EDA Lab Assignment JVGiannantonio

2023-11-19

EDA LAB Assignment

J. Vincent Giannantonio

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
ames <- read.csv("~/Datasets/ames.csv")
head(ames)
```

	Id	MSSubClass	MSZoni…	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	▶
	<int>	<int>	<chr>	<int>	<int>	<chr>	<chr>	<chr>	<chr>	
1	1	60	RL	65	8450	Pave	*NA*	Reg	Lvl	
2	2	20	RL	80	9600	Pave	*NA*	Reg	Lvl	
3	3	60	RL	68	11250	Pave	*NA*	IR1	Lvl	
4	4	70	RL	60	9550	Pave	*NA*	IR1	Lvl	
5	5	60	RL	84	14260	Pave	*NA*	IR1	Lvl	
6	6	50	RL	85	14115	Pave	*NA*	IR1	Lvl	

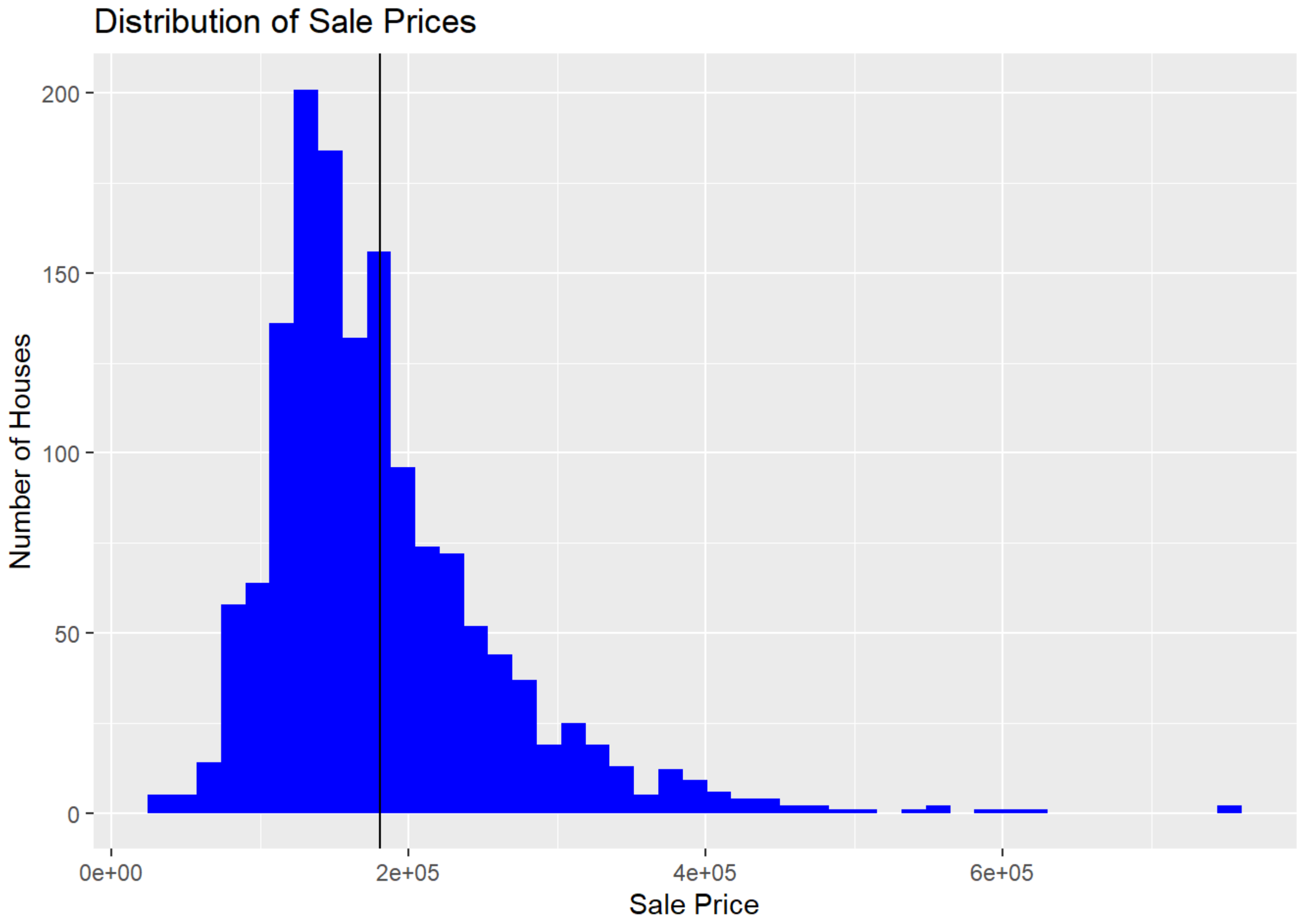
6 rows | 1-10 of 82 columns

