COMPSCIX 415.2 Homework 7

Michelle Gomez

7/18/2018

Contents

Exercise 1

```
train <- file.path("/Users/michellegomez/Downloads/train.csv")
train_data <- read_csv(train)
glimpse(train_data)
```

```
## Observations: 1,460
## Variables: 81
## $ Id            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...
## $ MSSubClass    <int> 60, 20, 60, 70, 60, 50, 20, 60, 50, 190, 20, 60,...
## $ MSZoning      <chr> "RL", "RL", "RL", "RL", "RL", "RL", "RL", "RL", ...
## $ LotFrontage   <int> 65, 80, 68, 60, 84, 85, 75, NA, 51, 50, 70, 85, ...
## $ LotArea       <int> 8450, 9600, 11250, 9550, 14260, 14115, 10084, 10...
## $ Street        <chr> "Pave", "Pave", "Pave", "Pave", "Pave", "Pave", ...
## $ Alley         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ LotShape      <chr> "Reg", "Reg", "IR1", "IR1", "IR1", "IR1", "Reg",...
## $ LandContour   <chr> "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl",...
## $ Utilities     <chr> "AllPub", "AllPub", "AllPub", "AllPub", "AllPub"...
## $ LotConfig     <chr> "Inside", "FR2", "Inside", "Corner", "FR2", "Ins...
## $ LandSlope     <chr> "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl",...
## $ Neighborhood  <chr> "CollgCr", "Veenker", "CollgCr", "Crawfor", "NoR...
## $ Condition1    <chr> "Norm", "Feedr", "Norm", "Norm", "Norm", "Norm",...
## $ Condition2    <chr> "Norm", "Norm", "Norm", "Norm", "Norm", "Norm", ...
## $ BldgType      <chr> "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", ...
## $ HouseStyle    <chr> "2Story", "1Story", "2Story", "2Story", "2Story"...
## $ OverallQual   <int> 7, 6, 7, 7, 8, 5, 8, 7, 7, 5, 5, 9, 5, 7, 6, 7, ...
## $ OverallCond   <int> 5, 8, 5, 5, 5, 5, 5, 6, 5, 6, 5, 5, 6, 5, 5, 8, ...
## $ YearBuilt     <int> 2003, 1976, 2001, 1915, 2000, 1993, 2004, 1973, ...
## $ YearRemodAdd  <int> 2003, 1976, 2002, 1970, 2000, 1995, 2005, 1973, ...
## $ RoofStyle     <chr> "Gable", "Gable", "Gable", "Gable", "Gable", "Ga...
## $ RoofMatl      <chr> "CompShg", "CompShg", "CompShg", "CompShg", "Com...
## $ Exterior1st   <chr> "VinylSd", "MetalSd", "VinylSd", "Wd Sdng", "Vin...
## $ Exterior2nd   <chr> "VinylSd", "MetalSd", "VinylSd", "Wd Shng", "Vin...
```

