Final Project Draft

Peter Antonaros

Packages & Setup

```
#Memory allocation for Java ~10gb and Garbage Collection
options(java.parameters = c("-XX:+UseConcMarkSweepGC", "-Xmx10000m"))

#Packages to load
pacman::p_load(
  ggplot2,
  tidyverse,
  data.table,
  R.utils,
  magrittr,
  dplyr,
  testthat,
  YARF,
  lubridate,
  missForest
)

set_YARF_num_cores(8)
```

```
## YARF can now make use of 8 cores.
```

```
library(rJava)
gc()
```

```
##           used (Mb) gc trigger (Mb) max used  (Mb)
## Ncells 1261169 67.4    2190651  117  2190651 117.0
## Vcells 2131581 16.3    8388608   64  3460265  26.4
```

```
.jinit()
```

```
## [1] 0
```

The Data

```
housingDataFilePath = "/home/peterjr/RepoCollections/MATH_342W_FinalProject/Datasets/housing_data_2016_
housingData = data.table(read.csv(housingDataFilePath))

housingData
```

```
##                               HITId                      HITTypeId
##    1: 3OID399FXG7F26JWONXF0Y86J90FD4 36BILMLQB75QQNBTYKGYCZWDN8TVAU
##    2: 3MQY1YVHS3K2MF90MWR2LPQH7KJ2B0 36BILMLQB75QQNBTYKGYCZWDN8TVAU
##    3: 3DGDV62G7O94Q9AA5193G9V6OOY2PL 36BILMLQB75QQNBTYKGYCZWDN8TVAU
##    4: 3O87LXLJ6MGL3MI2CB9KLRONPKRF0B 36BILMLQB75QQNBTYKGYCZWDN8TVAU
##    5: 3FULMHZ7OUX88KSKHZ0ZSKY93XJ4MN 36BILMLQB75QQNBTYKGYCZWDN8TVAU
##   ---
## 2226:                           <NA>                          <NA>
## 2227:                           <NA>                          <NA>
## 2228:                           <NA>                          <NA>
## 2229:                           <NA>                          <NA>
## 2230:                           <NA>                          <NA>
##                                                                  Title
##    1: Find Information about Housing To Help a Student Project -- Very easy
##    2: Find Information about Housing To Help a Student Project -- Very easy
##    3: Find Information about Housing To Help a Student Project -- Very easy
##    4: Find Information about Housing To Help a Student Project -- Very easy
##    5: Find Information about Housing To Help a Student Project -- Very easy
##   ---
## 2226:                                                              <NA>
## 2227:                                                              <NA>
## 2228:                                                              <NA>
## 2229:                                                              <NA>
## 2230:                                                              <NA>
##                                        Description Keywords Reward
##    1: Go to a link and copy information into the HIT       NA $0.05
##    2: Go to a link and copy information into the HIT       NA $0.05
##    3: Go to a link and copy information into the HIT       NA $0.05
##    4: Go to a link and copy information into the HIT       NA $0.05
##    5: Go to a link and copy information into the HIT       NA $0.05
##   ---
## 2226:                                     <NA>       NA   <NA>
## 2227:                                     <NA>       NA   <NA>
## 2228:                                     <NA>       NA   <NA>
## 2229:                                     <NA>       NA   <NA>
## 2230:                                     <NA>       NA   <NA>
##                      CreationTime MaxAssignments
##    1: Wed Feb 15 22:13:37 PST 2017              1
##    2: Wed Feb 15 22:13:37 PST 2017              1
##    3: Wed Feb 15 22:13:41 PST 2017              1
##    4: Wed Feb 15 22:13:33 PST 2017              1
##    5: Wed Feb 15 22:13:38 PST 2017              1
##   ---
## 2226:                          <NA>            NA
## 2227:                          <NA>            NA
## 2228:                          <NA>            NA
## 2229:                          <NA>            NA
## 2230:                          <NA>            NA
##                                  RequesterAnnotation
##    1: BatchId:2689947;OriginalHitTemplateId:920937336;
##    2: BatchId:2689947;OriginalHitTemplateId:920937336;
##    3: BatchId:2689947;OriginalHitTemplateId:920937336;
##    4: BatchId:2689947;OriginalHitTemplateId:920937336;
##    5: BatchId:2689947;OriginalHitTemplateId:920937336;
```

```
##   ---
## 2226:                                          <NA>
## 2227:                                          <NA>
## 2228:                                          <NA>
## 2229:                                          <NA>
## 2230:                                          <NA>
##       AssignmentDurationInSeconds AutoApprovalDelayInSeconds
##    1:                         900                        60
##    2:                         900                        60
##    3:                         900                        60
##    4:                         900                        60
##    5:                         900                        60
##   ---
## 2226:                          NA                        NA
## 2227:                          NA                        NA
## 2228:                          NA                        NA
## 2229:                          NA                        NA
## 2230:                          NA                        NA
##                        Expiration NumberOfSimilarHITs LifetimeInSeconds
##    1: Wed Feb 22 22:13:37 PST 2017                 NA               NA
##    2: Wed Feb 22 22:13:37 PST 2017                 NA               NA
##    3: Wed Feb 22 22:13:41 PST 2017                 NA               NA
##    4: Wed Feb 22 22:13:33 PST 2017                 NA               NA
##    5: Wed Feb 22 22:13:38 PST 2017                 NA               NA
##   ---
## 2226:                         <NA>                 NA               NA
## 2227:                         <NA>                 NA               NA
## 2228:                         <NA>                 NA               NA
## 2229:                         <NA>                 NA               NA
## 2230:                         <NA>                 NA               NA
##                        AssignmentId      WorkerId AssignmentStatus
##    1: 32KTQ2V7RDFCSAWQOW1SXC5AZIC9MB A231MNJJDDF3LS         Approved
##    2: 35LDD5557A4W96FHSTSHNLJQAB7MKZ A394B5QVCVKU7A         Approved
##    3: 3FFJ6VRIL1O8OXIM3LK7C8XOF5UOI6 A231MNJJDDF3LS         Approved
##    4: 3S4AW7T8OBIRPM8T7P4MGRF5DL74L7  AHXBZXWIZJSVG         Approved
##    5: 3JMSRU9HQIUCDTHGAZI5CMPYH7REVS A231MNJJDDF3LS         Approved
##   ---
## 2226:                           <NA>          <NA>             <NA>
## 2227:                           <NA>          <NA>             <NA>
## 2228:                           <NA>          <NA>             <NA>
## 2229:                           <NA>          <NA>             <NA>
## 2230:                           <NA>          <NA>             <NA>
##                        AcceptTime                    SubmitTime
##    1: Thu Feb 16 05:32:36 PST 2017 Thu Feb 16 05:35:37 PST 2017
##    2: Wed Feb 15 22:19:51 PST 2017 Wed Feb 15 22:21:52 PST 2017
##    3: Thu Feb 16 03:17:01 PST 2017 Thu Feb 16 03:19:01 PST 2017
##    4: Thu Feb 16 04:54:24 PST 2017 Thu Feb 16 04:57:04 PST 2017
##    5: Wed Feb 15 23:54:29 PST 2017 Wed Feb 15 23:56:45 PST 2017
##   ---
## 2226:                         <NA>                         <NA>
## 2227:                         <NA>                         <NA>
## 2228:                         <NA>                         <NA>
## 2229:                         <NA>                         <NA>
## 2230:                         <NA>                         <NA>
```

```
##                   AutoApprovalTime            ApprovalTime RejectionTime
##    1: Thu Feb 16 05:36:37 PST 2017 2017-02-16 13:37:11 UTC            NA
##    2: Wed Feb 15 22:22:52 PST 2017 2017-02-16 06:23:11 UTC            NA
##    3: Thu Feb 16 03:20:01 PST 2017 2017-02-16 11:20:11 UTC            NA
##    4: Thu Feb 16 04:58:04 PST 2017 2017-02-16 12:58:11 UTC            NA
##    5: Wed Feb 15 23:57:45 PST 2017 2017-02-16 07:58:11 UTC            NA
##   ---
## 2226:                         <NA>                    <NA>            NA
## 2227:                         <NA>                    <NA>            NA
## 2228:                         <NA>                    <NA>            NA
## 2229:                         <NA>                    <NA>            NA
## 2230:                         <NA>                    <NA>            NA
##       RequesterFeedback WorkTimeInSeconds LifetimeApprovalRate
##    1:                NA               181       100% (187/187)
##    2:                NA               121           100% (8/8)
##    3:                NA               120       100% (187/187)
##    4:                NA               160       100% (115/115)
##    5:                NA               136       100% (187/187)
##   ---
## 2226:                NA                NA                 <NA>
## 2227:                NA                NA                 <NA>
## 2228:                NA                NA                 <NA>
## 2229:                NA                NA                 <NA>
## 2230:                NA                NA                 <NA>
##       Last30DaysApprovalRate Last7DaysApprovalRate
##    1:         100% (187/187)        100% (187/187)
##    2:             100% (8/8)            100% (8/8)
##    3:         100% (187/187)        100% (187/187)
##    4:         100% (115/115)        100% (103/103)
##    5:         100% (187/187)        100% (187/187)
##   ---
## 2226:                   <NA>                  <NA>
## 2227:                   <NA>                  <NA>
## 2228:                   <NA>                  <NA>
## 2229:                   <NA>                  <NA>
## 2230:                   <NA>                  <NA>
##                                                                                                            U
##    1: http://www.mlsli.com/homes-for-sale/address-not-available-from-broker-Flushing-NY-11355-1492383
##    2:                http://www.mlsli.com/homes-for-sale/30-11-Parsons-Blvd-Flushing-NY-11354-1552428
##    3:                    http://www.mlsli.com/homes-for-sale/102-14-Lewis-Ave-Corona-NY-11368-1570845
##    4:              http://www.mlsli.com/homes-for-sale/144-48-Roosevelt-Ave-Flushing-NY-11354-1553227
##    5:                  http://www.mlsli.com/homes-for-sale/245-27-76th-Ave-Bellerose-NY-11426-1612800
##   ---
## 2226:                                                                                                       <
## 2227:                                                                                                       <
## 2228:                                                                                                       <
## 2229:                                                                                                       <
## 2230:                                                                                                       <
##       approx_year_built cats_allowed common_charges community_district_num
##    1:              1955           no           $767                     25
##    2:              1955           no           <NA>                     25
##    3:              2004           no           $167                     24
##    4:              2002           no           $275                     25
##    5:              1949          yes           <NA>                     26
```

```
##   ---
## 2226:              1987           no          $480                      25
## 2227:              1983          yes          $956                      25
## 2228:              2010           no          $250                      24
## 2229:              2010           no          $250                      24
## 2230:              1982           no          $792                      25
##       coop_condo date_of_sale dining_room_type dogs_allowed fuel_type
##    1:      co-op    2/16/2016            combo           no       gas
##    2:      co-op    2/16/2016           formal           no       oil
##    3:      condo    2/17/2016            combo           no      <NA>
##    4:      condo    2/17/2016            combo           no       gas
##    5:      co-op    2/18/2016            combo          yes       gas
##   ---
## 2226:      condo         <NA>            combo           no       gas
## 2227:      condo         <NA>           formal           no       gas
## 2228:      condo         <NA>           formal           no       gas
## 2229:      condo         <NA>           formal           no       gas
## 2230:      condo         <NA>           formal           no       gas
##                                      full_address_or_zip_code garage_exists
##    1:                                        Flushing NY, 11355          <NA>
##    2: 30-11 Parsons Blvd,  Flushing NY, 11354 ( Sold )      Share          <NA>
##    3:                          102-14 Lewis Ave,  Corona NY, 11368          <NA>
##    4:                     144-48 Roosevelt Ave,  Flushing NY, 11354          <NA>
##    5:                         245-27 76th Ave,  Bellerose NY, 11426          <NA>
##   ---
## 2226:                             Not AvailableFlushing NY, 11355          <NA>
## 2227:                           One Bay Club Dr,  Bayside NY, 11360           yes
## 2228:                                          Ridgewood NY, 11385          <NA>
## 2229:                                          Ridgewood NY, 11385          <NA>
## 2230:                       Two Bay Club Drive,  Bayside NY, 11360           yes
##       kitchen_type maintenance_cost          model_type num_bedrooms
##    1:       eat in             <NA>   Mitchell Garden 3            2
##    2:       eat in             $604           Jr-4 Model            1
##    3:   efficiency             <NA>         Apt In Bldg             1
##    4:       eat in             <NA>    144-48 Roosevelt             3
##    5:       eat in             $660                 C-1             2
##   ---
## 2226:        combo             <NA> Colden Luxury Condo             2
## 2227:        eatin             <NA>         2 Br Deluxe             2
## 2228:        combo             <NA>              Modern             3
## 2229:        combo             <NA>               Condo             3
## 2230:        combo             <NA>           2 Bedroom             2
##       num_floors_in_building num_full_bathrooms num_half_bathrooms
##    1:                      6                  1                 NA
##    2:                      7                  1                 NA
##    3:                      1                  1                 NA
##    4:                     NA                  2                 NA
##    5:                      2                  1                 NA
##   ---
## 2226:                      7                  1                 NA
## 2227:                     NA                  2                 NA
## 2228:                     NA                  2                 NA
## 2229:                      4                  2                 NA
## 2230:                     NA                  2                 NA
```

```
##       num_total_rooms parking_charges pct_tax_deductibl sale_price sq_footage
##    1:               5            <NA>                NA  $228,000         NA
##    2:               4            <NA>                NA  $235,500        890
##    3:               3            <NA>                NA  $137,550        550
##    4:               5            <NA>                NA  $545,000         NA
##    5:               4            <NA>                39  $241,700        675
##   ---
## 2226:               4            <NA>                NA      <NA>         NA
## 2227:               5            $99                 NA      <NA>         NA
## 2228:               6            <NA>                NA      <NA>       1500
## 2229:               6            <NA>                NA      <NA>       1600
## 2230:               5            <NA>                NA      <NA>       1134
##       total_taxes walk_score listing_price_to_nearest_1000
##    1:        <NA>         82                          <NA>
##    2:        <NA>         89                          <NA>
##    3:      $5,500         90                          <NA>
##    4:      $2,260         94                          <NA>
##    5:        <NA>         71                          <NA>
##   ---
## 2226:      $3,588         97                          $628
## 2227:      $5,100         82                          $988
## 2228:        $250         96                          $850
## 2229:        $250         96                          $850
## 2230:      $3,785         82                          $899
##
##    1:
##    2:
##    3:
##    4:
##    5:
##   ---
## 2226: http://www.mlsli.com/homes-for-sale/address-not-available-from-broker-Flushing-NY-11355-169427
## 2227:                   http://www.mlsli.com/homes-for-sale/One-Bay-Club-Dr-Bayside-NY-11360-196274
## 2228: http://www.mlsli.com/homes-for-sale/address-not-available-from-broker-Ridgewood-NY-11385-92169
## 2229: http://www.mlsli.com/homes-for-sale/address-not-available-from-broker-Ridgewood-NY-11385-92101
## 2230:                http://www.mlsli.com/homes-for-sale/Two-Bay-Club-Drive-Bayside-NY-11360-140297
```

```
#Relevant columns begin at the column labeled (URL)
```

