Project3

Project Housing Price in Ames

```
library(mosaic)
```

This project will analyze the housing price in Ames. In the previous project, we have discussed that the model SalePrice~Gr.Liv.Area is not really reasonable because the value of the R^2 is around 0.5 which shows that the fitted model is not really close to the actual data so the model is not reasonably good. But when the model was added with Neighborhood explanatory variable (SalePrice ~ Gr.Liv.Area + Neighborhood model), the value of R^2 has increase a lot from 0.5 to 0.7 which is pretty strong. So SalePrice ~ Gr.Liv.Area + Neighborhood model seems to be the best model for the ames dataset. So now, this project will use regression assumptions to clarify if the SalePrice ~ Gr.Liv.Area + Neighborhood model reasonable or not.

First, let's look at the graph of the ames SalePrice ~ Gr.Liv.Area + Neighborhood model:

There are a few unusual points in the graph which may make the data not accurate. So I create a new dataset called **ames1** that eliminate all unusual points which is when Gr.Liv.Area is larger than 4000 sq ft:

Now I will look at 4 conditions of regression assumptions which is linear relationship, independent errors, normal errors, and equal variances. If the model above fits with all the conditions, the model is reasonable.

1. **Linear relationship**: From the scatter plot, we can tell that the residual is not distributed linearly because in the first half of the graph, it decreases while the second half, it starts to increase to the right.
2. **Independent errors**: The requirement that the residuals be independent simply means that the residuals should not be related to each other which means that the shape of the graph shouldn't have any pattern. However, according to the scatter plot above, we can see that the residual starts to spread out from the left to the right which creates a trumpet shape. So the residuals are not independent.
3. **Normal errors:** To qualify this requirement, the shape of the distribution of these errors should be a normal distribution (centered at 0) which means that histogram of the residuals should curve into a straight line. Let's check:

From the graph, I can tell that the model doesn't has normal distribution because it doesn't lie on the straight line. It curves up at the end of the graph (on the right) and there is some data that is under the straight line on the left of the graph.

4. **Equal variance (homoskedasticity)**: means that the standard deviation of that normal distribution can be anything, but it should be the same everywhere. In this cases, we can see that the standard deviation on the right tends to be higher than the standard deviation on the left (according the the scatter plot) So it means that this model doesn't qualify the equal variance requirement.

Therefore, from these 4 condition, the SalePrice ~ Gr.Liv.Area + Neighborhood model seems to violate with the regression assumptions.

So to find a reasonable model, we need to modify the model to fit with the regression assumptions by using transformations: `log(`SalePrice`)` ~Gr.Liv.Area + Neighborhood.

According to the new scatter plot, there is a **linear relationship, independent errors** (the shape of the graph doesn't have any pattern which means that the plot doesn't have a trumpet shape) and **equal variance** (the standard deviation is the same everywhere) And the new model **distribute normally** because it lies mostly on a straight line (the graph below)

Therefore, the new model has fulfill the regression assumptions.

Now we will compare the R^2 of the 2 models to see which model is more precise.

```
[1] 0.7555103
```

```
[1] 0.759322
```

So we see that the R^2 of model 1 = 0.7555103 and the R^2 of model 2 = 0.759322. The R^2 of model 2 is higher than the R^2 of model 1. Therefore, the model 2 provides a better R^2.

Now we calculate the precision of the coefficience of 2 models:

- The confident interval of model 1:

```
                           2.5 %      97.5 %
(Intercept)          67429.90437  97956.3864
Gr.Liv.Area             77.64986     84.5953
NeighborhoodBlueste  -61386.45642  -4975.5464
NeighborhoodBrDale   -87689.84826 -47420.9864
NeighborhoodBrkSide  -74355.84404 -41875.6966
NeighborhoodClearCr  -34078.19429   2997.8407
NeighborhoodCollgCr  -17464.04290  12946.4166
NeighborhoodCrawfor  -31247.53651   1447.5841
NeighborhoodEdwards  -71733.67582 -40748.4763
NeighborhoodGilbert  -39197.44801  -7878.3794
NeighborhoodGreens   -13732.64418  47650.6420
NeighborhoodGrnHill   27843.90676 139869.9462
NeighborhoodIDOTRR   -93224.17976 -60201.8475
NeighborhoodLandmrk -130658.44619  25108.5442
NeighborhoodMeadowV  -92675.26408 -54265.4044
NeighborhoodMitchel  -44339.96969 -12053.2072
NeighborhoodNAmes    -57328.26130 -27492.8662
NeighborhoodNoRidge   21219.81841  56239.4944
NeighborhoodNPkVill  -64449.03118 -21362.6078
NeighborhoodNridgHt   65987.41664  97479.1197
NeighborhoodNWAmes   -47247.48395 -15319.9099
NeighborhoodOldTown  -90153.25788 -59580.2465
NeighborhoodSawyer   -57677.39595 -26146.9123
NeighborhoodSawyerW  -44764.23171 -12733.3114
NeighborhoodSomerst    1275.53201  32376.9790
NeighborhoodStoneBr   65311.69287 101509.5977
NeighborhoodSWISU    -98592.62676 -62160.6132
NeighborhoodTimber     7732.33932  41888.5924
NeighborhoodVeenker   -3321.34226  39352.3894
```

