AML Project: AirBnB (New York) Dataset

Radhika Rajeevan

Sunit Nair

12/12/2019

The project aims to analyze the AirBnB data set available at OpenDataSoft and apply techniques learned as part of the course and some additional methods to predict the price that a new property should expect to charge based on its features. The data for New York City was extracted and used for the purpose of this project. Analysis, filtering, and extraction of categorical variables from the data set was done using Python. The combined, cleansed data file was loaded into a database (MySQL) from where it is read for regression. Extraction of derived columns and running the regression models was performed in R and the following methods were used- Linear Regression, Tree, Ridge/Lasso Regression, and XGBoost. Validation set and k-fold cross validation techniques are used. The output of the predicted prices (for data without prices in data set) is written to a CSV file at the end.

The required packages need to installed. Please uncomment any packages that are not installed to make sure the program runs successfully.

```r
print("Installing required packages")

## [1] "Installing required packages"

#install.packages("RMySQL")
#install.packages("MASS")
#install.packages("tidyverse")
#install.packages("glmnet")
#install.packages("tree")
#install.packages("xgboost")
#install.packages("caret")

library(RMySQL)

## Loading required package: DBI

library(MASS)
library(tidyverse)

## -- Attaching packages ---------------------------------------------------
--------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------
---------------------------------- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```r
library(glmnet)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 3.0-2
```

```r
library(tree)
```

```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

```r
library(xgboost)
```

```
##
## Attaching package: 'xgboost'

## The following object is masked from 'package:dplyr':
##
##     slice
```

```r
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

The data is available as a CSV file and was loaded into MySQL database for this project. The user can choose to read from either source. For the purpose of this markdown, we read from the CSV file.

The regression models can be run with price or log of price (which has distribution closer to normal). The variable log_regression controls this feature and is set to FALSE for the purpose of this markdown.

```r
print("NOTE: The data set is available as a cleaned CSV file and was loaded
into MySQL database for this project")
```

```
## [1] "NOTE: The data set is available as a cleaned CSV file and was loaded
into MySQL database for this project"
```

```
read_from <- 0
r_user <- "r_user"
r_password <- "r_password"
db_name <- "aml"
log_regression <- FALSE
if(as.integer(read_from)==1){
  print(paste("Connecting to database with user",r_user))
  mydb <- dbConnect(MySQL(), user=r_user, password=r_password,
dbname=db_name, host="localhost")
  print(paste("Showing list of tables available in schema",db_name))
  tableNames <- dbListTables(mydb)
  print(tableNames)
  print(paste("Checking columns in table",tableNames[1]))
  colNames <- dbListFields(mydb, tableNames[1])
  print(colNames)
  print(paste("Fetching all data from ",tableNames[1]))
  tableQuery <- paste("SELECT * FROM ",db_name,".",tableNames[1],sep="")
  resultSet <- dbSendQuery(mydb, tableQuery)
  airData <- fetch(resultSet,n=-1)
  dbDisconnect(mydb)
} else {
  if(as.integer(read_from)==0){
    print("Reading from CSV")
  } else{
    print("Invalid input, defaulting to reading from CSV")
  }
  airData <- read.csv("final_project.csv")
}
```

```
## [1] "Reading from CSV"
```

Dimensions of airData.

```
print("Dimensions of airData:")
```

```
## [1] "Dimensions of airData:"
```

```
print(dim(airData))
```

```
## [1] 19273   130
```

Changing boolean columns to integer (0/1) columns so as to later convert them to factors.

```
print("Converting boolean columns to integers")
```

```
## [1] "Converting boolean columns to integers"
```

```
airData[,31:130] <- lapply(airData[,31:130],as.integer)
```

Deriving new features: 1. featureCount: Number of 0/1 features provided by each listing. 2. yearsAsHost: (2019 - first year as host)

```
print("Deriving new features from data set")

## [1] "Deriving new features from data set"

print("Deriving number of features as numerical feature")

## [1] "Deriving number of features as numerical feature"

airData$featureCount <- apply(airData[,31:130],1,sum)
print("Deriving yearAsHost as numerical feature")

## [1] "Deriving yearAsHost as numerical feature"

airData$yearsAsHost <- (2019 - airData$hostYear)
```

Converting categorical columns to factors.

```
print("Converting categorical columns to factors")

## [1] "Converting categorical columns to factors"

airData[,31:130] <- lapply(airData[,31:130],as.factor)
airData$neighbourhoodCleansed <- as.factor(airData$neighbourhoodCleansed)
airData$neighbourhoodGroupCleansed <-
as.factor(airData$neighbourhoodGroupCleansed)
airData$bedType <- as.factor(airData$bedType)
airData$cancellationPolicy <- as.factor(airData$cancellationPolicy)
airData$propertyType <- as.factor(airData$propertyType)
airData$roomType <- as.factor(airData$roomType)
airData <- as.data.frame(airData)
```

Summary of airData after above manipulations.

```
print("Summary of airData:")

## [1] "Summary of airData:"

print(summary(airData))

##        id               hostYear     hostResponseHours
##  Min.   :    2515   Min.   :2008    Min.   : 1.000
##  1st Qu.: 4941141   1st Qu.:2013    1st Qu.: 1.000
##  Median : 9906178   Median :2014    Median : 1.000
##  Mean   : 9832847   Mean   :2014    Mean   : 9.231
##  3rd Qu.:14834468   3rd Qu.:2015    3rd Qu.:12.000
##  Max.   :18516103   Max.   :2017    Max.   :72.000
##
##      neighbourhoodCleansed neighbourhoodGroupCleansed  accommodates
##  Harlem         :2481      Bronx        :   15         Min.   : 1.000
##  East Village   :1828      Brooklyn     :  212         1st Qu.: 2.000
```

```
##  Upper West Side:1750      Manhattan    :18902        Median : 2.000
##  Hell's Kitchen :1557      Queens       :  140        Mean   : 2.803
##  Upper East Side:1520      Staten Island:    4        3rd Qu.: 4.000
##  Chelsea        :1056                                 Max.   :16.000
##  (Other)        :9081
##    bathrooms             bedType          bedrooms          beds
##  Min.   :0.000   Airbed       :  133   Min.   :0.000   Min.   : 0.000
##  1st Qu.:1.000   Couch        :   67   1st Qu.:1.000   1st Qu.: 1.000
##  Median :1.000   Futon        :  186   Median :1.000   Median : 1.000
##  Mean   :1.098   Pull-out Sofa:  205   Mean   :1.093   Mean   : 1.506
##  3rd Qu.:1.000   Real Bed     :18682   3rd Qu.:1.000   3rd Qu.: 2.000
##  Max.   :6.500                         Max.   :6.000   Max.   :14.000
##
##        TV              cancellationPolicy  cleaningFee
##  Min.   :0.0000   flexible       :5799     Min.   :  0.00
##  1st Qu.:0.0000   moderate       :4250     1st Qu.:  0.00
##  Median :1.0000   strict         :9220     Median : 40.00
##  Mean   :0.6886   super_strict_30:   4     Mean   : 47.86
##  3rd Qu.:1.0000                            3rd Qu.: 75.00
##  Max.   :2.0000                            Max.   :600.00
##
##   extraPeople     guestsIncluded   maximumNights    minimumNights
##  Min.   :  0.00   Min.   : 1.000   Min.   :   1.0   Min.   :   1.00
##  1st Qu.:  0.00   1st Qu.: 1.000   1st Qu.:  30.0   1st Qu.:   1.00
##  Median :  0.00   Median : 1.000   Median :1125.0   Median :   2.00
##  Mean   : 14.05   Mean   : 1.435   Mean   : 696.8   Mean   :   4.21
##  3rd Qu.: 25.00   3rd Qu.: 2.000   3rd Qu.:1125.0   3rd Qu.:   3.00
##  Max.   :300.00   Max.   :16.000   Max.   :1125.0   Max.   :1250.00
##
##      price           propertyType             roomType
##  Min.   :  0.0   Apartment  :18092   Entire home/apt:10818
##  1st Qu.: 88.0   House      :  294   Private room   : 7880
##  Median :133.0   Loft       :  264   Shared room    :  575
##  Mean   :165.1   Condominium:  191
##  3rd Qu.:200.0   Townhouse  :  178
##  Max.   :999.0   Other      :  101
##                  (Other)    :  153
##  securityDeposit numberOfReviews  reviewScoresAccuracy reviewScoresCheckin
##  Min.   :  0.0   Min.   :  0.00   Min.   : 0.000       Min.   : 0.000
##  1st Qu.:  0.0   1st Qu.:  1.00   1st Qu.: 6.000       1st Qu.: 7.000
##  Median :  0.0   Median :  5.00   Median : 9.000       Median :10.000
##  Mean   :122.9   Mean   : 16.91   Mean   : 7.212       Mean   : 7.358
##  3rd Qu.:200.0   3rd Qu.: 19.00   3rd Qu.:10.000       3rd Qu.:10.000
##  Max.   :999.0   Max.   :432.00   Max.   :10.000       Max.   :10.000
##
##  reviewScoresCleanliness reviewScoresCommunication reviewScoresLocation
##  Min.   : 0.000          Min.   : 0.000            Min.   : 0.000
##  1st Qu.: 6.000          1st Qu.: 8.000            1st Qu.: 7.000
##  Median : 9.000          Median :10.000            Median :10.000
##  Mean   : 6.957          Mean   : 7.394            Mean   : 7.265
```

```
##  3rd Qu.:10.000          3rd Qu.:10.000            3rd Qu.:10.000
##  Max.   :10.000          Max.   :10.000            Max.   :10.000
##
##  reviewScoresRating reviewScoresValue reviewsperMonth   checkIn24Hours
##  Min.   :  0.00     Min.   : 0.000    Min.   :  0.000   0:14476
##  1st Qu.: 60.00     1st Qu.: 6.000    1st Qu.:  0.060   1: 4797
##  Median : 92.00     Median : 9.000    Median :  0.420
##  Mean   : 70.64     Mean   : 7.063    Mean   :  1.028
##  3rd Qu.: 98.00     3rd Qu.:10.000    3rd Qu.:  1.460
##  Max.   :100.00     Max.   :10.000    Max.   :125.920
##
##  accessibleHeightToilet airConditioning BBQgrill  babyBath  babyMonitor
##  0:19272                0: 2732         0:19272   0:19255   0:19264
##  1:    1                1:16541         1:    1   1:   18   1:    9
##
##
##
##
##
##  babySitterRecommendations bathTub   bedLinens breakfast
##  0:19211                   0:18825   0:19265   0:17882
##  1:   62                   1:  448   1:    8   1: 1391
##
##
##
##
##
##  buzzerOrWirelessIntercom cableTV   carbonMonoxideDetector cats
##  0:9731                   0:12835   0: 8174                0:18596
##  1:9542                   1: 6438   1:11099                1:  677
##
##
##
##
##
##  changingTable childrenBooksAndToys childrenDinnerware
##  0:19253       0:19210              0:19231
##  1:   20       1:   63              1:   42
##
##
##
##
##
##  cleaningBeforeCheckout coffeemaker cookingBasics crib
##  0:19272                0:19266     0:19265       0:19237
##  1:    1                1:    7     1:    8       1:   36
##
##
##
##
```

```
##
##  dishesAndSilverware dishwasher dogs      doorman   doormanEntry dryer
##  0:19265             0:19271    0:18526   0:16544   0:18933      0:11162
##  1:    8             1:    2    1:  747   1: 2729   1:  340      1: 8111
##
##
##
##
##
##  elevator  essentials extraPillowsBlankets familyAndKidFriendly
##  0:11421   0: 3232    0:19269              0:10810
##  1: 7852   1:16041    1:    4              1: 8463
##
##
##
##
##
##  fireExtinguisher fireplaceGuards firstAidKit freeParkingOnPremises
##  0:14376          0:19261         0:14181     0:18402
##  1: 4897          1:   12         1: 5092     1:  871
##
##
##
##
##
##  freeParkingOnStreet gameConsole gardenOrBackyard railsInShowerToilet
##  0:19269             0:19220     0:19270          0:19272
##  1:    4             1:   53     1:    3          1:    1
##
##
##
##
##
##  gym       hairdryer hangers   heating   highchair hottub    hotwater
##  0:17535   0: 9072   0: 7893   0: 1215   0:19217   0:18135   0:19267
##  1: 1738   1:10201   1:11380   1:18058   1:   56   1: 1138   1:    6
##
##
##
##
##
##  indoorFireplace internet  iron     keypad    kitchen
##  0:18442         0: 5891   0:9734   0:19131   0: 1004
##  1:  831         1:13382   1:9539   1:  142   1:18269
##
##
##
##
##
##  laptopFriendlyWorkspace lockOnBedroomDoor lockbox   longTermStaysAllowed
```

```
##  0:9335                  0:15388          0:18630   0:19271
##  1:9938                  1: 3885          1:  643   1:    2
##
##
##
##
##
##  luggageDropOffAllowed microwave otherPets outletCovers oven
##  0:19271               0:19266   0:19244   0:19251      0:19264
##  1:    2               1:    7   1:   29   1:   22      1:    9
##
##
##
##
##
##  packNPlayTravelCrib pathToEntranceLitAtNight patioOrBalcony petsAllowed
##  0:19180             0:19272                  0:19270        0:16973
##  1:   93             1:    1                  1:    3        1: 2300
##
##
##
##
##
##  petsLiveOnThisProperty pool      privateBathroom privatEentrance
##  0:17756                0:18984   0:19271         0:18734
##  1: 1517                1:  289   1:    2         1:  539
##
##
##
##
##
##  privateLivingRoom refrigerator shades    safetyCard selfCheckIn shampoo
##  0:18986           0:19264      0:19088   0:17077    0:18102     0: 7196
##  1:  287           1:    9      1:  185   1: 2196    1: 1171     1:12077
##
##
##
##
##
##  smartlock smokeDetector smokingAllowed stairGates stepFreeaccess
##  0:19207   0: 4141       0:18355        0:19250    0:19271
##  1:   66   1:15132       1:  918        1:   23    1:    2
##
##
##
##
##
##  stove     suitableForEvents tableCornerGuards washer    washerDryer
##  0:19265   0:18587           0:19266           0:11093   0:19267
##  1:    8   1:  686           1:    7           1: 8180   1:    6
```

```
##
##
##
##
##
##  wheelchairAccessible wideClearanceToBed wideClearanceShowerToilet
##  0:17632              0:19272            0:19272
##  1: 1641              1:    1            1:    1
##
##
##
##
##
##  wideDoorway wideHallwayClearance windowGuards wirelessInternet
##  0:19272     0:19272              0:19153      0:  533
##  4:    1     1:    1              1:  120      1:18740
##
##
##
##
##
##  hosting_amenity_49 hosting_amenity_50 hostHasProfilePic
##  0:12997            0:11648            0:   79
##  1: 6276            1: 7625            1:19194
##
##
##
##
##
##  hostIdentityVerified hostIsSuperhost instantBookable isLocationExact
##  0: 6586              0:17569         0:15645         0: 3368
##  1:12687              1: 1704         1: 3628         1:15905
##
##
##
##
##
##  requireGuestPhoneVerification requireGuestProfilePicture  featureCount
##  0:18595                       0:18680                    Min.   : 2.00
##  1:  678                       1:  593                    1st Qu.:14.00
##                                                            Median :17.00
##                                                            Mean   :17.65
##                                                            3rd Qu.:21.00
##                                                            Max.   :51.00
##
##   yearsAsHost
##  Min.   : 2.000
##  1st Qu.: 4.000
##  Median : 5.000
##  Mean   : 5.087
```

```
##  3rd Qu.: 6.000
##  Max.   :11.000
##
```

Dropping columns that have been used to derive other columns, have repeated information or have low count which may cause issues when training and test data sets have different levels.

```
print("Dropping columns from airData that have been used to derive other
columns, have repeated information, or have low count for certain levels")

## [1] "Dropping columns from airData that have been used to derive other
columns, have repeated information, or have low count for certain levels"

dropColumns <-
c("id","hostYear","accessibleHeightToilet","BBQgrill","babyBath","babyMonitor
","babySitterRecommendations","bathTub","bedLinens","breakfast","changingTabl
e","cats","childrenBooksAndToys","childrenDinnerware","cleaningBeforeCheckout
","coffeemaker","cookingBasics","crib","dishesAndSilverware","dishwasher","do
gs","doormanEntry","extraPillowsBlankets","fireplaceGuards","freeParkingOnStr
eet","gameConsole","gardenOrBackyard","railsInShowerToilet","highchair","hotw
ater","indoorFireplace","keypad","lockbox","longTermStaysAllowed","luggageDro
pOffAllowed","microwave","otherPets","outletCovers","oven","packNPlayTravelCr
ib","pathToEntranceLitAtNight","patioOrBalcony","privateBathroom","privatEent
rance","privateLivingRoom","refrigerator","shades","safetyCard","smartlock","
stairGates","stepFreeaccess","stove","suitableForEvents","tableCornerGuards",
"washerDryer","wideClearanceToBed","wideClearanceShowerToilet","wideDoorway",
"wideHallwayClearance","windowGuards","hosting_amenity_49","hosting_amenity_5
0")
airDataClean <- airData[,!(names(airData) %in% dropColumns)]
airDataClean <- as.data.frame(airDataClean)
```

