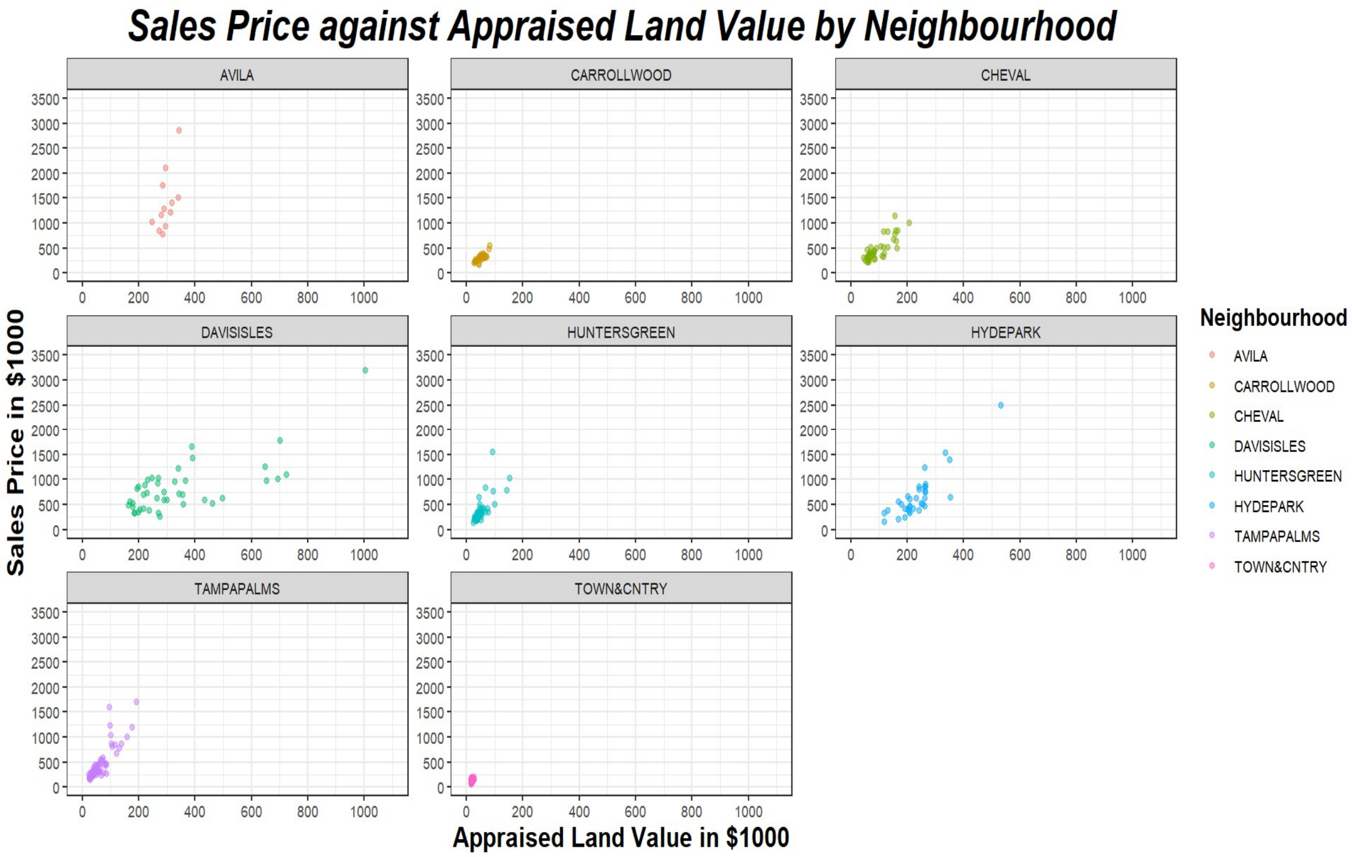
Question 1: Produce the scatterplots of

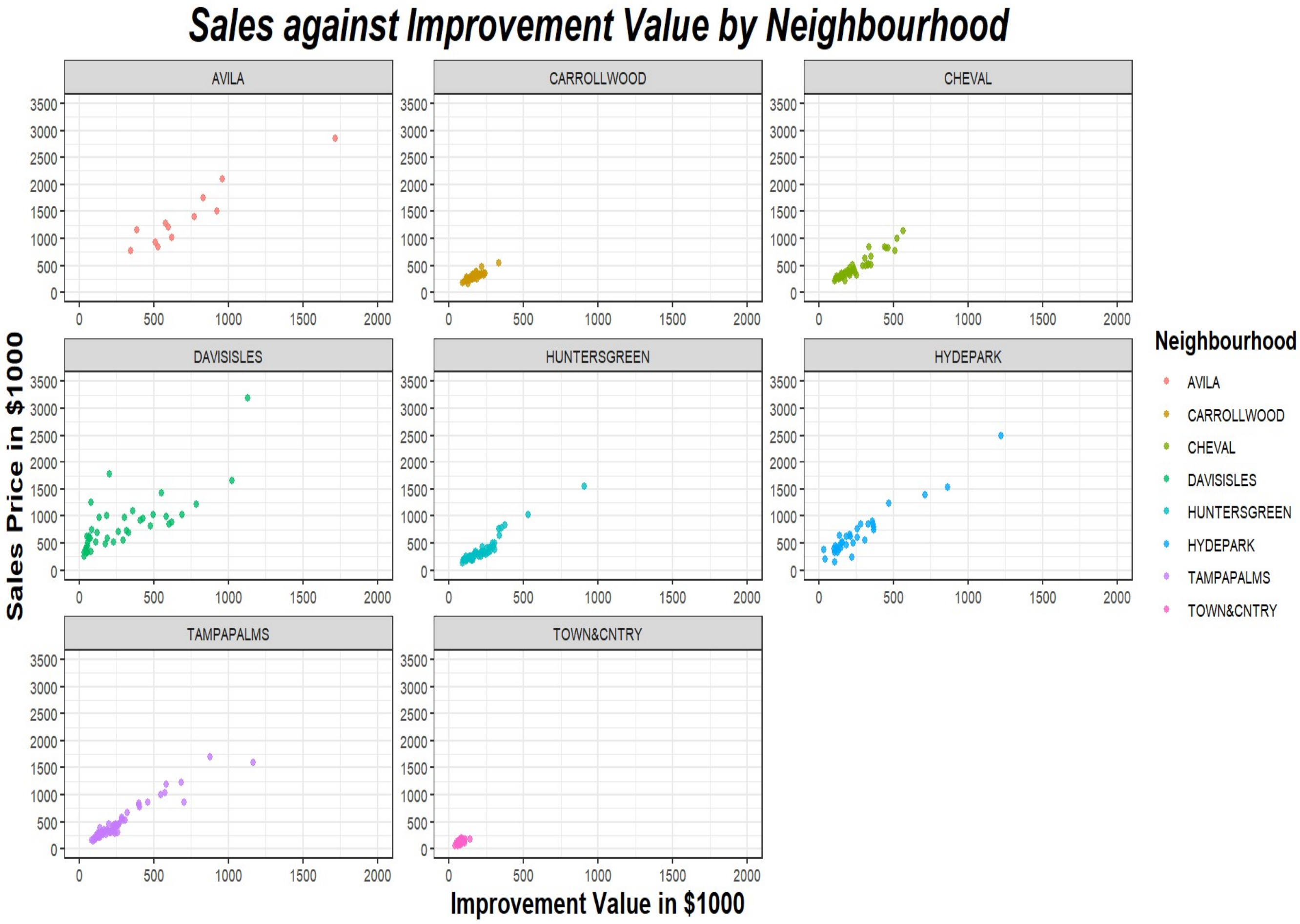
(i) SALES against LAND by NBHD

```
  library(ggplot2)
  ggplot(TamSales,aes(LAND,SALES,color=NBHD))+
    geom_point(alpha=0.5)+
    theme_bw()+
    facet_wrap(~NBHD,scales='free')+
    labs(x="Appraised Land Value in $1000", y="Sales Price in
$1000", colour="Neighbourhood")+
    ggtitle("Sales Price against Appraised Land Value by
Neighbourhood")+
    scale_x_continuous(limits = c(0,1100), breaks=seq(0,1100,200))+
    scale_y_continuous(limits = c(0,3500), breaks=seq(0,3500,500))+
    theme(
      plot.title = element_text(color="black", size=24,
face="bold.italic", hjust=0.5),
      axis.title.x = element_text(color="black", size=16,
face="bold"),
      axis.title.y = element_text(color="black", size=16,
face="bold"),
      legend.title = element_text(color="black", size=14,
face="bold")
    )
```