Initial Data Preparation I (Dropping Irrelevant Columns & Storing Possible Ones for Later Use)

```
#Dropping Mturk columns that are not relevant to our housing model
housingData[,c(1:27):=NULL]

#Save the urls for later and remove from data frame (might be useful but not immediately)
housingURLS = housingData[,.(URL)]

#Dropping URL from the data table
housingData[,URL:=NULL]
#Dropping other useless url column from data table (ALL NA's)
housingData[,url:=NULL]
#Dropping model_type because similar information is contained in other columns
housingData[,model_type:=NULL]
```

```
housingData
```

```
##       approx_year_built cats_allowed common_charges community_district_num
##    1:              1955           no          $767                      25
##    2:              1955           no           <NA>                     25
##    3:              2004           no          $167                      24
##    4:              2002           no          $275                      25
##    5:              1949          yes           <NA>                     26
##   ---
## 2226:              1987           no          $480                      25
## 2227:              1983          yes          $956                      25
## 2228:              2010           no          $250                      24
## 2229:              2010           no          $250                      24
## 2230:              1982           no          $792                      25
##       coop_condo date_of_sale dining_room_type dogs_allowed fuel_type
##    1:      co-op    2/16/2016            combo           no       gas
##    2:      co-op    2/16/2016           formal           no       oil
##    3:      condo    2/17/2016            combo           no      <NA>
##    4:      condo    2/17/2016            combo           no       gas
##    5:      co-op    2/18/2016            combo          yes       gas
##   ---
## 2226:      condo         <NA>            combo           no       gas
## 2227:      condo         <NA>           formal           no       gas
## 2228:      condo         <NA>           formal           no       gas
## 2229:      condo         <NA>           formal           no       gas
## 2230:      condo         <NA>           formal           no       gas
##                                       full_address_or_zip_code garage_exists
##    1:                                        Flushing NY, 11355          <NA>
##    2: 30-11 Parsons Blvd,  Flushing NY, 11354 ( Sold )      Share          <NA>
##    3:                          102-14 Lewis Ave,  Corona NY, 11368          <NA>
##    4:                    144-48 Roosevelt Ave,  Flushing NY, 11354          <NA>
##    5:                        245-27 76th Ave,  Bellerose NY, 11426          <NA>
##   ---
## 2226:                             Not AvailableFlushing NY, 11355          <NA>
## 2227:                          One Bay Club Dr,  Bayside NY, 11360           yes
## 2228:                                        Ridgewood NY, 11385          <NA>
## 2229:                                        Ridgewood NY, 11385          <NA>
## 2230:                       Two Bay Club Drive,  Bayside NY, 11360           yes
##       kitchen_type maintenance_cost num_bedrooms num_floors_in_building
##    1:       eat in             <NA>            2                      6
##    2:       eat in            $604             1                      7
##    3:   efficiency             <NA>            1                      1
##    4:       eat in             <NA>            3                     NA
##    5:       eat in            $660             2                      2
##   ---
## 2226:        combo             <NA>            2                      7
## 2227:        eatin             <NA>            2                     NA
## 2228:        combo             <NA>            3                     NA
## 2229:        combo             <NA>            3                      4
## 2230:        combo             <NA>            2                     NA
##       num_full_bathrooms num_half_bathrooms num_total_rooms parking_charges
##    1:                  1                NA               5            <NA>
##    2:                  1                NA               4            <NA>
```

```
##    3:                  1                NA              3            <NA>
##    4:                  2                NA              5            <NA>
##    5:                  1                NA              4            <NA>
##   ---
## 2226:                  1                NA              4            <NA>
## 2227:                  2                NA              5            $99
## 2228:                  2                NA              6            <NA>
## 2229:                  2                NA              6            <NA>
## 2230:                  2                NA              5            <NA>
##       pct_tax_deductibl sale_price sq_footage total_taxes walk_score
##    1:                NA  $228,000          NA        <NA>         82
##    2:                NA  $235,500         890        <NA>         89
##    3:                NA  $137,550         550     $5,500          90
##    4:                NA  $545,000          NA     $2,260          94
##    5:                39  $241,700         675        <NA>         71
##   ---
## 2226:                NA       <NA>         NA     $3,588          97
## 2227:                NA       <NA>         NA     $5,100          82
## 2228:                NA       <NA>       1500       $250          96
## 2229:                NA       <NA>       1600       $250          96
## 2230:                NA       <NA>       1134     $3,785          82
##       listing_price_to_nearest_1000
##    1:                          <NA>
##    2:                          <NA>
##    3:                          <NA>
##    4:                          <NA>
##    5:                          <NA>
##   ---
## 2226:                         $628
## 2227:                         $988
## 2228:                         $850
## 2229:                         $850
## 2230:                         $899
```