```
attach(ames)
```

2

Histogram of Sale Price

```
ggplot(ames, aes(x=SalePrice))+
  geom_histogram(fill="blue",bins=45)+
  ggtitle("Distribution of Sale Prices")+
  xlab("Sale Price")+
  ylab("Number of Houses")+
  geom_vline(xintercept = mean(SalePrice, na.rm=TRUE))
```



```
mean(SalePrice)
```

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

```
median(SalePrice)
```

```
## [1] 163000
```

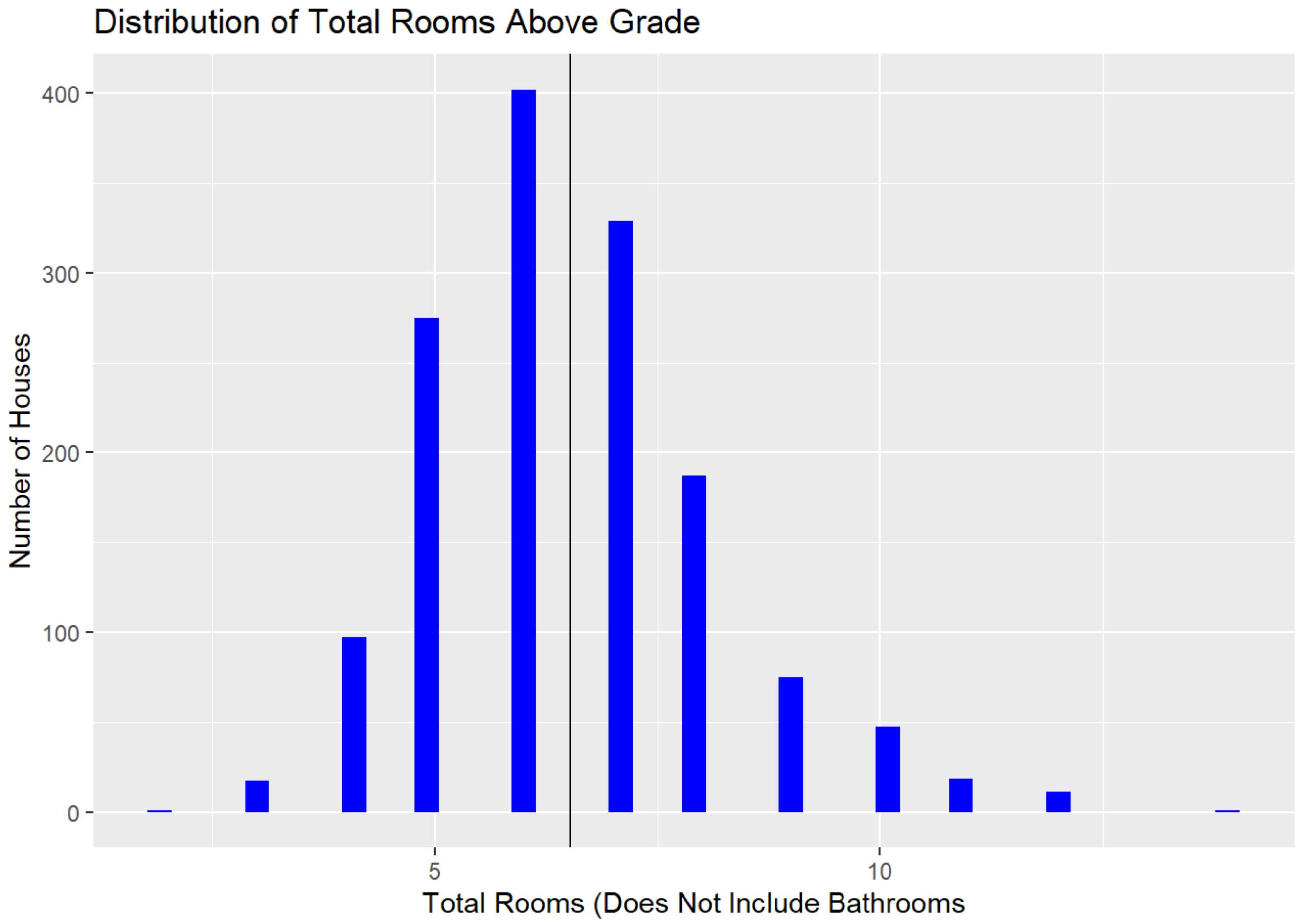
```
sd(SalePrice)
```

```
## [1] 79442.5
```

Looks like a log normal distribution. Most houses in this sample are clustered around the median value of $163,000, but the higher-end homes are pulling the mean up to over $180,000

Histogram of Rooms Above Grade