(ii) SALES against IMP by NBHD

```
ggplot(TamSales,aes(IMP,SALES,color=NBHD))+
    geom_point(alpha=0.8)+
    theme_bw()+
    facet_wrap(~NBHD,scales='free')+
    labs(x="Improvement Value in $1000", y="Sales Price in $1000",
colour="Neighbourhood")+
    ggtitle("Sales against Improvement Value by Neighbourhood")+
    scale_x_continuous(limits = c(0,2000), breaks=seq(0,2000,500))+
    scale_y_continuous(limits = c(0,3500),breaks=seq(0,3500,500))+
    theme(
      plot.title = element_text(color="black", size=24,
face="bold.italic", hjust=0.5),
      axis.title.x = element_text(color="black", size=16,
face="bold"),
      axis.title.y = element_text(color="black", size=16,
face="bold"),
      legend.title = element_text(color="black", size=14,
face="bold")
    )
```



Question 2: Comment on the plots produced in part (1).

- *For the plot of SALES against LAND by NBHD*: It can be observed that although there are some outliers in each neighbourhood, we can fit all the values of all the neighbourhoods using a single line. AVILA and TAMPAPALMS have many outlier values, that is, the appraised land value (LAND) is not linearly correlated to sales for these neighbourhoods.

- *For the plot of SALES against IMP by NBHD*: Similarly, it can also be observed that although there are some outliers in each neighbourhood, we can fit all the values of all the neighbourhoods using a single line. Improved value (IMP) seems to be more positively correlated to sales price.

Question 3: Fit Model 1 using R (Report your R-code and R-output). Report also the fitted line.

R-Code

```
M1=lm(SALES~LAND+IMP,TamSales)

M1
```

R-Output

```
Call:
lm(formula = SALES ~ LAND + IMP, data = TamSales)

Coefficients:
(Intercept)         LAND          IMP  
     -6.445        1.338        1.371  
```

The fitted line for **M1** is:

$\widehat{SALES}$= -6.445 + 1.338 LAND + 1.371 IMP

Question 4: Fit Model 2 using R (Report your R-code and R-output). Report also the fitted line.

R-Code

```
M2=lm(SALES~LAND+IMP+AVILA+CARROLLWOOD+CHEVAL+DAVISISLES+HUNTERSGREE
N+HYDEPARK+TAMPAPALMS,TamSales)

M2
```

R-Output

```
Call:
lm(formula = SALES ~ LAND + IMP + AVILA + CARROLLWOOD + CHEVAL + 
    DAVISISLES + HUNTERSGREEN + HYDEPARK + TAMPAPALMS, data = TamSales)

Coefficients:
 (Intercept)          LAND           IMP         AVILA   CARROLLWOOD  
      -5.146         1.588         1.338       -48.553        -6.129  
      CHEVAL    DAVISISLES  HUNTERSGREEN      HYDEPARK    TAMPAPALMS  
     -20.214      -103.041        -9.149       -67.410        12.590  
```

To obtain the filled line for each neighbourhood, we plug in 1 for that neighbourhood dummy variable and the remaining are kept as 0.