Column names in cleaned data set.

```
print("Columns in cleaned airData set")

## [1] "Columns in cleaned airData set"

print(names(airDataClean))

##  [1] "hostResponseHours"             "neighbourhoodCleansed"
##  [3] "neighbourhoodGroupCleansed"    "accommodates"
##  [5] "bathrooms"                     "bedType"
##  [7] "bedrooms"                      "beds"
##  [9] "TV"                            "cancellationPolicy"
## [11] "cleaningFee"                   "extraPeople"
## [13] "guestsIncluded"                "maximumNights"
## [15] "minimumNights"                 "price"
## [17] "propertyType"                  "roomType"
## [19] "securityDeposit"               "numberOfReviews"
## [21] "reviewScoresAccuracy"          "reviewScoresCheckin"
## [23] "reviewScoresCleanliness"       "reviewScoresCommunication"
```

```
## [25] "reviewScoresLocation"           "reviewScoresRating"
## [27] "reviewScoresValue"              "reviewsperMonth"
## [29] "checkIn24Hours"                 "airConditioning"
## [31] "buzzerOrWirelessIntercom"       "cableTV"
## [33] "carbonMonoxideDetector"         "doorman"
## [35] "dryer"                          "elevator"
## [37] "essentials"                     "familyAndKidFriendly"
## [39] "fireExtinguisher"               "firstAidKit"
## [41] "freeParkingOnPremises"          "gym"
## [43] "hairdryer"                      "hangers"
## [45] "heating"                        "hottub"
## [47] "internet"                       "iron"
## [49] "kitchen"                        "laptopFriendlyWorkspace"
## [51] "lockOnBedroomDoor"              "petsAllowed"
## [53] "petsLiveOnThisProperty"         "pool"
## [55] "selfCheckIn"                    "shampoo"
## [57] "smokeDetector"                  "smokingAllowed"
## [59] "washer"                         "wheelchairAccessible"
## [61] "wirelessInternet"               "hostHasProfilePic"
## [63] "hostIdentityVerified"           "hostIsSuperhost"
## [65] "instantBookable"                "isLocationExact"
## [67] "requireGuestPhoneVerification" "requireGuestProfilePicture"
## [69] "featureCount"                   "yearsAsHost"
```

Clubbing infrequent neighborhoods into ‘Other’ cateory so as to reduce the probability of issues caused due to different levels in training and test data sets.

```r
print("Clubbing infrequent neighborhoods into 'Other' category")

## [1] "Clubbing infrequent neighborhoods into 'Other' category"

neighborhoodList <- unique(airDataClean$neighbourhoodCleansed)

print("Converting column as character to replace with Other")

## [1] "Converting column as character to replace with Other"

airDataClean$neighbourhoodCleansed <-
as.character(airDataClean$neighbourhoodCleansed)
print("Neighborhood summary before cleansing")

## [1] "Neighborhood summary before cleansing"

print(summary(airDataClean$neighbourhoodCleansed))

##    Length     Class      Mode
##     19273 character character

print("Clubbing infrequent neighborhoods into 'Other'")

## [1] "Clubbing infrequent neighborhoods into 'Other'"
```

```r
for(n in neighborhoodList)
{
  tempList <- airDataClean$neighbourhoodCleansed == n
  rcount <- length(which(tempList))
  print(paste(n,":",rcount))
  if(rcount <= 300){
    print(paste("Changing ",n," to Other"))
    airDataClean$neighbourhoodCleansed[tempList] <- "Other"
  }
}
```

```
## [1] "Long Island City : 28"
## [1] "Changing  Long Island City  to Other"
## [1] "Lower East Side : 947"
## [1] "Midtown : 946"
## [1] "Kips Bay : 409"
## [1] "Little Italy : 89"
## [1] "Changing  Little Italy  to Other"
## [1] "Murray Hill : 253"
## [1] "Changing  Murray Hill  to Other"
## [1] "Morningside Heights : 388"
## [1] "NoHo : 77"
## [1] "Changing  NoHo  to Other"
## [1] "Nolita : 284"
## [1] "Changing  Nolita  to Other"
## [1] "Roosevelt Island : 57"
## [1] "Changing  Roosevelt Island  to Other"
## [1] "SoHo : 350"
## [1] "Stuyvesant Town : 38"
## [1] "Changing  Stuyvesant Town  to Other"
## [1] "Hell's Kitchen : 1557"
## [1] "Greenwich Village : 373"
## [1] "Harlem : 2481"
## [1] "Inwood : 223"
## [1] "Changing  Inwood  to Other"
## [1] "East Harlem : 1038"
## [1] "East Village : 1828"
## [1] "Financial District : 387"
## [1] "Flatiron District : 88"
## [1] "Changing  Flatiron District  to Other"
## [1] "Flatbush : 6"
## [1] "Changing  Flatbush  to Other"
## [1] "Gramercy : 302"
## [1] "Upper West Side : 1750"
## [1] "Theater District : 171"
## [1] "Changing  Theater District  to Other"
## [1] "Tribeca : 155"
## [1] "Changing  Tribeca  to Other"
## [1] "Two Bridges : 54"
## [1] "Changing  Two Bridges  to Other"
```

```
## [1] "Upper East Side : 1520"
## [1] "Washington Heights : 851"
## [1] "West Village : 769"
## [1] "Williamsburg : 62"
## [1] "Changing  Williamsburg  to Other"
## [1] "Ditmars Steinway : 15"
## [1] "Changing  Ditmars Steinway  to Other"
## [1] "Astoria : 37"
## [1] "Changing  Astoria  to Other"
## [1] "Battery Park City : 63"
## [1] "Changing  Battery Park City  to Other"
## [1] "Bedford-Stuyvesant : 31"
## [1] "Changing  Bedford-Stuyvesant  to Other"
## [1] "Prospect-Lefferts Gardens : 9"
## [1] "Changing  Prospect-Lefferts Gardens  to Other"
## [1] "Bushwick : 30"
## [1] "Changing  Bushwick  to Other"
## [1] "Chinatown : 354"
## [1] "Chelsea : 1056"
## [1] "Civic Center : 41"
## [1] "Changing  Civic Center  to Other"
## [1] "Concourse : 2"
## [1] "Changing  Concourse  to Other"
## [1] "Jamaica Hills : 1"
## [1] "Changing  Jamaica Hills  to Other"
## [1] "Greenpoint : 6"
## [1] "Changing  Greenpoint  to Other"
## [1] "Crown Heights : 13"
## [1] "Changing  Crown Heights  to Other"
## [1] "East Flatbush : 3"
## [1] "Changing  East Flatbush  to Other"
## [1] "Elmhurst : 8"
## [1] "Changing  Elmhurst  to Other"
## [1] "Sunset Park : 4"
## [1] "Changing  Sunset Park  to Other"
## [1] "Mariners Harbor : 1"
## [1] "Changing  Mariners Harbor  to Other"
## [1] "Sunnyside : 13"
## [1] "Changing  Sunnyside  to Other"
## [1] "Ridgewood : 18"
## [1] "Changing  Ridgewood  to Other"
## [1] "Far Rockaway : 1"
## [1] "Changing  Far Rockaway  to Other"
## [1] "Fort Greene : 9"
## [1] "Changing  Fort Greene  to Other"
## [1] "Fresh Meadows : 1"
## [1] "Changing  Fresh Meadows  to Other"
## [1] "Gravesend : 2"
## [1] "Changing  Gravesend  to Other"
## [1] "Kensington : 3"
```

```
## [1] "Changing  Kensington  to Other"
## [1] "Rockaway Beach : 1"
## [1] "Changing  Rockaway Beach  to Other"
## [1] "Marble Hill : 3"
## [1] "Changing  Marble Hill  to Other"
## [1] "Port Morris : 1"
## [1] "Changing  Port Morris  to Other"
## [1] "Kingsbridge : 2"
## [1] "Changing  Kingsbridge  to Other"
## [1] "Midwood : 2"
## [1] "Changing  Midwood  to Other"
## [1] "Red Hook : 2"
## [1] "Changing  Red Hook  to Other"
## [1] "Bensonhurst : 1"
## [1] "Changing  Bensonhurst  to Other"
## [1] "Concourse Village : 1"
## [1] "Changing  Concourse Village  to Other"
## [1] "Corona : 2"
## [1] "Changing  Corona  to Other"
## [1] "Highbridge : 2"
## [1] "Changing  Highbridge  to Other"
## [1] "Carroll Gardens : 3"
## [1] "Changing  Carroll Gardens  to Other"
## [1] "Flushing : 3"
## [1] "Changing  Flushing  to Other"
## [1] "South Slope : 1"
## [1] "Changing  South Slope  to Other"
## [1] "Prospect Heights : 4"
## [1] "Changing  Prospect Heights  to Other"
## [1] "Stapleton : 1"
## [1] "Changing  Stapleton  to Other"
## [1] "Fort Hamilton : 1"
## [1] "Changing  Fort Hamilton  to Other"
## [1] "Brooklyn Heights : 4"
## [1] "Changing  Brooklyn Heights  to Other"
## [1] "Jamaica : 2"
## [1] "Changing  Jamaica  to Other"
## [1] "Riverdale : 1"
## [1] "Changing  Riverdale  to Other"
## [1] "Park Slope : 3"
## [1] "Changing  Park Slope  to Other"
## [1] "Rego Park : 2"
## [1] "Changing  Rego Park  to Other"
## [1] "Morrisania : 2"
## [1] "Changing  Morrisania  to Other"
## [1] "Throgs Neck : 1"
## [1] "Changing  Throgs Neck  to Other"
## [1] "Jackson Heights : 2"
## [1] "Changing  Jackson Heights  to Other"
## [1] "Mott Haven : 1"
```

```
## [1] "Changing  Mott Haven  to Other"
## [1] "Cypress Hills : 1"
## [1] "Changing  Cypress Hills  to Other"
## [1] "Arrochar : 1"
## [1] "Changing  Arrochar  to Other"
## [1] "Bay Ridge : 2"
## [1] "Changing  Bay Ridge  to Other"
## [1] "Borough Park : 1"
## [1] "Changing  Borough Park  to Other"
## [1] "Forest Hills : 2"
## [1] "Changing  Forest Hills  to Other"
## [1] "New Brighton : 1"
## [1] "Changing  New Brighton  to Other"
## [1] "Maspeth : 1"
## [1] "Changing  Maspeth  to Other"
## [1] "Vinegar Hill : 1"
## [1] "Changing  Vinegar Hill  to Other"
## [1] "Brownsville : 1"
## [1] "Changing  Brownsville  to Other"
## [1] "Bellerose : 1"
## [1] "Changing  Bellerose  to Other"
## [1] "Longwood : 1"
## [1] "Changing  Longwood  to Other"
## [1] "Clinton Hill : 4"
## [1] "Changing  Clinton Hill  to Other"
## [1] "Boerum Hill : 1"
## [1] "Changing  Boerum Hill  to Other"
## [1] "Morris Heights : 1"
## [1] "Changing  Morris Heights  to Other"
## [1] "Canarsie : 1"
## [1] "Changing  Canarsie  to Other"
## [1] "Ozone Park : 1"
## [1] "Changing  Ozone Park  to Other"
## [1] "Woodside : 1"
## [1] "Changing  Woodside  to Other"
## [1] "Bath Beach : 1"
## [1] "Changing  Bath Beach  to Other"
```

```r
print("Converting neighborhood as factor post-clubbing")
```

```
## [1] "Converting neighborhood as factor post-clubbing"
```

```r
airDataClean$neighbourhoodCleansed <-
as.factor(airDataClean$neighbourhoodCleansed)

print("Summary of neighborhood column post clubbing")
```

```
## [1] "Summary of neighborhood column post clubbing"
```

```r
print(summary(airDataClean$neighbourhoodCleansed))
```

```
##             Chelsea           Chinatown         East Harlem 
##                1056                 354                1038 
##        East Village  Financial District            Gramercy 
##                1828                 387                 302 
##    Greenwich Village              Harlem      Hell's Kitchen 
##                 373                2481                1557 
##            Kips Bay     Lower East Side             Midtown 
##                 409                 947                 946 
## Morningside Heights               Other                SoHo 
##                 388                1967                 350 
##      Upper East Side     Upper West Side  Washington Heights 
##                1520                1750                 851 
##        West Village 
##                 769
```

Extract data without price information as data set to predict on at the end.

```r
print("Extract data without price as predict data set")

## [1] "Extract data without price as predict data set"

airDataPredict <- airDataClean[airDataClean$price == 0,]
print(paste("Number of rows in airDataPredict:",nrow(airDataPredict)))

## [1] "Number of rows in airDataPredict: 131"
```

Extract data set with valid price information to use for biulding training and test data sets.

```r
print("Extract valid data with price as actual data set")

## [1] "Extract valid data with price as actual data set"

airDataValid <- airDataClean[airDataClean$price > 0,]
print(paste("Number of rows in airDataValid:",nrow(airDataValid)))

## [1] "Number of rows in airDataValid: 19142"
```

Log transform price if required.

```r
if(log_regression){
  airDataValid$price <- log(airDataValid$price)
}
```

Create training and sample tests by sampling 95% of data for training set and 5% for test set.

```r
print("Creating sample training set and test sets")

## [1] "Creating sample training set and test sets"

train <- sample(1:nrow(airDataValid),0.95*nrow(airDataValid))
airDataTrain <- airDataValid[train,]
airDataTest <- airDataValid[-train,]
```