```
ggplot(ames, aes(x=TotRmsAbvGrd))+
  geom_histogram(fill="blue",bins=45)+
  ggtitle("Distribution of Total Rooms Above Grade")+
  xlab("Total Rooms (Does Not Include Bathrooms")+
  ylab("Number of Houses")+
  geom_vline(xintercept = mean(TotRmsAbvGrd, na.rm=TRUE))
```



```
mean(TotRmsAbvGrd)
```

```
## [1] 6.517808
```

```
median(TotRmsAbvGrd)
```

```
## [1] 6
```

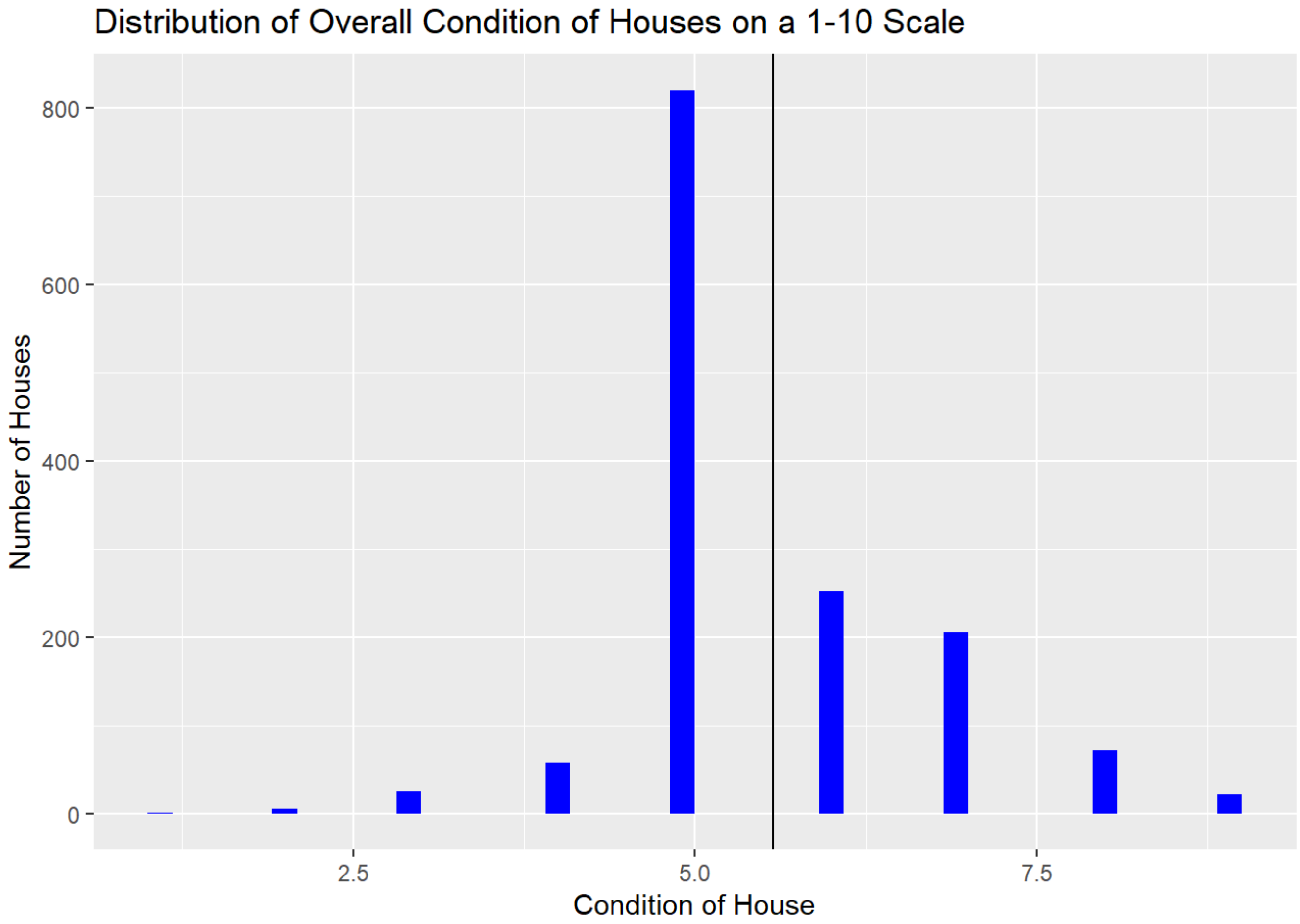
```
sd(TotRmsAbvGrd)
```

```
## [1] 1.625393
```

The number of rooms in houses is approximately normally distributed, with a mean and median around 6 rooms. There are some houses with twice as many rooms as the average, but overall the distribution is less skewed than the sale price distribution

Histogram of Overall Condition

```
ggplot(ames, aes(x=OverallCond))+
  geom_histogram(fill="blue",bins=45)+
  ggtitle("Distribution of Overall Condition of Houses on a 1-10 Scale")+
  xlab("Condition of House")+
  ylab("Number of Houses")+
  geom_vline(xintercept = mean(OverallCond, na.rm=TRUE))
```