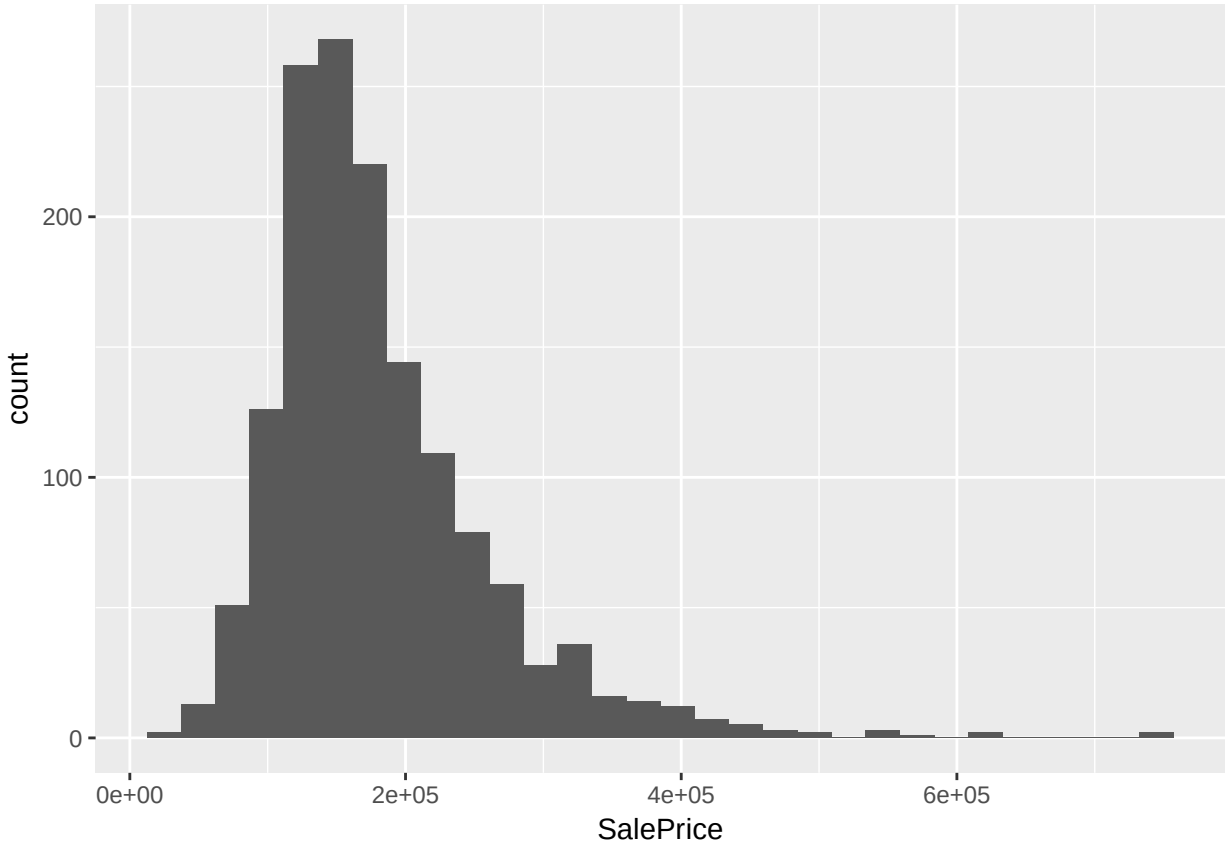
```
## $ MasVnrType    <chr> "BrkFace", "None", "BrkFace", "None", "BrkFace",...
## $ MasVnrArea    <int> 196, 0, 162, 0, 350, 0, 186, 240, 0, 0, 0, 286, ...
## $ ExterQual     <chr> "Gd", "TA", "Gd", "TA", "Gd", "TA", "Gd", "TA", ...
## $ ExterCond     <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", ...
## $ Foundation    <chr> "PConc", "CBlock", "PConc", "BrkTil", "PConc", "...
## $ BsmtQual      <chr> "Gd", "Gd", "Gd", "TA", "Gd", "Gd", "Ex", "Gd", ...
## $ BsmtCond      <chr> "TA", "TA", "TA", "Gd", "TA", "TA", "TA", "TA", ...
## $ BsmtExposure  <chr> "No", "Gd", "Mn", "No", "Av", "No", "Av", "Mn", ...
## $ BsmtFinType1  <chr> "GLQ", "ALQ", "GLQ", "ALQ", "GLQ", "GLQ", "GLQ",...
## $ BsmtFinSF1    <int> 706, 978, 486, 216, 655, 732, 1369, 859, 0, 851,...
## $ BsmtFinType2  <chr> "Unf", "Unf", "Unf", "Unf", "Unf", "Unf", "Unf",...
## $ BsmtFinSF2    <int> 0, 0, 0, 0, 0, 0, 0, 32, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ BsmtUnfSF     <int> 150, 284, 434, 540, 490, 64, 317, 216, 952, 140,...
## $ TotalBsmtSF   <int> 856, 1262, 920, 756, 1145, 796, 1686, 1107, 952,...
## $ Heating       <chr> "GasA", "GasA", "GasA", "GasA", "GasA", "GasA", ...
## $ HeatingQC     <chr> "Ex", "Ex", "Ex", "Gd", "Ex", "Ex", "Ex", "Ex", ...
## $ CentralAir    <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y"...
## $ Electrical    <chr> "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SB...
## $ `1stFlrSF`    <int> 856, 1262, 920, 961, 1145, 796, 1694, 1107, 1022...
## $ `2ndFlrSF`    <int> 854, 0, 866, 756, 1053, 566, 0, 983, 752, 0, 0, ...
## $ LowQualFinSF  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ GrLivArea     <int> 1710, 1262, 1786, 1717, 2198, 1362, 1694, 2090, ...
## $ BsmtFullBath  <int> 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, ...
## $ BsmtHalfBath  <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ FullBath      <int> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 3, 1, 2, 1, 1, ...
## $ HalfBath      <int> 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, ...
## $ BedroomAbvGr  <int> 3, 3, 3, 3, 4, 1, 3, 3, 2, 2, 3, 4, 2, 3, 2, 2, ...
## $ KitchenAbvGr  <int> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, ...
