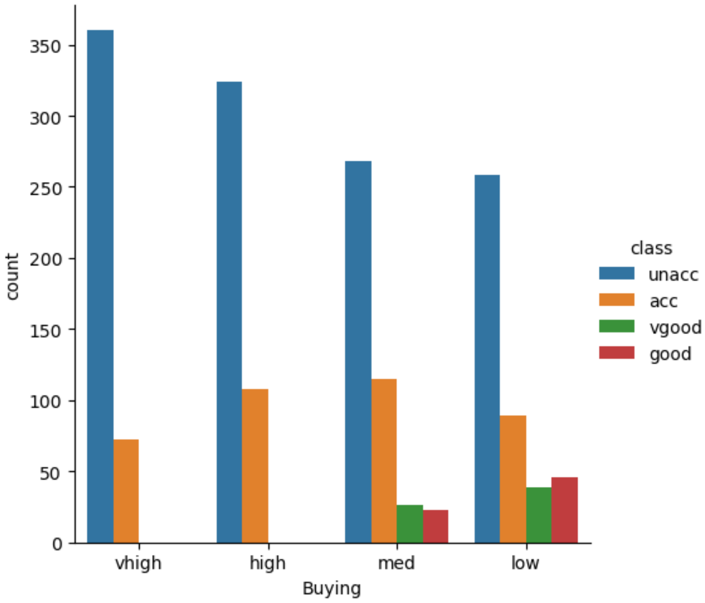
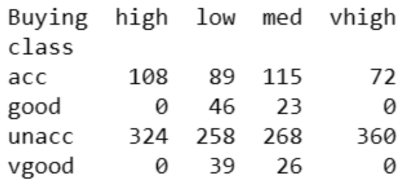
Comp615 – Lab 11

Jessica Wong – 14877422

Ling Bin - 21152215

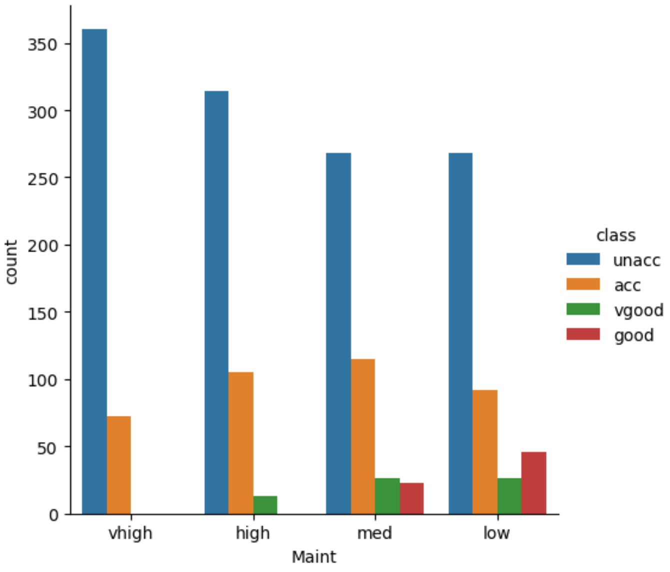
Q1: Perform a similar ‘breakdown’ analysis for the rest of the features. Provide the plots and explain your findings

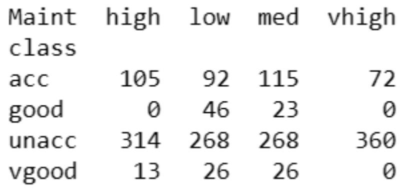




```
Buying  high  low  med  vhigh
class
acc      108   89  115     72
good       0   46   23      0
unacc    324  258  268    360
vgood      0   39   26      0
```