```
mean(OverallCond)
```

```
## [1] 5.575342
```

```
median(OverallCond)
```

```
## [1] 5
```

```
sd(OverallCond)
```

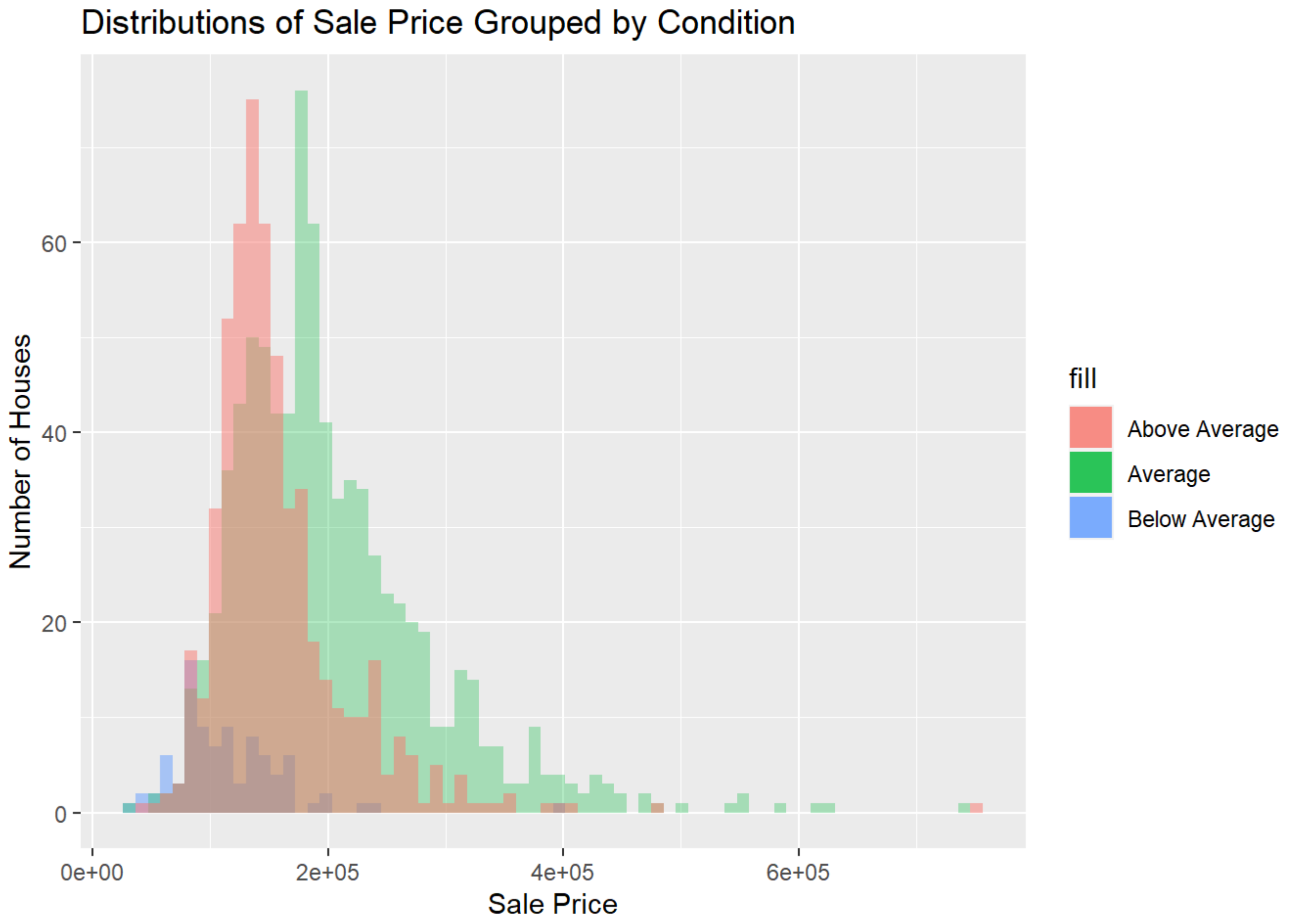
```
## [1] 1.112799
```

Most homes have a condition of 5. It seems like we should treat this as a categorical rather than numeric variable, since the difference between conditions is so abrupt

3

```
below_average_condition <- subset(ames, OverallCond < 5)
average_condition <- subset(ames, OverallCond == 5)
above_average_condition <- subset(ames, OverallCond > 5)
```

