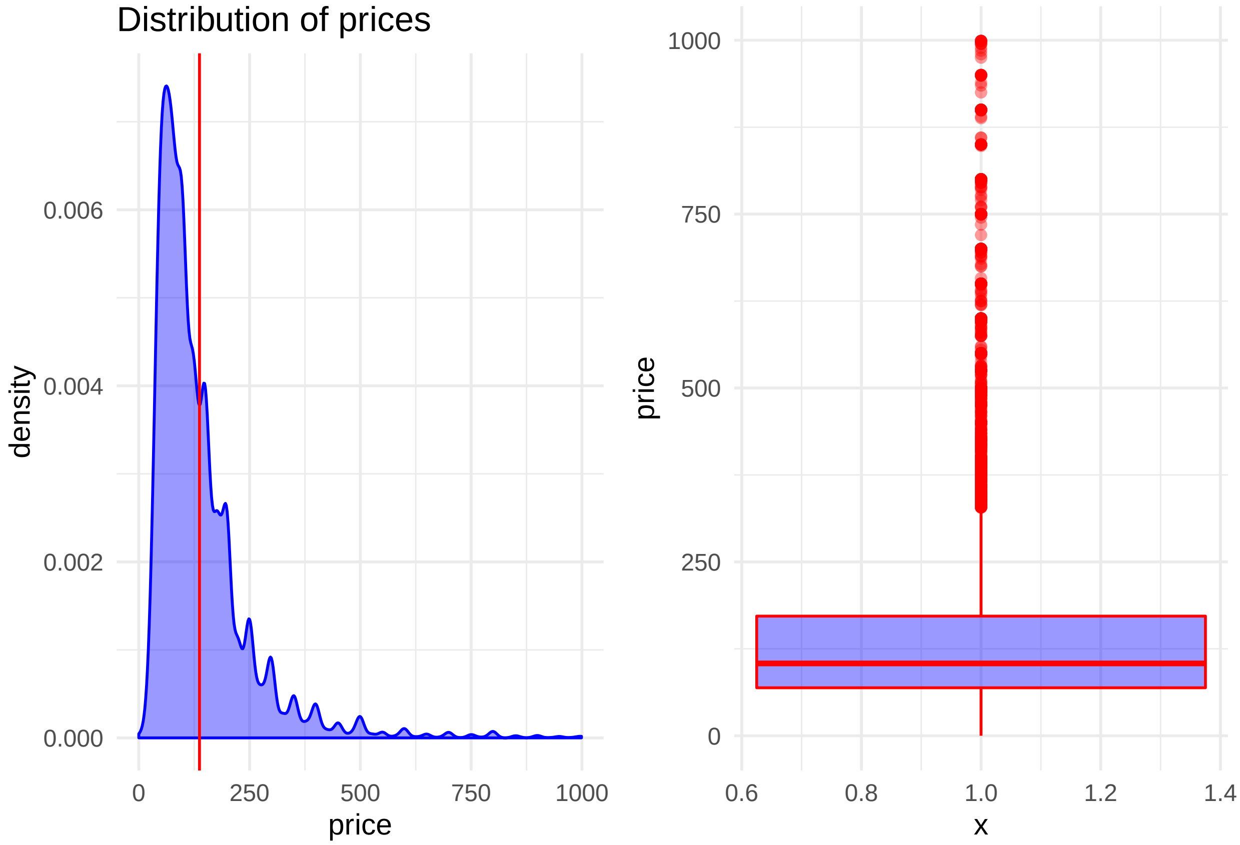
Predicting Airbnb Prices

Charlie Mei (cm3947)

1. Understanding the distribution of Airbnb prices

Airbnb prices in our dataset are right-skewed, indicating that there some listings on Airbnb are extremely expensive. There also appears to be a significant number of outliers in the data. Additionally, there are instances where the price is zero. These have all been removed before conducting further analysis.

```r
# Density plot
p1 <- ggplot(train, aes(price)) +
    geom_density(col = "blue", fill = "blue", alpha = 0.4) +
    geom_vline(xintercept = mean(mydata$price), col = "red") +
    labs(title = "Distribution of prices") +
    theme_minimal()

# Boxplot to identify outliers
p2 <- ggplot(train, aes(1, price)) +
    geom_boxplot(col = "red", fill = "blue", alpha = 0.4) +
    theme_minimal()

grid.arrange(p1, p2, ncol = 2)
```



```r
# Create LB (lower bound) and UB (upper bound) to remove outliers, also remove price = 0
LB <- quantile(train$price, probs = 0.25 ) - 1.5*IQR(train$price)
```

```
UB <- quantile(train$price, probs = 0.75) + 1.5*IQR(train$price)

train <- train %>% filter(price >= LB & price <= UB, price != 0)
```