- The confident interval of model 2:

```
                           2.5 %        97.5 %
(Intercept)         11.5119556630 11.6684991217
Gr.Liv.Area          0.0004012132  0.0004368302
NeighborhoodBlueste -0.3645543404 -0.0752724328
NeighborhoodBrDale  -0.5998704079 -0.3933668583
NeighborhoodBrkSide -0.5001146102 -0.3335525214
NeighborhoodClearCr -0.1982545949 -0.0081242438
NeighborhoodCollgCr -0.1135123179  0.0424361631
NeighborhoodCrawfor -0.1999721984 -0.0323077013
NeighborhoodEdwards -0.4843423806 -0.3254465638
NeighborhoodGilbert -0.2041987693 -0.0435908318
NeighborhoodGreens  -0.0653286984  0.2494521500
NeighborhoodGrnHill  0.0628705020  0.6373534589
NeighborhoodIDOTRR  -0.7058733465 -0.5365308689
NeighborhoodLandmrk -0.7149957676  0.0837960282
NeighborhoodMeadowV -0.6856979904 -0.4887276264
NeighborhoodMitchel -0.2630759539 -0.0975055661
NeighborhoodNAmes   -0.3444100782 -0.1914105963
NeighborhoodNoRidge -0.0475538224  0.1320312731
NeighborhoodNPkVill -0.3697363536 -0.1487840104
NeighborhoodNridgHt  0.1540435510  0.3155367786
NeighborhoodNWAmes  -0.2526878524 -0.0889594259
NeighborhoodOldTown -0.5960602822 -0.4392782154
NeighborhoodSawyer  -0.3563962187 -0.1947041198
NeighborhoodSawyerW -0.2580773460 -0.0938189475
NeighborhoodSomerst -0.0269093934  0.1325825541
NeighborhoodStoneBr  0.1218679231  0.3074951173
NeighborhoodSWISU   -0.5832953638 -0.3964676319
NeighborhoodTimber  -0.0179376087  0.1572197507
NeighborhoodVeenker -0.0732320701  0.1456039401
```

From the calculation above, we can see that the confident interval of model 1 is wider than the confident interval of model 2. For example, the confident interval of Gr.Liv.Area in model 1 is between 77.64986 and 84.5953 (6.94544 difference)while the confident interval of Gr.Liv.Area in model 2 is between 0.0004012132 and 0.0004368302(3.5617e-05 difference). From this example, we can see that the confident interval in model 1 is wider than model 2 which means that the model 2 is more precise than the model 1.

Now, I will pick Gr.Liv.Area = 2000 and Neighborhood = Somerst and use regression equation to predict its sale price.

```
       fit      lwr      upr
1 12.48111 12.08752 12.87469
```

```
[1] 263316
```

```
[1] 177641
```

```
[1] 390307.4
```

Using the regression equation, the sale price will be $263316. And we are 95% confidence that the price of the house having 2000 square feet ground living area in Somerst neighborhood are between $177641 and $390307.4. The range of the price is quite wide but it is not so wide as to be useless.

To make the model become more useful, I can add more variable to the model like Year.Built variable (in project 2, it increases the value of R^2 a lot). And we can 95% confidence that the price of the house that built in 1990 with 2000 square feet in Somerst is between $198957.1 and $335430.4. The gap between the price is much lesser than the model am2 which is $136473.3 while the am2 is $212666.4 which means that the new model can be more useful than the log(SalePrice)~Gr.Liv.Area + Neighborhood model.

```
       fit      lwr      upr
1 267193.7 198957.1 335430.4
```

```
[1] 136473.3
```

```
[1] 212666.4
```

```
mod<-(lm(length~sex+birthmonth, data=KidsFeet))
summary(mod)
```

```
Call:
lm(formula = length ~ sex + birthmonth, data = KidsFeet)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.75065 -0.71966 -0.09623  0.77202  2.41133 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.143094   0.487602  51.565   <2e-16 ***
sexG        -0.786398   0.414238  -1.898   0.0657 .  
birthmonth  -0.006047   0.062385  -0.097   0.9233    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.291 on 36 degrees of freedom
Multiple R-squared:  0.09101,	Adjusted R-squared:  0.04051 
F-statistic: 1.802 on 2 and 36 DF,  p-value: 0.1795
```