The fitted line for M2 **AVILA** neighborhood is:
$\widehat{SALES}$= -53.699 + 1.588 LAND + 1.338 IMP

The fitted line for M2 **CARROLLWOOD** neighborhood is:
$\widehat{SALES}$= -11.275 + 1.588 LAND + 1.338 IMP

The fitted line for M2 **CHEVAL** neighborhood is:
$\widehat{SALES}$= -25.36 + 1.588 LAND + 1.338 IMP

The fitted line for M2 **DAVIDISLES** neighborhood is:
$\widehat{SALES}$= -108.187 + 1.588 LAND + 1.338 IMP

The fitted line for M2 **HUNTERSGREEN** neighborhood is:
$\widehat{SALES}$= -14.295 + 1.588 LAND + 1.338 IMP

The fitted line for M2 **HYDEPARK** neighborhood is:
$\widehat{SALES}$= -14.295 + 1.588 LAND + 1.338 IMP

The fitted line for M2 **TAMPAPALMS** neighborhood is:
$\widehat{SALES}$= 7.444 + 1.588 LAND + 1.338 IMP

For the last neighborhood, we can find the line by plugging in zeroes for all neighbourhood dummy variables.
The fitted line for M2 **TOWN&CNTRY** neighborhood is:
$\widehat{SALES}$= -5.146 + 1.588 LAND + 1.338 IMP

Question 5: Fit Model 3 using R (Report your R-code and R-output). Report also the fitted line.

R-Code

```
M3=lm(SALES~LAND+IMP+AVILA+CARROLLWOOD+CHEVAL+DAVISISLES+HUNTERSGREE
N+HYDEPARK+TAMPAPALMS+AVILA*LAND+CARROLLWOOD*LAND+CHEVAL*LAND+DAVISI
SLES*LAND+HUNTERSGREEN*LAND+HYDEPARK*LAND+TAMPAPALMS*LAND+AVILA*IMP+
CARROLLWOOD*IMP+CHEVAL*IMP+DAVISISLES*IMP+HUNTERSGREEN*IMP+HYDEPARK*
IMP+TAMPAPALMS*IMP,TamSales)

M3
```

R-Output

```
Call:
lm(formula = SALES ~ LAND + IMP + AVILA + CARROLLWOOD + CHEVAL +
    DAVISISLES + HUNTERSGREEN + HYDEPARK + TAMPAPALMS + AVILA *
    LAND + CARROLLWOOD * LAND + CHEVAL * LAND + DAVISISLES *
    LAND + HUNTERSGREEN * LAND + HYDEPARK * LAND + TAMPAPALMS *
    LAND + AVILA * IMP + CARROLLWOOD * IMP + CHEVAL * IMP + DAVISISLES *
    IMP + HUNTERSGREEN * IMP + HYDEPARK * IMP + TAMPAPALMS *
    IMP, data = TamSales)

Coefficients:
      (Intercept)              LAND                IMP              AVILA
          2.11776           3.03220            0.85731          468.77444
      CARROLLWOOD            CHEVAL         DAVISISLES       HUNTERSGREEN
         38.55479           8.03556          -63.05388          -67.69716
         HYDEPARK        TAMPAPALMS         LAND:AVILA   LAND:CARROLLWOOD
       -110.43919         -23.87393           -3.79233           -0.98193
      LAND:CHEVAL   LAND:DAVISISLES  LAND:HUNTERSGREEN      LAND:HYDEPARK
         -2.44460          -1.43587           -1.43978           -1.33290
  LAND:TAMPAPALMS         IMP:AVILA    IMP:CARROLLWOOD         IMP:CHEVAL
         -0.62972           0.71976            0.04089            0.72385
   IMP:DAVISISLES  IMP:HUNTERSGREEN       IMP:HYDEPARK     IMP:TAMPAPALMS
          0.30445           0.71856            0.51468            0.39308
```

To obtain the filled line for each neighbourhood, we plug in 1 for that neighbourhood dummy variable and the remaining are kept as 0.

The values below have been rounded to the 4th nearest decimal place.