Initial Data Preparation II (Writing some notes about Columns)

```
#Getting the column names to write some notes about each column
names(housingData)
```

```
##  [1] "approx_year_built"            "cats_allowed"
##  [3] "common_charges"               "community_district_num"
##  [5] "coop_condo"                   "date_of_sale"
##  [7] "dining_room_type"             "dogs_allowed"
##  [9] "fuel_type"                    "full_address_or_zip_code"
## [11] "garage_exists"                "kitchen_type"
## [13] "maintenance_cost"             "num_bedrooms"
## [15] "num_floors_in_building"       "num_full_bathrooms"
## [17] "num_half_bathrooms"           "num_total_rooms"
## [19] "parking_charges"              "pct_tax_deductibl"
## [21] "sale_price"                   "sq_footage"
## [23] "total_taxes"                  "walk_score"
## [25] "listing_price_to_nearest_1000"
```

```
#Getting some general information about the table
summary(housingData)
```

```
##  approx_year_built cats_allowed       common_charges     community_district_num
##  Min.   :1893      Length:2230        Length:2230        Min.   : 3.00
##  1st Qu.:1950      Class :character   Class :character   1st Qu.:25.00
##  Median :1958      Mode  :character   Mode  :character   Median :26.00
##  Mean   :1963                                            Mean   :26.33
##  3rd Qu.:1970                                            3rd Qu.:28.00
##  Max.   :2017                                            Max.   :32.00
##  NA's   :40                                              NA's   :19
##   coop_condo        date_of_sale       dining_room_type   dogs_allowed
##  Length:2230        Length:2230        Length:2230        Length:2230
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   fuel_type         full_address_or_zip_code garage_exists
##  Length:2230        Length:2230              Length:2230
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
##  kitchen_type       maintenance_cost    num_bedrooms   num_floors_in_building
##  Length:2230        Length:2230        Min.   :0.000   Min.   : 1.000
##  Class :character   Class :character   1st Qu.:1.000   1st Qu.: 3.000
##  Mode  :character   Mode  :character   Median :2.000   Median : 6.000
##                                        Mean   :1.653   Mean   : 7.785
##                                        3rd Qu.:2.000   3rd Qu.: 7.000
##                                        Max.   :6.000   Max.   :34.000
##                                        NA's   :115     NA's   :650
##  num_full_bathrooms num_half_bathrooms num_total_rooms  parking_charges
##  Min.   :1.000      Min.   :0.0000     Min.   : 0.000   Length:2230
##  1st Qu.:1.000      1st Qu.:1.0000     1st Qu.: 3.000   Class :character
##  Median :1.000      Median :1.0000     Median : 4.000   Mode  :character
##  Mean   :1.231      Mean   :0.9535     Mean   : 4.139
##  3rd Qu.:1.000      3rd Qu.:1.0000     3rd Qu.: 5.000
##  Max.   :3.000      Max.   :2.0000     Max.   :14.000
##                     NA's   :2058       NA's   :2
##  pct_tax_deductibl  sale_price          sq_footage     total_taxes
##  Min.   :20.0       Length:2230        Min.   : 100.0   Length:2230
##  1st Qu.:40.0       Class :character   1st Qu.: 743.0   Class :character
##  Median :50.0       Mode  :character   Median : 881.0   Mode  :character
##  Mean   :45.4                          Mean   : 955.4
##  3rd Qu.:50.0                          3rd Qu.:1100.0
##  Max.   :75.0                          Max.   :6215.0
##  NA's   :1754                          NA's   :1210
##    walk_score    listing_price_to_nearest_1000
##  Min.   : 7.00   Length:2230
```

```
##  1st Qu.:77.00   Class :character
##  Median :89.00   Mode  :character
##  Mean   :83.92
##  3rd Qu.:95.00
##  Max.   :99.00
##
```

Column Name | Information | Notes to Self about column

"approx_year_built" | Integer representing the year the house was built | 40 NA's

"cats_allowed" | Binary decision (0,1) are cats allowed in the home or not | Check for NA's & Factor

"common_charges" | Some sort of charges in dollars ($) | Remove the dollar symbol & Convert to integer & Check for NA's

"community_district_num" | Integer representing the district number of community home is a part of | 19 NA's

"coop_condo" | String representing "Co-op" or "Condo" | Lowercase everything | Check for levels & Factor

"date_of_sale" | String representing the date the home was sold |

"dining_room_type" | String representing "formal" or "combo" dining room type | Lowercase everything & Check for NA's & Factor

"dogs_allowed" | Binary decision (0,1) are dogs allowed in the home or not | Factor this & Check for NA's

"fuel_type" | String representing "gas", "oil", or "other" energy source for the home | Lowercase everything & Check for NA's & factor

"full_address_or_zip_code" | String representing the address of the home |

"garage_exists" | String representing "Yes" if the home has a garage | Check for NA's & Factor this & Missingness column

"kitchen_type" | String representing "Eat-In", "Efficiency", or "Combo" kitchen type | Lowercase everything & Factor this & Check for NA's

"maintenance_cost" | Cost of maintenece for the home in dollars ($) | Remove the dollar symbol & Convert to integer & Check for NA's

"num_bedrooms" | Integer representing number of bedrooms present in the home | 115 NA's

"num_floors_in_building" | Integer representing number of floors present in building containing home | 650 NA's

"num_full_bathrooms" | Integer representing the number of full bathrooms present in the home | No NA's

"num_half_bathrooms" | Integer representing the number of half bathrooms present in the home | 2058 NA's

"num_total_rooms" | Integer representing the number of total rooms present in the home | 2 NA's

"parking_charges" | Parking charges in dollars ($) | Remove the dollar symbol & Convert to integer & Check for NA's

