EAS506 HW3
Matthew Sah
msah #72

1)

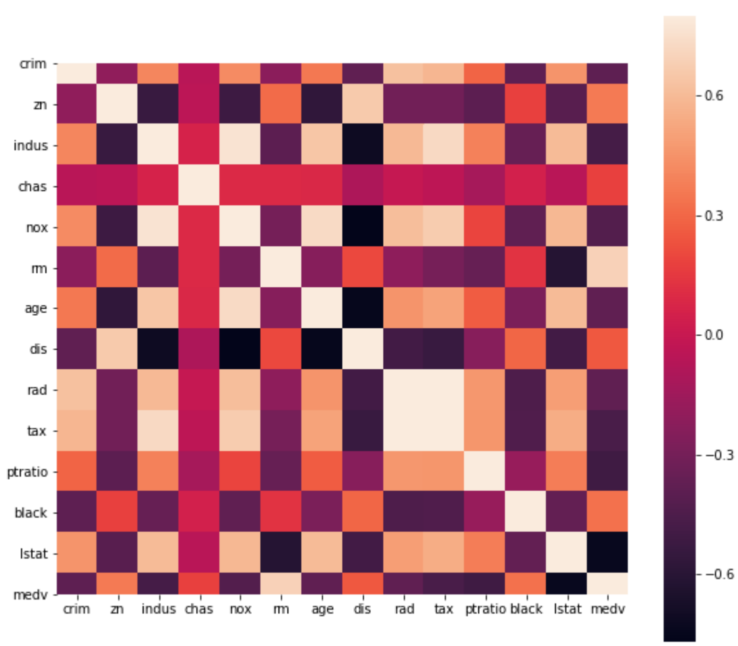
For data preparing, I created a whole new column that will either be 1 or 0 depending on whether the crime rate at an area is greater than 0.5, this is to make verification easier in later stages. Out of all my results, LDA has recieved the highest test accuracy with KNN as the second.

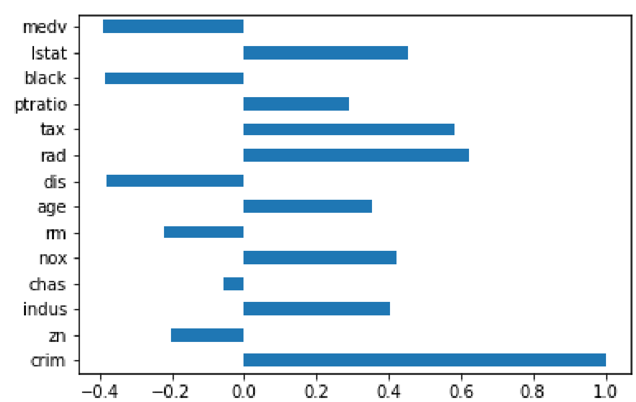
	crim	greater_half
10	0.22489	0
11	0.11747	0
12	0.09378	0
13	0.62976	1
14	0.63796	1

```
Accuracy of logistic regression classifier on test set: 0.89
Accuracy of LDA on test set: 0.91
KNN training error :  0.9483292079207921
KNN testing error :  0.9056372549019608
```

Going through a heatmap and looking at the correlation of variables, i removed the three variables with lowers correlation to achieve significantly higher accuracy. The three variables are rn, zn, and chas. LDA still has higher accuracy but difference was not significant.