Linear Regression: Price vs Everything.

```r
print("Running Linear Regression Model: Price vs Everything")

## [1] "Running Linear Regression Model: Price vs Everything"

lm1 <- lm(price~.,data=airDataTrain)
print("Summary of Linear Regression")

## [1] "Summary of Linear Regression"

print(summary(lm1))

##
## Call:
## lm(formula = price ~ ., data = airDataTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -514.04  -39.75   -4.77   26.99  904.86
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 -1.386103  23.661112  -0.059
## hostResponseHours                           -0.047165   0.040678  -1.159
## neighbourhoodCleansedChinatown             -31.028809   4.946381  -6.273
## neighbourhoodCleansedEast Harlem           -70.362687   3.602816 -19.530
## neighbourhoodCleansedEast Village          -27.316559   3.132938  -8.719
## neighbourhoodCleansedFinancial District    -37.997174   4.897226  -7.759
## neighbourhoodCleansedGramercy              -29.522401   5.230841  -5.644
## neighbourhoodCleansedGreenwich Village       8.277170   4.850873   1.706
## neighbourhoodCleansedHarlem                -75.597872   3.075542 -24.580
## neighbourhoodCleansedHell's Kitchen        -15.283612   3.226681  -4.737
## neighbourhoodCleansedKips Bay              -27.640052   4.679288  -5.907
## neighbourhoodCleansedLower East Side       -33.244748   3.618505  -9.187
## neighbourhoodCleansedMidtown                -2.493174   3.706684  -0.673
## neighbourhoodCleansedMorningside Heights -67.348538   4.844746 -13.901
## neighbourhoodCleansedOther                 -21.555978   3.205641  -6.724
## neighbourhoodCleansedSoHo                   24.843834   4.975777   4.993
## neighbourhoodCleansedUpper East Side       -33.938207   3.232933 -10.498
## neighbourhoodCleansedUpper West Side       -29.619841   3.148623  -9.407
## neighbourhoodCleansedWashington Heights    -79.477200   3.812532 -20.846
## neighbourhoodCleansedWest Village           10.456850   3.804379   2.749
## neighbourhoodGroupCleansedBrooklyn          31.530886  20.787084   1.517
## neighbourhoodGroupCleansedManhattan         81.615634  20.211702   4.038
## neighbourhoodGroupCleansedQueens            11.855518  21.107152   0.562
## neighbourhoodGroupCleansedStaten Island    -33.181092  49.133333  -0.675
## accommodates                                15.867516   0.689854  23.001
## bathrooms                                   59.336945   1.976715  30.018
## bedTypeCouch                                11.331045  12.245096   0.925
## bedTypeFuton                                19.920931   9.146332   2.178
```

```
## bedTypePull-out Sofa                     16.440314   8.914900   1.844
## bedTypeReal Bed                          16.412054   7.039755   2.331
## bedrooms                                 30.420432   1.221160  24.911
## beds                                      2.634716   1.123543   2.345
## TV                                        8.964447   1.419963   6.313
## cancellationPolicymoderate               -5.264101   1.717846  -3.064
## cancellationPolicystrict                 -7.126372   1.562511  -4.561
## cancellationPolicysuper_strict_30        64.939468  44.924193   1.446
## cleaningFee                               0.283760   0.015419  18.404
## extraPeople                               0.015059   0.026381   0.571
## guestsIncluded                            2.600795   0.782856   3.322
## maximumNights                             0.001010   0.001127   0.896
## minimumNights                            -0.116080   0.038184  -3.040
## propertyTypeBed & Breakfast              19.046868  10.572002   1.802
## propertyTypeBoat                        187.108591  77.991940   2.399
## propertyTypeBoutique hotel               49.649615  26.088743   1.903
## propertyTypeBungalow                    -52.494227  77.512013  -0.677
## propertyTypeCabin                        78.369087  77.610233   1.010
## propertyTypeCastle                      111.624567  77.519928   1.440
## propertyTypeCondominium                  51.461825   5.971218   8.618
## propertyTypeDorm                        -32.909670  44.806708  -0.734
## propertyTypeGuest suite                 171.753594  54.901057   3.128
## propertyTypeGuesthouse                   16.775622  21.767825   0.771
## propertyTypeHostel                       -8.334471  22.836205  -0.365
## propertyTypeHouse                        23.529344   4.996240   4.709
## propertyTypeHut                         -32.138453  77.877729  -0.413
## propertyTypeLighthouse                   29.887377  77.587456   0.385
## propertyTypeLoft                         52.360398   5.084780  10.297
## propertyTypeOther                        41.084368   8.341148   4.926
## propertyTypeServiced apartment           94.799199  38.845659   2.440
## propertyTypeTimeshare                   102.976647  13.607063   7.568
## propertyTypeTownhouse                    22.138879   6.204668   3.568
## propertyTypeVacation home               185.882999  77.598523   2.395
## propertyTypeVilla                       -15.844696  38.811528  -0.408
## roomTypePrivate room                    -58.723827   1.602457 -36.646
## roomTypeShared room                     -72.898881   3.777639 -19.297
## securityDeposit                          -0.013277   0.003501  -3.792
## numberOfReviews                          -0.065135   0.024643  -2.643
## reviewScoresAccuracy                      0.950880   1.037882   0.916
## reviewScoresCheckin                      -1.881277   1.086623  -1.731
## reviewScoresCleanliness                   4.220609   0.824670   5.118
## reviewScoresCommunication                -2.268170   1.135158  -1.998
## reviewScoresLocation                      0.422377   0.916654   0.461
## reviewScoresRating                        0.326521   0.117344   2.783
## reviewScoresValue                        -6.306832   1.063625  -5.930
## reviewsperMonth                          -1.610618   0.440165  -3.659
## checkIn24Hours1                          -1.299334   1.641293  -0.792
## airConditioning1                         -1.122820   1.877579  -0.598
## buzzerOrWirelessIntercom1                -1.741814   1.428461  -1.219
## cableTV1                                 10.770992   1.543699   6.977
```

```
## carbonMonoxideDetector1                 1.165293   1.601707   0.728
## doorman1                                7.950986   2.190256   3.630
## dryer1                                 -0.310951   3.407495  -0.091
## elevator1                               8.353844   1.610309   5.188
## essentials1                           -10.588924   1.847402  -5.732
## familyAndKidFriendly1                   2.668674   1.330909   2.005
## fireExtinguisher1                       2.719266   1.632304   1.666
## firstAidKit1                            2.847427   1.659956   1.715
## freeParkingOnPremises1                 -1.741520   2.932166  -0.594
## gym1                                   19.095249   2.529974   7.548
## hairdryer1                             -2.547332   1.713685  -1.486
## hangers1                               -5.790698   1.741407  -3.325
## heating1                               -3.832088   2.593456  -1.478
## hottub1                                -4.918711   2.585287  -1.903
## internet1                              -1.805301   1.546928  -1.167
## iron1                                  -3.459870   1.686572  -2.051
## kitchen1                               -7.318994   2.862023  -2.557
## laptopFriendlyWorkspace1               -1.306653   1.609633  -0.812
## lockOnBedroomDoor1                     -2.804050   1.750691  -1.602
## petsAllowed1                           -6.967597   1.946391  -3.580
## petsLiveOnThisProperty1                -2.961472   2.442083  -1.213
## pool1                                  17.905360   5.109975   3.504
## selfCheckIn1                            9.143911   2.860155   3.197
## shampoo1                                3.912367   1.500874   2.607
## smokeDetector1                         -6.340369   1.853096  -3.422
## smokingAllowed1                        -0.300668   2.847226  -0.106
## washer1                                 3.188479   3.414901   0.934
## wheelchairAccessible1                  12.373064   2.370779   5.219
## wirelessInternet1                      -1.964830   3.781742  -0.520
## hostHasProfilePic1                    -13.621731   8.815305  -1.545
## hostIdentityVerified1                  -1.429651   1.434768  -0.996
## hostIsSuperhost1                       15.260998   2.264191   6.740
## instantBookable1                       -4.451783   1.691223  -2.632
## isLocationExact1                       -0.131177   1.696189  -0.077
## requireGuestPhoneVerification1        -12.652511   5.390762  -2.347
## requireGuestProfilePicture1             5.663180   5.732609   0.988
## featureCount                            0.760911   0.537129   1.417
## yearsAsHost                             0.316785   0.387069   0.818
##                                       Pr(>|t|)
## (Intercept)                           0.953286
## hostResponseHours                     0.246274
## neighbourhoodCleansedChinatown        3.62e-10 ***
## neighbourhoodCleansedEast Harlem       < 2e-16 ***
## neighbourhoodCleansedEast Village      < 2e-16 ***
## neighbourhoodCleansedFinancial District 9.02e-15 ***
## neighbourhoodCleansedGramercy         1.69e-08 ***
## neighbourhoodCleansedGreenwich Village 0.087965 .
## neighbourhoodCleansedHarlem            < 2e-16 ***
## neighbourhoodCleansedHell's Kitchen   2.19e-06 ***
## neighbourhoodCleansedKips Bay         3.55e-09 ***
```

```
## neighbourhoodCleansedLower East Side        < 2e-16 ***
## neighbourhoodCleansedMidtown               0.501200
## neighbourhoodCleansedMorningside Heights    < 2e-16 ***
## neighbourhoodCleansedOther                 1.82e-11 ***
## neighbourhoodCleansedSoHo                  6.00e-07 ***
## neighbourhoodCleansedUpper East Side        < 2e-16 ***
## neighbourhoodCleansedUpper West Side        < 2e-16 ***
## neighbourhoodCleansedWashington Heights     < 2e-16 ***
## neighbourhoodCleansedWest Village          0.005990 **
## neighbourhoodGroupCleansedBrooklyn         0.129322
## neighbourhoodGroupCleansedManhattan        5.41e-05 ***
## neighbourhoodGroupCleansedQueens           0.574339
## neighbourhoodGroupCleansedStaten Island    0.499476
## accommodates                                < 2e-16 ***
## bathrooms                                   < 2e-16 ***
## bedTypeCouch                               0.354794
## bedTypeFuton                               0.029417 *
## bedTypePull-out Sofa                       0.065179 .
## bedTypeReal Bed                            0.019746 *
## bedrooms                                    < 2e-16 ***
## beds                                       0.019037 *
## TV                                         2.80e-10 ***
## cancellationPolicymoderate                 0.002185 **
## cancellationPolicystrict                   5.13e-06 ***
## cancellationPolicysuper_strict_30          0.148325
## cleaningFee                                 < 2e-16 ***
## extraPeople                                0.568139
## guestsIncluded                             0.000895 ***
## maximumNights                              0.370270
## minimumNights                              0.002369 **
## propertyTypeBed & Breakfast                0.071620 .
## propertyTypeBoat                           0.016447 *
## propertyTypeBoutique hotel                 0.057043 .
## propertyTypeBungalow                       0.498262
## propertyTypeCabin                          0.312615
## propertyTypeCastle                         0.149900
## propertyTypeCondominium                     < 2e-16 ***
## propertyTypeDorm                           0.462665
## propertyTypeGuest suite                    0.001760 **
## propertyTypeGuesthouse                     0.440918
## propertyTypeHostel                         0.715140
## propertyTypeHouse                          2.50e-06 ***
## propertyTypeHut                            0.679847
## propertyTypeLighthouse                     0.700087
## propertyTypeLoft                            < 2e-16 ***
## propertyTypeOther                          8.49e-07 ***
## propertyTypeServiced apartment             0.014680 *
## propertyTypeTimeshare                      3.98e-14 ***
## propertyTypeTownhouse                      0.000361 ***
## propertyTypeVacation home                  0.016610 *
```

```
## propertyTypeVilla                              0.683097
## roomTypePrivate room                           < 2e-16 ***
## roomTypeShared room                            < 2e-16 ***
## securityDeposit                               0.000150 ***
## numberOfReviews                               0.008220 **
## reviewScoresAccuracy                          0.359588
## reviewScoresCheckin                           0.083414 .
## reviewScoresCleanliness                       3.12e-07 ***
## reviewScoresCommunication                     0.045720 *
## reviewScoresLocation                          0.644961
## reviewScoresRating                            0.005398 **
## reviewScoresValue                             3.09e-09 ***
## reviewsperMonth                               0.000254 ***
## checkIn24Hours1                               0.428573
## airConditioning1                              0.549837
## buzzerOrWirelessIntercom1                     0.222722
## cableTV1                                      3.11e-12 ***
## carbonMonoxideDetector1                       0.466909
## doorman1                                      0.000284 ***
## dryer1                                        0.927291
## elevator1                                     2.15e-07 ***
## essentials1                                   1.01e-08 ***
## familyAndKidFriendly1                         0.044962 *
## fireExtinguisher1                             0.095749 .
## firstAidKit1                                  0.086296 .
## freeParkingOnPremises1                        0.552562
## gym1                                          4.64e-14 ***
## hairdryer1                                    0.137174
## hangers1                                      0.000885 ***
## heating1                                      0.139533
## hottub1                                       0.057111 .
## internet1                                     0.243216
## iron1                                         0.040240 *
## kitchen1                                      0.010558 *
## laptopFriendlyWorkspace1                      0.416934
## lockOnBedroomDoor1                            0.109243
## petsAllowed1                                  0.000345 ***
## petsLiveOnThisProperty1                       0.225267
## pool1                                         0.000459 ***
## selfCheckIn1                                  0.001391 **
## shampoo1                                      0.009149 **
## smokeDetector1                                0.000624 ***
## smokingAllowed1                               0.915901
## washer1                                       0.350473
## wheelchairAccessible1                         1.82e-07 ***
## wirelessInternet1                             0.603379
## hostHasProfilePic1                            0.122307
## hostIdentityVerified1                         0.319053
## hostIsSuperhost1                              1.63e-11 ***
## instantBookable1                              0.008488 **
```

```
## isLocationExact1                                  0.938357
## requireGuestPhoneVerification1                    0.018932 *
## requireGuestProfilePicture1                       0.323220
## featureCount                                      0.156609
## yearsAsHost                                       0.413128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.4 on 18068 degrees of freedom
## Multiple R-squared:  0.5992, Adjusted R-squared:  0.5966
## F-statistic: 234.9 on 115 and 18068 DF,  p-value: < 2.2e-16
```

```
print("Using Linear Regression Model for prediction")
```

```
## [1] "Using Linear Regression Model for prediction"
```

```
print("WARNING!!!: This step may fail in case the training set and test set
have different levels. If this occurs, re-running the code usually fixes
it.")
```

```
## [1] "WARNING!!!: This step may fail in case the training set and test set
have different levels. If this occurs, re-running the code usually fixes it."
```

```
lm1Pred <- predict(lm1, airDataTest)
print(paste("Test MSE:",mean((airDataTest$price - lm1Pred) ^ 2)))
```

```
## [1] "Test MSE: 6130.21599833604"
```

Running Principal Component Analysis in numerical features to analyze factor loading of each against first 3 PCs.

```
print("Running PCA on numerical features")
```

```
## [1] "Running PCA on numerical features"
```

```
pcaDF <-
data.frame(cbind(hostResponseHours=airDataValid$hostResponseHours,accommodate
s=airDataValid$accommodates,bathrooms=airDataValid$bathrooms,bedrooms=airData
Valid$bedrooms,beds=airDataValid$beds,TV=airDataValid$TV,cleaningFee=airDataV
alid$cleaningFee,extraPeople=airDataValid$extraPeople,guestsIncluded=airDataV
alid$guestsIncluded,maximumNights=airDataValid$maximumNights,minimumNights=ai
rDataValid$minimumNights,price=airDataValid$price,securityDeposit=airDataVali
d$securityDeposit,numberOfReviews=airDataValid$numberOfReviews,reviewScoresAc
curacy=airDataValid$reviewScoresAccuracy,reviewScoresCheckin=airDataValid$rev
iewScoresCheckin,reviewScoresCleanliness=airDataValid$reviewScoresCleanliness
,reviewScoresCommunication=airDataValid$reviewScoresCommunication,reviewScore
sLocation=airDataValid$reviewScoresLocation,reviewScoresLocation=airDataValid
$reviewScoresLocation,reviewScoresRating=airDataValid$reviewScoresRating,revi
ewScoresValue=airDataValid$reviewScoresValue,reviewsperMonth=airDataValid$rev
iewsperMonth))
pcaModel <- prcomp(pcaDF,scale.=TRUE,center=TRUE)
print("Summary of PCA Model")
```

```
## [1] "Summary of PCA Model"
```

```r
print(summary(pcaModel))
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5    PC6
## Standard deviation     2.8700 1.9640 1.16666 1.13561 1.00818 1.0003
## Proportion of Variance 0.3581 0.1677 0.05918 0.05607 0.04419 0.0435
## Cumulative Proportion  0.3581 0.5258 0.58502 0.64109 0.68528 0.7288
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.99088 0.97523 0.91901 0.87159 0.77786 0.70400
## Proportion of Variance 0.04269 0.04135 0.03672 0.03303 0.02631 0.02155
## Cumulative Proportion  0.77147 0.81282 0.84954 0.88257 0.90888 0.93043
##                           PC13    PC14    PC15    PC16   PC17    PC18
## Standard deviation     0.68109 0.66843 0.60992 0.42687 0.2300 0.16019
## Proportion of Variance 0.02017 0.01943 0.01617 0.00792 0.0023 0.00112
## Cumulative Proportion  0.95060 0.97002 0.98620 0.99412 0.9964 0.99754
##                           PC19    PC20   PC21    PC22      PC23
## Standard deviation     0.13649 0.12205 0.1177 0.09645 1.072e-16
## Proportion of Variance 0.00081 0.00065 0.0006 0.00040 0.000e+00
## Cumulative Proportion  0.99835 0.99899 0.9996 1.00000 1.000e+00
```

```r
print("Rotation Matrix of PCA Model")
```

```
## [1] "Rotation Matrix of PCA Model"
```

```r
print(pcaModel$rotation)
```

```
##                                      PC1          PC2         PC3
## hostResponseHours            0.010623765 -0.007706717 -0.18499252
## accommodates                 0.063732043  0.433984113  0.02134112
## bathrooms                    0.009975546  0.266840841 -0.12242621
## bedrooms                     0.025473654  0.364391600 -0.02061117
## beds                         0.051321852  0.418914350  0.02779839
## TV                           0.022849329  0.141432471 -0.06684028
## cleaningFee                  0.075166481  0.308855500 -0.11693451
## extraPeople                  0.060479077  0.133465184  0.16490221
## guestsIncluded               0.076333846  0.327590430  0.15867777
## maximumNights               -0.005899567  0.037846809 -0.09498251
## minimumNights               -0.015018446  0.006629710 -0.15811710
## price                        0.015448283  0.377032375 -0.14213289
## securityDeposit              0.044306385  0.158417754 -0.09346542
## numberOfReviews              0.127160608  0.028293594  0.63398279
## reviewScoresAccuracy         0.342630272 -0.053178947 -0.06501591
## reviewScoresCheckin          0.342944200 -0.052538004 -0.05923336
## reviewScoresCleanliness      0.340897084 -0.045027804 -0.05775156
## reviewScoresCommunication    0.342973450 -0.053822732 -0.06104353
## reviewScoresLocation         0.342565870 -0.049776293 -0.06655161
## reviewScoresLocation.1       0.342565870 -0.049776293 -0.06655161
## reviewScoresRating           0.342773576 -0.052045062 -0.06853238
## reviewScoresValue            0.342486951 -0.054990316 -0.06611126
```

```
## reviewsperMonth            0.140758080  0.029555194  0.61725736
##                                    PC4           PC5          PC6
## hostResponseHours          0.11468152  0.4768823601 -0.530426737
## accommodates              -0.14393054  0.0109854904  0.006117475
## bathrooms                 -0.25578886  0.1485500307 -0.034180704
## bedrooms                  -0.26634366  0.1826668956  0.041951750
## beds                      -0.20003260  0.0871710705  0.037126329
## TV                         0.14091974 -0.4058946367 -0.505656909
## cleaningFee                0.31160668 -0.1201743935 -0.001795881
## extraPeople                0.50121519  0.0301580023  0.153196526
## guestsIncluded             0.18692646  0.0524654193  0.134475346
## maximumNights             -0.23814466 -0.5597261480  0.332404369
## minimumNights              0.14008885  0.4257551554  0.526300789
## price                     -0.04362853 -0.0834066517 -0.115507220
## securityDeposit            0.54328882 -0.1433795872  0.080135986
## numberOfReviews            0.02866579  0.0664847294 -0.094188157
## reviewScoresAccuracy      -0.01938522 -0.0002550280  0.005479408
## reviewScoresCheckin       -0.02261918  0.0036082138  0.008543965
## reviewScoresCleanliness   -0.01323716 -0.0050488529  0.001222884
## reviewScoresCommunication -0.02145754  0.0023628008  0.010235701
## reviewScoresLocation      -0.02127709 -0.0038897114  0.007649949
## reviewScoresLocation.1    -0.02127709 -0.0038897114  0.007649949
## reviewScoresRating        -0.01851883  0.0007029437  0.004302680
## reviewScoresValue         -0.02792812  0.0019930679  0.007730111
## reviewsperMonth           -0.09412678 -0.0008205797 -0.023099828
##                                     PC7          PC8          PC9
## hostResponseHours          0.3355708341  0.566864405 -0.058980363
## accommodates               0.0091476687  0.001937219 -0.001244235
## bathrooms                 -0.0455993059 -0.076933701 -0.007965630
## bedrooms                   0.0632854178 -0.085536980 -0.006094501
## beds                       0.0369257938 -0.028320443 -0.023427949
## TV                        -0.4934633492  0.141484043  0.427113854
## cleaningFee               -0.0711042933  0.077608445 -0.242276495
## extraPeople                0.3741437985 -0.065394366  0.520768713
## guestsIncluded             0.2064639975 -0.045557635  0.247246049
## maximumNights              0.3038405190  0.629564261  0.011941815
## minimumNights             -0.5532262039  0.395300794  0.176931144
## price                     -0.1266365811  0.013591642 -0.047679061
## securityDeposit           -0.0380049983 -0.018977744 -0.593988262
## numberOfReviews           -0.1444883131  0.194542361 -0.148224489
## reviewScoresAccuracy       0.0002082373 -0.016575063  0.010217946
## reviewScoresCheckin        0.0032382472 -0.016638715  0.010205570
## reviewScoresCleanliness   -0.0036408610 -0.019763040  0.007612665
## reviewScoresCommunication  0.0015953378 -0.019287792  0.009535240
## reviewScoresLocation       0.0016900906 -0.008994315  0.004260073
## reviewScoresLocation.1     0.0016900906 -0.008994315  0.004260073
## reviewScoresRating        -0.0010108374 -0.022504100  0.012530115
## reviewScoresValue          0.0017354049 -0.020319219  0.014349402
## reviewsperMonth           -0.1091575553  0.197053494 -0.114655031
##                                    PC10         PC11          PC12
```

```
## hostResponseHours          3.759995e-02 -0.037598424  0.0338447302
## accommodates               2.622319e-01 -0.014876237  0.0686369642
## bathrooms                 -8.345226e-01 -0.165817479  0.2248775596
## bedrooms                   8.141674e-02 -0.247236504 -0.6201060152
## beds                       2.337677e-01 -0.115920020 -0.0542707355
## TV                         1.891869e-02 -0.270775108 -0.0747216557
## cleaningFee               -2.996961e-02  0.567650450 -0.1452802011
## extraPeople               -2.817751e-01  0.073145154 -0.3150466887
## guestsIncluded             2.328287e-01 -0.176930136  0.6200777645
## maximumNights             -8.414694e-02 -0.072539985 -0.0314629879
## minimumNights              2.425072e-02 -0.023691289 -0.0021928475
## price                     -7.726768e-02  0.468409522  0.1510952722
## securityDeposit           -7.340979e-02 -0.485405975 -0.0082187877
## numberOfReviews           -9.361614e-02  0.038311411 -0.1243650213
## reviewScoresAccuracy       1.752297e-03 -0.007825300  0.0052071997
## reviewScoresCheckin        7.899421e-03 -0.010212244 -0.0051972929
## reviewScoresCleanliness    3.205107e-05  0.002544175  0.0031547747
## reviewScoresCommunication  7.240649e-03 -0.011483686 -0.0051808577
## reviewScoresLocation       3.745000e-03  0.010434534 -0.0016693948
## reviewScoresLocation.1     3.745000e-03  0.010434534 -0.0016693948
## reviewScoresRating        -2.671173e-03 -0.005094931  0.0017325495
## reviewScoresValue          1.675310e-03 -0.012593420 -0.0007386662
## reviewsperMonth           -9.646931e-02  0.032892007  0.0787243978
##                                   PC13         PC14         PC15
## hostResponseHours          0.045101492 -0.012740088  0.007086225
## accommodates               0.072403628 -0.103406172 -0.352800607
## bathrooms                 -0.034763136  0.171523004 -0.097766490
## bedrooms                  -0.056973377  0.102198222  0.531035757
## beds                       0.113309722 -0.021585964 -0.538164823
## TV                         0.024254560  0.088634032  0.006772006
## cleaningFee                0.028518542  0.598589310 -0.016195704
## extraPeople                0.109420941 -0.203691672 -0.149120045
## guestsIncluded            -0.187766959  0.243271962  0.356949320
## maximumNights             -0.077714150 -0.004726907  0.018666802
## minimumNights              0.016966797 -0.041238963 -0.003976291
## price                     -0.064308811 -0.658400561  0.301017205
## securityDeposit            0.056459886 -0.206969824  0.017121997
## numberOfReviews           -0.663876370 -0.053969721 -0.129153419
## reviewScoresAccuracy      -0.006904114 -0.005897608  0.003729282
## reviewScoresCheckin       -0.014617176 -0.003279478 -0.003228543
## reviewScoresCleanliness    0.002580730 -0.007527287  0.003826496
## reviewScoresCommunication -0.018740070 -0.005639855 -0.003315280
## reviewScoresLocation      -0.004317425 -0.019230525  0.002515401
## reviewScoresLocation.1    -0.004317425 -0.019230525  0.002515401
## reviewScoresRating        -0.008794778 -0.009399030  0.005968689
## reviewScoresValue          0.002457585 -0.003116225  0.001319768
## reviewsperMonth            0.686787221  0.006509022  0.190514409
##                                   PC16         PC17         PC18
## hostResponseHours          0.002734262 -0.004788514  0.001006052
## accommodates               0.756267545  0.001130522 -0.010100455
```

```
## bathrooms                    0.046495875  0.008956450 -0.001478290
## bedrooms                     0.025332154  0.001303373 -0.008052979
## beds                        -0.628735525 -0.004784490  0.004868690
## TV                          -0.030228371  0.009471549  0.001767942
## cleaningFee                 -0.012491736 -0.001071584  0.010774217
## extraPeople                  0.023758038  0.001362257  0.002031904
## guestsIncluded              -0.089961501  0.009515004 -0.004503719
## maximumNights               -0.017558622 -0.011758880  0.001654749
## minimumNights                0.006294102 -0.003616091 -0.002267408
## price                       -0.139161548 -0.012502452  0.023535595
## securityDeposit              0.003306071  0.003311821  0.000919025
## numberOfReviews              0.001741909 -0.002996495 -0.009993529
## reviewScoresAccuracy         0.007765761 -0.194494077  0.227839361
## reviewScoresCheckin          0.002514669  0.079240819  0.486962404
## reviewScoresCleanliness     -0.009142728 -0.580828194 -0.542864587
## reviewScoresCommunication  0.001366255 -0.008312346  0.482666536
## reviewScoresLocation        -0.009414401  0.519060136 -0.297066376
## reviewScoresLocation.1      -0.009414401  0.519060136 -0.297066376
## reviewScoresRating          -0.002456447 -0.273115020  0.006505995
## reviewScoresValue           -0.000288166 -0.066142665 -0.074406047
## reviewsperMonth             -0.012221994  0.003012105  0.014578591
##                                      PC19          PC20          PC21
## hostResponseHours           -0.0006866173 -0.0027474425  0.0001815162
## accommodates                -0.0014902581 -0.0036621723  0.0038389343
## bathrooms                   -0.0073052603  0.0013471827 -0.0010296893
## bedrooms                    -0.0018601975  0.0052103999  0.0023966721
## beds                         0.0056933800  0.0055737511  0.0043828780
## TV                          -0.0013165148 -0.0002538783  0.0002024053
## cleaningFee                  0.0076092030 -0.0037412572 -0.0007715564
## extraPeople                  0.0022983650  0.0014848696 -0.0006644462
## guestsIncluded              -0.0003912556 -0.0016889240  0.0001572761
## maximumNights                0.0001558799 -0.0023745095  0.0005642854
## minimumNights               -0.0010420589 -0.0001813681 -0.0006859989
## price                        0.0062781012 -0.0054291511 -0.0101092127
## securityDeposit              0.0030645341 -0.0023187785 -0.0023312803
## numberOfReviews              0.0122447795 -0.0004897797 -0.0012296908
## reviewScoresAccuracy         0.3194417507  0.8219935057  0.0783886251
## reviewScoresCheckin         -0.3258322228 -0.0746034398 -0.5087331332
## reviewScoresCleanliness     -0.4376800242  0.0480382177 -0.2009758423
## reviewScoresCommunication -0.2971905427 -0.2231266337  0.2740977725
## reviewScoresLocation        -0.0692621582  0.0724232630  0.0821402199
## reviewScoresLocation.1      -0.0692621582  0.0724232630  0.0821402199
## reviewScoresRating           0.1883482593 -0.3350103459  0.6367131792
## reviewScoresValue            0.6829913042 -0.3792036725 -0.4477004784
## reviewsperMonth             -0.0008793112 -0.0030579705  0.0063010504
##                                      PC22          PC23
## hostResponseHours            0.0019034908 -2.372948e-18
## accommodates                -0.0013923189  1.307853e-16
## bathrooms                    0.0016417702  9.466267e-17
## bedrooms                    -0.0003170390 -4.764127e-17
```