2. Understanding the factors that affect Airbnb prices

There are around 90 potential variables that influence Airbnb prices. Many of these contain text data which could potentially contain useful features for prediction. A number of steps have been taken to first understand and prepare our data for modeling

Understanding missingness of numeric data

Let's first consider just the numeric data in our dataset and the extent of missingness.

```
# Tabularize counts of missingness
tab.missing <- function(df){
    locs <- lapply(df, is.numeric) %>% unlist()
    ndata <- df[, locs] %>% select(-id, -price, -weekly_price, -monthly_price)
    lapply(ndata, function(x){sum(is.na(x))}) %>% unlist()
}

tab.missing(train)
```

```
##                          host_listings_count
##                                            1
##                    host_total_listings_count
##                                            1
##                                 accommodates
##                                            0
##                                    bathrooms
##                                            0
##                                     bedrooms
##                                            0
##                                         beds
##                                           13
##                                  square_feet
##                                        27691
##                             security_deposit
##                                         9816
##                                 cleaning_fee
##                                         4640
##                              guests_included
##                                            0
##                                 extra_people
##                                            0
##                               minimum_nights
##                                            0
##                               maximum_nights
##                                            0
##                       minimum_minimum_nights
##                                            0
##                       maximum_minimum_nights
##                                            0
##                       minimum_maximum_nights
##                                            0
##                       maximum_maximum_nights
##                                            0
##                       minimum_nights_avg_ntm
##                                            0
##                       maximum_nights_avg_ntm
##                                            0
##                              availability_30
##                                            0
##                              availability_60
##                                            0
##                              availability_90
##                                            0
##                             availability_365
##                                            0
##                            number_of_reviews
##                                            0
##                        number_of_reviews_ltm
##                                            0
##                         review_scores_rating
##                                            0
##                       review_scores_accuracy
##                                            0
##                    review_scores_cleanliness
##                                            0
##                        review_scores_checkin
##                                            0
##                  review_scores_communication
##                                            0
##                       review_scores_location
##                                            0
##                          review_scores_value
##                                            0
##               calculated_host_listings_count
##                                            0
##  calculated_host_listings_count_entire_homes
##                                            0
## calculated_host_listings_count_private_rooms
##                                            0
##  calculated_host_listings_count_shared_rooms
##                                            0
##                            reviews_per_month
##                                            2
```

Some obervations:

- There are a small number of missing values (less than 10) for beds, reviews per month, and host listings counts. The missing values have been removed for further analysis.
- Square feet has over 29,000 missing values and does not provide any additional influence to Airbnb prices.
- Needing to pay a security deposit or cleaning fee may be significant and have been recoded to factor variables for further analysis.

```
train <- train %>%
    filter(!is.na(beds), !is.na(reviews_per_month), !is.na(host_listings_count), !is.na(host_total_listi
    mutate(security_deposit2 = ifelse(security_deposit > 0, 1, 0),
           cleaning_fee2 = ifelse(cleaning_fee > 0, 1, 0)) %>%
    select(-security_deposit, -cleaning_fee, -square_feet)

train$security_deposit2[is.na(train$security_deposit2)] <- 0
train$cleaning_fee2[is.na(train$cleaning_fee2)] <- 0

# Make corresponding additional feature additions to scoringdata and test data
scoring <- scoring %>%
    mutate(security_deposit2 = ifelse(security_deposit > 0, 1, 0),
           cleaning_fee2 = ifelse(cleaning_fee > 0, 1, 0))

scoring$security_deposit2[is.na(scoring$security_deposit2)] <- 0
scoring$cleaning_fee2[is.na(scoring$cleaning_fee2)] <- 0

test <- test %>%
    mutate(security_deposit2 = ifelse(security_deposit > 0, 1, 0),
           cleaning_fee2 = ifelse(cleaning_fee > 0, 1, 0))

test$security_deposit2[is.na(test$security_deposit2)] <- 0
test$cleaning_fee2[is.na(test$cleaning_fee2)] <- 0
```

Understanding correlations in the numeric data

There are a number of numerical variables that are quite similar to one another which may result in multicollinearity concerns if used together for modeling. These relate to:

- total listings by host
- the various variables for maximum and minimum nights are highly correlated with each other
- availability over 30, 60, 90 days are all highly correlated with each other
- number of reviews
- review scores

```
train <- train %>%
    select(-host_total_listings_count, -minimum_minimum_nights, -minimum_nights, -maximum_nights,
           -maximum_minimum_nights, -minimum_maximum_nights, -maximum_maximum_nights, -availability_30,
          -availability_60, -availability_90, -number_of_reviews_ltm)
```

3. Model selection

Let's first create the final dataset for modeling and see the results of modeling

```
# mdata is a placeholder atm with just numeric variables with no NAs
mdata <- train

# Placeholder to just get numeric variables
locs <- lapply(mdata, is.numeric) %>% unlist()
mdata <- mdata[, locs] %>% select(-contains("_price"))
```

Modeling Framework

Step 1: Choose the model that results in the best RMSE first.