## $ KitchenQual   <chr> "Gd", "TA", "Gd", "Gd", "Gd", "TA", "Gd", "TA", ...
## $ TotRmsAbvGrd  <int> 8, 6, 6, 7, 9, 5, 7, 7, 8, 5, 5, 11, 4, 7, 5, 5,...
## $ Functional    <chr> "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ",...
## $ Fireplaces    <int> 0, 1, 1, 1, 1, 0, 1, 2, 2, 2, 0, 2, 0, 1, 1, 0, ...
## $ FireplaceQu   <chr> NA, "TA", "TA", "Gd", "TA", NA, "Gd", "TA", "TA"...
## $ GarageType    <chr> "Attchd", "Attchd", "Attchd", "Detchd", "Attchd"...
## $ GarageYrBlt   <int> 2003, 1976, 2001, 1998, 2000, 1993, 2004, 1973, ...
## $ GarageFinish  <chr> "RFn", "RFn", "RFn", "Unf", "RFn", "Unf", "RFn",...
## $ GarageCars    <int> 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 3, 1, 3, 1, 2, ...
## $ GarageArea    <int> 548, 460, 608, 642, 836, 480, 636, 484, 468, 205...
## $ GarageQual    <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", ...
## $ GarageCond    <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", ...
## $ PavedDrive    <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y"...
## $ WoodDeckSF    <int> 0, 298, 0, 0, 192, 40, 255, 235, 90, 0, 0, 147, ...
## $ OpenPorchSF   <int> 61, 0, 42, 35, 84, 30, 57, 204, 0, 4, 0, 21, 0, ...
## $ EnclosedPorch <int> 0, 0, 0, 272, 0, 0, 0, 228, 205, 0, 0, 0, 0, 0, ...
## $ `3SsnPorch`   <int> 0, 0, 0, 0, 0, 320, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ ScreenPorch   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 176, 0, 0, 0...
## $ PoolArea      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ PoolQC        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ Fence         <chr> NA, NA, NA, NA, NA, "MnPrv", NA, NA, NA, NA, NA,...
## $ MiscFeature   <chr> NA, NA, NA, NA, NA, "Shed", NA, "Shed", NA, NA, ...
## $ MiscVal       <int> 0, 0, 0, 0, 0, 700, 0, 350, 0, 0, 0, 0, 0, 0, 0,...
## $ MoSold        <int> 2, 5, 9, 2, 12, 10, 8, 11, 4, 1, 2, 7, 9, 8, 5, ...
## $ YrSold        <int> 2008, 2007, 2008, 2006, 2008, 2009, 2007, 2009, ...
## $ SaleType      <chr> "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD", ...
```

```
## $ SaleCondition <chr> "Normal", "Normal", "Normal", "Abnorml", "Normal...
## $ SalePrice     <int> 208500, 181500, 223500, 140000, 250000, 143000, ...
```