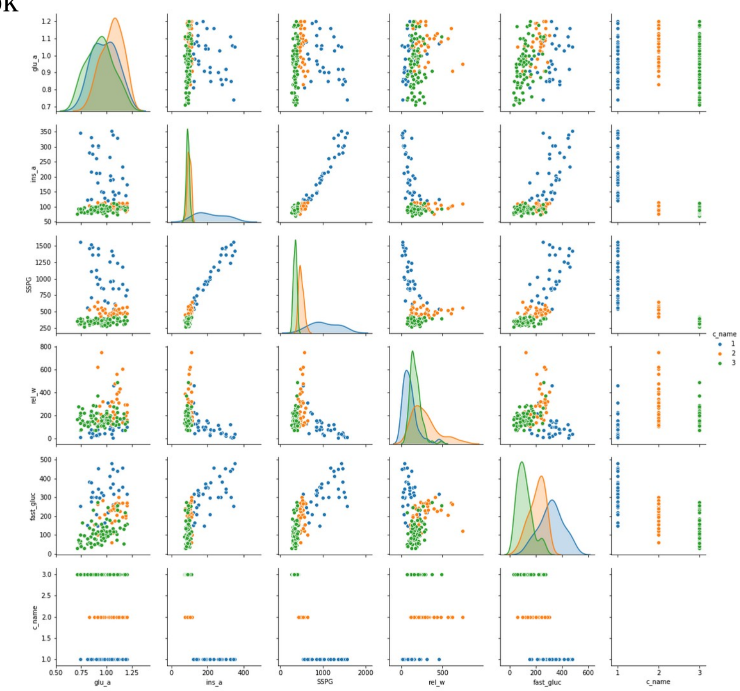
```
Accuracy of logistic regression classifier on test set: 0.92
Accuracy of LDA on test set: 0.96
testing error :  0.9375
```





2)

a)Looking at insulin area and SSPG, the charts look almost identical through out the whole area while being verified with different variables, the area distribution of classes is almost identical.



b) QDA has slightly higher accuracy than LDA

LDA

	precision	recall	f1-score	support
1	1.000	0.962	0.980	26
2	0.931	0.964	0.947	28
3	0.984	0.984	0.984	62
accuracy			0.974	116
macro avg	0.972	0.970	0.971	116
weighted avg	0.975	0.974	0.974	116

QDA

	precision	recall	f1-score	support
1	1.000	0.857	0.923	7
2	0.727	1.000	0.842	8
3	1.000	0.857	0.923	14
accuracy			0.897	29
macro avg	0.909	0.905	0.896	29
weighted avg	0.925	0.897	0.901	29

c)

LDA : 3

QDA : 1

3)

a)

Q3

a) $\log \dfrac{\Pr(G=1 \mid X=x)}{\Pr(G=k \mid X=x)} = \beta_{10} + \beta_1^T x.$

$\log \dfrac{\Pr(G=2 \mid X=x)}{\Pr(G=k \mid X=x)} = \beta_{20} + \beta_2^T x.$

to generalize $\log \dfrac{\Pr(G=k-1 \mid X=x)}{\Pr(G=k \mid X=x)} = \beta_{(k-1)0} + \beta_{k-1}^T x$

therefore

$$\Pr(G=k \mid X=x) = \frac{1}{1+\sum_{1}^{k-1} \exp(D_0 + \beta_l^T x)}$$

3)
b)

Q3.

b)

$$1-P(X) = \frac{1-\exp(B_0+B_1X)}{1+\exp(B_0+B_1X)}$$

$$= \frac{1}{1+\exp(B_0+B_1X)}.$$

$$\frac{1}{1-P(X)} = 1+\exp(B_0+B_1X)$$

$$P(x) \times \frac{1}{1-P(x)} = \frac{\exp(B_0+B_1X)}{1+\exp(B_0+B_1X)} \; (1+\exp(B_0+B_1X))$$

$$= \frac{P(x)}{1-P(x)} = e^{B_0+B_1X} = \exp(B_0+B_1X)$$

4)
a)

```
x1 :  15.247909406047418
x2 :  9.051634889078853
x3 :  11.673970471331907
x4 :  9.952449740354645
```

b)
I received the least error on the second model X2, “$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$”.
where the second least error occured with X4. Originally i thought that either X or X^4 would have the least error.
Firstly for X, i originally thought that in the scenario of X with good accuracy itself would suffice, adding additional edits to the model would make the mode more inaccurate.
And for X^4 i thought that if X was off target for all values by a bit, X^4 could possibly make the minor edits to bring the model closer to accuracy.
However it is also fair to say X^2 has the lowest error. Since each value in Y was generated based off of X^2.

c)

Yes, since based on the coefficients estimates, we can see that X^2 is the one that ends of statistically stagnant.