Name: Parth Pareek
UNI: PP2547
Assignment: HW3
Date: 3/3/2016

Question 1.

a. Summary of training set

```
       X          Purchase WeekofPurchase     StoreID         PriceCH         PriceMM          DiscCH       
 Min.   :   1.0   CH:322   Min.   :227.0   Min.   :1.000   Min.   :1.690   Min.   :1.690   Min.   :0.00000  
 1st Qu.: 288.5   MM:213   1st Qu.:240.0   1st Qu.:2.000   1st Qu.:1.790   1st Qu.:2.090   1st Qu.:0.00000  
 Median : 526.0            Median :256.0   Median :3.000   Median :1.860   Median :2.130   Median :0.00000  
 Mean   : 537.8            Mean   :254.1   Mean   :3.935   Mean   :1.864   Mean   :2.087   Mean   :0.04862  
 3rd Qu.: 799.5            3rd Qu.:267.0   3rd Qu.:7.000   3rd Qu.:1.990   3rd Qu.:2.180   3rd Qu.:0.00000  
 Max.   :1070.0            Max.   :278.0   Max.   :7.000   Max.   :2.090   Max.   :2.290   Max.   :0.50000  
     DiscMM         SpecialCH        SpecialMM          LoyalCH          SalePriceMM     SalePriceCH   
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000011   Min.   :1.190   Min.   :1.390  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.321920   1st Qu.:1.690   1st Qu.:1.750  
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.589079   Median :2.090   Median :1.860  
 Mean   :0.1197   Mean   :0.1495   Mean   :0.1607   Mean   :0.560161   Mean   :1.967   Mean   :1.816  
 3rd Qu.:0.2000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.836900   3rd Qu.:2.180   3rd Qu.:1.890  
 Max.   :0.8000   Max.   :1.0000   Max.   :1.0000   Max.   :0.999947   Max.   :2.290   Max.   :2.090  
   PriceDiff          PctDiscMM          PctDiscCH        ListPriceDiff        STORE      
 Min.   :-0.6700   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.000  
 1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.1400   1st Qu.:0.000  
 Median : 0.2400   Median :0.00000   Median :0.00000   Median :0.2400   Median :2.000  
 Mean   : 0.1515   Mean   :0.05752   Mean   :0.02568   Mean   :0.2225   Mean   :1.619  
 3rd Qu.: 0.3200   3rd Qu.:0.11268   3rd Qu.:0.00000   3rd Qu.:0.3000   3rd Qu.:3.000  
 Max.   : 0.6400   Max.   :0.40201   Max.   :0.25269   Max.   :0.4400   Max.   :4.000  
```

b. Summary of logistic regression

```
Call:
glm(formula = Purchase ~ . - X, family = "binomial", data = dat.train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7041  -0.5335  -0.2410   0.5408   2.6176  

Coefficients: (4 not defined because of singularities)
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)      6.70266    2.90294   2.309 0.020948 *  
WeekofPurchase  -0.01267    0.01471  -0.861 0.389093    
StoreID         -0.02676    0.07213  -0.371 0.710615    
PriceCH          4.39824    2.55035   1.725 0.084607 .  
PriceMM         -4.32873    1.30691  -3.312 0.000926 ***
DiscCH          35.86001   27.98880   1.281 0.200114    
DiscMM          26.95865   12.47994   2.160 0.030760 *  
SpecialCH       -0.25188    0.48802  -0.516 0.605762    
SpecialMM        0.23946    0.37427   0.640 0.522309    
LoyalCH         -6.19063    0.57210 -10.821  < 2e-16 ***
SalePriceMM           NA         NA      NA       NA    
SalePriceCH           NA         NA      NA       NA    
PriceDiff             NA         NA      NA       NA    
PctDiscMM      -51.49017   26.20453  -1.965 0.049422 *  
PctDiscCH      -74.81096   52.94465  -1.413 0.157655    
ListPriceDiff         NA         NA      NA       NA    
STORE            0.05450    0.14220   0.383 0.701523    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 719.30  on 534  degrees of freedom
Residual deviance: 411.11  on 522  degrees of freedom
AIC: 437.11

Number of Fisher Scoring iterations: 5
```