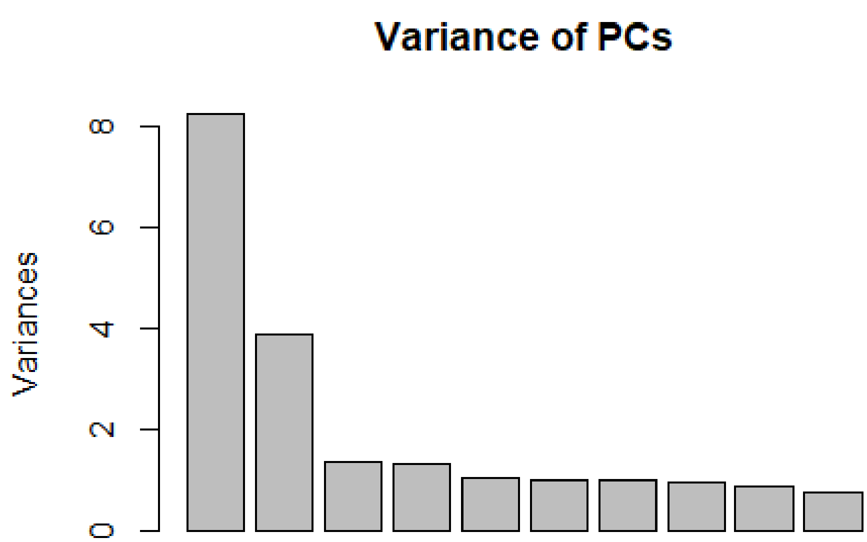
```
## beds                        -0.0022960187  5.444345e-17
## TV                           0.0006178135 -5.343026e-17
## cleaningFee                  0.0013760036  4.970285e-17
## extraPeople                  0.0009916549 -5.814551e-18
## guestsIncluded               0.0006935043 -5.588022e-17
## maximumNights               -0.0010728195 -7.502198e-17
## minimumNights               -0.0003127416  2.841532e-17
## price                        0.0040251978 -1.542974e-16
## securityDeposit             -0.0005599607 -3.346111e-17
## numberOfReviews             -0.0009298338 -5.033139e-17
## reviewScoresAccuracy         0.0302726502 -1.458515e-16
## reviewScoresCheckin         -0.5106302395  2.902152e-16
## reviewScoresCleanliness      0.1061053514  1.322417e-17
## reviewScoresCommunication    0.6540866676 -4.796392e-16
## reviewScoresLocation        -0.0065733661  7.071068e-01
## reviewScoresLocation.1      -0.0065733661 -7.071068e-01
## reviewScoresRating          -0.4960707889 -1.112250e-16
## reviewScoresValue            0.2303120943 -1.192665e-16
## reviewsperMonth             -0.0005533699  4.983270e-17
```

Screeplot

```r
print("PCA Screeplot")
```

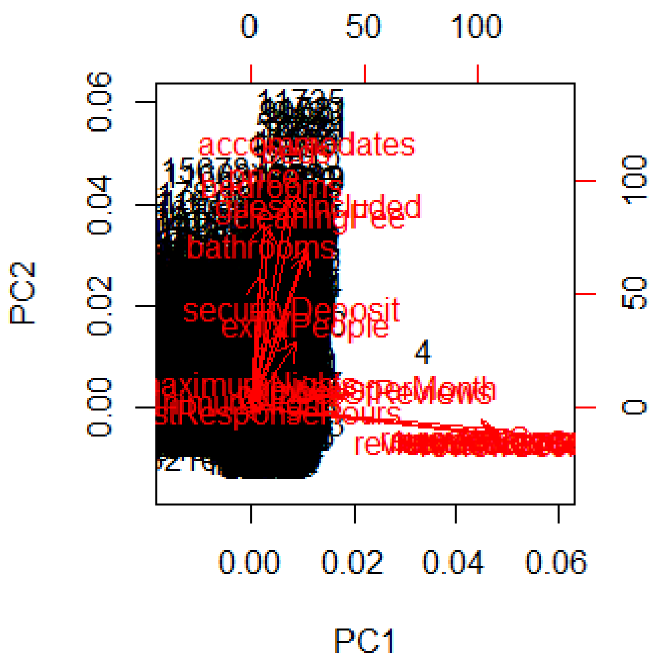
```
## [1] "PCA Screeplot"
```

```r
screeplot(pcaModel, main="Variance of PCs")
```



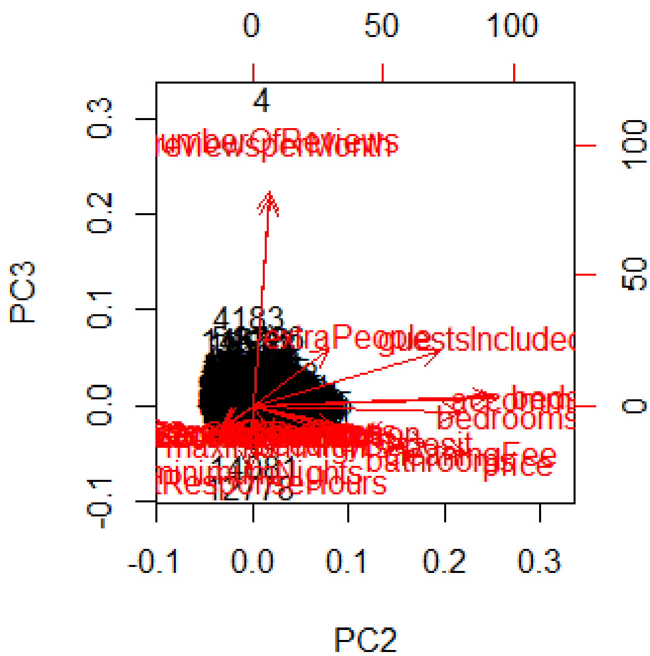
Biplot: PC1 vs PC2

```r
print("Plotting Biplot: PC1 vs PC2 (Please wait till plot appears to continue)")

## [1] "Plotting Biplot: PC1 vs PC2 (Please wait till plot appears to continue)"

biplot(pcaModel,choices = c(1,2))
```



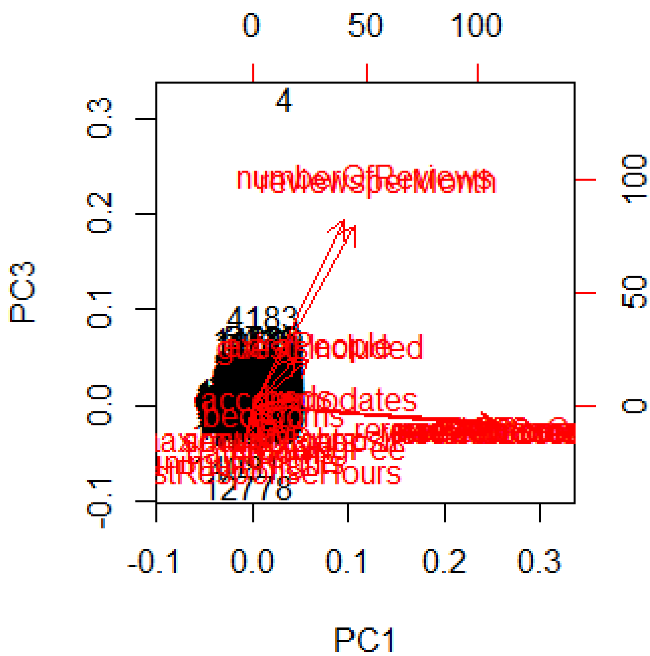
Biplot: PC2 vs PC3

```r
print("Plotting Biplot: PC2 vs PC3 (Please wait till plot appears to continue)")

## [1] "Plotting Biplot: PC2 vs PC3 (Please wait till plot appears to continue)"

biplot(pcaModel,choices = c(2,3))
```



Biplot: PC1 vs PC3

```
print("Plotting Biplot: PC1 vs PC3 (Please wait till plot appears to
continue)")

## [1] "Plotting Biplot: PC1 vs PC3 (Please wait till plot appears to
continue)"

biplot(pcaModel,choices = c(1,3))
```



Removing features with similar factor laodings to avoid issues like collinearity.

```
print("Reviews per month and number of reviews have similar factor loadings as do all the individual review columns")

## [1] "Reviews per month and number of reviews have similar factor loadings as do all the individual review columns"

print("Removing Reviews per month, the indiviudal review scores except Review score rating")

## [1] "Removing Reviews per month, the indiviudal review scores except Review score rating"

dropColumns <- c("reviewsperMonth","reviewScoresAccuracy","reviewScoresCheckin","reviewScoresCleanliness","reviewScoresCommunication","reviewScoresLocation","reviewScoresValue")
airDataValid <- airDataValid[,!(names(airDataValid) %in% dropColumns)]
airDataValid <- as.data.frame(airDataValid)
```