```
ggplot()+
  geom_histogram(aes(x=below_average_condition$SalePrice, fill="Below Average"), alpha=0.5, bins
=70)+
  geom_histogram(aes(x=average_condition$SalePrice, fill="Average"), alpha=0.3, bins=70)+
  geom_histogram(aes(x=above_average_condition$SalePrice, fill="Above Average"), alpha=0.5, bins
=70)+
  ggtitle("Distributions of Sale Price Grouped by Condition")+
  xlab("Sale Price")+
  ylab("Number of Houses")
```



First, we note again that the majority of the houses have average condition, then about 1/3 have above average condition, then less than 10% have below average condition.

As we might expect, the average condition therefore contains houses across a broader spectrum of the sale price range than either the below-average or above-average houses.

Another unsurprising finding is that below-average condition houses have a price distribution that is much lower than average or above-average condition houses.

But what might be surprising is that above-average condition houses do not seem to have higher average sale prices than average condition houses. In fact, above-average condition houses seem more clustered around a particular price range, especially the $100,000 to $200,000 range, whereas average condition houses are more frequent above $200,000. We might want to investigate further to understand what kinds of houses are rated as above-average condition, since this goes against a standard assumption that better condition would mean higher cost.

4

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
ames_num <- ames %>% select_if(is.numeric)

sort(cor(ames_num$SalePrice, ames_num[,]))
```

```
##  [1] -0.13590737 -0.12857796 -0.08428414 -0.07785589 -0.02892259 -0.02560613
##  [7] -0.02191672 -0.02118958 -0.01684415 -0.01137812  0.04458367  0.04643225
## [13]  0.09240355  0.11144657  0.16821315  0.21447911  0.22712223  0.26384335
## [19]  0.28410768  0.31585623  0.31933380  0.32441344  0.38641981  0.46692884
## [25]  0.50710097  0.52289733  0.53372316  0.56066376  0.60585218  0.61358055
## [31]  0.62343144  0.64040920  0.70862448  0.79098160  1.00000000
```

Use the highest and lowest values above to match with the their corresponding variable names below, since this will be faster than me having to figure out how to code this without ripping off ChatGPT.

```
cor(ames_num$SalePrice, ames_num[,])
```

```
##               Id  MSSubClass LotFrontage   LotArea OverallQual OverallCond
## [1,] -0.02191672 -0.08428414          NA 0.2638434   0.7909816 -0.07785589
##      YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1  BsmtFinSF2 BsmtUnfSF
## [1,] 0.5228973     0.507101         NA  0.3864198 -0.01137812 0.2144791
##      TotalBsmtSF X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath
## [1,]   0.6135806 0.6058522 0.3193338  -0.02560613 0.7086245    0.2271222
##      BsmtHalfBath  FullBath  HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd
## [1,]  -0.01684415 0.5606638 0.2841077    0.1682132   -0.1359074    0.5337232
##      Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF
## [1,]  0.4669288          NA  0.6404092  0.6234314  0.3244134   0.3158562
##      EnclosedPorch X3SsnPorch ScreenPorch   PoolArea     MiscVal     MoSold
## [1,]     -0.128578 0.04458367   0.1114466 0.09240355 -0.02118958 0.04643225
##           YrSold SalePrice
## [1,] -0.02892259         1
```