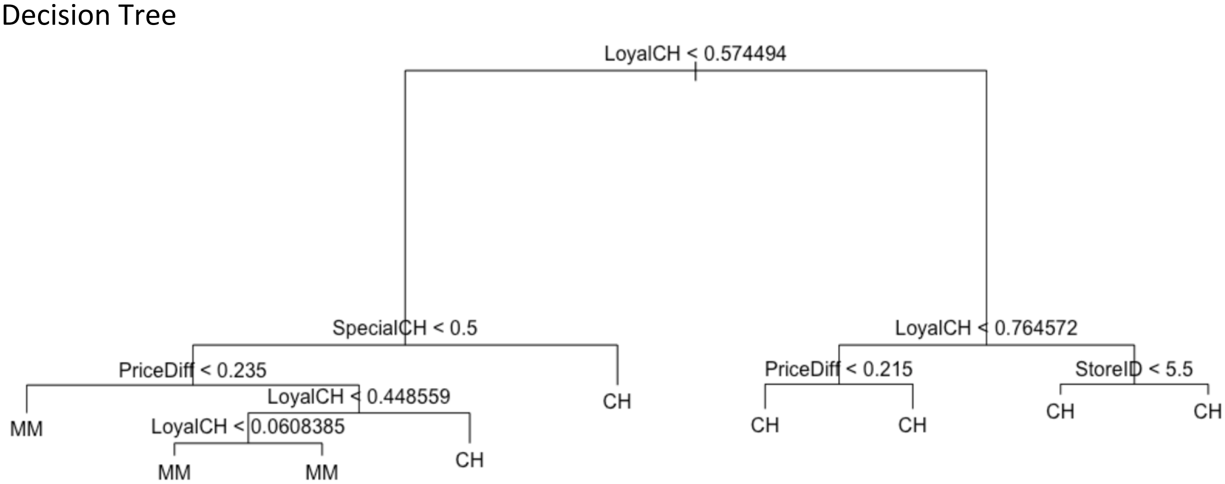
Interpretation:
Some variables are linearly dependent on others and are ignores in the regression. For example, Sale Price can be calculated using discount and Price Diff from Sales Prices of MM and CH.
Pct Disc seems to be driving the maximum impact on Purchase since they have the highest absolute coefficients.
Price MM and Loyal CH seem to be the most important variables since they have the least p-values and would be most impactful on regression
Week of Purchase and Store details make no difference at all (low coeff. and high p-values)
I would recheck how PctDiscMM is calculated since MM purchase should increase with more discount (evident in DiscMM), however, in this case there is a negative correlation

c. Price MM and Loyal CH have the lowest p-values and would be most impactful on regression covariates. Logically it makes sense to have these two variables since Loyal CH will gauge loyalty of customers to CH and Price MM will drive.

```
Call:
glm(formula = Purchase ~ PriceMM + LoyalCH, family = "binomial",
    data = dat.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3197  -0.6535  -0.2993   0.6512   2.5440

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.6780     1.7087   3.908 9.29e-05 ***
PriceMM      -1.9337     0.8170  -2.367   0.0179 *
LoyalCH      -5.8743     0.5052 -11.628  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 719.30  on 534  degrees of freedom
Residual deviance: 467.68  on 532  degrees of freedom
AIC: 473.68

Number of Fisher Scoring iterations: 5
```

PriceMM and LoyalCH are both negatively correlated to Purchase, which is logical since increase in CH loyalty will reduce MM sales. Also, decrease in prices increase demand (demand-supply reference)
AIC has increased from previous model, indicating this is more relevant

d. Decision Tree



e. SVM is trained on training set and validated on validation set. Best cost parameter: 0.1

