Week 2 Quiz 2

Christopher Graham

October 3, 2015

1. Selection and summary statistics: In the notebook we covered in the module, we discovered which neighborhood (zip code) of Seattle had the highest average house sale price. Now, take the sales data, select only the houses with this zip code, and compute the average price. Save this result to answer the quiz at the end.

```
sales <- read.csv('home_data.csv', stringsAsFactors = F)
require(lubridate)
```

```
## Loading required package: lubridate
```

```
sales$date <- ymd(substr(sales$date, 1, 8))
sales$zipcode <- as.factor(sales$zipcode)
zip_avg <- tapply(sales$price, sales$zipcode, mean)
top_zip_avg <- zip_avg[zip_avg==max(zip_avg)]
top_zip_avg
```

```
##   98039
## 2160607
```

2. Filtering data: One of the key features we used in our model was the number of square feet of living space ('sqft_living') in the house. For this part, we are going to use the idea of filtering (selecting) data.

In particular, we are going to use logical filters to select rows of an SFrame.

Using such filters, first select the houses that have 'sqft_living' higher than 2000 sqft but no larger than 4000 sqft.

What fraction of the all houses have 'sqft_living' in this range? Save this result to answer the quiz at the end.

```
len1 <- nrow(sales[sales$sqft_living > 2000 & sales$sqft_living <= 4000,])
len1 / nrow(sales)
```

```
## [1] 0.4218757
```

3. Building a regression model with several more features: In the sample notebook, we built two regression models to predict house prices, one using just 'sqft_living' and the other one using a few more features, we called this set my_features

Compute the RMSE (root mean squared error) on the test_data for the model using just my_features, and for the one using advanced_features.

```r
set.seed(0)
train_crit <- sample(nrow(sales), floor(nrow(sales) * 0.8))
train_set <- sales[train_crit,]
test_set <- sales[-train_crit,]

model1 <- lm(price ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors +
             zipcode, data=train_set)
test_pred <- predict(model1, test_set)
test_price <- test_set$price
SS_residual <- sum((test_price - test_pred)^2)
mod1_rmse <- sqrt(SS_residual/nrow(test_set))

model2 <- lm(price ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors +
             zipcode + condition + grade + waterfront + view + sqft_above +
             sqft_basement + yr_built + yr_renovated + lat + long +
             sqft_living15 + sqft_lot15, data=train_set)
test_pred2 <- predict(model2, test_set)
```

```
## Warning in predict.lm(model2, test_set): prediction from a rank-deficient
## fit may be misleading
```

```r
SS_resid2 <- sum((test_price - test_pred2)^2)
mod2_rmse <- sqrt(SS_resid2/nrow(test_set))
mod1_rmse - mod2_rmse
```

```
## [1] 21904.77
```