Exercise 2

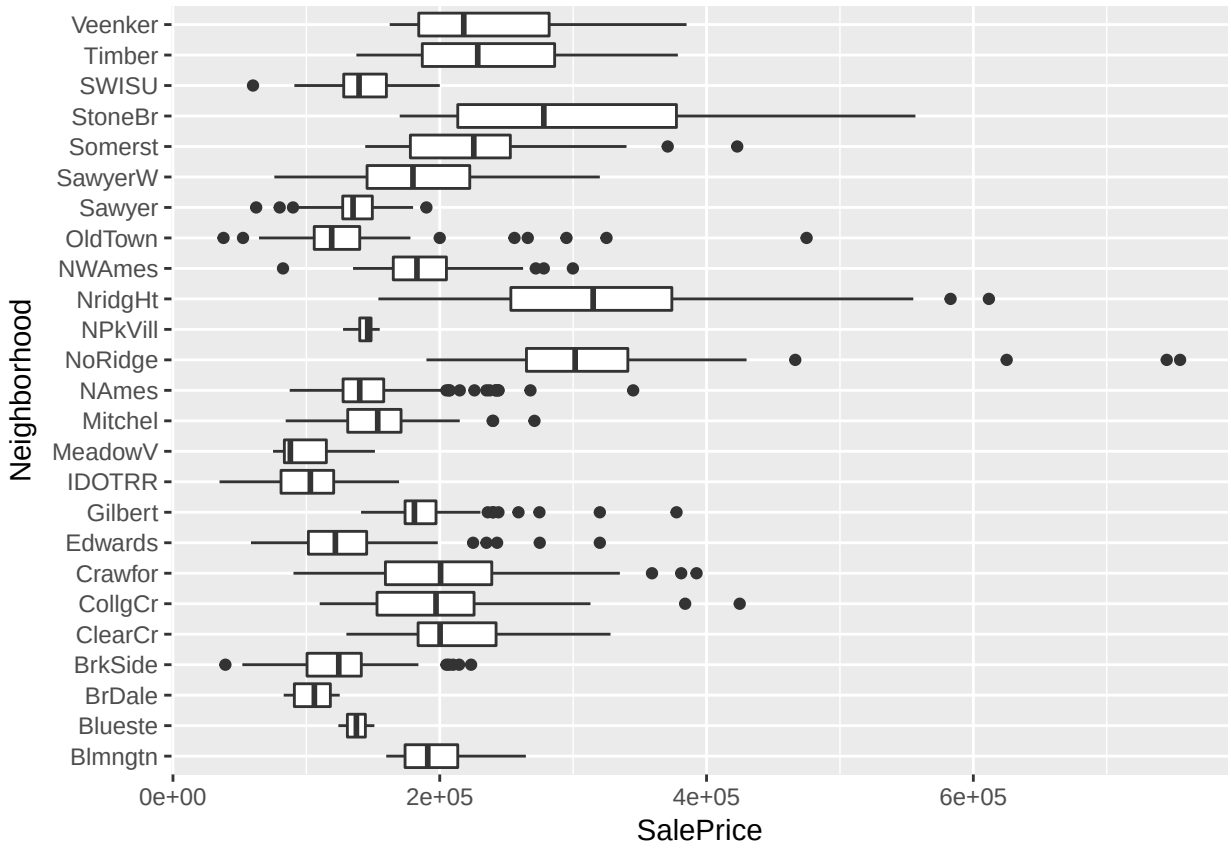
```
train_data %>%
  group_by(SalePrice) %>%
  ggplot() + geom_histogram(aes(x = SalePrice))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



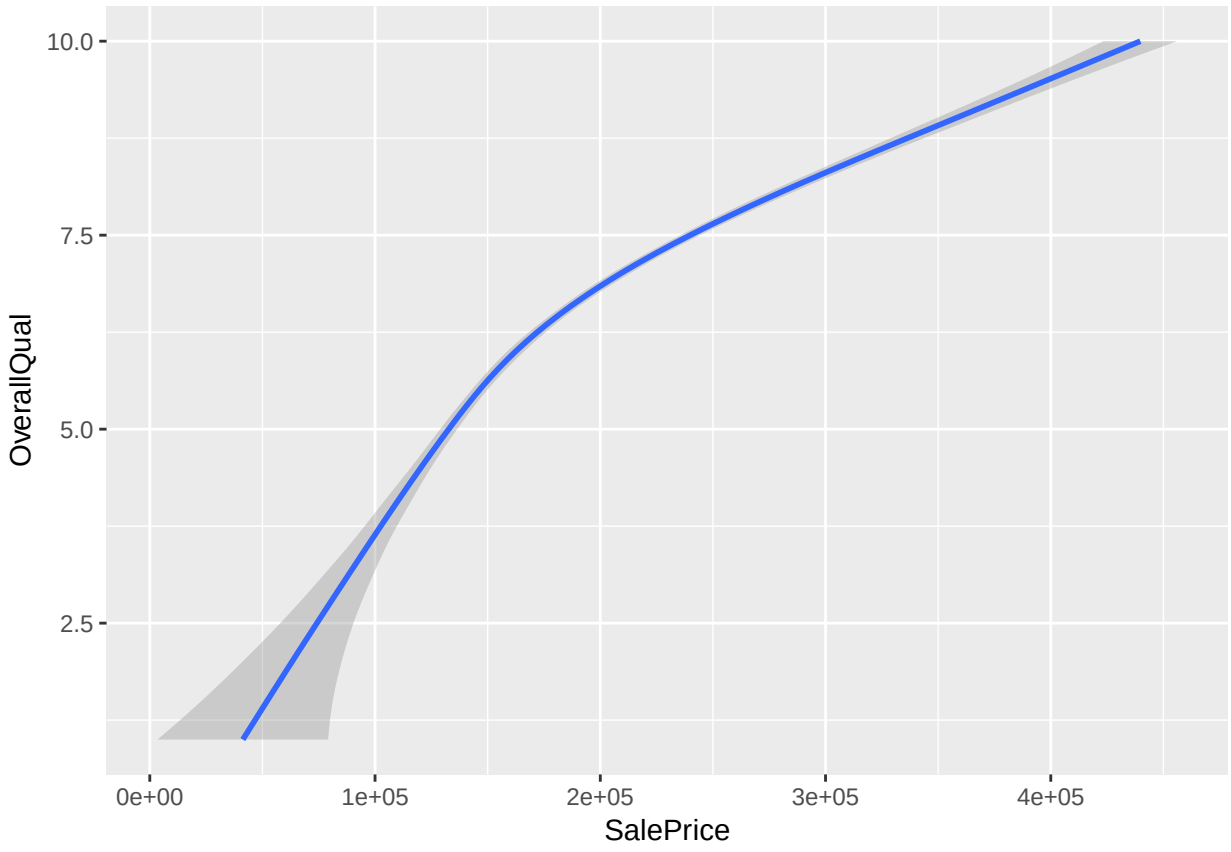