1. Forward stepwise model
2. Lasso regression
3. Regression tree

4. Random forest
5. Gradient boosting

   Step 2: Refine best model for making predictions

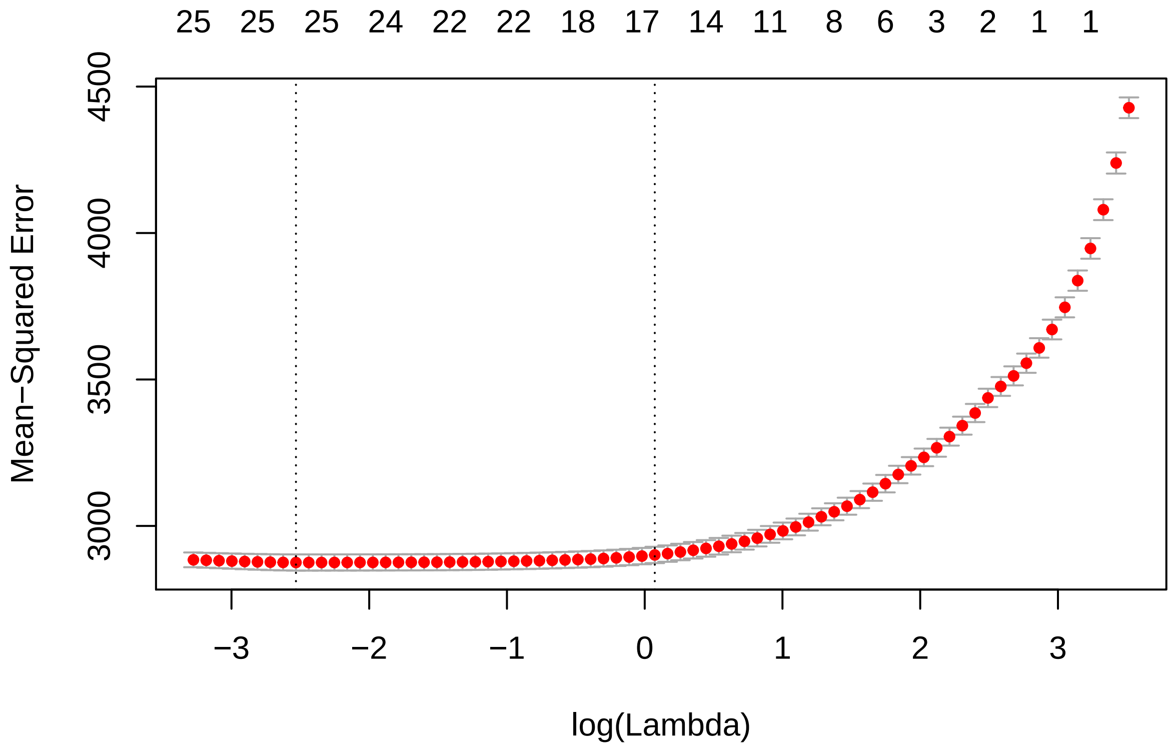
Forward stepwise model outputs

```
# Forward stepwise model
start_model <- lm(price ~ 1, mdata)
end_model <- lm(price ~ ., mdata)

model <- step(start_model,
              scope = list(upper = end_model, lower = start_model),
              direction = "forward",
              trace = F)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ accommodates + review_scores_location +
##     calculated_host_listings_count_private_rooms + reviews_per_month +
##     review_scores_value + review_scores_cleanliness + guests_included +
##     calculated_host_listings_count_entire_homes + calculated_host_listings_count_shared_rooms +
##     cleaning_fee2 + review_scores_rating + review_scores_checkin +
##     security_deposit2 + id + beds + bedrooms + bathrooms + number_of_reviews +
##     review_scores_communication + extra_people + minimum_nights_avg_ntm +
##     availability_365, data = mdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -279.73  -36.44  -10.97   27.45  267.59
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                  -4.320e+01  5.356e+00  -8.065
## accommodates                                  1.826e+01  3.334e-01  54.763
## review_scores_location                        1.852e+01  4.677e-01  39.601
## calculated_host_listings_count_private_rooms -2.036e+00  9.587e-02 -21.235
## reviews_per_month                            -3.190e+00  2.599e-01 -12.271
## review_scores_value                          -1.251e+01  5.616e-01 -22.269
## review_scores_cleanliness                     3.598e+00  4.298e-01   8.370
## guests_included                               5.203e+00  4.026e-01  12.921
## calculated_host_listings_count_entire_homes   5.365e-01  3.615e-02  14.841
## calculated_host_listings_count_shared_rooms  -3.506e+00  3.155e-01 -11.112
## cleaning_fee2                                 6.492e+00  8.310e-01   7.812
## review_scores_rating                          7.288e-01  6.965e-02  10.464
## review_scores_checkin                        -3.358e+00  5.910e-01  -5.681
## security_deposit2                             2.951e+00  6.969e-01   4.234
## id                                           -2.024e-07  4.053e-08  -4.992
## beds                                         -2.408e+00  5.243e-01  -4.592
## bedrooms                                      3.234e+00  6.583e-01   4.914
## bathrooms                                    -2.968e+00  9.292e-01  -3.194
## number_of_reviews                            -3.098e-02  1.032e-02  -3.002
## review_scores_communication                  -1.766e+00  6.420e-01  -2.750
## extra_people                                  3.952e-02  1.453e-02   2.720
## minimum_nights_avg_ntm                       -2.270e-02  1.087e-02  -2.089
## availability_365                             -4.160e-03  2.534e-03  -1.642
##                                              Pr(>|t|)    
## (Intercept)                                  7.60e-16 ***
## accommodates                                  < 2e-16 ***
## review_scores_location                        < 2e-16 ***
## calculated_host_listings_count_private_rooms  < 2e-16 ***
## reviews_per_month                             < 2e-16 ***
## review_scores_value                           < 2e-16 ***
## review_scores_cleanliness                     < 2e-16 ***
## guests_included                               < 2e-16 ***
## calculated_host_listings_count_entire_homes   < 2e-16 ***
## calculated_host_listings_count_shared_rooms   < 2e-16 ***
## cleaning_fee2                                5.84e-15 ***
## review_scores_rating                          < 2e-16 ***
## review_scores_checkin                        1.35e-08 ***
## security_deposit2                            2.30e-05 ***
## id                                           6.00e-07 ***
## beds                                         4.41e-06 ***
## bedrooms                                     8.99e-07 ***
## bathrooms                                     0.00141 ** 
## number_of_reviews                             0.00268 ** 
## review_scores_communication                   0.00596 ** 
## extra_people                                  0.00654 ** 
## minimum_nights_avg_ntm                        0.03676 *  
## availability_365                              0.10063    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 53.57 on 27923 degrees of freedom
## Multiple R-squared:  0.3527, Adjusted R-squared:  0.3522 
## F-statistic: 691.6 on 22 and 27923 DF,  p-value: < 2.2e-16
```