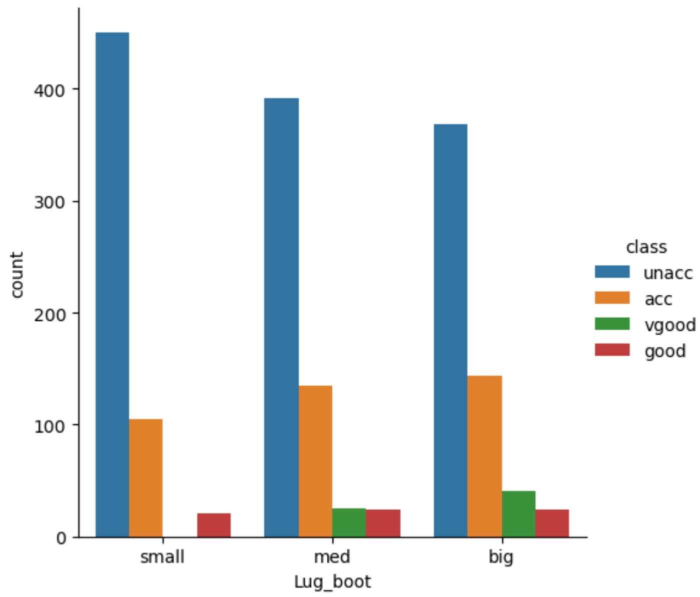
The buying feature shows that cars with a higher buying price are more likely to be in the 'unacc' class. There is no occurrence of 'good' or 'vgood' classes for any buying price category.

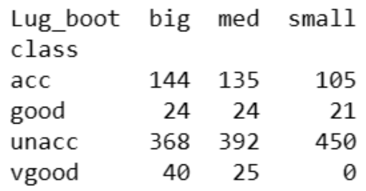




```
Maint  high  low  med  vhigh
class
acc     105   92  115     72
good      0   46   23      0
unacc   314  268  268    360
vgood    13   26   26      0
```

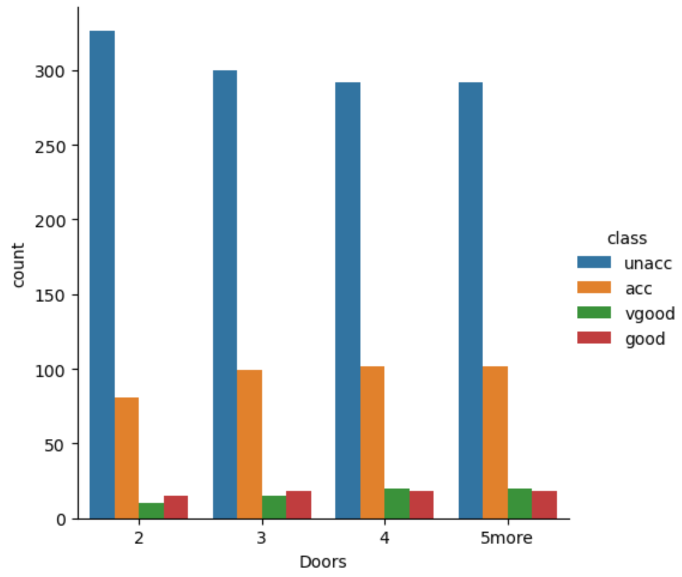
Cars with higher maintenance costs are more likely to be in the 'unacc' class. There is no occurrence of 'good' or 'vgood' classes for any maintenance cost category.

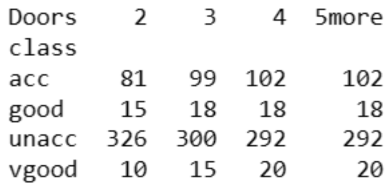




```
Lug_boot  big  med  small
class
acc       144  135    105
good       24   24     21
unacc     368  392    450
vgood      40   25      0
```

The size of the luggage boot doesn't have a strong association with the car class. However, cars with a 'big' size luggage boot tend to have a higher likelihood of being in the 'acc' class.



