "class")
accuracy = mean(test.data[, 1] == pred)

results = data.frame("Best lambda" = best_lambda, "Accuracy" = accuracy)
```

```
kable(results, caption = "Ridge Regression Results")
```

Table 2: Ridge Regression Results

Best.lambda	Accuracy
1.198349	0.7