The fitted line for M3 **AVILA** neighborhood is:
$\widehat{SALES}$= 470.8922 -0.7601 LAND + 1.5771 IMP

The fitted line for M3 **CARROLLWOOD** neighborhood is:
$\widehat{SALES}$= 40.6726 + 2.0503 LAND + 0.8982 IMP

The fitted line for M3 **CHEVAL** neighborhood is:
$\widehat{SALES}$= 10.1533 + 0.5876 LAND + 1.5812 IMP

The fitted line for M3 **DAVIDISLES** neighborhood is:

$\widehat{SALES}$= -60.9361 + 1.5963 LAND + 1.1618 IMP

The fitted line for M3 **HUNTERSGREEN** neighborhood is:
$\widehat{SALES}$= -65.5794 + 1.5924 LAND + 1.5759 IMP

The fitted line for M3 **HYDEPARK** neighborhood is:
$\widehat{SALES}$= -108.3214 + 1.6993 LAND + 1.372 IMP

The fitted line for M3 **TAMPAPALMS** neighborhood is:
$\widehat{SALES}$= -21.7562+ 2.4025 LAND + 1.2504 IMP

For the last neighborhood, we can find the line by plugging in zeroes for all neighbourhood dummy variables.
The fitted line for M3 **TOWN&CNTRY** neighborhood is:
$\widehat{SALES}$= 2.1178+ 3.0322 LAND + 0.8573 IMP

Question 6: Fit Model 4 using R (Report your R-code and R-output). Report also the fitted line.

R-Code

```
options(scipen=999)

M4=lm(SALES~LAND+IMP+AVILA+CARROLLWOOD+CHEVAL+DAVISISLES+HUNTERSGREE
N+HYDEPARK+TAMPAPALMS+AVILA*LAND+CARROLLWOOD*LAND+CHEVAL*LAND+DAVISI
SLES*LAND+HUNTERSGREEN*LAND+HYDEPARK*LAND+TAMPAPALMS*LAND+AVILA*IMP+
CARROLLWOOD*IMP+CHEVAL*IMP+DAVISISLES*IMP+HUNTERSGREEN*IMP+HYDEPARK*
IMP+TAMPAPALMS*IMP+LAND*IMP,TamSales)

M4
```

R-Output

```
Call:
lm(formula = SALES ~ LAND + IMP + AVILA + CARROLLWOOD + CHEVAL + 
    DAVISISLES + HUNTERSGREEN + HYDEPARK + TAMPAPALMS + AVILA * 
    LAND + CARROLLWOOD * LAND + CHEVAL * LAND + DAVISISLES * 
    LAND + HUNTERSGREEN * LAND + HYDEPARK * LAND + TAMPAPALMS * 
    LAND + AVILA * IMP + CARROLLWOOD * IMP + CHEVAL * IMP + DAVISISLES * 
    IMP + HUNTERSGREEN * IMP + HYDEPARK * IMP + TAMPAPALMS * 
    IMP + LAND * IMP, data = TamSales)

Coefficients:
      (Intercept)                LAND                 IMP               AVILA  
        3.3951306           2.9707346           0.8396422         655.5683219  
      CARROLLWOOD              CHEVAL          DAVISISLES        HUNTERSGREEN  
       45.1558580          30.2942388          81.5084895         -53.9244937  
         HYDEPARK          TAMPAPALMS          LAND:AVILA    LAND:CARROLLWOOD  
      -20.5796432          -4.7283617          -4.3204709          -1.0582065  
      LAND:CHEVAL     LAND:DAVISISLES   LAND:HUNTERSGREEN       LAND:HYDEPARK  
       -2.6126362          -1.7588288          -1.6309938          -1.5716270  
  LAND:TAMPAPALMS           IMP:AVILA     IMP:CARROLLWOOD          IMP:CHEVAL  
       -0.8591368           0.4634049           0.0082974           0.6409069  
   IMP:DAVISISLES    IMP:HUNTERSGREEN        IMP:HYDEPARK      IMP:TAMPAPALMS  
       -0.0504063           0.6744960           0.2150342           0.3315336  
         LAND:IMP  
        0.0008396  
```

To obtain the filled line for each neighbourhood, we plug in 1 for that neighbourhood dummy variable and the remaining are kept as 0.