"pct_tax_deductibl" | Integer representing percent of tax deduction | 1754 NA's

"sale_price" | Sale price of the home in dollars ($) | Remove the dollar symbol & Convert to integer & Check for NA's

"sq_footage" | Integer representing the total square footage of the home | 1210 NA's

"total_taxes" | Taxes on the home in dollars ($) | Remove the dollar symbol & Convert to integer & Check for NA's

"walk_score" | Integer representing a walking score for the home |

"listing_price_to_nearest_1000" | Listing price to the nearest 1000 for the home in dollars ($) | Remove the dollar symbol & Convert to integer & Check for NA's

Data Cleaning I (Fixing column types)

```
#First lets deal with the String columns that have $ symbols and convert to integer

#Extract dollar sign columns as subset to operate on
dollarSymbolSubset = housingData[,.(common_charges,maintenance_cost,parking_charges,sale_price,total_ta
#Remove dollar signs based on pattern matching
dollarSymbolSubset[] = lapply(dollarSymbolSubset,gsub,pattern="$",fixed=TRUE,replacement="")
#Also Remove any commas that may appear for large values
dollarSymbolSubset[] = lapply(dollarSymbolSubset,gsub,pattern=",",fixed=TRUE,replacement="")

#Replace the columns in housing Data with the new dollarSymbolSubset
housingData[,c("common_charges","maintenance_cost","parking_charges","sale_price","total_taxes","listing
            dollarSymbolSubset[,c("common_charges","maintenance_cost","parking_charges","sale_price","t

#Now we need to convert these columns in housing data to integer type
housingData[,c("common_charges","maintenance_cost","parking_charges","sale_price","total_taxes","listing


##############################################################################################
#Second lets deal with changing cats_allowed and dogs_allowed to factors

housingData[,sum(is.na(cats_allowed))] # No NA values for cats_allowed
```

```
## [1] 0
```

```
housingData[,sum(is.na(dogs_allowed))] # No NA values for dogs_allowed
```

```
## [1] 0
```

```
#Changing to factors for cats and dogs allowed
unique(housingData[,cats_allowed]) # 3 "unique" values
```

```
## [1] "no"  "yes" "y"
```

```
#Lets deal with the y instead of a yes
housingData$cats_allowed[grepl("y", housingData$cats_allowed)] = "yes"
length(unique(housingData[,cats_allowed])) # 2 unique values
```

```
## [1] 2
```

```r
#Lets do the same for dogs
unique(housingData[,dogs_allowed]) # 3 "unique" values"
```

```
## [1] "no"    "yes"   "yes89"
```

```r
housingData$dogs_allowed[grepl("yes89", housingData$dogs_allowed)] = "yes"
length(unique(housingData[,cats_allowed])) # 2 unique values
```

```
## [1] 2
```

```r
#Factor them
housingData[,c("cats_allowed","dogs_allowed")] = lapply(housingData[,c("cats_allowed","dogs_allowed")],

levels(housingData$cats_allowed) #Check levels
```

```
## [1] "no"  "yes"
```

```r
levels(housingData$dogs_allowed) #Check levels
```

```
## [1] "no"  "yes"
```

```r
##############################################################################################
#Third lets deal with the other String columns that need to be factored (track NA's for later)

housingData[,sum(is.na(coop_condo))] # No NA values for coop_condo
```

```
## [1] 0
```

```r
length(unique(housingData[,coop_condo])) # 2 unique values
```

```
## [1] 2
```

```r
#Factor it
housingData[,coop_condo := factor(coop_condo)]
levels(housingData$coop_condo)
```

```
## [1] "co-op" "condo"
```

```r
housingData[,sum(is.na(dining_room_type))] # 448 NA values for dining_room_type
```

```
## [1] 448
```

```r
length(unique(housingData[,dining_room_type])) # 6 unique values including NA
```

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

```
length(which(housingData$dining_room_type == "none")) #none occurs 2 times
```

```
## [1] 2
```

```
length(which(housingData$dining_room_type == "dining area")) #dining area occurs 2 times
```

```
## [1] 2
```

```
#Lets deal with the issue of "dining area" as the room type and consider it as type other
housingData$dining_room_type[grepl("dining area", housingData$dining_room_type)] = "other"
length(unique(housingData[,dining_room_type])) # 5 unique values including NA
```

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

```
housingData[,dining_room_type := factor(dining_room_type)]
levels(housingData$dining_room_type)
```

```
## [1] "combo"  "formal" "none"   "other"
```

```
housingData[,sum(is.na(fuel_type))] # 112 NA values for dining_room_type
```

```
## [1] 112
```

```
length(unique(housingData[,fuel_type])) # 7 "unique" values including NA
```

```
## [1] 7
```

```
#Lets deal with the capitalization issues for fuel_typenone
housingData[,fuel_type := tolower(fuel_type)]
length(unique(housingData[,fuel_type])) # 6 unique values including NA
```

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

```
housingData[,fuel_type := factor(fuel_type)]
levels(housingData$fuel_type)
```

```
## [1] "electric" "gas"      "none"     "oil"      "other"
```

```
housingData[,sum(is.na(kitchen_type))]# 16 NA values for dining_room_type
```

```
## [1] 16
```

```
length(unique(housingData[,kitchen_type])) # 14 "unique" values including NA
```

```
## [1] 14
```

```
#Lets deal with the upper case lower case kitchen type differences
housingData[,kitchen_type:=tolower(kitchen_type)] # Lowercase everything to pattern match
length(unique(housingData[,kitchen_type])) # 11 "unique" values including NA
```

```
## [1] 11
```

```
#Lets now deal with spaces creating more unique values
housingData[,kitchen_type := lapply(kitchen_type,gsub,pattern="eat in",fixed=TRUE,replacement="eatin")]
length(unique(housingData[,kitchen_type])) # 10 "unique" values including NA
```

```
## [1] 10
```

```
#Lets lets deal with the misspellings of efficiency kitchen
housingData$kitchen_type[grepl("effic", housingData$kitchen_type)] = "efficiency"
length(unique(housingData[,kitchen_type])) # 6 unique values including NA
```

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

```
#Finally lets deal with 1955 and replace that with NA -> I am assuming here 1955 is wrong and not a typ
housingData[, kitchen_type := sapply(kitchen_type, function(x) replace(x, which(x=="1955"), NA))]
length(unique(housingData[,kitchen_type])) # t unique values including NA (no 1955 -> NA)
```

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

```
housingData[,kitchen_type := factor(kitchen_type)]
levels(housingData$kitchen_type)
```

```
## [1] "combo"      "eatin"      "efficiency" "none"
```

```
##############################################################################################
#Fourth lets deal with the Garage column (track NA's for later)

housingData[,sum(is.na(garage_exists))] # 1826 NA values for garage exists
```

```
## [1] 1826
```

```
length(unique(housingData[,garage_exists])) # 7 "unique" values
```

```
## [1] 7
```

```
#Lets deal with the capitalization and misspelling of yes
housingData[,garage_exists := tolower(garage_exists)]
housingData$garage_exists[grepl("y", housingData$garage_exists)] = "yes"
length(unique(housingData[,garage_exists])) # 5 unique values including NA
```

