```
> # Homework 2 Question 1
> Retension <- read.csv("/Users/yankeyu/Desktop/Retention.csv")
> str(Retension)
'data.frame':   16362 obs. of  3 variables:
 $ user_id      : int  5295 8357 137 8085 5987 8015 8245 3906 5038 7274 ...
 $ play_duration: num  3.71 5.59 4.91 7.55 5.37 ...
 $ day          : int  1 1 1 1 1 1 1 1 1 1 ...
> summary(Retension)
    user_id      play_duration        day
 Min.   :    1   Min.   :1.048   Min.   :1.000
 1st Qu.: 2501   1st Qu.:4.243   1st Qu.:2.000
 Median : 4998   Median :4.953   Median :3.000
 Mean   : 5004   Mean   :4.959   Mean   :2.569
 3rd Qu.: 7517   3rd Qu.:5.673   3rd Qu.:3.000
 Max.   :10000   Max.   :9.108   Max.   :4.000
> #1)Retention rate of day 1 and day 2
> Retension_1 <- subset(Retension[Retension$day == 1,])
> Retension_2 <- subset(Retension[Retension$day == 2,])
> Retension_3 <- subset(Retension[Retension$day == 3,])
>
> #Using left merge to find who those leave the app after using (finding NAs)
> Retension_i1 <- merge (Retension_1, Retension_2,
+                        by.x = "user_id",
+                        by.y = "user_id",
+                        all = F,
+                        all.x = T,
+                        all.y = F)
> Retension_i2 <- merge (Retension_2, Retension_3,
+                        by.x = "user_id",
+                        by.y = "user_id",
+                        all = F,
+                        all.x = T,
+                        all.y = F)
```

```
> RetensionRate_1 <- 1-sum(is.na(Retension_i1$play_duration.y))/nrow(Retension_i1)
> RetensionRate_2 <- 1-sum(is.na(Retension_i2$play_duration.y))/nrow(Retension_i2)
>
> RetensionRate_1
[1] 0.4457232
> RetensionRate_2
[1] 0.4770663
> #2) Very active and marginally active users
> very_active_user_1 <- subset(Retension_i1, Retension_i1$play_duration.x > 6)
> marginally_active_user_1 <- subset(Retension_i1, Retension_i1$play_duration.x <= 6)
```

```
> RetensionRate_veryactive_1 <- 1-mean(is.na(very_active_user_1$play_duration.y))
> RetensionRate_marginallyactive_1 <- 1-mean(is.na(marginally_active_user_1$play_duration.y))
>
> RetensionRate_veryactive_1
[1] 0.4580777
> RetensionRate_marginallyactive_1
[1] 0.443507
>
> very_active_user_2 <- subset(Retension_i2, Retension_i2$play_duration.x >6)
> marginally_active_user_2 <- subset(Retension_i2, Retension_i2$play_duration.x <= 6)
>
> RetensionRate_veryactive_2 <- 1-mean(is.na(very_active_user_2$play_duration.y))
> RetensionRate_marginallyactive_2 <- 1-mean(is.na(marginally_active_user_2$play_duration.y))
>
> RetensionRate_veryactive_2
[1] 0.4894837
> RetensionRate_marginallyactive_2
[1] 0.4754386
>
```

```
> #winequality
>
> winequality <- read.csv2("/Users/yankeyu/Desktop/winequality-white.csv",sep = ";")
> head (winequality)
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
1             7             0.27        0.36           20.7     0.045                  45                  170   1.001    3      0.45
2           6.3              0.3        0.34            1.6     0.049                  14                  132   0.994  3.3      0.49
3           8.1             0.28         0.4            6.9      0.05                  30                   97  0.9951 3.26      0.44
4           7.2             0.23        0.32            8.5     0.058                  47                  186  0.9956 3.19       0.4
5           7.2             0.23        0.32            8.5     0.058                  47                  186  0.9956 3.19       0.4
6           8.1             0.28         0.4            6.9      0.05                  30                   97  0.9951 3.26      0.44
  alcohol quality
1     8.8       6
2     9.5       6
3    10.1       6
4     9.9       6
5     9.9       6
6    10.1       6
> summary(winequality)
 fixed.acidity      volatile.acidity   citric.acid       residual.sugar     chlorides         free.sulfur.dioxide total.sulfur.dioxide
 Length:4898        Length:4898        Length:4898        Length:4898        Length:4898        Length:4898         Length:4898
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character    Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character    Mode  :character


   density               pH            sulphates           alcohol            quality
 Length:4898        Length:4898        Length:4898        Length:4898        Min.   :3.000
 Class :character   Class :character   Class :character   Class :character   1st Qu.:5.000
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :6.000
                                                                            Mean   :5.878
                                                                            3rd Qu.:6.000
                                                                            Max.   :9.000
>
> #a)Training set and Validation set
> training.rows <- sample(1:nrow(winequality), 3500)
> training.data <- winequality[training.rows,]
> validation.data <- winequality[-training.rows,]
>
> write.csv(training.data, "/Users/yankeyu/Desktop/training_wine.csv")
> write.csv(validation.data, "/Users/yankeyu/Desktop/validation_wine.csv")
```

```
> #b)
> model_wine <- lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcohol, data=training.data)
> #R^2 for training set
> SSE=sum(model_wine$residuals ^2)
> SSE
[1] 482.989
> SST=sum((training.data$quality-mean(training.data$quality))^2)
> SST
[1] 2756.929
> 1-SSE/SST
[1] 0.8248091
>
> #R^2 for validation set
> quality_predict <- predict(model_wine, newdata = validation.data)

> OoS_SSE = sum((quality_predict-validation.data$quality)^2)
> OoS_SSE
[1] 1314.582
>
> OoS_SST = sum((mean(training.data$quality)-validation.data$quality)^2)
> OoS_SST
[1] 1084.159
>
> 1 - OoS_SSE/OoS_SST
[1] -0.2125356
>
```

```
> #c)
> model1 <-  lm(quality ~ volatile.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcohol, data=training.data)
> model2 <-  lm(quality ~ fixed.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcohol, data=training.data)
> model3 <-  lm(quality ~ fixed.acidity+volatile.acidity+residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcohol, data=training.data)
> model4 <-  lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcohol, data=training.data)
> model5 <-  lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcohol, data=training.data)
> model6 <-  lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+total.sulfur.dioxide+density+pH+sulphates+alcohol, data=training.data)
> model7 <-  lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dioxide+density+pH+sulphates+alcohol, data=training.data)
> model8 <-  lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+pH+sulphates+alcohol, data=training.data)
> model9 <-  lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+sulphates+alcohol, data=training.data)
> model10 <- lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+alcohol, data=training.data)
> model11 <- lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates, data=training.data)
>
> wine_list = list(model1,model2,model3,model4,model5,model6,model7,model8,model9,model10,model11)
>
> for (val in wine_list) {
+   SSE = sum(val$residuals^2)
+   SST = sum((training.data$quality-mean(training.data$quality))^2)
+   result <- 1-SSE/SST
+   print(result)
+ }
[1] 0.8108431
[1] 0.8011931
[1] 0.8111129
[1] 0.7691161
[1] 0.803426
[1] 0.7962769
[1] 0.7848073
[1] 0.6751962
[1] 0.8011765
[1] 0.81247
[1] 0.8083217
> # d)Through comparing with previous R^2, we can find that the out of sample R^2 of alcohol remain same, and so the feature of alcohol provide the least information for us.
> # However, for the R^2 of resudual.sugar we got the most additional information because it has the most difference with the previous R^2 we calculated before.
>
```