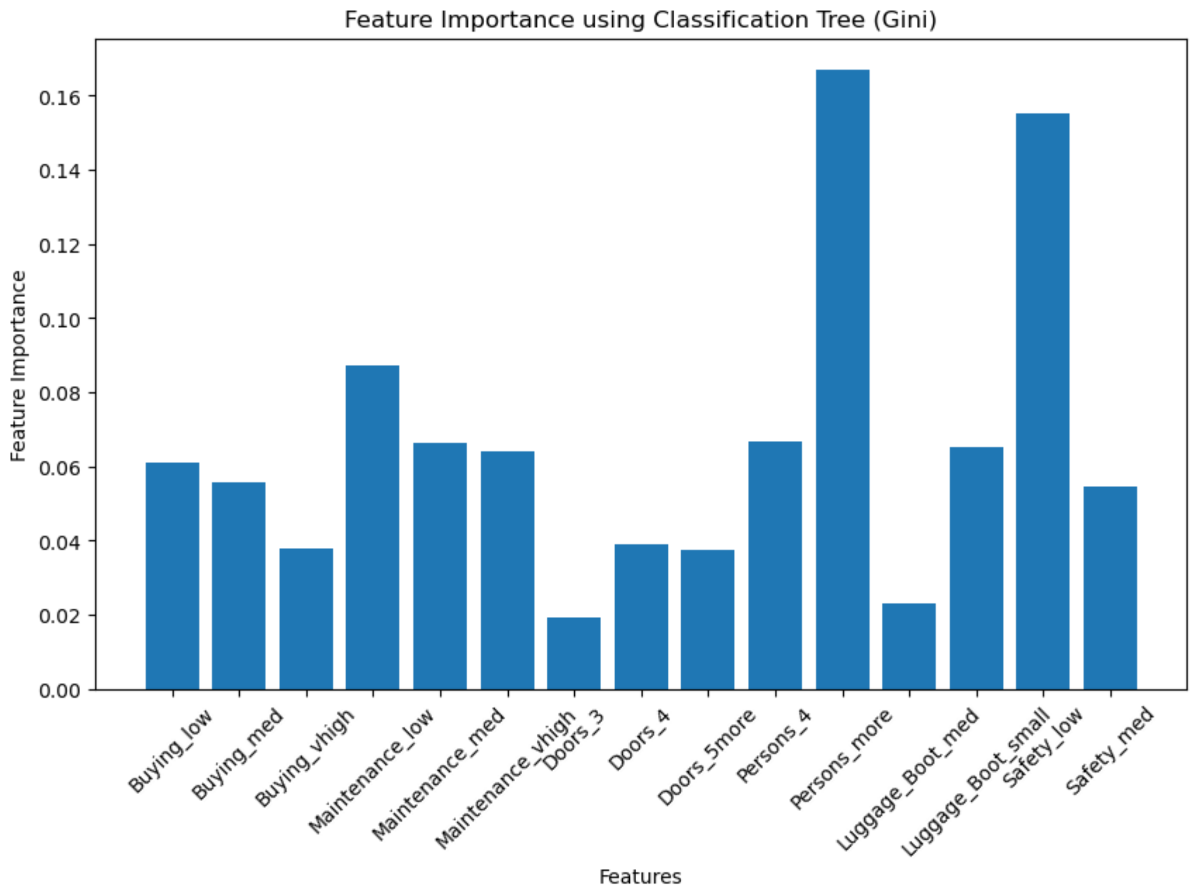


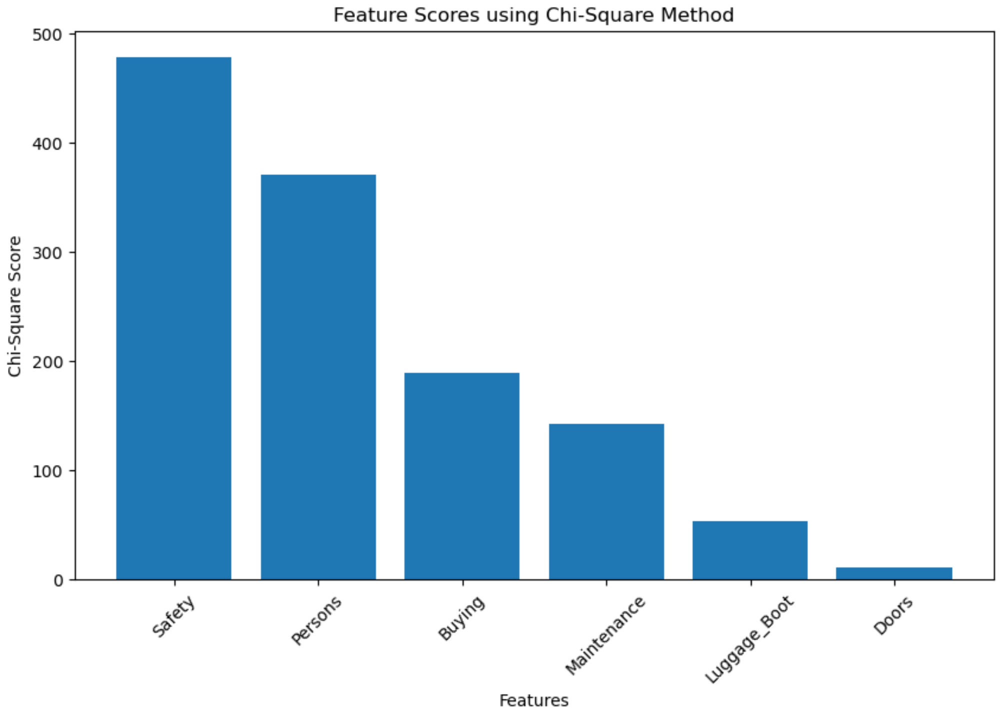
```
Doors    2    3    4  5more
class
acc     81   99  102    102
good    15   18   18     18
unacc  326  300  292    292
vgood   10   15   20     20
```

The number of doors doesn't show a strong association with the car class. However, cars with 5 or more doors have a higher likelihood of being in the 'unacc' class.