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

```
#Lets treat underground and ug as yes
housingData$garage_exists[grepl("u", housingData$garage_exists)] = "yes"
length(unique(housingData[,garage_exists])) # 3 unique values including NA
```

```
## [1] 3
```

```
#Lets treat 1 as a yes
housingData$garage_exists[grepl("1", housingData$garage_exists)] = "yes"
length(unique(housingData[,garage_exists])) # 2 unique values including NA
```

```
## [1] 2
```

```
housingData[,garage_exists := factor(garage_exists)]
setattr(housingData$garage_exists,"levels",c("yes","no"))
levels(housingData$garage_exists)
```

```
## [1] "yes" "no"
```

```
#Fill NA's in garage with No's -> Use 1s in missingness to indicate this later om.
housingData[, c("garage_exists")][is.na(housingData[, c("garage_exists")])] = "no"


##############################################################################################
#Fifth lets take the date column and treat it as an ordinal factor

#In order to limit the total number of levels in Date, lets just grabs the months
#We sacrifice some granularity, but hopefully this generalize better

housingData$date_of_sale = format(as.Date(housingData$date_of_sale, format="%m/%d/%Y"),"%m")
housingData[,date_of_sale:= factor(date_of_sale,ordered=TRUE)]
length(unique(housingData[,date_of_sale])) #13 including NA which is what we want
```

```
## [1] 13
```

```
#Lets take a look at our data set now

housingData
```

```
##       approx_year_built cats_allowed common_charges community_district_num
##    1:              1955           no            767                     25
##    2:              1955           no             NA                     25
##    3:              2004           no            167                     24
##    4:              2002           no            275                     25
##    5:              1949          yes             NA                     26
##   ---
## 2226:              1987           no            480                     25
## 2227:              1983          yes            956                     25
## 2228:              2010           no            250                     24
## 2229:              2010           no            250                     24
## 2230:              1982           no            792                     25
```

```
##       coop_condo date_of_sale dining_room_type dogs_allowed fuel_type
##    1:      co-op           02            combo           no       gas
##    2:      co-op           02           formal           no       oil
##    3:      condo           02            combo           no      <NA>
##    4:      condo           02            combo           no       gas
##    5:      co-op           02            combo          yes       gas
##   ---
## 2226:      condo         <NA>            combo           no       gas
## 2227:      condo         <NA>           formal           no       gas
## 2228:      condo         <NA>           formal           no       gas
## 2229:      condo         <NA>           formal           no       gas
## 2230:      condo         <NA>           formal           no       gas
##                                      full_address_or_zip_code garage_exists
##    1:                                        Flushing NY, 11355           no
##    2: 30-11 Parsons Blvd,  Flushing NY, 11354 ( Sold )      Share           no
##    3:                          102-14 Lewis Ave,  Corona NY, 11368           no
##    4:                  144-48 Roosevelt Ave,  Flushing NY, 11354           no
##    5:                      245-27 76th Ave,  Bellerose NY, 11426           no
##   ---
## 2226:                           Not AvailableFlushing NY, 11355           no
## 2227:                        One Bay Club Dr,  Bayside NY, 11360          yes
## 2228:                                      Ridgewood NY, 11385           no
## 2229:                                      Ridgewood NY, 11385           no
## 2230:                     Two Bay Club Drive,  Bayside NY, 11360          yes
##       kitchen_type maintenance_cost num_bedrooms num_floors_in_building
##    1:        eatin              NA            2                     6
##    2:        eatin             604            1                     7
##    3:   efficiency              NA            1                     1
##    4:        eatin              NA            3                    NA
##    5:        eatin             660            2                     2
##   ---
## 2226:        combo              NA            2                     7
## 2227:        eatin              NA            2                    NA
## 2228:        combo              NA            3                    NA
## 2229:        combo              NA            3                     4
## 2230:        combo              NA            2                    NA
##       num_full_bathrooms num_half_bathrooms num_total_rooms parking_charges
##    1:                  1                 NA               5              NA
##    2:                  1                 NA               4              NA
##    3:                  1                 NA               3              NA
##    4:                  2                 NA               5              NA
##    5:                  1                 NA               4              NA
##   ---
## 2226:                  1                 NA               4              NA
## 2227:                  2                 NA               5              99
## 2228:                  2                 NA               6              NA
## 2229:                  2                 NA               6              NA
## 2230:                  2                 NA               5              NA
##       pct_tax_deductibl sale_price sq_footage total_taxes walk_score
##    1:                NA     228000         NA          NA         82
##    2:                NA     235500        890          NA         89
##    3:                NA     137550        550        5500         90
##    4:                NA     545000         NA        2260         94
##    5:                39     241700        675          NA         71
```

```
##   ---
## 2226:                NA         NA         NA        3588         97
## 2227:                NA         NA         NA        5100         82
## 2228:                NA         NA       1500         250         96
## 2229:                NA         NA       1600         250         96
## 2230:                NA         NA       1134        3785         82
##       listing_price_to_nearest_1000
##    1:                            NA
##    2:                            NA
##    3:                            NA
##    4:                            NA
##    5:                            NA
##   ---
## 2226:                           628
## 2227:                           988
## 2228:                           850
## 2229:                           850
## 2230:                           899
```

```
summary(housingData)
```

```
##  approx_year_built cats_allowed common_charges   community_district_num
##  Min.   :1893      no :1402     Min.   :  70.0   Min.   : 3.00
##  1st Qu.:1950      yes: 828     1st Qu.: 280.0   1st Qu.:25.00
##  Median :1958                   Median : 390.0   Median :26.00
##  Mean   :1963                   Mean   : 441.8   Mean   :26.33
##  3rd Qu.:1970                   3rd Qu.: 551.5   3rd Qu.:28.00
##  Max.   :2017                   Max.   :2499.0   Max.   :32.00
##  NA's   :40                     NA's   :1684     NA's   :19
##  coop_condo    date_of_sale  dining_room_type dogs_allowed    fuel_type
##  co-op:1661   12     :  58   combo :957       no :1684     electric:  62
##  condo: 569   06     :  53   formal:620       yes: 546     gas     :1348
##               01     :  50   none  :  2                    none    :   3
##               11     :  47   other :203                    oil     : 664
##               05     :  46   NA's  :448                    other   :  41
##               (Other): 274                                 NA's    : 112
##               NA's   :1702
##  full_address_or_zip_code garage_exists     kitchen_type maintenance_cost
##  Length:2230              yes: 404      combo     :399   Min.   : 155.0
##  Class :character         no :1826      eatin     :942   1st Qu.: 630.5
##  Mode  :character                       efficiency:849   Median : 767.0
##                                         none      : 23   Mean   : 858.9
##                                         NA's      : 17   3rd Qu.: 985.5
##                                                          Max.   :4659.0
##                                                          NA's   :623
##   num_bedrooms   num_floors_in_building num_full_bathrooms num_half_bathrooms
##  Min.   :0.000   Min.   : 1.000         Min.   :1.000      Min.   :0.0000
##  1st Qu.:1.000   1st Qu.: 3.000         1st Qu.:1.000      1st Qu.:1.0000
##  Median :2.000   Median : 6.000         Median :1.000      Median :1.0000
##  Mean   :1.653   Mean   : 7.785         Mean   :1.231      Mean   :0.9535
##  3rd Qu.:2.000   3rd Qu.: 7.000         3rd Qu.:1.000      3rd Qu.:1.0000
##  Max.   :6.000   Max.   :34.000         Max.   :3.000      Max.   :2.0000
##  NA's   :115     NA's   :650                               NA's   :2058
##  num_total_rooms  parking_charges pct_tax_deductibl   sale_price
```

```
##  Min.   : 0.000   Min.   :  6.0   Min.   :20.0      Min.   : 55000
##  1st Qu.: 3.000   1st Qu.: 60.0   1st Qu.:40.0      1st Qu.:171500
##  Median : 4.000   Median : 99.0   Median :50.0      Median :259500
##  Mean   : 4.139   Mean   :107.6   Mean   :45.4      Mean   :314957
##  3rd Qu.: 5.000   3rd Qu.:149.0   3rd Qu.:50.0      3rd Qu.:428875
##  Max.   :14.000   Max.   :837.0   Max.   :75.0      Max.   :999999
##  NA's   :2        NA's   :1671    NA's   :1754      NA's   :1702
##    sq_footage      total_taxes     walk_score    listing_price_to_nearest_1000
##  Min.   : 100.0   Min.   :  11   Min.   : 7.00   Min.   :  65.0
##  1st Qu.: 743.0   1st Qu.: 281   1st Qu.:77.00   1st Qu.: 229.8
##  Median : 881.0   Median :2411   Median :89.00   Median : 329.5
##  Mean   : 955.4   Mean   :2226   Mean   :83.92   Mean   : 385.6
##  3rd Qu.:1100.0   3rd Qu.:3500   3rd Qu.:95.00   3rd Qu.: 525.0
##  Max.   :6215.0   Max.   :9300   Max.   :99.00   Max.   :1000.0
##  NA's   :1210     NA's   :1646                   NA's   :534
```