Lasso model outputs

```
# Lasso model
X <- model.matrix(price ~ ., mdata)
y <- mdata$price

set.seed(1431)
cv.lasso <- cv.glmnet(X, y, alpha = 1)
plot(cv.lasso)
```



```r
coef(cv.lasso, s = cv.lasso$lambda.min)
```

```
## 28 x 1 sparse Matrix of class "dgCMatrix"
##                                                          1
## (Intercept)                                  -4.405788e+01
## (Intercept)                                   .           
## id                                           -1.892210e-07
## host_listings_count                           3.859036e-03
## accommodates                                  1.813568e+01
## bathrooms                                    -2.781067e+00
## bedrooms                                      2.978001e+00
## beds                                         -2.056520e+00
## guests_included                               5.153349e+00
## extra_people                                  3.697629e-02
## minimum_nights_avg_ntm                       -2.146556e-02
## maximum_nights_avg_ntm                        7.073410e-09
## availability_365                             -3.622961e-03
## number_of_reviews                            -2.792780e-02
## review_scores_rating                          6.944142e-01
## review_scores_accuracy                       -2.596453e-01
## review_scores_cleanliness                     3.535041e+00
## review_scores_checkin                        -3.176918e+00
## review_scores_communication                  -1.496668e+00
## review_scores_location                        1.836639e+01
## review_scores_value                          -1.207461e+01
## calculated_host_listings_count                .           
## calculated_host_listings_count_entire_homes   5.224319e-01
## calculated_host_listings_count_private_rooms -2.028138e+00
## calculated_host_listings_count_shared_rooms  -3.489482e+00
## reviews_per_month                            -3.203559e+00
## security_deposit2                             2.864221e+00
## cleaning_fee2                                 6.368585e+00
```

4. Model Assessment

```
# Calculate RMSEs
pred <- predict(model)

rmse <- mean((train$price - pred)^2) %>% sqrt()

test_pred <- predict(model, newdata = test)
t_rmse <- mean((test$price - test_pred)^2) %>% sqrt()

paste("Training RMSE:", rmse)
```

```
## [1] "Training RMSE: 53.5519085806217"
```

```
paste("Test RMSE:", t_rmse)
```

```
## [1] "Test RMSE: NA"
```

5. Save modeling outputs

```
pred <- predict(model, newdata = scoring)
submission <- data.frame(id = scoring$id, price = pred)
write_csv(submission, "submission.csv")
```

6. Conclusion