Highest: OverallQual

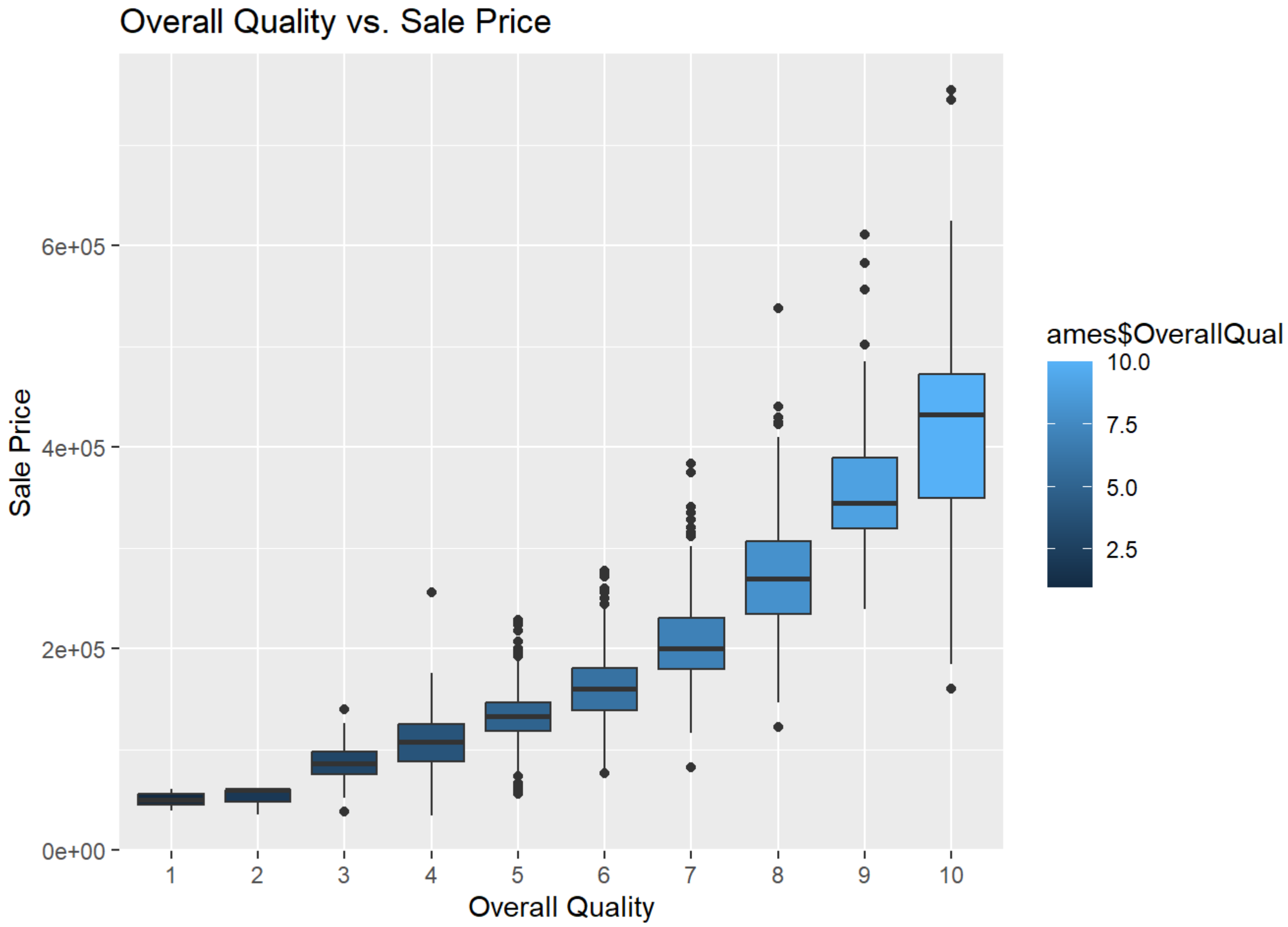
Maximum Correlation Value: 0.7909816

Lowest: KitchenAbvGr

Minimum Correlation Value: -0.1359074

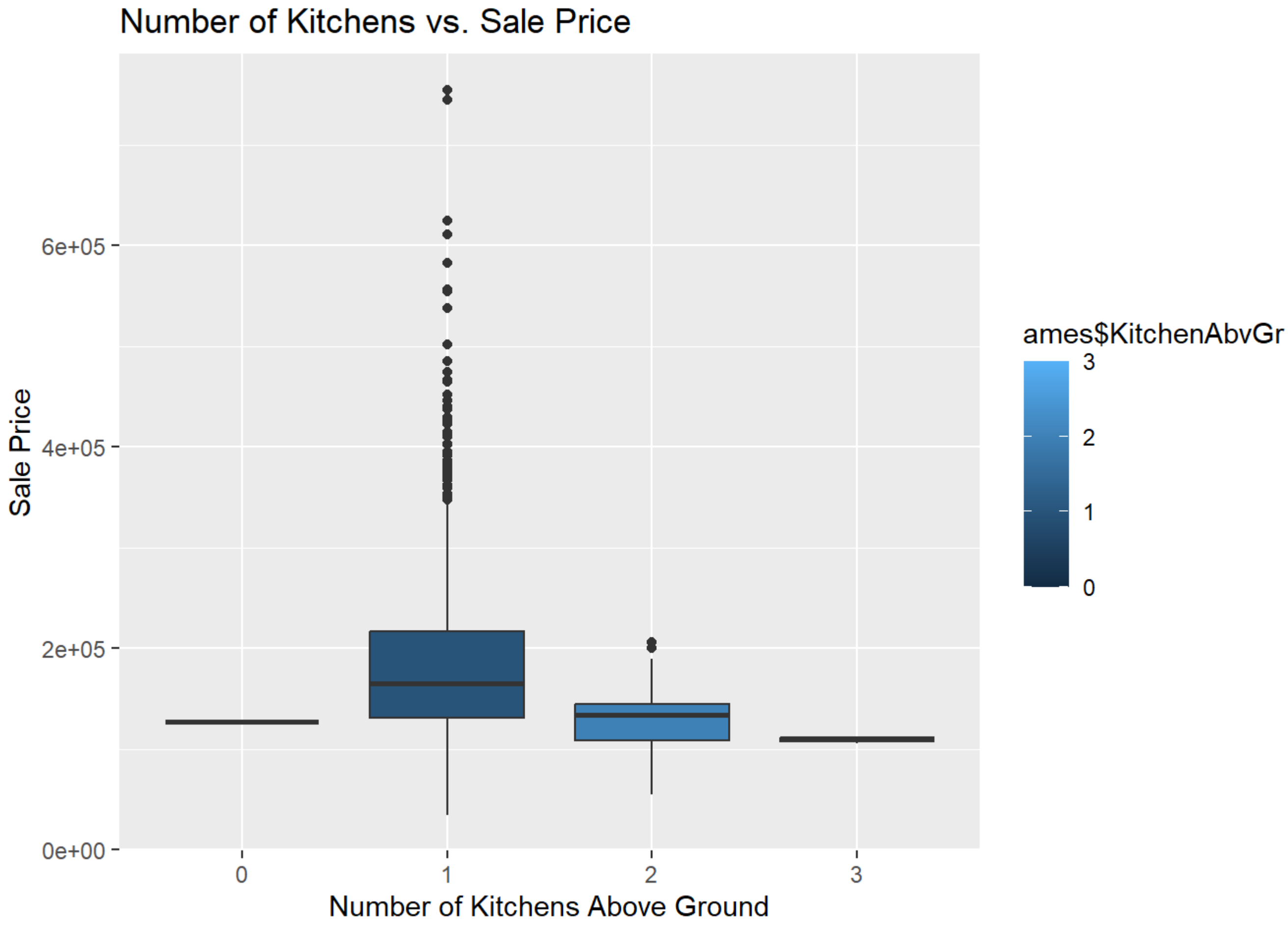
```
ggplot(ames, aes(x=factor(OverallQual), y=SalePrice, fill=ames$OverallQual)) +
  geom_boxplot()+
  ggtitle("Overall Quality vs. Sale Price")+
  xlab("Overall Quality")+
  ylab("Sale Price")
```

```
## Warning: Use of `ames$OverallQual` is discouraged.
## ℹ Use `OverallQual` instead.
```



```
ggplot(ames, aes(x=factor(KitchenAbvGr), y=SalePrice, fill=ames$KitchenAbvGr)) +
  geom_boxplot()+
  ggtitle("Number of Kitchens vs. Sale Price")+
  xlab("Number of Kitchens Above Ground")+
  ylab("Sale Price")
```

```
## Warning: Use of `ames$KitchenAbvGr` is discouraged.
## ℹ Use `KitchenAbvGr` instead.
```