Data Manipulation I (Creating new columns)

```
#First lets just add up all the charges into a single column
#Assign new column totalCharges to be the row sum of the chargeCols ignoring NA's
housingData[, totalCharges := rowSums(.SD,na.rm=TRUE), .SDcols = c("common_charges","maintenance_cost",
```

```
##       approx_year_built cats_allowed common_charges community_district_num
##    1:              1955          no            767                     25
##    2:              1955          no             NA                     25
##    3:              2004          no            167                     24
##    4:              2002          no            275                     25
##    5:              1949         yes             NA                     26
##   ---
## 2226:              1987          no            480                     25
## 2227:              1983         yes            956                     25
## 2228:              2010          no            250                     24
## 2229:              2010          no            250                     24
## 2230:              1982          no            792                     25
##       coop_condo date_of_sale dining_room_type dogs_allowed fuel_type
##    1:      co-op           02            combo           no       gas
##    2:      co-op           02           formal           no       oil
##    3:      condo           02            combo           no      <NA>
##    4:      condo           02            combo           no       gas
##    5:      co-op           02            combo          yes       gas
##   ---
## 2226:      condo         <NA>            combo           no       gas
## 2227:      condo         <NA>           formal           no       gas
## 2228:      condo         <NA>           formal           no       gas
## 2229:      condo         <NA>           formal           no       gas
## 2230:      condo         <NA>           formal           no       gas
##                                      full_address_or_zip_code garage_exists
##    1:                                      Flushing NY, 11355            no
##    2: 30-11 Parsons Blvd,  Flushing NY, 11354 ( Sold )      Share           no
##    3:                          102-14 Lewis Ave,  Corona NY, 11368           no
##    4:                   144-48 Roosevelt Ave,  Flushing NY, 11354           no
##    5:                       245-27 76th Ave,  Bellerose NY, 11426           no
##   ---
```

```
## 2226:                            Not AvailableFlushing NY, 11355           no
## 2227:                        One Bay Club Dr,  Bayside NY, 11360          yes
## 2228:                                        Ridgewood NY, 11385           no
## 2229:                                        Ridgewood NY, 11385           no
## 2230:                     Two Bay Club Drive,  Bayside NY, 11360          yes
##       kitchen_type maintenance_cost num_bedrooms num_floors_in_building
##    1:        eatin               NA            2                      6
##    2:        eatin              604            1                      7
##    3:   efficiency               NA            1                      1
##    4:        eatin               NA            3                     NA
##    5:        eatin              660            2                      2
##   ---
## 2226:        combo               NA            2                      7
## 2227:        eatin               NA            2                     NA
## 2228:        combo               NA            3                     NA
## 2229:        combo               NA            3                      4
## 2230:        combo               NA            2                     NA
##       num_full_bathrooms num_half_bathrooms num_total_rooms parking_charges
##    1:                  1                 NA               5              NA
##    2:                  1                 NA               4              NA
##    3:                  1                 NA               3              NA
##    4:                  2                 NA               5              NA
##    5:                  1                 NA               4              NA
##   ---
## 2226:                  1                 NA               4              NA
## 2227:                  2                 NA               5              99
## 2228:                  2                 NA               6              NA
## 2229:                  2                 NA               6              NA
## 2230:                  2                 NA               5              NA
##       pct_tax_deductibl sale_price sq_footage total_taxes walk_score
##    1:                NA     228000         NA          NA         82
##    2:                NA     235500        890          NA         89
##    3:                NA     137550        550        5500         90
##    4:                NA     545000         NA        2260         94
##    5:                39     241700        675          NA         71
##   ---
## 2226:                NA         NA         NA        3588         97
## 2227:                NA         NA         NA        5100         82
## 2228:                NA         NA       1500         250         96
## 2229:                NA         NA       1600         250         96
## 2230:                NA         NA       1134        3785         82
##       listing_price_to_nearest_1000 totalCharges
##    1:                            NA          767
##    2:                            NA          604
##    3:                            NA         5667
##    4:                            NA         2535
##    5:                            NA          660
##   ---
## 2226:                           628         4068
## 2227:                           988         6155
## 2228:                           850          500
## 2229:                           850          500
## 2230:                           899         4577
```

```
housingData[,sum(is.na(totalCharges))] # No NA's here which is good since
```

```
## [1] 0
```

```
##############################################################################################
#Second lets extract the zip codes and assign them to their own column

#Lets use a regular expression to extract the zip code out of this field
housingData[,zip_code := substr(str_extract(full_address_or_zip_code,"[0-9]{5}"),1,5)]
housingData[,zip_code := as.numeric(zip_code)]
#We can now drop the full_address column since we wont need that
housingData[,full_address_or_zip_code := NULL]


##############################################################################################
#Third lets add up full and half bathrooms
#Lets divide the half bathroom columns by 2 so that when we add them it is more granular
housingData[,num_half_bathrooms:=num_half_bathrooms/2]
#Assign a new column to represent the total number of bathrooms
housingData[,totalBathrooms :=rowSums(.SD,na.rm=TRUE), .SDcols = c("num_full_bathrooms","num_half_bathr
```

```
##       approx_year_built cats_allowed common_charges community_district_num
##    1:              1955           no            767                     25
##    2:              1955           no             NA                     25
##    3:              2004           no            167                     24
##    4:              2002           no            275                     25
##    5:              1949          yes             NA                     26
##   ---
## 2226:              1987           no            480                     25
## 2227:              1983          yes            956                     25
## 2228:              2010           no            250                     24
## 2229:              2010           no            250                     24
## 2230:              1982           no            792                     25
##       coop_condo date_of_sale dining_room_type dogs_allowed fuel_type
##    1:      co-op           02            combo           no       gas
##    2:      co-op           02           formal           no       oil
##    3:      condo           02            combo           no      <NA>
##    4:      condo           02            combo           no       gas
##    5:      co-op           02            combo          yes       gas
##   ---
## 2226:      condo         <NA>            combo           no       gas
## 2227:      condo         <NA>           formal           no       gas
## 2228:      condo         <NA>           formal           no       gas
## 2229:      condo         <NA>           formal           no       gas
## 2230:      condo         <NA>           formal           no       gas
##       garage_exists kitchen_type maintenance_cost num_bedrooms
##    1:            no        eatin               NA            2
##    2:            no        eatin              604            1
##    3:            no   efficiency               NA            1
##    4:            no        eatin               NA            3
##    5:            no        eatin              660            2
##   ---
```

```
## 2226:            no        combo               NA           2
## 2227:           yes        eatin               NA           2
## 2228:            no        combo               NA           3
## 2229:            no        combo               NA           3
## 2230:           yes        combo               NA           2
##       num_floors_in_building num_full_bathrooms num_half_bathrooms
##    1:                      6                 1                NA
##    2:                      7                 1                NA
##    3:                      1                 1                NA
##    4:                     NA                 2                NA
##    5:                      2                 1                NA
##   ---
## 2226:                      7                 1                NA
## 2227:                     NA                 2                NA
## 2228:                     NA                 2                NA
## 2229:                      4                 2                NA
## 2230:                     NA                 2                NA
##       num_total_rooms parking_charges pct_tax_deductibl sale_price sq_footage
##    1:               5              NA                NA     228000         NA
##    2:               4              NA                NA     235500        890
##    3:               3              NA                NA     137550        550
##    4:               5              NA                NA     545000         NA
##    5:               4              NA                39     241700        675
##   ---
## 2226:               4              NA                NA         NA         NA
## 2227:               5              99                NA         NA         NA
## 2228:               6              NA                NA         NA       1500
## 2229:               6              NA                NA         NA       1600
## 2230:               5              NA                NA         NA       1134
##       total_taxes walk_score listing_price_to_nearest_1000 totalCharges
##    1:          NA         82                            NA          767
##    2:          NA         89                            NA          604
##    3:        5500         90                            NA         5667
##    4:        2260         94                            NA         2535
##    5:          NA         71                            NA          660
##   ---
## 2226:        3588         97                           628         4068
## 2227:        5100         82                           988         6155
## 2228:         250         96                           850          500
## 2229:         250         96                           850          500
## 2230:        3785         82                           899         4577
##       zip_code totalBathrooms
##    1:    11355              1
##    2:    11354              1
##    3:    11368              1
##    4:    11354              2
##    5:    11426              1
##   ---
## 2226:    11355              1
## 2227:    11360              2
## 2228:    11385              2
## 2229:    11385              2
## 2230:    11360              2
```

```
#################################################################################################
#Fourth lets bring in some extra data that shows median income by zipcode
queensIncomeDataFilePath = "/home/peterjr/RepoCollections/MATH_342W_FinalProject/Datasets/income_queens_
queensIncomeData = data.table(read.csv(queensIncomeDataFilePath))

#Grab columns we want and remove the first row description of columns
queensIncomeData = queensIncomeData[-1,.(GEO_ID,S1901_C01_012E)]

#Change Data Type
queensIncomeData[,zip_code := as.numeric(GEO_ID)]


#Rename median income column
setnames(queensIncomeData, "S1901_C01_012E", "median_income")

queensIncomeData[,median_income := as.numeric(median_income)]
```

```
## Warning in eval(jsub, SDenv, parent.frame()): NAs introduced by coercion
```

```
#Drop the geo_id column
queensIncomeData[,GEO_ID := NULL]


#################################################################################################
#Fifth lets join this to our housing data on the zipcode
#We are doing a left join because I want everything in housing preserved -> median income can be impute

housingData = left_join(housingData,queensIncomeData,by.x = "zip_code",by.y = "zip_code")
```

```
## Joining, by = "zip_code"
```

```
housingData[,sum(is.na(median_income))] # 64 NA values, not bad since most are getting filled, should b
```

```
## [1] 64
```