Summary after removing features with similar factor loading.

```
print("Summary after deletion of columns")

## [1] "Summary after deletion of columns"

print(summary(airDataValid))
```

```
##  hostResponseHours     neighbourhoodCleansed neighbourhoodGroupCleansed
##  Min.   : 1.000    Harlem         :2475      Bronx        :   15
##  1st Qu.: 1.000    Other          :1948      Brooklyn     :  212
##  Median : 1.000    East Village   :1819      Manhattan    :18771
##  Mean   : 9.246    Upper West Side:1738      Queens       :  140
##  3rd Qu.:12.000    Hell's Kitchen :1549      Staten Island:    4
##  Max.   :72.000    Upper East Side:1507
##                    (Other)        :8106
##   accommodates      bathrooms             bedType         bedrooms
##  Min.   : 1.000   Min.   :0.000   Airbed       :  132   Min.   :0.000
##  1st Qu.: 2.000   1st Qu.:1.000   Couch        :   65   1st Qu.:1.000
##  Median : 2.000   Median :1.000   Futon        :  185   Median :1.000
##  Mean   : 2.779   Mean   :1.092   Pull-out Sofa:  204   Mean   :1.085
##  3rd Qu.: 4.000   3rd Qu.:1.000   Real Bed     :18556   3rd Qu.:1.000
##  Max.   :16.000   Max.   :5.000                         Max.   :6.000
##
##       beds              TV              cancellationPolicy
##  Min.   : 0.000   Min.   :0.0000   flexible       :5741
##  1st Qu.: 1.000   1st Qu.:0.0000   moderate       :4234
##  Median : 1.000   Median :1.0000   strict         :9163
##  Mean   : 1.494   Mean   :0.6877   super_strict_30:   4
##  3rd Qu.: 2.000   3rd Qu.:1.0000
##  Max.   :12.000   Max.   :2.0000
##
##   cleaningFee      extraPeople     guestsIncluded   maximumNights
##  Min.   :  0.00   Min.   :  0.00   Min.   : 1.000   Min.   :   1
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.: 1.000   1st Qu.:  30
##  Median : 40.00   Median :  0.00   Median : 1.000   Median :1125
##  Mean   : 47.39   Mean   : 14.01   Mean   : 1.427   Mean   : 696
##  3rd Qu.: 75.00   3rd Qu.: 25.00   3rd Qu.: 2.000   3rd Qu.:1125
##  Max.   :600.00   Max.   :300.00   Max.   :16.000   Max.   :1125
##
##  minimumNights          price           propertyType
##  Min.   :   1.000   Min.   : 10.0   Apartment  :18001
##  1st Qu.:   1.000   1st Qu.: 89.0   House      :  280
##  Median :   2.000   Median :135.0   Loft       :  253
##  Mean   :   4.142   Mean   :166.3   Condominium:  187
##  3rd Qu.:   4.000   3rd Qu.:200.0   Townhouse  :  171
##  Max.   :1250.000   Max.   :999.0   Other      :   99
##                                     (Other)    :  151
##            roomType     securityDeposit numberOfReviews
##  Entire home/apt:10705   Min.   :  0.0   Min.   :  0.00
##  Private room   : 7864   1st Qu.:  0.0   1st Qu.:  1.00
##  Shared room    :  573   Median :  0.0   Median :  5.00
##                          Mean   :123.1   Mean   : 16.97
##                          3rd Qu.:200.0   3rd Qu.: 19.00
##                          Max.   :999.0   Max.   :432.00
##
##  reviewScoresRating checkIn24Hours airConditioning
##  Min.   :  0.00     0:14387        0: 2721
```

```
##  1st Qu.: 67.00     1: 4755        1:16421
##  Median : 92.00
##  Mean   : 70.83
##  3rd Qu.: 98.00
##  Max.   :100.00
##
##  buzzerOrWirelessIntercom cableTV   carbonMonoxideDetector doorman
##  0:9681                   0:12787   0: 8110                0:16449
##  1:9461                   1: 6355   1:11032                1: 2693
##
##
##
##
##
##  dryer     elevator  essentials familyAndKidFriendly fireExtinguisher
##  0:11122   0:11362   0: 3182    0:10766              0:14298
##  1: 8020   1: 7780   1:15960    1: 8376              1: 4844
##
##
##
##
##
##  firstAidKit freeParkingOnPremises gym       hairdryer hangers   heating
##  0:14085     0:18278               0:17424   0: 8997   0: 7828   0: 1206
##  1: 5057     1:  864               1: 1718   1:10145   1:11314   1:17936
##
##
##
##
##
##  hottub    internet  iron     kitchen   laptopFriendlyWorkspace
##  0:18021   0: 5862   0:9660   0: 1002   0:9270
##  1: 1121   1:13280   1:9482   1:18140   1:9872
##
##
##
##
##
##  lockOnBedroomDoor petsAllowed petsLiveOnThisProperty pool
##  0:15274           0:16871     0:17632                0:18857
##  1: 3868           1: 2271     1: 1510                1:  285
##
##
##
##
##
##  selfCheckIn shampoo   smokeDetector smokingAllowed washer
##  0:17978     0: 7129   0: 4090       0:18229        0:11051
##  1: 1164     1:12013   1:15052       1:  913        1: 8091
##
```

```
##
##
##
##
##  wheelchairAccessible wirelessInternet hostHasProfilePic
##  0:17526              0:  526          0:   79
##  1: 1616              1:18616          1:19063
##
##
##
##
##
##  hostIdentityVerified hostIsSuperhost instantBookable isLocationExact
##  0: 6524              0:17447         0:15530         0: 3346
##  1:12618              1: 1695         1: 3612         1:15796
##
##
##
##
##
##  requireGuestPhoneVerification requireGuestProfilePicture  featureCount
##  0:18472                       0:18558                    Min.   : 2.00
##  1:  670                       1:  584                    1st Qu.:14.00
##                                                           Median :17.00
##                                                           Mean   :17.64
##                                                           3rd Qu.:21.00
##                                                           Max.   :51.00
##
##   yearsAsHost
##  Min.   : 2.000
##  1st Qu.: 4.000
##  Median : 5.000
##  Mean   : 5.086
##  3rd Qu.: 6.000
##  Max.   :11.000
##
```

Using K-fold validation on Linear Model to check if model still performs comparably after removing features based on PCA. The aim is to achieve close to same R-squared model with a muich smaller set of features.

```
print("Retraining linear model with repeated k-fold cross validation")

## [1] "Retraining linear model with repeated k-fold cross validation"

train.control <- trainControl(method="repeatedcv",number=10,repeats=3)
lm2 <- train(price~.,data=airDataValid,method="lm",trControl=train.control)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Summary of K-fold model.

```
print("Summary of Linear Regression")

## [1] "Summary of Linear Regression"

print(summary(lm2))
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -514.11  -39.94   -5.08   27.23  899.36
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 5.665e+00  2.364e+01   0.240
## hostResponseHours                          -2.667e-02  3.949e-02  -0.675
## neighbourhoodCleansedChinatown             -3.274e+01  4.837e+00  -6.769
## `neighbourhoodCleansedEast Harlem`         -7.058e+01  3.480e+00 -20.283
## `neighbourhoodCleansedEast Village`        -2.733e+01  3.065e+00  -8.919
## `neighbourhoodCleansedFinancial District`  -3.623e+01  4.792e+00  -7.560
## neighbourhoodCleansedGramercy              -2.964e+01  5.142e+00  -5.765
## `neighbourhoodCleansedGreenwich Village`    7.824e+00  4.745e+00   1.649
## neighbourhoodCleansedHarlem                -7.552e+01  2.970e+00 -25.426
## `neighbourhoodCleansedHell's Kitchen`      -1.488e+01  3.156e+00  -4.716
## `neighbourhoodCleansedKips Bay`            -2.885e+01  4.582e+00  -6.296
## `neighbourhoodCleansedLower East Side`     -3.225e+01  3.540e+00  -9.109
## neighbourhoodCleansedMidtown               -2.622e+00  3.610e+00  -0.726
## `neighbourhoodCleansedMorningside Heights` -6.683e+01  4.733e+00 -14.120
## neighbourhoodCleansedOther                 -2.161e+01  3.142e+00  -6.879
## neighbourhoodCleansedSoHo                   2.442e+01  4.869e+00   5.016
## `neighbourhoodCleansedUpper East Side`     -3.386e+01  3.164e+00 -10.702
## `neighbourhoodCleansedUpper West Side`     -3.074e+01  3.077e+00  -9.991
## `neighbourhoodCleansedWashington Heights`  -7.903e+01  3.698e+00 -21.371
## `neighbourhoodCleansedWest Village`         9.385e+00  3.732e+00   2.515
## neighbourhoodGroupCleansedBrooklyn          2.936e+01  2.081e+01   1.410
## neighbourhoodGroupCleansedManhattan         8.031e+01  2.026e+01   3.963
## neighbourhoodGroupCleansedQueens            1.146e+01  2.114e+01   0.542
## `neighbourhoodGroupCleansedStaten Island`  -4.091e+01  4.386e+01  -0.933
## accommodates                                1.565e+01  6.766e-01  23.136
## bathrooms                                   5.933e+01  1.928e+00  30.777
## bedTypeCouch                                5.684e+00  1.184e+01   0.480
## bedTypeFuton                                1.551e+01  8.887e+00   1.745
## `bedTypePull-out Sofa`                      1.347e+01  8.713e+00   1.546
## `bedTypeReal Bed`                           1.261e+01  6.872e+00   1.835
## bedrooms                                    3.069e+01  1.195e+00  25.679
## beds                                        2.470e+00  1.099e+00   2.248
## TV                                          9.528e+00  1.387e+00   6.872
## cancellationPolicymoderate                 -4.676e+00  1.676e+00  -2.790
## cancellationPolicystrict                   -7.381e+00  1.523e+00  -4.846
## cancellationPolicysuper_strict_30           7.576e+01  3.905e+01   1.940
## cleaningFee                                 2.903e-01  1.492e-02  19.457
## extraPeople                                 2.075e-02  2.576e-02   0.805
## guestsIncluded                              2.267e+00  7.665e-01   2.958
## maximumNights                               5.594e-04  1.100e-03   0.509
```

```
## minimumNights                                  -1.196e-01  3.707e-02  -3.226
## `propertyTypeBed & Breakfast`                   2.229e+01  1.050e+01   2.123
## propertyTypeBoat                                1.894e+02  7.821e+01   2.422
## `propertyTypeBoutique hotel`                    2.728e+01  2.368e+01   1.152
## propertyTypeBungalow                           -5.078e+01  7.776e+01  -0.653
## propertyTypeCabin                               8.068e+01  7.785e+01   1.036
## propertyTypeCastle                              1.138e+02  7.776e+01   1.463
## propertyTypeCondominium                         5.173e+01  5.842e+00   8.855
## propertyTypeDorm                               -3.372e+01  4.494e+01  -0.750
## `propertyTypeGuest suite`                       1.718e+02  5.507e+01   3.119
## propertyTypeGuesthouse                          1.828e+01  2.181e+01   0.838
## propertyTypeHostel                             -1.339e+01  2.286e+01  -0.586
## propertyTypeHouse                               2.305e+01  4.920e+00   4.684
## propertyTypeHut                                -4.032e+01  7.810e+01  -0.516
## propertyTypeLighthouse                          1.948e+01  7.781e+01   0.250
## propertyTypeLoft                                5.226e+01  5.007e+00  10.436
## propertyTypeOther                               3.885e+01  8.122e+00   4.784
## `propertyTypeServiced apartment`                1.344e+02  3.486e+01   3.854
## propertyTypeTimeshare                           1.007e+02  1.328e+01   7.585
## propertyTypeTownhouse                           2.317e+01  6.107e+00   3.793
## `propertyTypeVacation home`                     1.895e+02  7.784e+01   2.435
## propertyTypeVilla                              -2.490e+01  3.485e+01  -0.715
## `roomTypePrivate room`                         -5.959e+01  1.563e+00 -38.122
## `roomTypeShared room`                          -7.367e+01  3.691e+00 -19.959
## securityDeposit                                -1.092e-02  3.415e-03  -3.197
## numberOfReviews                                -1.146e-01  2.135e-02  -5.368
## reviewScoresRating                             -1.929e-01  1.588e-02 -12.147
## checkIn24Hours1                                -1.017e+00  1.604e+00  -0.634
## airConditioning1                               -5.668e-01  1.836e+00  -0.309
## buzzerOrWirelessIntercom1                      -1.190e+00  1.397e+00  -0.852
## cableTV1                                        1.087e+01  1.510e+00   7.199
## carbonMonoxideDetector1                         1.233e+00  1.565e+00   0.788
## doorman1                                        9.324e+00  2.139e+00   4.360
## dryer1                                          1.328e+00  3.307e+00   0.401
## elevator1                                       8.761e+00  1.573e+00   5.570
## essentials1                                    -1.076e+01  1.807e+00  -5.957
## familyAndKidFriendly1                           2.486e+00  1.302e+00   1.908
## fireExtinguisher1                               2.506e+00  1.593e+00   1.573
## firstAidKit1                                    3.677e+00  1.620e+00   2.270
## freeParkingOnPremises1                         -1.743e+00  2.859e+00  -0.610
## gym1                                            1.881e+01  2.466e+00   7.627
## hairdryer1                                     -2.180e+00  1.676e+00  -1.301
## hangers1                                       -5.750e+00  1.703e+00  -3.375
## heating1                                       -3.964e+00  2.539e+00  -1.561
## hottub1                                        -5.525e+00  2.518e+00  -2.194
## internet1                                      -1.990e+00  1.513e+00  -1.315
## iron1                                          -3.102e+00  1.649e+00  -1.881
## kitchen1                                       -8.075e+00  2.798e+00  -2.886
## laptopFriendlyWorkspace1                       -1.008e+00  1.572e+00  -0.641
## lockOnBedroomDoor1                             -3.277e+00  1.707e+00  -1.920
```

```
## petsAllowed1                               -6.981e+00  1.903e+00  -3.669
## petsLiveOnThisProperty1                    -3.356e+00  2.385e+00  -1.407
## pool1                                       1.629e+01  4.996e+00   3.260
## selfCheckIn1                                9.135e+00  2.800e+00   3.262
## shampoo1                                    3.542e+00  1.465e+00   2.418
## smokeDetector1                             -5.792e+00  1.814e+00  -3.193
## smokingAllowed1                            -1.438e+00  2.783e+00  -0.517
## washer1                                     1.410e+00  3.315e+00   0.425
## wheelchairAccessible1                       1.250e+01  2.310e+00   5.412
## wirelessInternet1                          -8.504e-01  3.711e+00  -0.229
## hostHasProfilePic1                         -1.472e+01  8.836e+00  -1.666
## hostIdentityVerified1                      -1.561e+00  1.403e+00  -1.113
## hostIsSuperhost1                            1.632e+01  2.201e+00   7.414
## instantBookable1                           -5.241e+00  1.646e+00  -3.184
## isLocationExact1                            1.712e-01  1.658e+00   0.103
## requireGuestPhoneVerification1             -9.042e+00  5.316e+00  -1.701
## requireGuestProfilePicture1                 5.374e+00  5.641e+00   0.953
## featureCount                                6.047e-01  5.254e-01   1.151
## yearsAsHost                                 4.283e-01  3.736e-01   1.146
##                                            Pr(>|t|)
## (Intercept)                                0.810638
## hostResponseHours                          0.499429
## neighbourhoodCleansedChinatown             1.34e-11 ***
## `neighbourhoodCleansedEast Harlem`          < 2e-16 ***
## `neighbourhoodCleansedEast Village`         < 2e-16 ***
## `neighbourhoodCleansedFinancial District`  4.20e-14 ***
## neighbourhoodCleansedGramercy              8.30e-09 ***
## `neighbourhoodCleansedGreenwich Village`   0.099193 .
## neighbourhoodCleansedHarlem                 < 2e-16 ***
## `neighbourhoodCleansedHell's Kitchen`      2.43e-06 ***
## `neighbourhoodCleansedKips Bay`            3.12e-10 ***
## `neighbourhoodCleansedLower East Side`      < 2e-16 ***
## neighbourhoodCleansedMidtown               0.467631
## `neighbourhoodCleansedMorningside Heights`  < 2e-16 ***
## neighbourhoodCleansedOther                 6.21e-12 ***
## neighbourhoodCleansedSoHo                  5.32e-07 ***
## `neighbourhoodCleansedUpper East Side`      < 2e-16 ***
## `neighbourhoodCleansedUpper West Side`      < 2e-16 ***
## `neighbourhoodCleansedWashington Heights`   < 2e-16 ***
## `neighbourhoodCleansedWest Village`        0.011916 *
## neighbourhoodGroupCleansedBrooklyn         0.158411
## neighbourhoodGroupCleansedManhattan        7.42e-05 ***
## neighbourhoodGroupCleansedQueens           0.587807
## `neighbourhoodGroupCleansedStaten Island`  0.350991
## accommodates                                < 2e-16 ***
## bathrooms                                   < 2e-16 ***
## bedTypeCouch                               0.631243
## bedTypeFuton                               0.081039 .
## `bedTypePull-out Sofa`                     0.122114
## `bedTypeReal Bed`                          0.066476 .
```

```
## bedrooms                                        < 2e-16 ***
## beds                                           0.024593 *
## TV                                             6.55e-12 ***
## cancellationPolicymoderate                     0.005281 **
## cancellationPolicystrict                       1.27e-06 ***
## cancellationPolicysuper_strict_30              0.052395 .
## cleaningFee                                     < 2e-16 ***
## extraPeople                                    0.420684
## guestsIncluded                                 0.003098 **
## maximumNights                                  0.611069
## minimumNights                                  0.001257 **
## `propertyTypeBed & Breakfast`                  0.033781 *
## propertyTypeBoat                               0.015450 *
## `propertyTypeBoutique hotel`                   0.249475
## propertyTypeBungalow                           0.513692
## propertyTypeCabin                              0.300041
## propertyTypeCastle                             0.143445
## propertyTypeCondominium                         < 2e-16 ***
## propertyTypeDorm                               0.453056
## `propertyTypeGuest suite`                      0.001817 **
## propertyTypeGuesthouse                         0.401960
## propertyTypeHostel                             0.558194
## propertyTypeHouse                              2.83e-06 ***
## propertyTypeHut                                0.605661
## propertyTypeLighthouse                         0.802309
## propertyTypeLoft                                < 2e-16 ***
## propertyTypeOther                              1.73e-06 ***
## `propertyTypeServiced apartment`               0.000117 ***
## propertyTypeTimeshare                          3.49e-14 ***
## propertyTypeTownhouse                          0.000149 ***
## `propertyTypeVacation home`                    0.014895 *
## propertyTypeVilla                              0.474914
## `roomTypePrivate room`                          < 2e-16 ***
## `roomTypeShared room`                           < 2e-16 ***
## securityDeposit                                0.001391 **
## numberOfReviews                                8.05e-08 ***
## reviewScoresRating                              < 2e-16 ***
## checkIn24Hours1                                0.526227
## airConditioning1                               0.757506
## buzzerOrWirelessIntercom1                      0.394082
## cableTV1                                       6.28e-13 ***
## carbonMonoxideDetector1                        0.430972
## doorman1                                       1.31e-05 ***
## dryer1                                         0.688076
## elevator1                                      2.58e-08 ***
## essentials1                                    2.61e-09 ***
## familyAndKidFriendly1                          0.056355 .
## fireExtinguisher1                              0.115840
## firstAidKit1                                   0.023231 *
## freeParkingOnPremises1                         0.542138
```

```
## gym1                                         2.52e-14 ***
## hairdryer1                                   0.193193
## hangers1                                     0.000739 ***
## heating1                                     0.118523
## hottub1                                      0.028234 *
## internet1                                    0.188424
## iron1                                        0.060003 .
## kitchen1                                     0.003901 **
## laptopFriendlyWorkspace1                     0.521430
## lockOnBedroomDoor1                           0.054906 .
## petsAllowed1                                 0.000244 ***
## petsLiveOnThisProperty1                      0.159463
## pool1                                        0.001117 **
## selfCheckIn1                                 0.001107 **
## shampoo1                                     0.015627 *
## smokeDetector1                               0.001409 **
## smokingAllowed1                              0.605261
## washer1                                      0.670602
## wheelchairAccessible1                        6.32e-08 ***
## wirelessInternet1                            0.818755
## hostHasProfilePic1                           0.095655 .
## hostIdentityVerified1                        0.265760
## hostIsSuperhost1                             1.28e-13 ***
## instantBookable1                             0.001453 **
## isLocationExact1                             0.917736
## requireGuestPhoneVerification1               0.088954 .
## requireGuestProfilePicture1                  0.340768
## featureCount                                 0.249781
## yearsAsHost                                  0.251705
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.65 on 19033 degrees of freedom
## Multiple R-squared:  0.5951, Adjusted R-squared:  0.5928
## F-statistic:   259 on 108 and 19033 DF,  p-value: < 2.2e-16
```