The column with the highest correlation is overall quality. According to the data description:

OverallQual: Rates the overall material and finish of the house

```
       10	Very Excellent
       9	Excellent
       8	Very Good
       7	Good
       6	Above Average
       5	Average
       4	Below Average
       3	Fair
       2	Poor
       1	Very Poor
```

It is somewhat difficult to understand how this is different from OverallCond, which has similar values.

There is a clear positive relationship between overall quality and sale price, although it looks like potentially an exponential relationship rather than a linear one. For example, the minimum “non-outlier” (Q1 - 1.5*IQR) home with quality 10 (Very Excellent) sells for about the same price as the median home with quality 6 (Above Average).

The column with the most negative correlation is the number of kitchens above ground. According to the data description:

KitchenAbvGr: Kitchens above grade

From the plot, it is clear that almost all houses have 1 or 2 kitchens above grade, although there are some with 0 or 3.

Somewhat similar to the earlier OverallCond discussion, it seems that more kitchens are associated with lower price, which is somewhat counterintuitive. Essentially all of the houses with 2 kitchens sold for less than $200,000, whereas homes with 1 kitchen sometimes sold for much more.

One thing we might want to investigate is what kinds of homes have two kitchens. Are they also homes with low quality, possibly student housing at Iowa State University?

5

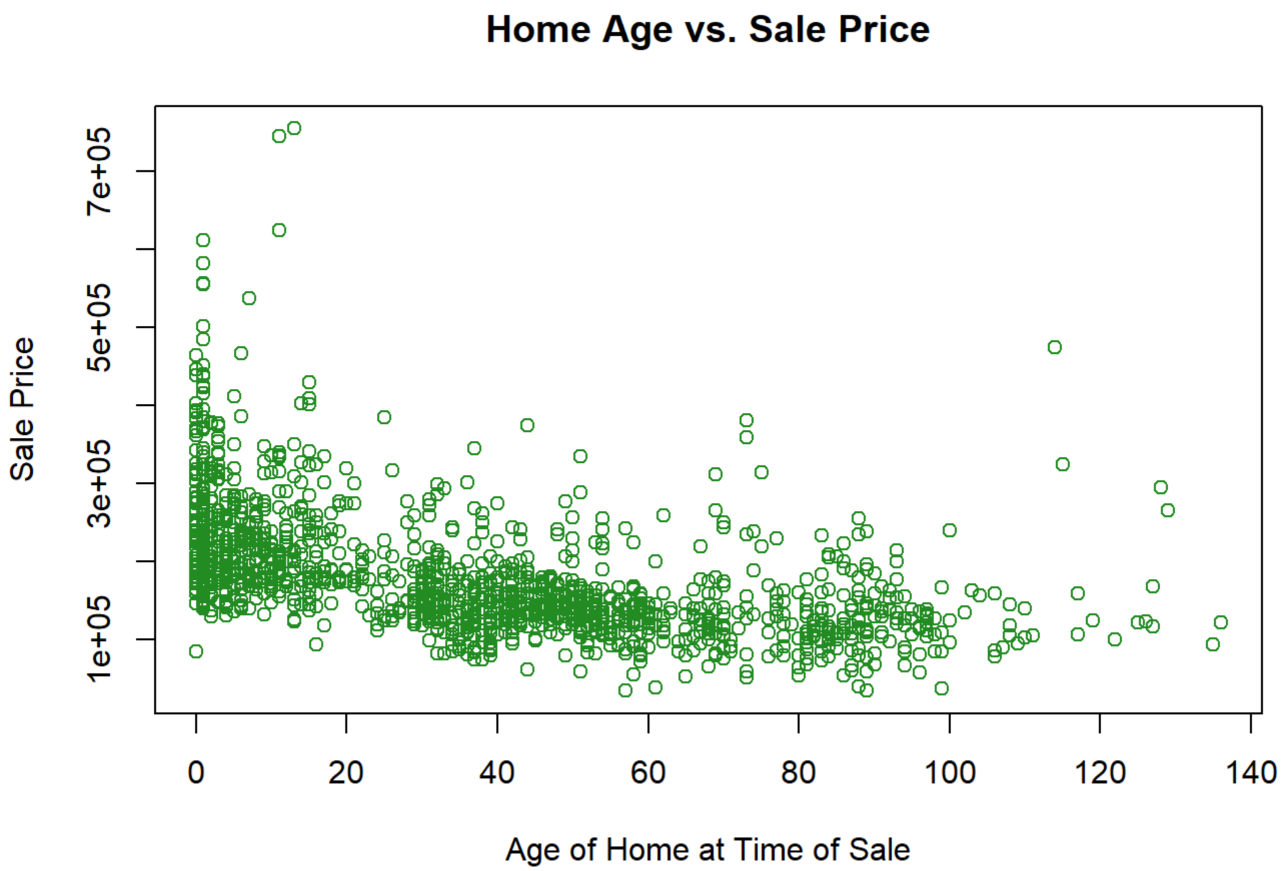
```
table(YrSold)
```

```
## YrSold
## 2006 2007 2008 2009 2010 
##  314  329  304  338  175
```

```
ames$Age = YrSold - YearBuilt
attach(ames)
```

```
## The following objects are masked from ames (pos = 4):
## 
##     Alley, BedroomAbvGr, BldgType, BsmtCond, BsmtExposure, BsmtFinSF1,
##     BsmtFinSF2, BsmtFinType1, BsmtFinType2, BsmtFullBath, BsmtHalfBath,
##     BsmtQual, BsmtUnfSF, CentralAir, Condition1, Condition2,
##     Electrical, EnclosedPorch, ExterCond, Exterior1st, Exterior2nd,
##     ExterQual, Fence, FireplaceQu, Fireplaces, Foundation, FullBath,
##     Functional, GarageArea, GarageCars, GarageCond, GarageFinish,
##     GarageQual, GarageType, GarageYrBlt, GrLivArea, HalfBath, Heating,
##     HeatingQC, HouseStyle, Id, KitchenAbvGr, KitchenQual, LandContour,
##     LandSlope, LotArea, LotConfig, LotFrontage, LotShape, LowQualFinSF,
##     MasVnrArea, MasVnrType, MiscFeature, MiscVal, MoSold, MSSubClass,
##     MSZoning, Neighborhood, OpenPorchSF, OverallCond, OverallQual,
##     PavedDrive, PoolArea, PoolQC, RoofMatl, RoofStyle, SaleCondition,
##     SalePrice, SaleType, ScreenPorch, Street, TotalBsmtSF,
##     TotRmsAbvGrd, Utilities, WoodDeckSF, X1stFlrSF, X2ndFlrSF,
##     X3SsnPorch, YearBuilt, YearRemodAdd, YrSold
```

```
plot(Age,SalePrice,xlab="Age of Home at Time of Sale", ylab="Sale Price", col="forestgreen")
title("Home Age vs. Sale Price")
```



In general, newer houses appear to be more valuable, with value increasing as homes age. Interestingly the variance seems to increase once the home age goes over 100 years, with several above-average sale prices and fewer home sales in general.

We are also seeing potential housing booms and busts over the past decades, indicated by e.g. relatively few 20-year-old houses compared to 25-year-old houses being sold. We might find something interesting if we investigae this further.