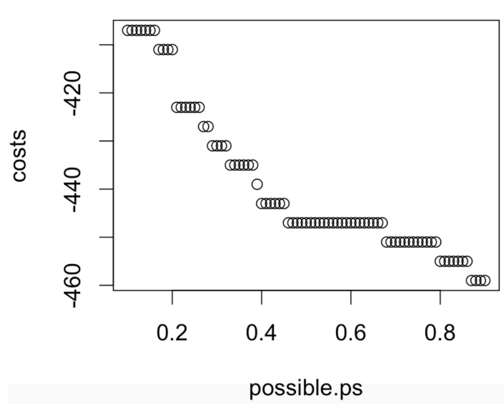
```
Cost:  0.01
    truth
pred  CH  MM
  CH 170  97
  MM   0   0
Misclassification Rate:  0.3632959

Cost:  0.1
    truth
pred  CH  MM
  CH 144  19
  MM  26  78
Misclassification Rate:  0.1685393

Cost:  1
    truth
pred  CH  MM
  CH 146  23
  MM  24  74
Misclassification Rate:  0.17603

Cost:  10
    truth
pred  CH  MM
  CH 147  22
  MM  23  75
Misclassification Rate:  0.1685393

Cost:  100
    truth
pred  CH  MM
  CH 147  23
  MM  23  74
Misclassification Rate:  0.1722846
```

f. The performance is tested on test set. Misclassification is noted below:
   Logistic: 0.157
   Tree: 0.184
   SVM: 0.169

   Logistic performs the best

g. Logistic regression's performance is tested on test set. Misclassification: 0.23 (23.13%)
   Performance of other models on test set:
   Tree: 22% misclassification
   SVM: 19.4% misclassification

h. Logistic regression was used. Probability threshold: 0.87
   Best Payoff: $459



Question 2.

a. With outlier

```
       X               X1                   X2            y
 Min.   :  1   Min.   :-995.7677   Min.   :-996.0271   N:121
 1st Qu.: 61   1st Qu.:  -2.7721   1st Qu.:  -3.4850   Y:120
 Median :121   Median :   0.3446   Median :  -1.7006
 Mean   :121   Mean   :  -3.4803   Mean   :  -5.5972
 3rd Qu.:181   3rd Qu.:   3.9347   3rd Qu.:   0.5794
 Max.   :241   Max.   :  11.1972   Max.   :   8.1020
```

Without outlier:

```
       X                 X1                 X2            y
 Min.   :  1.00   Min.   :-10.6529   Min.   :-9.2254   N:120
 1st Qu.: 60.75   1st Qu.: -2.7673   1st Qu.:-3.4291   Y:120
 Median :120.50   Median :  0.3587   Median :-1.6687
 Mean   :120.50   Mean   :  0.6543   Mean   :-1.4704
 3rd Qu.:180.25   3rd Qu.:  3.9391   3rd Qu.: 0.5937
 Max.   :240.00   Max.   : 11.1972   Max.   : 8.1020
```

b. With outlier

```
Call:
lda(y ~ X1 + X2, data = dat1)

Prior probabilities of groups:
        N         Y
0.5020747 0.4979253

Group means:
          X1          X2
N -10.508874 -10.2788474
Y   3.606944  -0.8765869

Coefficients of linear discriminants:
          LD1
X1  0.2498175
X2 -0.2472988
```

Without outlier

```
Call:
lda(y ~ X1 + X2, data = dat2)

Prior probabilities of groups:
  N   Y
0.5 0.5

Group means:
         X1         X2
N -2.298383 -2.0642783
Y  3.606944 -0.8765869

Coefficients of linear discriminants:
          LD1
X1 0.33001582
X2 0.02227882
```

The outlier is in the N group, and it can be seen that the group mean for N is extremely high in the LDA summary that includes the outlier. There is huge difference in the LDA model coefficients too.

c. SVM is tuned on the dataset that includes outlier to find the best cost (=100). The same cost is then used for both models