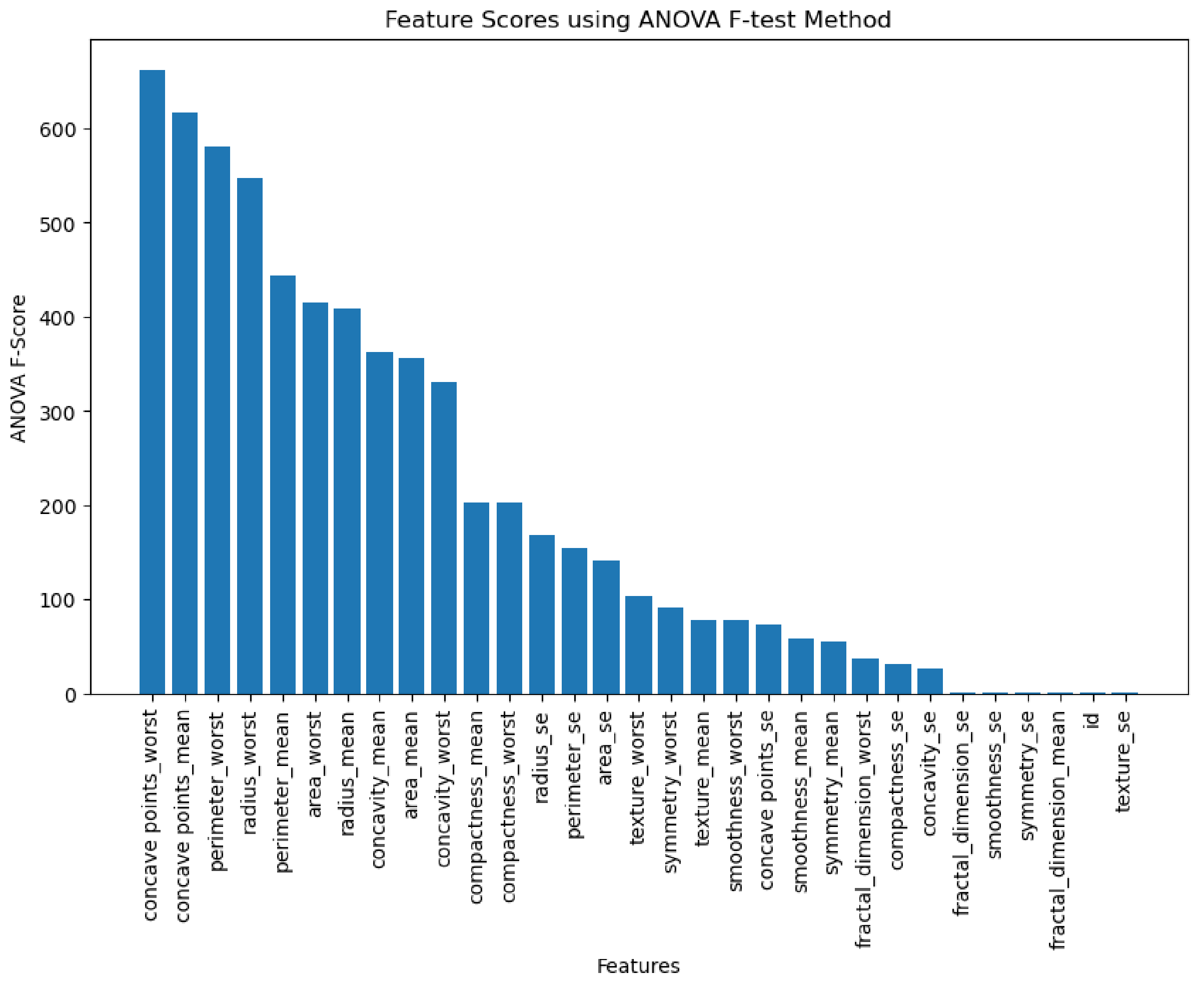
Q2: Compare the Chi-square selected features with Lab 9’s Classification Tree (‘mytree_gini’) feature importance result (‘mytree_gini.feature_importances_’). Provide the plots and explain your findings.

We can see from the two plots below from the chi-square selected features and the feature importance result from the Gini Classification tree that they are quite similar and the first plot shows more detail than the second plot. Safety and persons are more important features for the y variable class and luggage boot and doors are the lowest importance.