```
print(lm2)
```

```
## Linear Regression
##
## 19142 samples
##    62 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 17228, 17227, 17227, 17227, 17229, 17228, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   78.16914  0.5874553  50.29873
```

```
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Removing features from Linear Model with high p-values.

```
print("Removing features with high p-values in cross validated linear model")

## [1] "Removing features with high p-values in cross validated linear model"

dropColumns <-
c("hostResponseHours","neighbourhoodGroupCleansed","bedType","beds","extraPeo
ple","maximumNights","checkIn24Hours","airConditioning","buzzerOrWirelessInte
rcom","dryer","familyAndKidFriendly","fireExtinguisher","firstAidKit","freePa
rkingOnPremises","hairdryer","hangers","heating","hottub","internet","iron","
kitchen","laptopFriendlyWorkspace","lockOnBedroomDoor","petsLiveOnThisPropert
y","selfCheckIn","shampoo","smokingAllowed","washer","wirelessInternet","host
HasProfilePic","hostIdentityVerified","instantBookable","isLocationExact","re
quireGuestPhoneVerification","requireGuestProfilePicture","featureCount","yea
rsAsHost")
airDataValid <- airDataValid[,!(names(airDataValid) %in% dropColumns)]
airDataValid <- as.data.frame(airDataValid)
```

Summary of data set after removal of features.

```
print("Summary after deletion of columns")

## [1] "Summary after deletion of columns"

print(summary(airDataValid))

##     neighbourhoodCleansed  accommodates      bathrooms
##  Harlem           :2475      Min.   : 1.000   Min.   :0.000
##  Other            :1948      1st Qu.: 2.000   1st Qu.:1.000
##  East Village     :1819      Median : 2.000   Median :1.000
##  Upper West Side:1738      Mean   : 2.779   Mean   :1.092
##  Hell's Kitchen :1549      3rd Qu.: 4.000   3rd Qu.:1.000
##  Upper East Side:1507      Max.   :16.000   Max.   :5.000
##  (Other)          :8106
##     bedrooms             TV               cancellationPolicy
##  Min.   :0.000   Min.   :0.0000   flexible       :5741
##  1st Qu.:1.000   1st Qu.:0.0000   moderate       :4234
##  Median :1.000   Median :1.0000   strict         :9163
##  Mean   :1.085   Mean   :0.6877   super_strict_30:   4
##  3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :6.000   Max.   :2.0000
##
##   cleaningFee     guestsIncluded   minimumNights          price
##  Min.   :  0.00   Min.   : 1.000   Min.   :   1.000   Min.   : 10.0
##  1st Qu.:  0.00   1st Qu.: 1.000   1st Qu.:   1.000   1st Qu.: 89.0
##  Median : 40.00   Median : 1.000   Median :   2.000   Median :135.0
##  Mean   : 47.39   Mean   : 1.427   Mean   :   4.142   Mean   :166.3
##  3rd Qu.: 75.00   3rd Qu.: 2.000   3rd Qu.:   4.000   3rd Qu.:200.0
```

```
##  Max.   :600.00   Max.   :16.000   Max.   :1250.000   Max.   :999.0
##
##       propertyType              roomType     securityDeposit
##  Apartment  :18001   Entire home/apt:10705   Min.   :  0.0
##  House      :  280   Private room   : 7864   1st Qu.:  0.0
##  Loft       :  253   Shared room    :  573   Median :  0.0
##  Condominium:  187                           Mean   :123.1
##  Townhouse  :  171                           3rd Qu.:200.0
##  Other      :   99                           Max.   :999.0
##  (Other)    :  151
##  numberOfReviews  reviewScoresRating cableTV   carbonMonoxideDetector
##  Min.   :  0.00   Min.   :  0.00     0:12787   0: 8110
##  1st Qu.:  1.00   1st Qu.: 67.00     1: 6355   1:11032
##  Median :  5.00   Median : 92.00
##  Mean   : 16.97   Mean   : 70.83
##  3rd Qu.: 19.00   3rd Qu.: 98.00
##  Max.   :432.00   Max.   :100.00
##
##  doorman   elevator  essentials gym       petsAllowed pool
##  0:16449   0:11362   0: 3182    0:17424   0:16871     0:18857
##  1: 2693   1: 7780   1:15960    1: 1718   1: 2271     1:  285
##
##
##
##
##
##  smokeDetector wheelchairAccessible hostIsSuperhost
##  0: 4090       0:17526              0:17447
##  1:15052       1: 1616              1: 1695
##
##
##
##
##
```

Using K-fold validation on Linear Model to check if model still performs comparably after removing features based on p-values. The aim is to achieve close to same R-squared model with a muich smaller set of features.

```
print("Retraining linear model with repeated k-fold cross validation")

## [1] "Retraining linear model with repeated k-fold cross validation"

train.control <- trainControl(method="repeatedcv",number=10,repeats=3)
lm3 <- train(price~.,data=airDataValid,method="lm",trControl=train.control)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Summary of K-fold model

```r
print("Summary of Linear Regression")

## [1] "Summary of Linear Regression"
```

```r
print(summary(lm3))
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -521.33  -39.97   -5.71   27.14  895.89
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                  74.98239    3.76614  19.910
## neighbourhoodCleansedChinatown              -32.29262    4.85327  -6.654
## `neighbourhoodCleansedEast Harlem`          -70.72683    3.48853 -20.274
## `neighbourhoodCleansedEast Village`         -26.94199    3.07905  -8.750
## `neighbourhoodCleansedFinancial District`   -36.54804    4.81209  -7.595
## neighbourhoodCleansedGramercy               -30.27116    5.17033  -5.855
## `neighbourhoodCleansedGreenwich Village`      7.59746    4.77357   1.592
## neighbourhoodCleansedHarlem                 -74.95405    2.96750 -25.258
## `neighbourhoodCleansedHell's Kitchen`       -15.79077    3.16646  -4.987
## `neighbourhoodCleansedKips Bay`             -30.02821    4.60714  -6.518
## `neighbourhoodCleansedLower East Side`      -32.23106    3.55594  -9.064
## neighbourhoodCleansedMidtown                 -2.73585    3.61804  -0.756
## `neighbourhoodCleansedMorningside Heights` -66.61796    4.73675 -14.064
## neighbourhoodCleansedOther                  -32.80271    3.04366 -10.777
## neighbourhoodCleansedSoHo                    24.65450    4.89290   5.039
## `neighbourhoodCleansedUpper East Side`      -34.25711    3.17934 -10.775
## `neighbourhoodCleansedUpper West Side`      -31.13937    3.09165 -10.072
## `neighbourhoodCleansedWashington Heights`   -78.77315    3.70312 -21.272
## `neighbourhoodCleansedWest Village`           9.68040    3.75370   2.579
## accommodates                                 16.58471    0.55864  29.687
## bathrooms                                    60.12775    1.91176  31.451
## bedrooms                                     31.43630    1.15731  27.163
## TV                                            9.17734    1.36551   6.721
## cancellationPolicymoderate                   -5.01875    1.66583  -3.013
## cancellationPolicystrict                     -8.19915    1.50180  -5.460
## cancellationPolicysuper_strict_30            64.01483   39.23118   1.632
## cleaningFee                                   0.30057    0.01480  20.309
## guestsIncluded                                2.59663    0.72828   3.565
## minimumNights                                -0.11693    0.03726  -3.138
## `propertyTypeBed & Breakfast`                25.25500   10.50638   2.404
## propertyTypeBoat                            203.82930   78.41415   2.599
## `propertyTypeBoutique hotel`                 30.80298   23.71069   1.299
## propertyTypeBungalow                        -55.79836   78.25616  -0.713
## propertyTypeCabin                            89.24455   78.25960   1.140
## propertyTypeCastle                          117.84411   78.26117   1.506
## propertyTypeCondominium                      52.68935    5.85792   8.995
## propertyTypeDorm                            -25.64820   45.21056  -0.567
## `propertyTypeGuest suite`                   164.38148   55.37552   2.968
```

```
## propertyTypeGuesthouse                      17.76576   21.89237   0.812
## propertyTypeHostel                         -12.68207   22.76451  -0.557
## propertyTypeHouse                           12.49786    4.80199   2.603
## propertyTypeHut                            -22.88089   78.35598  -0.292
## propertyTypeLighthouse                      20.85363   78.29605   0.266
## propertyTypeLoft                            53.65344    5.02030  10.687
## propertyTypeOther                           46.05459    7.98749   5.766
## `propertyTypeServiced apartment`           132.88281   35.03730   3.793
## propertyTypeTimeshare                      105.15306   13.27982   7.918
## propertyTypeTownhouse                       24.66136    6.10609   4.039
## `propertyTypeVacation home`                181.35300   78.29776   2.316
## propertyTypeVilla                          -25.49723   35.06686  -0.727
## `roomTypePrivate room`                     -60.70211    1.48249 -40.946
## `roomTypeShared room`                      -73.27667    3.52769 -20.772
## securityDeposit                             -0.01025    0.00341  -3.006
## numberOfReviews                             -0.11046    0.02033  -5.433
## reviewScoresRating                          -0.19633    0.01561 -12.582
## cableTV1                                    12.10858    1.37094   8.832
## carbonMonoxideDetector1                      2.74906    1.41997   1.936
## doorman1                                    11.98476    2.04470   5.861
## elevator1                                   10.68763    1.40502   7.607
## essentials1                                -10.76535    1.60247  -6.718
## gym1                                        19.45066    2.39740   8.113
## petsAllowed1                                -6.46179    1.76779  -3.655
## pool1                                       17.80043    4.93093   3.610
## smokeDetector1                              -4.71966    1.70954  -2.761
## wheelchairAccessible1                       14.25975    2.22030   6.422
## hostIsSuperhost1                            16.87090    2.08539   8.090
##                                            Pr(>|t|)
## (Intercept)                                 < 2e-16 ***
## neighbourhoodCleansedChinatown             2.93e-11 ***
## `neighbourhoodCleansedEast Harlem`          < 2e-16 ***
## `neighbourhoodCleansedEast Village`         < 2e-16 ***
## `neighbourhoodCleansedFinancial District`  3.22e-14 ***
## neighbourhoodCleansedGramercy              4.85e-09 ***
## `neighbourhoodCleansedGreenwich Village`   0.111498
## neighbourhoodCleansedHarlem                 < 2e-16 ***
## `neighbourhoodCleansedHell's Kitchen`      6.19e-07 ***
## `neighbourhoodCleansedKips Bay`            7.32e-11 ***
## `neighbourhoodCleansedLower East Side`      < 2e-16 ***
## neighbourhoodCleansedMidtown               0.449558
## `neighbourhoodCleansedMorningside Heights`  < 2e-16 ***
## neighbourhoodCleansedOther                  < 2e-16 ***
## neighbourhoodCleansedSoHo                  4.73e-07 ***
## `neighbourhoodCleansedUpper East Side`      < 2e-16 ***
## `neighbourhoodCleansedUpper West Side`      < 2e-16 ***
## `neighbourhoodCleansedWashington Heights`   < 2e-16 ***
## `neighbourhoodCleansedWest Village`        0.009919 **
## accommodates                                < 2e-16 ***
## bathrooms                                   < 2e-16 ***
```

```
## bedrooms                                      < 2e-16 ***
## TV                                           1.86e-11 ***
## cancellationPolicymoderate                   0.002592 **
## cancellationPolicystrict                     4.83e-08 ***
## cancellationPolicysuper_strict_30            0.102752
## cleaningFee                                   < 2e-16 ***
## guestsIncluded                               0.000364 ***
## minimumNights                                0.001705 **
## `propertyTypeBed & Breakfast`                0.016236 *
## propertyTypeBoat                             0.009346 **
## `propertyTypeBoutique hotel`                 0.193919
## propertyTypeBungalow                         0.475841
## propertyTypeCabin                            0.254148
## propertyTypeCastle                           0.132140
## propertyTypeCondominium                       < 2e-16 ***
## propertyTypeDorm                             0.570513
## `propertyTypeGuest suite`                    0.002996 **
## propertyTypeGuesthouse                       0.417086
## propertyTypeHostel                           0.577467
## propertyTypeHouse                            0.009258 **
## propertyTypeHut                              0.770281
## propertyTypeLighthouse                       0.789978
## propertyTypeLoft                              < 2e-16 ***
## propertyTypeOther                            8.25e-09 ***
## `propertyTypeServiced apartment`             0.000150 ***
## propertyTypeTimeshare                        2.54e-15 ***
## propertyTypeTownhouse                        5.39e-05 ***
## `propertyTypeVacation home`                  0.020558 *
## propertyTypeVilla                            0.467172
## `roomTypePrivate room`                        < 2e-16 ***
## `roomTypeShared room`                         < 2e-16 ***
## securityDeposit                              0.002651 **
## numberOfReviews                              5.61e-08 ***
## reviewScoresRating                            < 2e-16 ***
## cableTV1                                      < 2e-16 ***
## carbonMonoxideDetector1                      0.052883 .
## doorman1                                     4.67e-09 ***
## elevator1                                    2.94e-14 ***
## essentials1                                  1.89e-11 ***
## gym1                                         5.23e-16 ***
## petsAllowed1                                 0.000258 ***
## pool1                                        0.000307 ***
## smokeDetector1                               0.005772 **
## wheelchairAccessible1                        1.37e-10 ***
## hostIsSuperhost1                             6.32e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.2 on 19076 degrees of freedom
```

```
## Multiple R-squared:  0.5885, Adjusted R-squared:  0.5871
## F-statistic: 419.6 on 65 and 19076 DF,  p-value: < 2.2e-16
```

```
print(lm3)
```

```
## Linear Regression
##
## 19142 samples
##    25 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 17228, 17228, 17227, 17228, 17228, 17228, ...
## Resampling results:
##
##   RMSE      Rsquared  MAE
##   78.52298  0.583468  50.50286
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Using Regression Tree to model price vs features.

```
print("Creating Regression Tree for prediction")
```

```
## [1] "Creating Regression Tree for prediction"
```

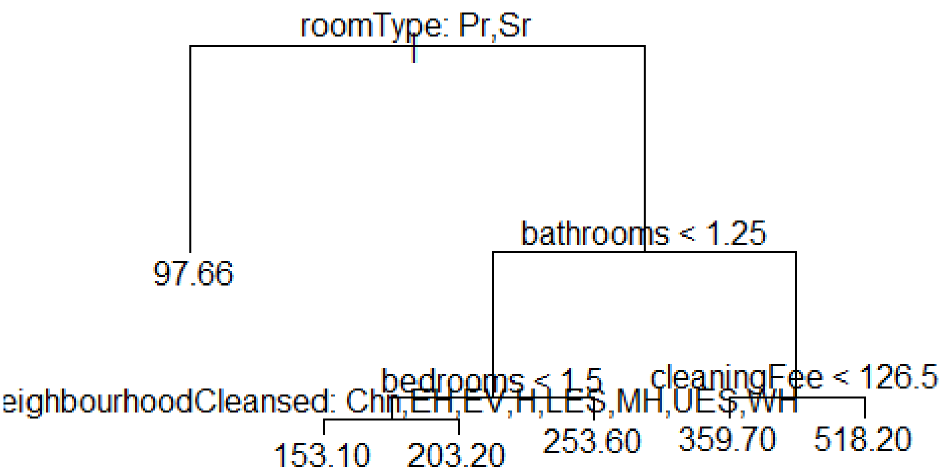
```
airTree <- tree(price~., data=airDataTrain)
```

Tree plot.

```
print("Plotting Tree")
```

```
## [1] "Plotting Tree"
```

```
plot(airTree)
text(airTree,pretty=TRUE)
```



Tree-based prediction.

```
print("Using Tree Model to predict for Test data")

## [1] "Using Tree Model to predict for Test data"

treePred <- predict(airTree, airDataTest)
print(paste("Test MSE:",mean((airDataTest$price - treePred) ^ 2)))

## [1] "Test MSE: 7471.89554988577"
```

Pruning tree to get best tree.

```
print("Using Cross Validation to get Pruned Tree")

## [1] "Using Cross Validation to get Pruned Tree"

cvtree <- cv.tree(airTree,FUN=prune.tree,K=10)
```

Pruning: Sizes considered

```
print("Sizes considered:")

## [1] "Sizes considered:"

print(cvtree$size)

## [1] 6 5 4 3 2 1
```

Pruning: k-values considered

```r
print("Various alphas (k's) considered:")
```

```
## [1] "Various alphas (k's) considered:"
```

```r
print(cvtree$k)
```

```
## [1]      -Inf  4457445  6423652  7416235 48021749 67434626
```

Pruning: misclassification rates for corresponding k-values

```r
print("Corresponding misclassification rates:")
```

```
## [1] "Corresponding misclassification rates:"
```

```r
print(cvtree$dev)
```

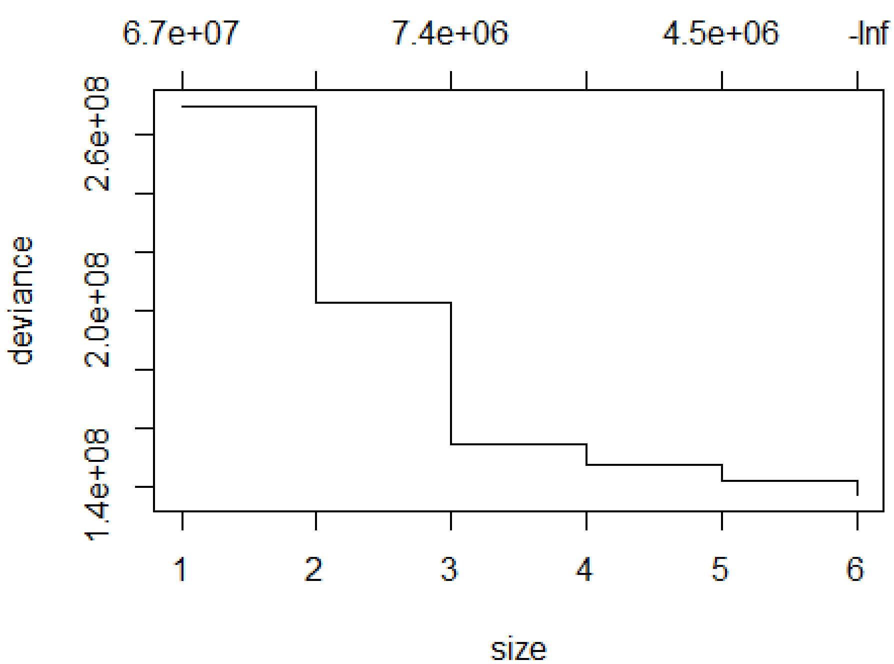
```
## [1] 137147120 141796176 147723058 154722397 202653593 270084492
```

