Jianqi Liu

Aver Mckay

Dylan Patel

Lab 2

**Exercise 1:**

1.

|  |
| --- |
| int\_type = lapply(ameslist, class)  #get name which type is int  Ames = ameslist[int\_type=='integer']  names(Ames)  #check the names we want to leave  Ames <-Ames[ , !(names(Ames) %in% c("MSSubClass", "MasVnrArea", "BsmtFinSF1", "BsmtFinSF2", "BsmntUnfSF", "LowQualFinSF", "X3SsnPorch", "MiscVal"))]  names(Ames)  #save the new select data  save(Ames, file = "Ames.txt") |

2.

|  |
| --- |
| #12 of the variables that are type = int in the data set.  features <-c("SalePrice","BedroomAbvGr","TotRmsAbvGrd","GarageCars", "YrSold", "GrLivArea", "TotalBsmtSF", "OverallCond","OverallQual","LotArea","YearBuilt","YearRemodAdd")  pairs(Ames[,features]) |

3.

A matrix of correlation between the 12 variables. The only two correlations we fond that confused us were that the data says as year sold increases the sales price decreases, along as if the condition of the house increases, the sales price decreases.

|  |
| --- |
| cor(Ames[,features]) |

4

Realtionship between SalePrice and GrLivArea.

The largest outlier is at x~4200, y~7e+05

|  |
| --- |
| attach(Ames)  lm.fit = lm(SalePrice ~ GrLivArea)  plot(Ames$GrLivArea, Ames$SalePrice, main = "price with living space", ylab = "price", xlab = "living space") + abline(lm.fit) |

**Exercise 2:**

1. simple linear regression

|  |
| --- |
|  |

2.

Most of the variables do have a correlation with the sales price. The statistically significant variables seem to be the lot area, overall quality, overall condition, year built, year remodeled, size of basement, number of kitchens, number of bedrooms, number of total rooms, how many cars the garage is able to fit, the deck size, and the area of the pool if it has one. The year sold variable is deemed statistically insignificant.

|  |
| --- |
|  |

3.

Let's first look at the residual plot and the fit plot. As we can see, it is a relatively straight line with a uniform distribution of residuals. This is a good thing because it reflects that there is a non-linear relationship. However, if we look closely, there will be a slight parabolic shape, which may reflect a slight non-linear relationship. In addition, as sales prices rose, we noticed that the data began to have larger residuals and more outliers.

|  |
| --- |
|  |

4.

When in relation to sales price, the overall quality, lot area, lot frontage, ground living area, and lot frontage \* lot area all are deemed statistically significant.

|  |
| --- |
|  |

5.

Taking the log of a finished basement could be useful when calculating

percent changes of a given price, however in this model keeping just

the normal basement would suffice. Square rooting and squaring the data

would also not make sense in this data when using quantitative units such

as the number of bathrooms and basements.

|  |
| --- |
|  |