```
train_data %>%
  select(SalePrice, Neighborhood) %>% arrange_all() %>% ggplot() + geom_boxplot(aes(x = Neighborhood, y
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```




```
train_data %>%
  select(SalePrice, OverallQual) %>% arrange_all() %>% ggplot() + geom_smooth(aes(x = OverallQual, y = S
```

```
## `geom_smooth()` using method = 'gam'
```



Exercise 3

```
(sale_lm <- lm(formula = SalePrice ~ OverallQual, data = train_data))
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual, data = train_data)
##
## Coefficients:
## (Intercept)  OverallQual
##      -96206        45436
```

```
tidy(sale_lm)
```

```
##          term  estimate std.error statistic       p.value
## 1 (Intercept) -96206.08 5756.4074 -16.71287  1.667971e-57
## 2 OverallQual  45435.80  920.4302  49.36366 2.185675e-313
```

```
glance(sale_lm)
```

```
##   r.squared adj.r.squared    sigma statistic       p.value df    logLik
## 1 0.6256519     0.6253951 48622.76  2436.771 2.185675e-313  2 -17826.75
##        AIC      BIC     deviance df.residual
## 1 35659.49 35675.35 3.446964e+12        1458
```

```
mean(train_data$SalePrice)
```

```
## [1] 180921.2
```

```
45435.80/180921.2
```

```
## [1] 0.2511359
```

- take a look at the coefficient
  The coefficient is 45435.80 increase in SalePrice by unit increase in OverallQual.

- compare the coefficient to the average value of SalePrice
  The coefficient seems to be 1/4 of the average value of Sale Price.
- take a look at the R-squared
  R-squaredis .62, which can be considered a good fit.

Exercise 4

```
sale_mult_lm <- lm(formula = SalePrice ~ GrLivArea + OverallQual + Neighborhood, data = train_data)
sale_mult_lm
```

```
##
## Call:
## lm(formula = SalePrice ~ GrLivArea + OverallQual + Neighborhood,
##     data = train_data)
##
## Coefficients:
##         (Intercept)            GrLivArea          OverallQual
##           -34829.24                55.56             20951.42
## NeighborhoodBlueste   NeighborhoodBrDale  NeighborhoodBrkSide
##           -30752.88            -43358.88            -13025.45
## NeighborhoodClearCr  NeighborhoodCollgCr  NeighborhoodCrawfor
##            24575.64             11414.31             14444.25
## NeighborhoodEdwards  NeighborhoodGilbert   NeighborhoodIDOTRR
##           -17842.95              -892.88            -28178.99
## NeighborhoodMeadowV  NeighborhoodMitchel    NeighborhoodNAmes
##           -19099.02              2030.61             -4430.10
## NeighborhoodNoRidge  NeighborhoodNPkVill  NeighborhoodNridgHt
##            64642.99            -17807.18             71587.84
##  NeighborhoodNWAmes  NeighborhoodOldTown   NeighborhoodSawyer
##            -4720.66            -32080.88             -1219.38
## NeighborhoodSawyerW  NeighborhoodSomerst  NeighborhoodStoneBr
##              303.10             17766.96             69954.47
##   NeighborhoodSWISU   NeighborhoodTimber  NeighborhoodVeenker
##           -36640.15             29905.81             47106.89
```