Pruning: Plot of misclassification rate

```r
print("plot of misclassification rate:")
```

```
## [1] "plot of misclassification rate:"
```

```r
plot.tree.sequence(cvtree)
```



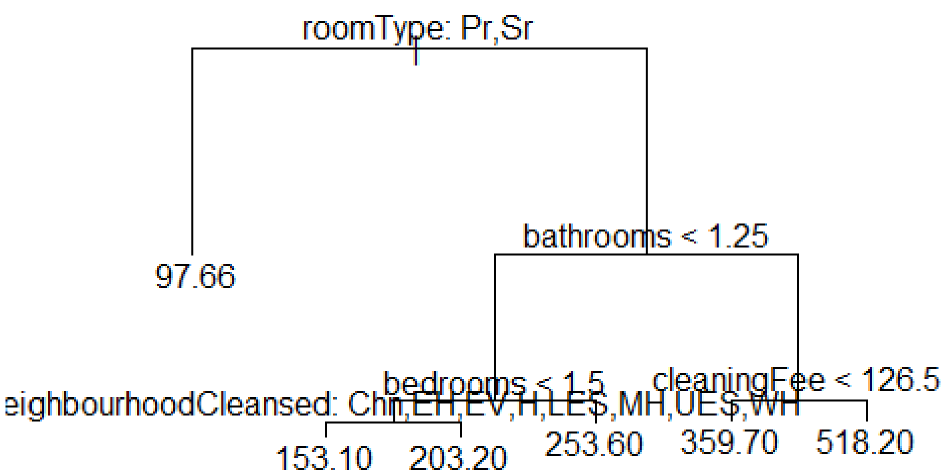
Extracting best tree model.

```r
print("Extracting best tree from CV model")
```

```
## [1] "Extracting best tree from CV model"

bestTree <- prune.tree(airTree, best=cvtree$size[which.min(cvtree$dev)])
```

Plotting best tree.

```
print("Plotting best tree")

## [1] "Plotting best tree"

plot(bestTree)
text(bestTree,pretty=TRUE)
```



Using best tree for prediction on test data.

```
print("Using Best Tree Model to predict for Test data")

## [1] "Using Best Tree Model to predict for Test data"

treePred2 <- predict(bestTree, airDataTest)
print(paste("Test MSE:",mean((airDataTest$price - treePred2) ^ 2)))

## [1] "Test MSE: 7471.89554988577"
```

Linear model with interaction. It is possible that the pricing of a listing in NYC is subjective and/or dependent on features not available in this set (like age of building, rent control etc.). It is also possible that the true relationship between price and some of the features in non-linear. Running a linear model with interaction or higher polynomial order might reveal these relationships. However, running a interaction model with even 27 variables for polynomial degree=2 takes prohibitively long time (~10 minutes). Running more complex models is not possible unless dedicated machines are available and/or time is not a constraint. The output of the quadratic model is pasted here for reference and there are indications that the relationship of price with at least some of these features may be non-linear.

```
print("Linear Regression with interaction")

## [1] "Linear Regression with interaction"

#lmi <- lm(price~(.)^2,data=airDataTrain)
#print(summary(lmi))
#lmiPred <- predict(lmi,airDataTest)
#print(mean((airDataTest$price - lmiPred) ^ 2))
print("Linear Regression with interaction runs prohibitively long (~10
minutes for polynomial degree 2)")

## [1] "Linear Regression with interaction runs prohibitively long (~10
minutes for polynomial degree 2)"

print("The output of the above commented code is pasted below for reference")

## [1] "The output of the above commented code is pasted below for reference"

print("***************************************************************")

## [1] "***************************************************************"

print("Residual standard error: 70.57 on 17072 degrees of freedom")

## [1] "Residual standard error: 70.57 on 17072 degrees of freedom"

print("Multiple R-squared:  0.684,  Adjusted R-squared:  0.6635")

## [1] "Multiple R-squared:  0.684,\tAdjusted R-squared:  0.6635"

print("F-statistic: 33.27 on 1111 and 17072 DF,  p-value: < 2.2e-16")

## [1] "F-statistic: 33.27 on 1111 and 17072 DF,  p-value: < 2.2e-16"

print("***************************************************************")

## [1] "***************************************************************"

print("This indicates that the relationship between price and othe features
may be non-linear and/or there may be an interaction effect between certain
features.")

## [1] "This indicates that the relationship between price and othe features
may be non-linear and/or there may be an interaction effect between certain
features."
```

Using data set with all columns for Ridge and Lasso regression. Recreating training and test data sets from this data set.

```
print("Recreating training and test data sets with all columns in cleaned dataSet")

## [1] "Recreating training and test data sets with all columns in cleaned dataSet"

airDataNew <- airDataClean[airDataClean$price > 0,]

if(log_regression){
  airDataNew$price <- log(airDataNew$price)
}

airDataTrain <- airDataNew[train,]
airDataTest <- airDataNew[-train,]
```

glm (Ridge/Lasso/Elastic Net) requires data to be split into two objects: a model matrix with all features and a label object with the response.

```
featureSet <- model.matrix(price~.,airDataTrain)[,-1]
responseSet <- airDataTrain$price
```

Running Ridge regression model.

```
print("Running Ridge Regression Model")

## [1] "Running Ridge Regression Model"

print("Using Cross Validation to determine value of Shrinkage Parameter")

## [1] "Using Cross Validation to determine value of Shrinkage Parameter"

cv <- cv.glmnet(featureSet, responseSet, alpha=0)
minLambda <- cv$lambda.min
print(paste("Min. Lambda determined: ",minLambda))

## [1] "Min. Lambda determined:  7.28207895394829"

ridgeModel <- glmnet(featureSet, responseSet, alpha=0, lambda=minLambda)
print(summary(ridgeModel))

##           Length Class     Mode
## a0          1    -none-    numeric
## beta      115    dgCMatrix S4
## df          1    -none-    numeric
## dim         2    -none-    numeric
## lambda      1    -none-    numeric
## dev.ratio   1    -none-    numeric
## nulldev     1    -none-    numeric
## npasses     1    -none-    numeric
## jerr        1    -none-    numeric
```

```
## offset      1    -none-    logical
## call        5    -none-    call
## nobs        1    -none-    numeric
```

```r
print(coef(ridgeModel))
```

```
## 116 x 1 sparse Matrix of class "dgCMatrix"
##                                                         s0
## (Intercept)                                   25.043539761
## hostResponseHours                             -0.039603960
## neighbourhoodCleansedChinatown               -13.654989883
## neighbourhoodCleansedEast Harlem             -51.720351044
## neighbourhoodCleansedEast Village            -10.335189986
## neighbourhoodCleansedFinancial District      -20.541644538
## neighbourhoodCleansedGramercy                -12.883093447
## neighbourhoodCleansedGreenwich Village        23.668712529
## neighbourhoodCleansedHarlem                  -56.451700566
## neighbourhoodCleansedHell's Kitchen            0.553424969
## neighbourhoodCleansedKips Bay                -10.849036469
## neighbourhoodCleansedLower East Side         -16.191388834
## neighbourhoodCleansedMidtown                  13.475026658
## neighbourhoodCleansedMorningside Heights -48.186784120
## neighbourhoodCleansedOther                    -5.552506097
## neighbourhoodCleansedSoHo                     39.335445730
## neighbourhoodCleansedUpper East Side         -16.863818214
## neighbourhoodCleansedUpper West Side         -12.558955419
## neighbourhoodCleansedWashington Heights      -60.176767465
## neighbourhoodCleansedWest Village             25.868568448
## neighbourhoodGroupCleansedBrooklyn            -6.962422080
## neighbourhoodGroupCleansedManhattan           41.412957060
## neighbourhoodGroupCleansedQueens             -25.279363666
## neighbourhoodGroupCleansedStaten Island      -64.990841678
## accommodates                                  14.376611974
## bathrooms                                     56.816726053
## bedTypeCouch                                   5.413123281
## bedTypeFuton                                  12.393776859
## bedTypePull-out Sofa                           9.845004385
## bedTypeReal Bed                               11.827860818
## bedrooms                                      28.020448692
## beds                                           5.642655602
## TV                                             9.824073249
## cancellationPolicymoderate                    -4.429111284
## cancellationPolicystrict                      -5.826180115
## cancellationPolicysuper_strict_30             64.501598706
## cleaningFee                                    0.292998222
## extraPeople                                    0.010252366
## guestsIncluded                                 3.148255736
## maximumNights                                  0.001111189
## minimumNights                                 -0.111734515
## propertyTypeBed & Breakfast                   18.296985749
```

```
## propertyTypeBoat                              187.582436507
## propertyTypeBoutique hotel                     46.919576463
## propertyTypeBungalow                          -50.633094209
## propertyTypeCabin                              77.846360392
## propertyTypeCastle                            104.297230049
## propertyTypeCondominium                        50.611500558
## propertyTypeDorm                              -34.422643201
## propertyTypeGuest suite                       162.115788503
## propertyTypeGuesthouse                         13.056811104
## propertyTypeHostel                              5.276471103
## propertyTypeHouse                              20.663935954
## propertyTypeHut                               -34.905016186
## propertyTypeLighthouse                         29.850676484
## propertyTypeLoft                               51.926833711
## propertyTypeOther                              37.824416922
## propertyTypeServiced apartment                 88.733575298
## propertyTypeTimeshare                          98.016744314
## propertyTypeTownhouse                          20.566627679
## propertyTypeVacation home                     183.360915282
## propertyTypeVilla                             -14.295025364
## roomTypePrivate room                          -55.252579318
## roomTypeShared room                           -67.266866825
## securityDeposit                                -0.010197838
## numberOfReviews                                -0.069674940
## reviewScoresAccuracy                            0.081488460
## reviewScoresCheckin                            -0.853682725
## reviewScoresCleanliness                         1.419621984
## reviewScoresCommunication                      -0.855844252
## reviewScoresLocation                           -0.462290473
## reviewScoresRating                              0.041035068
## reviewScoresValue                              -1.497753171
## reviewsperMonth                                -1.523936975
## checkIn24Hours1                                -0.857422781
## airConditioning1                                0.853827404
## buzzerOrWirelessIntercom1                      -1.152413942
## cableTV1                                       11.394115083
## carbonMonoxideDetector1                         1.233845313
## doorman1                                        8.912556864
## dryer1                                          1.374740077
## elevator1                                       7.896844295
## essentials1                                    -9.281113779
## familyAndKidFriendly1                           3.905972489
## fireExtinguisher1                               3.221546292
## firstAidKit1                                    2.579599060
## freeParkingOnPremises1                         -2.620506541
## gym1                                           19.267061221
## hairdryer1                                     -1.690504110
## hangers1                                       -5.304232986
## heating1                                       -3.389636633
## hottub1                                        -4.777520593
```

```
## internet1                                  -1.436545199
## iron1                                      -3.082232969
## kitchen1                                   -6.646908205
## laptopFriendlyWorkspace1                   -0.639189669
## lockOnBedroomDoor1                         -3.250060944
## petsAllowed1                               -6.078464179
## petsLiveOnThisProperty1                    -3.845922966
## pool1                                      17.052716545
## selfCheckIn1                               10.228325098
## shampoo1                                    4.127420520
## smokeDetector1                             -5.447946148
## smokingAllowed1                            -0.886073773
## washer1                                     2.844872414
## wheelchairAccessible1                      12.086076800
## wirelessInternet1                          -2.255280422
## hostHasProfilePic1                        -13.497110568
## hostIdentityVerified1                      -1.084456368
## hostIsSuperhost1                           15.260718439
## instantBookable1                           -4.306107639
## isLocationExact1                            0.371599873
## requireGuestPhoneVerification1             -9.282581713
## requireGuestProfilePicture1                 2.956270521
## featureCount                                0.420047543
## yearsAsHost                                 0.407190999
```

```
ridgeTest <- model.matrix(price~.,airDataTest)[,-1]
ridgePred <- predict(ridgeModel,ridgeTest)
print(paste("RMSE:",RMSE(ridgePred, airDataTest$price)))
```

```
## [1] "RMSE: 78.3890763404307"
```

```
print(paste("R-squared:",R2(ridgePred, airDataTest$price)))
```

```
## [1] "R-squared: 0.560859666021504"
```

Running LASSO regression model.

```
print("Running Lasso Regression Model")
```

```
## [1] "Running Lasso Regression Model"
```

```
print("Using Cross Validation to determine value of Shrinkage Parameter")
```

```
## [1] "Using Cross Validation to determine value of Shrinkage Parameter"
```

```
cv <- cv.glmnet(featureSet, responseSet, alpha=1)
minLambda <- cv$lambda.min
print(paste("Min. Lambda determined: ",minLambda))
```

```
## [1] "Min. Lambda determined:  0.0468097358807009"
```

```r
lassoModel <- glmnet(featureSet, responseSet, alpha=1, lambda=minLambda)
print(summary(lassoModel))
```

```
##           Length Class     Mode
## a0          1    -none-    numeric
## beta      115    dgCMatrix S4
## df          1    -none-    numeric
## dim         2    -none-    numeric
## lambda      1    -none-    numeric
## dev.ratio   1    -none-    numeric
## nulldev     1    -none-    numeric
## npasses     1    -none-    numeric
## jerr        1    -none-    numeric
## offset      1    -none-    logical
## call        5    -none-    call
## nobs        1    -none-    numeric
```

```r
print(coef(lassoModel))
```

```
## 116 x 1 sparse Matrix of class "dgCMatrix"
##                                                       s0
## (Intercept)                                  1.083313e+01
## hostResponseHours                           -4.120741e-02
## neighbourhoodCleansedChinatown              -2.853248e+01
## neighbourhoodCleansedEast Harlem            -6.820004e+01
## neighbourhoodCleansedEast Village           -2.495130e+01
## neighbourhoodCleansedFinancial District     -3.528480e+01
## neighbourhoodCleansedGramercy               -2.692041e+01
## neighbourhoodCleansedGreenwich Village       1.033748e+01
## neighbourhoodCleansedHarlem                 -7.332343e+01
## neighbourhoodCleansedHell's Kitchen         -1.283230e+01
## neighbourhoodCleansedKips Bay               -2.512141e+01
## neighbourhoodCleansedLower East Side        -3.079013e+01
## neighbourhoodCleansedMidtown                 .
## neighbourhoodCleansedMorningside Heights -6.469314e+01
## neighbourhoodCleansedOther                  -1.918424e+01
## neighbourhoodCleansedSoHo                    2.690373e+01
## neighbourhoodCleansedUpper East Side        -3.157926e+01
## neighbourhoodCleansedUpper West Side        -2.720527e+01
## neighbourhoodCleansedWashington Heights     -7.721025e+01
## neighbourhoodCleansedWest Village            1.260491e+01
## neighbourhoodGroupCleansedBrooklyn           1.886843e+01
## neighbourhoodGroupCleansedManhattan          6.957408e+01
## neighbourhoodGroupCleansedQueens             .
## neighbourhoodGroupCleansedStaten Island     -4.157853e+01
## accommodates                                 1.587934e+01
## bathrooms                                    5.949907e+01
## bedTypeCouch                                 7.166551e+00
## bedTypeFuton                                 1.580840e+01
## bedTypePull-out Sofa                         1.246381e+01
```

```
## bedTypeReal Bed                              1.319597e+01
## bedrooms                                     3.034575e+01
## beds                                         2.564823e+00
## TV                                           8.940970e+00
## cancellationPolicymoderate                  -5.046291e+00
## cancellationPolicystrict                    -6.930013e+00
## cancellationPolicysuper_strict_30            6.090565e+01
## cleaningFee                                  2.843693e-01
## extraPeople                                  1.341034e-02
## guestsIncluded                               2.551104e+00
## maximumNights                                8.629911e-04
## minimumNights                               -1.130853e-01
## propertyTypeBed & Breakfast                  1.849352e+01
## propertyTypeBoat                             1.823427e+02
## propertyTypeBoutique hotel                   4.750546e+01
## propertyTypeBungalow                        -4.609131e+01
## propertyTypeCabin                            7.278802e+01
## propertyTypeCastle                           1.055791e+02
## propertyTypeCondominium                      5.131636e+01
## propertyTypeDorm                            -2.914546e+01
## propertyTypeGuest suite                      1.666482e+02
## propertyTypeGuesthouse                       1.492195e+01
## propertyTypeHostel                          -5.657774e+00
## propertyTypeHouse                            2.301196e+01
## propertyTypeHut                             -2.556598e+01
## propertyTypeLighthouse                       2.386969e+01
## propertyTypeLoft                             5.226207e+01
## propertyTypeOther                            4.000240e+01
## propertyTypeServiced apartment               9.124248e+01
## propertyTypeTimeshare                        1.014449e+02
## propertyTypeTownhouse                        2.181640e+01
## propertyTypeVacation home                    1.799646e+02
## propertyTypeVilla                           -1.309771e+01
## roomTypePrivate room                        -5.879000e+01
## roomTypeShared room                         -7.253163e+01
## securityDeposit                             -1.286394e-02
## numberOfReviews                             -6.571702e-02
## reviewScoresAccuracy                         2.966535e-01
## reviewScoresCheckin                         -9.949108e-01
## reviewScoresCleanliness                      4.511431e+00
## reviewScoresCommunication                   -1.543589e+00
## reviewScoresLocation                         .
## reviewScoresRating                           1.834673e-01
## reviewScoresValue                           -5.742365e+00
## reviewsperMonth                             -1.592578e+00
## checkIn24Hours1                             -9.924278e-01
## airConditioning1                            -5.623663e-01
## buzzerOrWirelessIntercom1                   -1.324278e+00
## cableTV1                                     1.099844e+01
## carbonMonoxideDetector1                      1.258457e+00
```

```
## doorman1                                     8.245811e+00
## dryer1                                       .
## elevator1                                    8.622926e+00
## essentials1                                 -1.009085e+01
## familyAndKidFriendly1                        2.634717e+00
## fireExtinguisher1                            3.025391e+00
## firstAidKit1                                 3.074645e+00
## freeParkingOnPremises1                      -1.303091e+00
## gym1                                         1.944936e+01
## hairdryer1                                  -2.164501e+00
## hangers1                                    -5.425258e+00
## heating1                                    -3.509780e+00
## hottub1                                     -4.433442e+00
## internet1                                   -1.507958e+00
## iron1                                       -3.135484e+00
## kitchen1                                    -6.963046e+00
## laptopFriendlyWorkspace1                    -8.247768e-01
## lockOnBedroomDoor1                          -2.295287e+00
## petsAllowed1                                -6.507069e+00
## petsLiveOnThisProperty1                     -2.251488e+00
## pool1                                        1.771080e+01
## selfCheckIn1                                 9.676471e+00
## shampoo1                                     4.016852e+00
## smokeDetector1                              -5.815790e+00
## smokingAllowed1                              .
## washer1                                      3.447717e+00
## wheelchairAccessible1                        1.258175e+01
## wirelessInternet1                           -1.468102e+00
## hostHasProfilePic1                          -1.295706e+01
## hostIdentityVerified1                       -9.797188e-01
## hostIsSuperhost1                             1.562294e+01
## instantBookable1                            -4.046998e+00
## isLocationExact1                             9.913903e-02
## requireGuestPhoneVerification1              -1.086545e+01
## requireGuestProfilePicture1                  4.347151e+00
## featureCount                                 4.672462e-01
## yearsAsHost                                  2.825844e-01
```

```
lassoTest <- model.matrix(price~.,airDataTest)[,-1]
lassoPred <- predict(lassoModel,lassoTest)
print(paste("RMSE:",RMSE(lassoPred, airDataTest$price)))
```

```
## [1] "RMSE: 78.2508024632149"
```

```
print(paste("R-squared:",R2(lassoPred, airDataTest$price)))
```

```
## [1] "R-squared: 0.563446444064143"
```