Dealing with collinearity (Will cause issues later on especially with OLS)

```
#################################################################################################
#First lets grab the columns that are of interest to us
housingData = housingData[,.(approx_year_built,cats_allowed,community_district_num,coop_condo,date_of_sa
                             dogs_allowed,fuel_type,garage_exists,kitchen_type,num_bedrooms,num_floors_
                             sale_price,sq_footage,walk_score,totalCharges,zip_code,median_income)]



#################################################################################################
#Second lets build up our missing table O/` where 1 indicates a NA value in the original

#Create a missing data table and fill with zeros
colNames = names(housingData)
missRows = nrow(housingData)
```

```
missCols = ncol(housingData)
missingData = setNames(data.table(matrix(0,nrow = missRows, ncol = missCols)), colNames)
setnames(missingData,1:ncol(missingData), paste0(names(missingData)[1:ncol(missingData)], '_miss'))
#Data Set with 1s indicating missing in housingData
missingData[is.na(housingData)] = 1


#Let's get a correlation matrix on the numeric only data in our housing data
numericOnlyData = housingData[ , .SD, .SDcols = is.numeric]
ncol(numericOnlyData) # 12 total numeric columns
```

```
## [1] 12
```

```
#We expect there to be at most 12 1 values in the nxn correlation matrix for matching columns
#More than 12 values indicates that there is somewhere else where two different columns are perfectly c

correlationMatrix = as.matrix(cor(numericOnlyData))

length(which(correlationMatrix==1)) # 12 matches for perfect correlation, this is okay since it is colu
```

```
## [1] 12
```

```
#Remove missing columns where the sum is 0. Implies housingData did not have any NAs.
#Due to the nature of the construction of the missing table, all columns in housingData have a correspo
#This isn't fully accurate for the original columns without missingness
checkZero= function(x){
    if(sum(x)==0){
      TRUE
    }
}

length(missingData[,sapply(missingData,  checkZero)]) # 7 columns where no missingness, we will drop th
```

```
## [1] 7
```

```
missingData = missingData[, colSums(missingData != 0) > 0, with = FALSE]
```

Imputation Via MissForest on the Data

```
#############################################################################################
#Lets impute our data set including sale price
imputeSet = housingData

Ximp = missForest(imputeSet,verbose = TRUE)
```

```
##   missForest iteration 1 in progress...done!
##     estimated error(s): 0.3512139 0.1674269
##     difference(s): 0.117787 0.09641256
##     time: 11.959 seconds
##
```

```
##   missForest iteration 2 in progress...done!
##     estimated error(s): 0.3400797 0.1646671
##     difference(s): 0.001556426 0.05235426
##     time: 11.488 seconds
##
##   missForest iteration 3 in progress...done!
##     estimated error(s): 0.3406191 0.1618975
##     difference(s): 0.001717541 0.04871076
##     time: 13.108 seconds
##
##   missForest iteration 4 in progress...done!
##     estimated error(s): 0.3486926 0.1665229
##     difference(s): 0.001777233 0.04607623
##     time: 13.815 seconds
##
##   missForest iteration 5 in progress...done!
##     estimated error(s): 0.3514107 0.1634513
##     difference(s): 0.001573849 0.04461883
##     time: 12.246 seconds
##
##   missForest iteration 6 in progress...done!
##     estimated error(s): 0.3493235 0.163591
##     difference(s): 0.001348848 0.04310538
##     time: 11.844 seconds
##
##   missForest iteration 7 in progress...done!
##     estimated error(s): 0.34564 0.1612243
##     difference(s): 0.00114858 0.04316143
##     time: 11.81 seconds
##
##   missForest iteration 8 in progress...done!
##     estimated error(s): 0.3423238 0.1687926
##     difference(s): 0.001182839 0.04108744
##     time: 11.771 seconds
##
##   missForest iteration 9 in progress...done!
##     estimated error(s): 0.3435774 0.1614964
##     difference(s): 0.001243828 0.04439462
##     time: 11.79 seconds
```

```
#Get our final imputed Dataset and bind it to the missiningness table
finalHousingData = cbind(Ximp$ximp,missingData)


#Lets do the same check as in previous for our finalHousingData

numericOnlyData2 = finalHousingData[ , .SD, .SDcols = is.numeric]
ncol(numericOnlyData2) # 25 total numeric columns
```

```
## [1] 25
```

```
#We expect there to be at most 25 1 values in the nxn correlation matrix for matching columns
#More than 25 values indicates that there is somewhere else where two different columns are perfectly c
```

```
correlationMatrix2 = as.matrix(cor(numericOnlyData2))

length(which(correlationMatrix2==1)) # 27 matches for perfect correlation, 2 columns are perfectly corr
```

```
## [1] 27
```

```
cor(finalHousingData[,"sale_price_miss"],finalHousingData[,"date_of_sale_miss"]) # These are the 2 perf
```

```
##                 date_of_sale_miss
## sale_price_miss                 1
```

```
#Let's remove sale_price_miss -> Also it makes sense these two are perfectly correlated, a house with n

finalHousingData = finalHousingData[,!("sale_price_miss")]


finalHousingData
```

```
##       approx_year_built cats_allowed community_district_num coop_condo
##    1:              1955           no                     25      co-op
##    2:              1955           no                     25      co-op
##    3:              2004           no                     24      condo
##    4:              2002           no                     25      condo
##    5:              1949          yes                     26      co-op
##   ---
## 2226:              1987           no                     25      condo
## 2227:              1983          yes                     25      condo
## 2228:              2010           no                     24      condo
## 2229:              2010           no                     24      condo
## 2230:              1982           no                     25      condo
##       date_of_sale dining_room_type dogs_allowed fuel_type garage_exists
##    1:           02            combo           no       gas            no
##    2:           02           formal           no       oil            no
##    3:           02            combo           no       gas            no
##    4:           02            combo           no       gas            no
##    5:           02            combo          yes       gas            no
##   ---
## 2226:           02            combo           no       gas            no
## 2227:           02           formal           no       gas           yes
## 2228:           06           formal           no       gas            no
## 2229:           06           formal           no       gas            no
## 2230:           02           formal           no       gas           yes
##       kitchen_type num_bedrooms num_floors_in_building totalBathrooms
##    1:        eatin            2               6.000000              1
##    2:        eatin            1               7.000000              1
##    3:   efficiency            1               1.000000              1
##    4:        eatin            3               6.336250              2
##    5:        eatin            2               2.000000              1
##   ---
## 2226:        combo            2               7.000000              1
## 2227:        eatin            2              17.813333              2
```

```
## 2228:        combo            3              4.019375             2
## 2229:        combo            3              4.000000             2
## 2230:        combo            2             17.465000             2
##       num_total_rooms sale_price sq_footage walk_score totalCharges zip_code
##    1:               5   228000.0   887.5841         82          767    11355
##    2:               4   235500.0   890.0000         89          604    11354
##    3:               3   137550.0   550.0000         90         5667    11368
##    4:               5   545000.0  1054.3343         94         2535    11354
##    5:               4   241700.0   675.0000         71          660    11426
##   ---
## 2226:               4   496979.0   902.8475         97         4068    11355
## 2227:               5   641201.7  1216.1000         82         6155    11360
## 2228:               6   559530.0  1500.0000         96          500    11385
## 2229:               6   559760.0  1600.0000         96          500    11385
## 2230:               5   625057.5  1134.0000         82         4577    11360
##       median_income approx_year_built_miss community_district_num_miss
##    1:         38451                      0                           0
##    2:         43660                      0                           0
##    3:         45980                      0                           0
##    4:         43660                      0                           0
##    5:         77487                      0                           0
##   ---
## 2226:         38451                      0                           0
## 2227:         82982                      0                           0
## 2228:         60526                      0                           0
## 2229:         60526                      0                           0
## 2230:         82982                      0                           0
##       date_of_sale_miss dining_room_type_miss fuel_type_miss kitchen_type_miss
##    1:                 0                     0              0                 0
##    2:                 0                     0              0                 0
##    3:                 0                     0              1                 0
##    4:                 0                     0              0                 0
##    5:                 0                     0              0                 0
##   ---
## 2226:                 1                     0              0                 0
## 2227:                 1                     0              0                 0
## 2228:                 1                     0              0                 0
## 2229:                 1                     0              0                 0
## 2230:                 1                     0              0                 0
##       num_bedrooms_miss num_floors_in_building_miss num_total_rooms_miss
##    1:                 0                           0                    0
##    2:                 0                           0                    0
##    3:                 0                           0                    0
##    4:                 0                           1                    0
##    5:                 0                           0                    0
##   ---
## 2226:                 0                           0                    0
## 2227:                 0                           1                    0
## 2228:                 0                           1                    0
## 2229:                 0                           0                    0
## 2230:                 0                           1                    0
##       sq_footage_miss zip_code_miss median_income_miss
##    1:               1             0                  0
##    2:               0             0                  0
```

```
##    3:              0            0                 0
##    4:              1            0                 0
##    5:              0            0                 0
##   ---
## 2226:              1            0                 0
## 2227:              1            0                 0
## 2228:              0            0                 0
## 2229:              0            0                 0
## 2230:              0            0                 0
```

