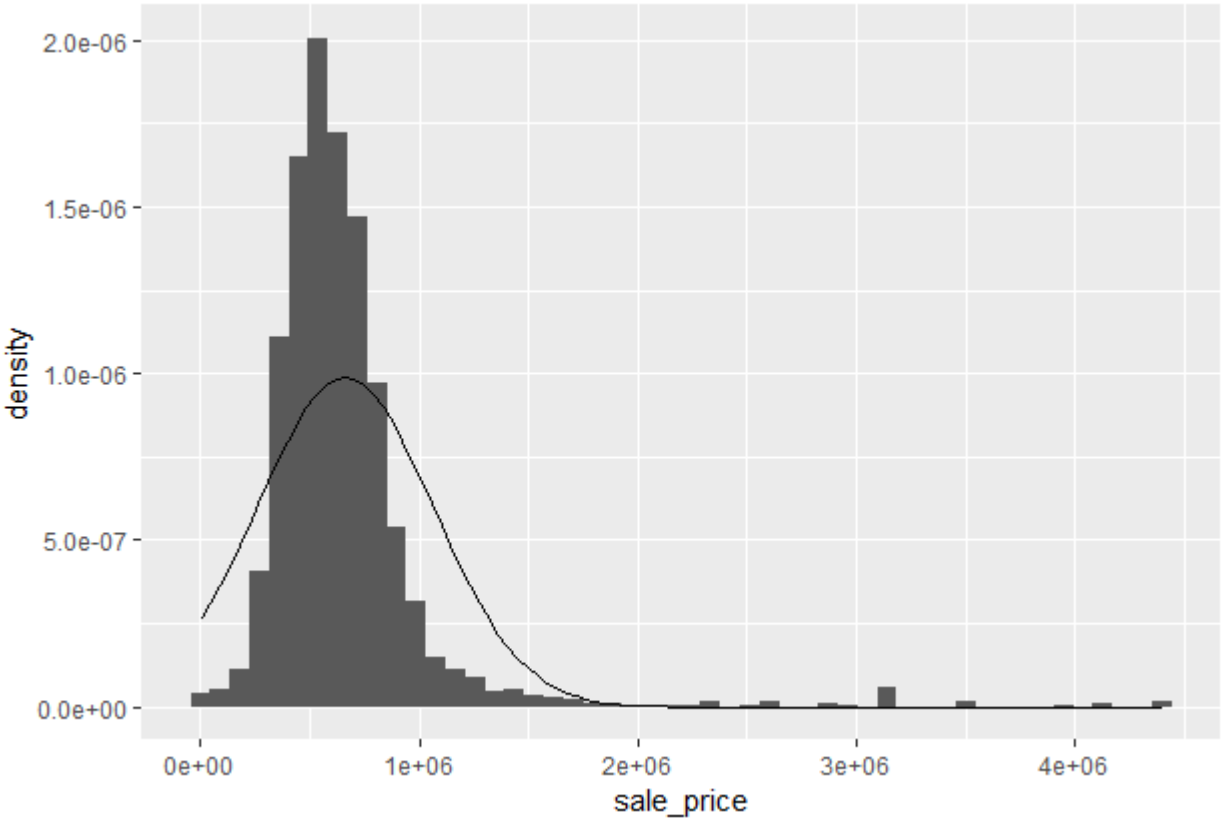
Housing Data

```
library("readxl")
library("pastecs")
library("ggplot2")
library("plyr")

housing_data <- read_xlsx("data\\week-7-housing.xlsx")
head(housing_data)

colnames(housing_data)[1:2] <- c("sale_date", "sale_price")

#Create at least two new variables
housing_data$row_num <- seq.int(nrow(housing_data))

sale_year <- format(housing_data$sale_date, format = "%Y")
sale_year <- matrix(sale_year)

sale_year <- apply(sale_year, 2, as.numeric)  # Use the apply function on a
variable in your dataset
sale_year <- sale_year[,1]
housing_data$sale_year <- sale_year


#Use the aggregate function on a variable in your dataset
aggregate(square_feet_total_living ~ year_built, housing_data, median)

#Use the plyr function on a variable in your dataset

sum_baths <- function(house_data){
  c(total_baths = house_data$bath_full_count +
                  house_data$bath_half_count * .5 +
                  house_data$bath_3qtr_count * .75)
}

total_baths <- ddply(housing_data, "row_num", sum_baths)
housing_data$total_baths <- total_baths$total_baths

#Check the distributions of the data
stat.desc(housing_data$sale_price[1:5000], basic=FALSE, norm=TRUE)
ggplot(housing_data, aes(sale_price)) + geom_histogram(bins=50,
aes(y=..density..)) +
  stat_function(fun=dnorm, args=list(mean=mean(housing_data$sale_price, na.rm
= TRUE),
                                     sd=sd(housing_data$sale_price,
na.rm=TRUE)))

#Indentify if there are any outliers
# There are some outliers in the sales price data.
# Specifically the prices exceeding two million dollars.
```


2.0e-06
1.5e-06
1.0e-06
5.0e-07
0.0e+00
density
0e+00
1e+06
2e+06
3e+06
4e+06
sale_price