XGBoost is a library designed and optimized for boosting trees algorithms. Gradient boosting trees model is originally proposed by Friedman et al. The underlying algorithm of XGBoost is similar, specifically it is an extension of the classic gbm algorithm. By employing multi-threads and imposing regularization, XGBoost is able to utilize more computational power and get more accurate prediction. XGBoost requires training and test data to be prepared similar to requirements of Ridge/Lasso regression models.

```
print("Using XGBoost")

## [1] "Using XGBoost"

print("Creating matrices for training and test data for XGBoost")

## [1] "Creating matrices for training and test data for XGBoost"

xgbTestData <- model.matrix(price~.,airDataTest)[,-1]
xgbPredData <- model.matrix(price~.,airDataPredict)[,-1]
xgbTestLabel <- airDataTest$price
```

XGBoost can be used with caret to set up training controls for the cross validation approach. The xgbGrid definees the values or list/range of values for each parameter that needs to be estimated. An average CV run for XGBoost (2 parameters with 3 values each) runs for ~45-50 minutes. The values below were estimated during earlier test runs and have been fixed to provide optimum results and reduce runtime for the markdown.

```
print("Setting up XGBoost Training Controls")

## [1] "Setting up XGBoost Training Controls"

xgb_trcontrol = trainControl(method="cv", number=5, allowParallel=TRUE,
verboseIter=TRUE, returnData=FALSE)
print("Setting up XGBoost Grid")

## [1] "Setting up XGBoost Grid"

xgbGrid <- expand.grid(nrounds=100, max_depth=10, colsample_bytree=0.5,
eta=0.1, gamma=0, min_child_weight=50, subsample=0.9)
xgb1 <- train(featureSet, responseSet, trControl=xgb_trcontrol,
tuneGrid=xgbGrid, method="xgbTree")

## + Fold1: nrounds=100, max_depth=10, colsample_bytree=0.5, eta=0.1,
gamma=0, min_child_weight=50, subsample=0.9
## - Fold1: nrounds=100, max_depth=10, colsample_bytree=0.5, eta=0.1,
gamma=0, min_child_weight=50, subsample=0.9
## + Fold2: nrounds=100, max_depth=10, colsample_bytree=0.5, eta=0.1,
gamma=0, min_child_weight=50, subsample=0.9
## - Fold2: nrounds=100, max_depth=10, colsample_bytree=0.5, eta=0.1,
gamma=0, min_child_weight=50, subsample=0.9
## + Fold3: nrounds=100, max_depth=10, colsample_bytree=0.5, eta=0.1,
gamma=0, min_child_weight=50, subsample=0.9
## - Fold3: nrounds=100, max_depth=10, colsample_bytree=0.5, eta=0.1,
gamma=0, min_child_weight=50, subsample=0.9
## + Fold4: nrounds=100, max_depth=10, colsample_bytree=0.5, eta=0.1,
gamma=0, min_child_weight=50, subsample=0.9
## - Fold4: nrounds=100, max_depth=10, colsample_bytree=0.5, eta=0.1,
```

```
gamma=0, min_child_weight=50, subsample=0.9
## + Fold5: nrounds=100, max_depth=10, colsample_bytree=0.5, eta=0.1,
gamma=0, min_child_weight=50, subsample=0.9
## - Fold5: nrounds=100, max_depth=10, colsample_bytree=0.5, eta=0.1,
gamma=0, min_child_weight=50, subsample=0.9
## Aggregating results
## Fitting final model on full training set
```

Best tuning parameters.

```
print("Best Tuning Parameters")

## [1] "Best Tuning Parameters"

print(xgb1$bestTune)

##   nrounds max_depth eta gamma colsample_bytree min_child_weight subsample
## 1     100        10 0.1     0              0.5               50       0.9
```

Summary of XGBoost model.

```
print("Summary of XGBoost model")

## [1] "Summary of XGBoost model"

print(xgb1)

## eXtreme Gradient Boosting
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 14547, 14549, 14547, 14547, 14546
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   70.25147  0.6679247  43.54998
##
## Tuning parameter 'nrounds' was held constant at a value of 100
##  0.5
## Tuning parameter 'min_child_weight' was held constant at a value
##  of 50
## Tuning parameter 'subsample' was held constant at a value of 0.9
```

Using XGBoost model to predict on test data.

```
print("Using XGBoost model to predict on Test Data")

## [1] "Using XGBoost model to predict on Test Data"

xgbPred <- predict(xgb1,xgbTestData)
print(paste("RMSE:",RMSE(xgbPred, airDataTest$price)))

## [1] "RMSE: 72.2505609015715"
```

```
print(paste("R-squared:",R2(xgbPred, airDataTest$price)))

## [1] "R-squared: 0.628885183117991"
```

Using XGBoost model for pure prediction.

```
print("Using XGBoost model to predict on Unknown Data")

## [1] "Using XGBoost model to predict on Unknown Data"

xgbPred <- predict(xgb1,xgbPredData)
airDataPredict$price <- xgbPred
print("Printing head and tail for predicted data")

## [1] "Printing head and tail for predicted data"

print(head(airDataPredict))

##      hostResponseHours neighbourhoodCleansed neighbourhoodGroupCleansed
## 140                 24                  SoHo                  Manhattan
## 424                  1              Gramercy                  Manhattan
## 462                  1       Upper West Side                  Manhattan
## 583                 72    Washington Heights                  Manhattan
## 692                 12               Chelsea                  Manhattan
## 1171                24               Chelsea                  Manhattan
##      accommodates bathrooms  bedType bedrooms beds TV cancellationPolicy
## 140             8         3 Real Bed        4    6  1           flexible
## 424            10         3 Real Bed        4    4  0             strict
## 462            10         6 Real Bed        6    6  1           flexible
## 583             2         1 Real Bed        1    1  1           flexible
## 692             6         2 Real Bed        2    2  1             strict
## 1171            2         1 Real Bed        1    2  1             strict
##      cleaningFee extraPeople guestsIncluded maximumNights minimumNights
## 140          250         100              6            12             2
## 424          450          12              4          1125             1
## 462          250           0              1          1125             4
## 583            0           0              1          1125             1
## 692          299           0              1          1125            10
## 1171          80           0              1          1125             3
##         price propertyType        roomType securityDeposit numberOfReviews
## 140  758.2528    Apartment Entire home/apt               0               9
## 424  642.8275    Apartment Entire home/apt             800              23
## 462  813.4902        House Entire home/apt               0               2
## 583  219.2037    Apartment Entire home/apt               0               0
## 692  527.8794    Apartment Entire home/apt               0               0
## 1171 205.9745    Apartment Entire home/apt               0              15
##      reviewScoresAccuracy reviewScoresCheckin reviewScoresCleanliness
## 140                    10                  10                      10
## 424                     9                  10                       9
## 462                    10                  10                      10
## 583                     0                   0                       0
## 692                     0                   0                       0
```

```
## 1171                   10                   10                        10
##      reviewScoresCommunication reviewScoresLocation reviewScoresRating
## 140                         10                   10                 98
## 424                          9                    9                 88
## 462                         10                   10                100
## 583                          0                    0                  0
## 692                          0                    0                  0
## 1171                        10                   10                 93
##      reviewScoresValue reviewsperMonth checkIn24Hours airConditioning
## 140                 10            0.82              0               1
## 424                  9            2.41              1               1
## 462                 10            0.21              1               1
## 583                  0            0.00              0               1
## 692                  0            0.00              0               1
## 1171                 9            1.88              0               1
##      buzzerOrWirelessIntercom cableTV carbonMonoxideDetector doorman dryer
## 140                         1       1                      1       0     1
## 424                         0       0                      1       0     1
## 462                         1       1                      1       0     1
## 583                         1       1                      0       0     0
## 692                         0       0                      1       0     1
## 1171                        0       0                      0       0     1
##      elevator essentials familyAndKidFriendly fireExtinguisher firstAidKit
## 140         0          1                    1                0           1
## 424         0          1                    1                0           1
## 462         0          1                    0                0           0
## 583         1          0                    0                0           0
## 692         1          1                    1                0           0
## 1171        0          0                    1                0           0
##      freeParkingOnPremises gym hairdryer hangers heating hottub internet
## 140                      0   0         0       1       1      0        1
## 424                      0   0         1       0       1      0        1
## 462                      0   0         1       1       1      0        1
## 583                      0   0         0       0       1      0        0
## 692                      0   0         1       1       1      0        0
## 1171                     0   0         0       0       1      0        1
##      iron kitchen laptopFriendlyWorkspace lockOnBedroomDoor petsAllowed
## 140     0       1                       1                 0           0
## 424     0       1                       0                 0           1
## 462     1       1                       1                 0           0
## 583     0       1                       0                 0           0
## 692     1       1                       1                 1           1
## 1171    0       1                       0                 0           0
##      petsLiveOnThisProperty pool selfCheckIn shampoo smokeDetector
## 140                       1    0           0       0             1
## 424                       0    0           0       0             1
## 462                       0    0           0       1             1
## 583                       0    0           0       0             0
## 692                       0    0           0       1             1
## 1171                      0    0           0       0             0
```

```
##      smokingAllowed washer wheelchairAccessible wirelessInternet
## 140               1      1                    0                1
## 424               0      1                    0                1
## 462               0      1                    0                1
## 583               0      0                    0                1
## 692               0      1                    0                1
## 1171              0      1                    0                1
##      hostHasProfilePic hostIdentityVerified hostIsSuperhost
## 140                  1                    1               0
## 424                  1                    1               0
## 462                  1                    0               0
## 583                  1                    0               0
## 692                  1                    0               0
## 1171                 1                    1               0
##      instantBookable isLocationExact requireGuestPhoneVerification
## 140                0               1                             0
## 424                1               0                             0
## 462                0               1                             0
## 583                0               1                             0
## 692                1               1                             0
## 1171               0               0                             0
##      requireGuestProfilePicture featureCount yearsAsHost
## 140                           0           23           4
## 424                           0           18           4
## 462                           0           24           4
## 583                           0            9           4
## 692                           0           22           3
## 1171                          0           10           8
```

```
print(tail(airDataPredict))
```

```
##       hostResponseHours neighbourhoodCleansed neighbourhoodGroupCleansed
## 17992                 1        Hell's Kitchen                  Manhattan
## 18377                 1        Hell's Kitchen                  Manhattan
## 18811                 1                 Other                  Manhattan
## 19050                 1       Upper East Side                  Manhattan
## 19055                 1       Upper West Side                  Manhattan
## 19154                 1               Chelsea                  Manhattan
##       accommodates bathrooms  bedType bedrooms beds TV cancellationPolicy
## 17992            4         1 Real Bed        2    2  1             strict
## 18377            1         1 Real Bed        1    1  1             strict
## 18811            3         1 Real Bed        1    2  0           flexible
## 19050            3         2 Real Bed        3    3  1           flexible
## 19055            2         1 Real Bed        1    1  1           flexible
## 19154           16         1 Real Bed        0    1  0             strict
##       cleaningFee extraPeople guestsIncluded maximumNights minimumNights
## 17992          70           0              1           365            30
## 18377           0           0              1          1125             1
## 18811           0          30              2          1125             1
## 19050           0           0              1          1125             1
```

```
## 19055        100         0             1         1125             1
## 19154          0         0             1         1125             1
##          price propertyType        roomType securityDeposit
## 17992 238.2175    Apartment Entire home/apt               0
## 18377 164.2854    Apartment    Private room               0
## 18811 144.5242    Apartment Entire home/apt               0
## 19050 318.6658    Apartment Entire home/apt               0
## 19055 282.7828    Apartment Entire home/apt               0
## 19154 317.8828        Other    Private room               0
##       numberOfReviews reviewScoresAccuracy reviewScoresCheckin
## 17992               6                    9                  10
## 18377               0                    0                   0
## 18811               2                   10                   7
## 19050               0                    0                   0
## 19055               0                    0                   0
## 19154               0                    0                   0
##       reviewScoresCleanliness reviewScoresCommunication
## 17992                      10                        10
## 18377                       0                         0
## 18811                       5                         8
## 19050                       0                         0
## 19055                       0                         0
## 19154                       0                         0
##       reviewScoresLocation reviewScoresRating reviewScoresValue
## 17992                   10                 97                10
## 18377                    0                  0                 0
## 18811                    9                 76                 7
## 19050                    0                  0                 0
## 19055                    0                  0                 0
## 19154                    0                  0                 0
##       reviewsperMonth checkIn24Hours airConditioning
## 17992            0.57              1               1
## 18377            0.00              0               0
## 18811            0.04              0               1
## 19050            0.00              1               1
## 19055            0.00              0               1
## 19154            0.00              0               1
##       buzzerOrWirelessIntercom cableTV carbonMonoxideDetector doorman
## 17992                        1       1                      1       0
## 18377                        1       1                      0       0
## 18811                        0       0                      0       0
## 19050                        1       1                      0       1
## 19055                        1       1                      0       0
## 19154                        0       0                      0       0
##       dryer elevator essentials familyAndKidFriendly fireExtinguisher
## 17992     1        0          1                    0                0
## 18377     0        0          0                    0                0
## 18811     0        0          0                    0                0
## 19050     1        1          1                    1                0
## 19055     1        1          0                    1                0
```

```
## 19154     0        0          0                    1               0
##       firstAidKit freeParkingOnPremises gym hairdryer hangers heating
## 17992           0                     0   0         1       1       1
## 18377           0                     0   0         0       0       1
## 18811           0                     0   0         0       0       1
## 19050           0                     0   0         1       0       1
## 19055           0                     0   0         0       0       1
## 19154           0                     0   0         0       0       0
##       hottub internet iron kitchen laptopFriendlyWorkspace
## 17992      0        1    1       1                       1
## 18377      0        1    0       1                       0
## 18811      1        0    0       1                       0
## 19050      0        1    1       1                       1
## 19055      0        1    0       1                       0
## 19154      0        0    0       1                       0
##       lockOnBedroomDoor petsAllowed petsLiveOnThisProperty pool
## 17992                 0           0                      0    0
## 18377                 0           0                      0    0
## 18811                 0           0                      0    0
## 19050                 0           0                      0    0
## 19055                 0           0                      1    0
## 19154                 0           1                      0    0
##       selfCheckIn shampoo smokeDetector smokingAllowed washer
## 17992           0       1             1              0      1
## 18377           0       0             0              0      0
## 18811           0       0             0              0      0
## 19050           0       0             0              0      1
## 19055           0       0             0              0      1
## 19154           0       0             0              0      0
##       wheelchairAccessible wirelessInternet hostHasProfilePic
## 17992                    0                1                 1
## 18377                    0                1                 1
## 18811                    0                1                 1
## 19050                    0                1                 1
## 19055                    0                1                 1
## 19154                    1                0                 1
##       hostIdentityVerified hostIsSuperhost instantBookable isLocationExact
## 17992                    1               0               0               0
## 18377                    0               0               0               0
## 18811                    0               0               0               1
## 19050                    0               0               0               1
## 19055                    0               0               0               1
## 19154                    0               0               0               0
##       requireGuestPhoneVerification requireGuestProfilePicture
## 17992                             0                          0
## 18377                             0                          0
## 18811                             1                          1
## 19050                             0                          0
## 19055                             0                          0
## 19154                             0                          0
```

```
##       featureCount yearsAsHost
## 17992           22           5
## 18377            7           4
## 18811            9           9
## 19050           19           4
## 19055           15           5
## 19154            8           2
```

```r
print("Writing predicted data to file predictedPrices.csv")
```

```
## [1] "Writing predicted data to file predictedPrices.csv"
```

```r
if(log_regression){
  airDataPredict$price <- exp(airDataPredict$price)
}

write.csv(airDataPredict,"predictedPrices.csv")

print("---------- End of Algorithmic Machine Learning Project ----------")
```

```
## [1] "---------- End of Algorithmic Machine Learning Project ----------"
```

End of Markdown.