With outlier

```
Call:
best.tune(method = svm, train.x = y ~ X1 + X2, data = dat1, ranges = list(cost = c(
0.1,
    1, 10, 100, 1000)), tunecontrol = tune.control(sampling = c("cross"),
    cross = 10), kernel = "linear")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  100
      gamma:  0.5

Number of Support Vectors:  103
```

Without outlier

```
Call:
best.tune(method = svm, train.x = y ~ X1 + X2, data = dat2, ranges = list(cost = c(
0.1,
    1, 10, 100, 1000)), tunecontrol = tune.control(sampling = c("cross"),
    cross = 10), kernel = "linear")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.1
      gamma:  0.5

Number of Support Vectors:  117
```

SVM model on data set containing outlier has 103 support vectors while the one without outlier has 117 support vectors. In general, more support vectors indicate more stability in the model. We can say that the cost is high in the first model (outlier included) since SVM is trying to accommodate maximum data points by increasing the cost

d. Model performance on test data set

```
        LDA (All) 0.666666666666667
LDA (No Outlier) 0.783333333333333
        SVM (All)              0.75
SVM (No Outlier) 0.766666666666667
```

Rank in terms of performance on test data set:

1. LDA (without outlier)
2. SVM (without outlier)
3. SVM (with outlier)
4. LDA (with outlier)

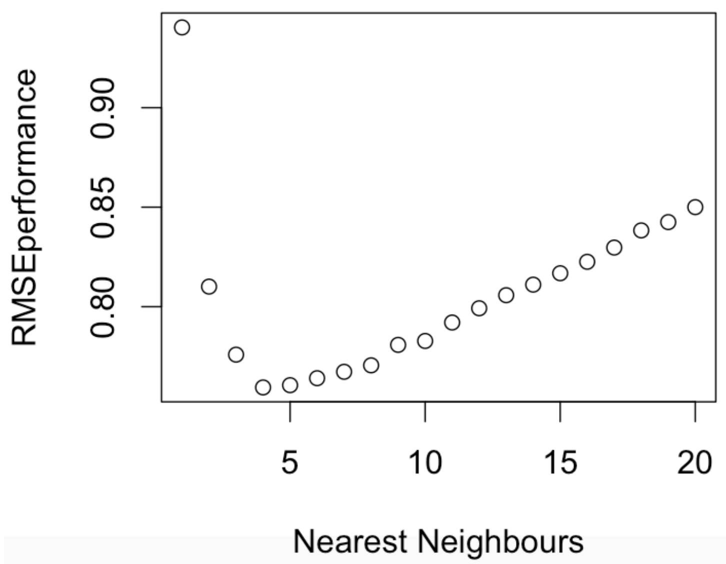
e. There is a huge variation in LDA's performance. This can be attributed to the jump in variance and mean when the outlier is included in the data set. SVM on the other hand is a more stable form of modeling and is not affected much by including the outlier

Question 3.

a. Data from section 1 loaded as training data
b. 5 closest students:
"Umair Mesiya" "Zhi Li" "Saaransh Jakhar" "Avinesh Vasudevan" "Nishant Jain"
c. Predictions for 5 students on 5 cuisines are noted below. The entire prediction matrix can be found in the code

	Italian	Mexican	Chinese...Cantonese	Chinese....Sichuan	Greek
Cedric Colle	NA	NA	NA	NA	NA
Yuan Zhong	NA	NA	NA	NA	4.000000
Xiao ran Ye	NA	NA	NA	NA	NA
Chaofan Da	NA	NA	NA	NA	NA
Ling Dong	NA	NA	NA	NA	3.666667

d. 4 nearest neighbors minimizes the RMSE



e. Choices of 3 students was predicted using 4-NN. Owing to limited space, I have printed only 5 cuisine preferences

	Italian	Mexican	Chinese...Cantonese	Chinese....Sichuan	Greek
Mufei Li	4.50	3.00	3.75	1.75	2.75
Pierre Laurent	4.75	4.50	4.00	3.50	3.50
Yao Wu	4.25	3.75	4.50	3.50	3.25

RMSE of predictions: 1.04