The values below have been rounded to the 4th nearest decimal place.

The fitted line for M4 **AVILA** neighborhood is:
$\widehat{SALES}$= 658.9635 -1.3497 LAND + 1.3031 IMP + 0.0008 LAND IMP

The fitted line for M4 **CARROLLWOOD** neighborhood is:
$\widehat{SALES}$= 48.551 + 1.9125 LAND + 0.8479 IMP + 0.0008 LAND IMP

The fitted line for M4 **CHEVAL** neighborhood is:

$\widehat{SALES}$= 33.6894 -0.3581 LAND + 1.4806 IMP + 0.0008 LAND IMP

The fitted line for M4 **DAVIDISLES** neighborhood is:
$\widehat{SALES}$= 84.9036 +1.2119 LAND + 0.7892 IMP + 0.0008 LAND IMP

The fitted line for M4 **HUNTERSGREEN** neighborhood is:
$\widehat{SALES}$= -50.5294 + 1.3397 LAND + 1.5141 IMP + 0.0008 LAND IMP

The fitted line for M4 **HYDEPARK** neighborhood is:
$\widehat{SALES}$= -17.1845 + 1.3991 LAND + 1.0547 IMP + 0.0008 LAND IMP

The fitted line for M4 **TAMPAPALMS** neighborhood is:
$\widehat{SALES}$= -1.3332 + 2.1116 LAND + 1.1712 IMP + 0.0008 LAND IMP

For the last neighborhood, we can find the line by plugging in zeroes for all neighbourhood dummy variables.
The fitted line for M4 **TOWN&CNTRY** neighborhood is:
$\widehat{SALES}$= 3.3951 + 2.9707 LAND + 0.8396 IMP + 0.0008 LAND IMP

Question 7: Compare Model 1 and Model 2 using F-test. Report your R-code, R-output and the pvalue, which model do you prefer?

For comparing Model 1 and Model 2, we have the following hypothesis:

H0 = None of the Neighbourhoods contribute positively to the Sales **(Reduced Model)**

H1 = Atleast one of the neighbourhoods contribute positively to the Sales **(Full Model)**

```
>   anova(M1,M2)
Analysis of Variance Table

Model 1: SALES ~ LAND + IMP
Model 2: SALES ~ LAND + IMP + AVILA + CARROLLWOOD + CHEVAL + DAVISISLES + 
    HUNTERSGREEN + HYDEPARK + TAMPAPALMS
  Res.Df     RSS Df Sum of Sq      F Pr(>F)   
1    347 3192256                              
2    340 3022927  7    169329 2.7207 0.0093 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On viewing the output of ANOVA, we can see that p-value, p = 0.0093

p < 0.05. This is less statistically significant. We reject H0.

Therefore, we prefer the **Model 2**.

Question 8: Compare Model 2 and Model 3 using F-test. Report your R-code, R-output and the pvalue, which model do you prefer?

For comparing Model 2 and Model 3, we have the following hypothesis:

H0 = There is no synergy between neighbourhoods and LAND/IMP **(Reduced Model)**

H1 = At least one of the neighbourhoods synergise with LAND/IMP **(Full Model)**

```
>   anova(M2,M3)
Analysis of Variance Table

Model 1: SALES ~ LAND + IMP + AVILA + CARROLLWOOD + CHEVAL + DAVISISLES + 
    HUNTERSGREEN + HYDEPARK + TAMPAPALMS
Model 2: SALES ~ LAND + IMP + AVILA + CARROLLWOOD + CHEVAL + DAVISISLES + 
    HUNTERSGREEN + HYDEPARK + TAMPAPALMS + AVILA * LAND + CARROLLWOOD * 
    LAND + CHEVAL * LAND + DAVISISLES * LAND + HUNTERSGREEN * 
    LAND + HYDEPARK * LAND + TAMPAPALMS * LAND + AVILA * IMP + 
    CARROLLWOOD * IMP + CHEVAL * IMP + DAVISISLES * IMP + HUNTERSGREEN * 
    IMP + HYDEPARK * IMP + TAMPAPALMS * IMP
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)   
1    340 3022927                                
2    326 2756960 14    265967 2.2464 0.006365 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On viewing the output of ANOVA, we can see that p-value, p = 0.006365

p < 0.05. This is less statistically significant. We reject H0.