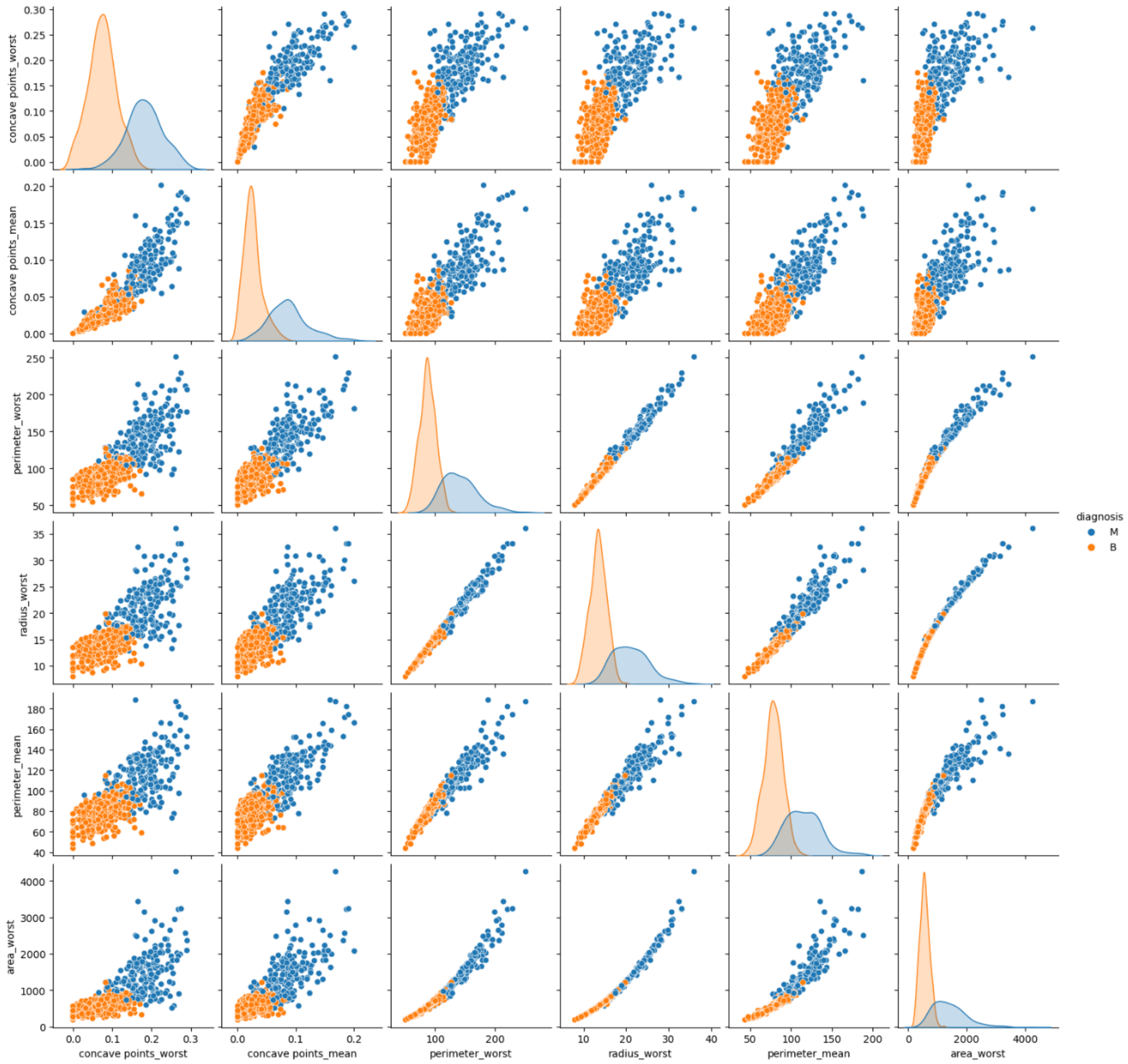
Q3: Summarise the ANOVA F-scores results in a 30 by 2 table with the first column holding the scores and the second column specifying the features name. Your table should be sorted based on F-scores in descending order. Present the first 10 rows of your table.



```
Top 10 rows of the feature scores table:
                 Feature     F-Score
28  concave points_worst  662.320337
8    concave points_mean  616.646516
23       perimeter_worst  580.041174
21          radius_worst  547.254347
3         perimeter_mean  443.061554
24            area_worst  414.736757
1            radius_mean  409.324586
7         concavity_mean  362.653784
4              area_mean  356.084616
27       concavity_worst  330.705030
```

Q4: Using the outcome of Q3, choose the top 6 features. Explain the distribution of their Class/features using proper visualization (e.g. factegrid or paired plots)

Looking at the scatter graph below to show the top 6 features in association with the diagnosis classes. We can see that the plots form a upwards line which shows its strength of correlation and relationships between the features and the classes. The strongest looks like perimeter_worst and radius_worst, having a strong relationship and potential significance in predicting the diagnosis of breast cancer.


concave points_worst
0.00
0.05
0.10
0.15
0.20
0.25
0.30
concave points_mean
0.00
0.05
0.10
0.15
0.20
perimeter_worst
50
100
150
200
250
radius_worst
10
15
20
25
30
35
perimeter_mean
40
60
80
100
120
140
160
180
area_worst
0
1000
2000
3000
4000
0.0
0.1
0.2
0.3
concave points_worst
0.0
0.1
0.2
concave points_mean
100
200
perimeter_worst
10
20
30
40
radius_worst
50
100
150
200
perimeter_mean
0
2000
4000
area_worst
diagnosis
M
B