```
tidy(sale_mult_lm)
```

```
##                   term    estimate    std.error   statistic      p.value
## 1          (Intercept) -34829.2399 11541.232315 -3.01780945 2.591005e-03
## 2            GrLivArea     55.5645     2.498787 22.23658695 2.375376e-94
## 3          OverallQual  20951.4249  1162.273747 18.02623948 1.238994e-65
## 4  NeighborhoodBlueste -30752.8759 27697.270091 -1.11032155 2.670468e-01
## 5   NeighborhoodBrDale -43358.8812 12978.523255 -3.34081778 8.568095e-04
## 6  NeighborhoodBrkSide -13025.4529 10450.444653 -1.24640179 2.128206e-01
## 7  NeighborhoodClearCr  24575.6351 11569.625542  2.12415129 3.382834e-02
```

```
## 8  NeighborhoodCollgCr  11414.3095  9496.581581  1.20193876 2.295858e-01
## 9  NeighborhoodCrawfor  14444.2502 10502.214763  1.37535278 1.692371e-01
## 10 NeighborhoodEdwards -17842.9513  9985.733698 -1.78684430 7.417398e-02
## 11 NeighborhoodGilbert   -892.8796  9954.350137 -0.08969743 9.285402e-01
## 12  NeighborhoodIDOTRR -28178.9866 11135.583978 -2.53053514 1.149517e-02
## 13 NeighborhoodMeadowV -19099.0203 12999.351487 -1.46922870 1.419903e-01
## 14 NeighborhoodMitchel   2030.6121 10555.017277  0.19238359 8.474690e-01
## 15   NeighborhoodNAmes  -4430.0994  9517.290569 -0.46547906 6.416592e-01
## 16 NeighborhoodNoRidge  64642.9895 10939.818250  5.90896375 4.294924e-09
## 17 NeighborhoodNPkVill -17807.1837 15304.007123 -1.16356347 2.447947e-01
## 18 NeighborhoodNridgHt  71587.8398  9994.775058  7.16252635 1.262091e-12
## 19  NeighborhoodNWAmes  -4720.6584 10079.514642 -0.46834183 6.396114e-01
## 20 NeighborhoodOldTown -32080.8775  9863.373223 -3.25252596 1.170466e-03
## 21  NeighborhoodSawyer  -1219.3805 10211.495047 -0.11941254 9.049653e-01
## 22 NeighborhoodSawyerW    303.0986 10264.231736  0.02952960 9.764463e-01
## 23 NeighborhoodSomerst  17766.9637  9829.881193  1.80744440 7.090264e-02
## 24 NeighborhoodStoneBr  69954.4718 11689.508047  5.98438117 2.740136e-09
## 25   NeighborhoodSWISU -36640.1517 11923.592410 -3.07291213 2.159781e-03
## 26  NeighborhoodTimber  29905.8114 10829.232809  2.76158172 5.825641e-03
## 27 NeighborhoodVeenker  47106.8929 14337.857486  3.28549039 1.042655e-03
```

```
glance(sale_mult_lm)
```

```
##   r.squared adj.r.squared    sigma statistic p.value df    logLik      AIC
## 1 0.7868484      0.782981 37008.52  203.4583       0 27 -17415.62 34887.25
##        BIC     deviance df.residual
## 1 35035.26 1.962681e+12        1433
```