Therefore, we prefer the **Model 3.**

Question 9: Compare Model 3 and Model 4 using F-test. Report your R-code, R-output and the pvalue, which model do you prefer?

For comparing Model 3 and Model 4, we have the following hypothesis:

H0 = There is no synergy between LAND and IMP when used together **(Reduced Model)**

H1 = There is some synergy between LAND and IMP when used together which contributes positively towards SALES **(Full Model)**

```
>   anova(M3,M4)
Analysis of Variance Table

Model 1: SALES ~ LAND + IMP + AVILA + CARROLLWOOD + CHEVAL + DAVISISLES +
    HUNTERSGREEN + HYDEPARK + TAMPAPALMS + AVILA * LAND + CARROLLWOOD *
    LAND + CHEVAL * LAND + DAVISISLES * LAND + HUNTERSGREEN *
    LAND + HYDEPARK * LAND + TAMPAPALMS * LAND + AVILA * IMP +
    CARROLLWOOD * IMP + CHEVAL * IMP + DAVISISLES * IMP + HUNTERSGREEN *
    IMP + HYDEPARK * IMP + TAMPAPALMS * IMP
Model 2: SALES ~ LAND + IMP + AVILA + CARROLLWOOD + CHEVAL + DAVISISLES +
    HUNTERSGREEN + HYDEPARK + TAMPAPALMS + AVILA * LAND + CARROLLWOOD *
    LAND + CHEVAL * LAND + DAVISISLES * LAND + HUNTERSGREEN *
    LAND + HYDEPARK * LAND + TAMPAPALMS * LAND + AVILA * IMP +
    CARROLLWOOD * IMP + CHEVAL * IMP + DAVISISLES * IMP + HUNTERSGREEN *
    IMP + HYDEPARK * IMP + TAMPAPALMS * IMP + LAND * IMP
  Res.Df     RSS Df Sum of Sq      F      Pr(>F)    
1    326 2756960                                   
2    325 2561820  1    195141 24.756 0.000001057 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On viewing the output of ANOVA, we can see that p-value, p = 0.000001057

$p < 0.05$. This is less statistically significant. We reject H0.

Therefore, we prefer the **Model 4.**

Question 10: Comment on the outcomes in part (7), (8), and (9).

- We have taken the reduced to full model approach while comparing the models, starting out from only LAND and IMP – M1.
- Then we added neighbourhoods using dummy variables – M2 and found that it positively contributes to sales. Based on the p-value we preferred M2 over M1.
- M3 also considered the interactions between neighbourhoods and LAND/IMP. Based on the p-value we preferred M3 over M2.
- Finally, we introduced a synergy term between LAND and IMP in M4. There seemed to be slightly better contribution to SALES by this term. Based on the p-value we preferred M4 over M3. We chose M4 over all others.

Question 11: Based on the above analysis, give your comments on the aims of this analysis.

Based on the above analysis, here are the comments of the aims of our experiments:

1) The data indicates that appraised value of land (LAND) and appraised value of improvements (IMP) are related to sale prices (SALES). The individual model coefficients and p-values supply sufficient evidence to indicate that these variables contribute information for the prediction of sale price (SALES).
2) From M4 we know that appraised value of land (LAND) and appraised value of improvements (IMP) to sale price (SALES) are interrelated. The relationship is NOT the same for a variety of neighborhoods (NBHD) because the coefficient for each fitted line is different. Therefore, the appraisers DO NOT use the same appraisal criteria for various types of neighborhoods (NBHD).