- How would you interpret the coefficients on GrLivArea and OverallQual?
  For every one unit increase in GrLivArea, the SalePrice increases, on average, by $55.57.
  For every one unit increase in OverallQual, the SalePrice increases, on average, by $20951.43.

- How would you interpret the coefficient on NeighborhoodBrkSide?
  The mean SalePrice difference between NeighborhoodBrkSide and Blmngtn is -$13025.45.

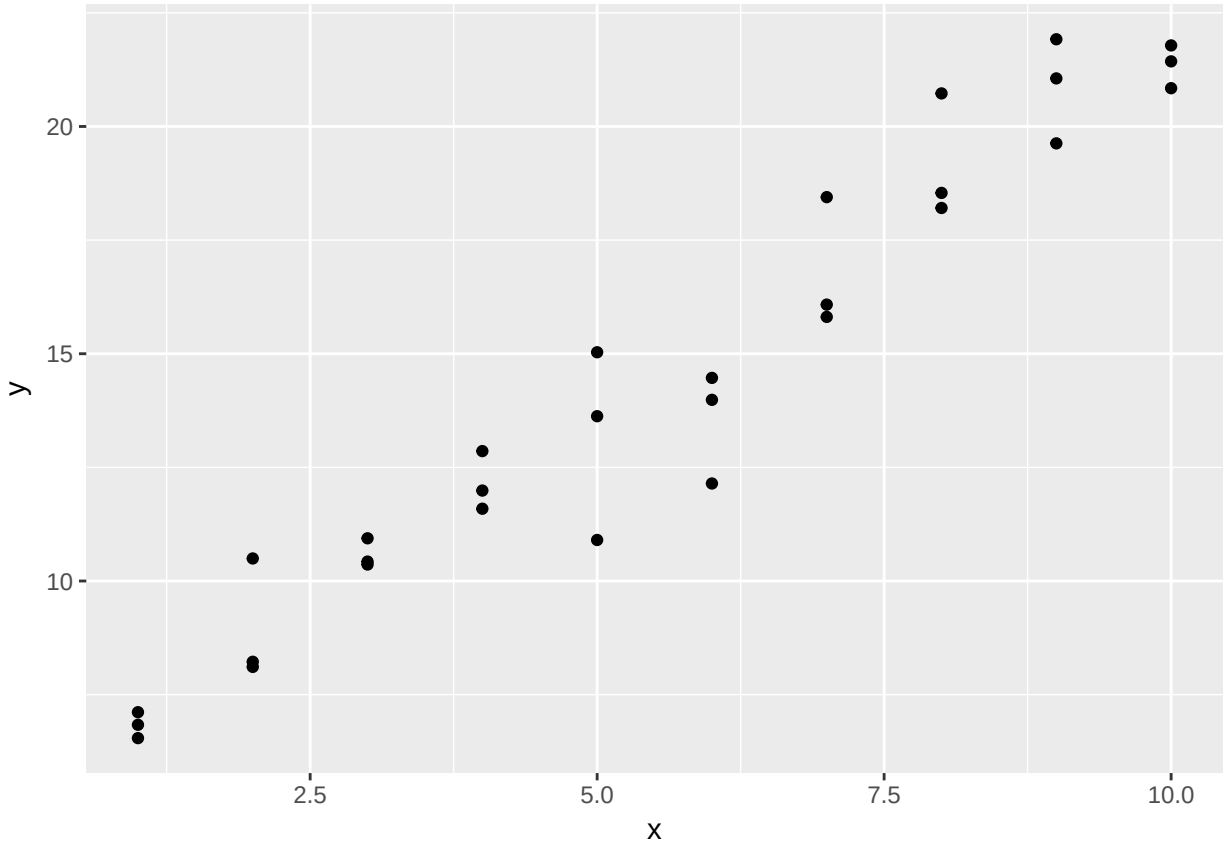
- Are the features significant?
  The p-values are all less than alpha=.05 therefore they are all significant prefictors of SalePrice.
- Are the features practically significant?
  I don't think that the features are practically siginificant because you can't increase them by a unit, rather they describe a relationship of something measurable like price to some more subjective features.

- Is the model a good fit?
  The model's adjusted R-squared is 0.78, which is closer to 1 so it's more a less a "good" fit but could definitely be better.

Exercise 6

One downside of the linear model is that it is sensitive to unusual values because the distance incorporates a squared term. Fit a linear model to the simulated data below (use y as the target and x as the feature), and look at the resulting coefficients and R-squared. Rerun it about 5-6 times to generate different simulated datasets. What do you notice about the model's coefficient on x and the R-squared values?

```
sim1a <- tibble(
x = rep(1:10, each = 3),
y = x * 1.5 + 6 + rt(length(x), df = 2)
)
ggplot(sim1a) + geom_point(aes(x = x, y = y))
```




```
sim1a_lm <- lm(formula = y ~ x , data = sim1a)
tidy(sim1a_lm)
```

```
##          term estimate  std.error statistic      p.value
## 1 (Intercept) 5.293484 0.47593039  11.12239 8.743360e-12
## 2           x 1.644370 0.07670313  21.43811 6.514247e-19
```

```
glance(sim1a_lm)
```

```
##   r.squared adj.r.squared    sigma statistic      p.value df    logLik
## 1  0.942575     0.9405241 1.206704  459.5924 6.514247e-19  2 -47.17004
##        AIC      BIC deviance df.residual
## 1 100.3401 104.5437 40.77175          28
```

I notice that the R-squared value keeps jumping up and down between .06-.99 and the x coefficient is also variable from 1.0-1.5. I think that the linear model can be improved by weighing outliers less than values that fall within range.