Breaking up our data into X and y

```
#Lets break X and y into X_train/_test and y_train/test
#Later we will implement K-fold, but for now we want to test oos performance of OLS
K=5
test_prop = 1 / K

#Training data
train_indices = sample(1 : nrow(finalHousingData), round((1 - test_prop) * nrow(finalHousingData)))
train_Data = finalHousingData[train_indices,]
X_train = train_Data[,!c("sale_price")]
y_train = train_Data$sale_price

#Testing data
test_indices = setdiff(1 : nrow(finalHousingData), train_indices)
test_Data = finalHousingData[test_indices, ]
X_test = test_Data[,!c("sale_price")]
y_test = test_Data$sale_price
```

Linear Regression Model

```
#To see if our correlation checks work, we should not receive the warning "prediction from a rank-defic

#Lets run a traditional OLS with all of our features

lin_mod = lm(y_train~.,X_train)

#OOS performance
yHats_OLS = predict(lin_mod,X_test)

oosRMSE_OLS = sqrt(sum((y_test-yHats_OLS)^2)/length(y_test))

oosRMSE_OLS
```

```
## [1] 63858.87
```

```
#We do not get the warning "prediction from a rank-deficient fit may be misleading" HOORAY
#I also would not trust this RMSE simply because it is highly variable (dependent on the split in the a
```

Linear Regression Model Cross Validated Lasso

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

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

```
## Loaded glmnet 4.1-4
```

```
lin_mod_lasso = cv.glmnet(data.matrix(X_train),y_train,nfolds=K,alpha = 1)
opt_Lambda = lin_mod_lasso$lambda.min

yHats_Lasso = predict(lin_mod_lasso, data.matrix(X_test),s = opt_Lambda)

oosRMSE_Lasso = sqrt(sum((y_test-yHats_Lasso)^2)/length(y_test))

oosRMSE_Lasso
```

```
## [1] 65470.36
```

Regression Tree Model

```
#Lets fit a regression tree with all features
regTree_mod_all = YARFCART(X_train, y_train, calculate_oob_error = FALSE)
```

```
## YARF initializing with a fixed 1 trees...
## YARF factors created...
## YARF after data preprocessed... 45 total features...
## Beginning YARF regression model construction...done.
```

```
#OOS performance
yHats_RegTree_all = predict(regTree_mod_all,X_test)

oosRMSE_RegTree_all = sqrt(sum((y_test-yHats_RegTree_all)^2)/length(y_test))

oosRMSE_RegTree_all
```

```
## [1] 53633.89
```

Random Forest Model

```
#Lets fit a random Forest on all features

rf_mod_all = YARF(X_train, y_train, calculate_oob_error = FALSE)
```

```
## YARF initializing with a fixed 500 trees...
## YARF factors created...
## YARF after data preprocessed... 45 total features...
## Beginning YARF regression model construction...done.
```

```
#OOS performance
yHats_rf_all = predict(rf_mod_all,X_test)


oosRMSE_rf_all = sqrt(sum((y_test-yHats_rf_all)^2)/length(y_test))
oosRMSE_rf_all
```

```
## [1] 40839.71
```

Bagged Random Forest Model

```
#Lets fit a bagged random forest on all features

rfBag_mod_all = YARFBAG(X_train, y_train, calculate_oob_error = FALSE)
```

```
## YARF initializing with a fixed 500 trees...
## YARF factors created...
## YARF after data preprocessed... 45 total features...
## Beginning YARF regression model construction...done.
```

```
#OOS performance
yHats_rfBag_all = predict(rfBag_mod_all,X_test)


oosRMSE_rfBag_all = sqrt(sum((y_test-yHats_rfBag_all)^2)/length(y_test))

oosRMSE_rfBag_all
```

```
## [1] 39854.59
```

At this point it is pretty obvious that a Random Forest model will perform the best. I am deciding to stick with this model and optimize for it's hyper parameters. I believe this will yield the best OOS performance and given what we know it is expected to perform better than OLS/Lasso and a single Regression Tree. OLS/Lasso suffer from misspecifcation and a single Regression tree is too variable.

Random Forest Model Optimization (Hyper-Parameter Tuning)

```
finalHousingData
```

```
##       approx_year_built cats_allowed community_district_num coop_condo
##    1:              1955           no                     25      co-op
##    2:              1955           no                     25      co-op
##    3:              2004           no                     24      condo
##    4:              2002           no                     25      condo
##    5:              1949          yes                     26      co-op
##   ---
```

```
## 2226:              1987           no                    25      condo
## 2227:              1983          yes                    25      condo
## 2228:              2010           no                    24      condo
## 2229:              2010           no                    24      condo
## 2230:              1982           no                    25      condo
##       date_of_sale dining_room_type dogs_allowed fuel_type garage_exists
##    1:           02            combo           no       gas            no
##    2:           02           formal           no       oil            no
##    3:           02            combo           no       gas            no
##    4:           02            combo           no       gas            no
##    5:           02            combo          yes       gas            no
##   ---
## 2226:           02            combo           no       gas            no
## 2227:           02           formal           no       gas           yes
## 2228:           06           formal           no       gas            no
## 2229:           06           formal           no       gas            no
## 2230:           02           formal           no       gas           yes
##       kitchen_type num_bedrooms num_floors_in_building totalBathrooms
##    1:        eatin            2               6.000000              1
##    2:        eatin            1               7.000000              1
##    3:   efficiency            1               1.000000              1
##    4:        eatin            3               6.336250              2
##    5:        eatin            2               2.000000              1
##   ---
## 2226:        combo            2               7.000000              1
## 2227:        eatin            2              17.813333              2
## 2228:        combo            3               4.019375              2
## 2229:        combo            3               4.000000              2
## 2230:        combo            2              17.465000              2
##       num_total_rooms sale_price sq_footage walk_score totalCharges zip_code
##    1:               5   228000.0   887.5841         82          767    11355
##    2:               4   235500.0   890.0000         89          604    11354
##    3:               3   137550.0   550.0000         90         5667    11368
##    4:               5   545000.0  1054.3343         94         2535    11354
##    5:               4   241700.0   675.0000         71          660    11426
##   ---
## 2226:               4   496979.0   902.8475         97         4068    11355
## 2227:               5   641201.7  1216.1000         82         6155    11360
## 2228:               6   559530.0  1500.0000         96          500    11385
## 2229:               6   559760.0  1600.0000         96          500    11385
## 2230:               5   625057.5  1134.0000         82         4577    11360
##       median_income approx_year_built_miss community_district_num_miss
##    1:         38451                      0                           0
##    2:         43660                      0                           0
##    3:         45980                      0                           0
##    4:         43660                      0                           0
##    5:         77487                      0                           0
##   ---
## 2226:         38451                      0                           0
## 2227:         82982                      0                           0
## 2228:         60526                      0                           0
## 2229:         60526                      0                           0
## 2230:         82982                      0                           0
##       date_of_sale_miss dining_room_type_miss fuel_type_miss kitchen_type_miss
```

```
##    1:                 0                    0             0                 0
##    2:                 0                    0             0                 0
##    3:                 0                    0             1                 0
##    4:                 0                    0             0                 0
##    5:                 0                    0             0                 0
##   ---
## 2226:                 1                    0             0                 0
## 2227:                 1                    0             0                 0
## 2228:                 1                    0             0                 0
## 2229:                 1                    0             0                 0
## 2230:                 1                    0             0                 0
##       num_bedrooms_miss num_floors_in_building_miss num_total_rooms_miss
##    1:                 0                           0                    0
##    2:                 0                           0                    0
##    3:                 0                           0                    0
##    4:                 0                           1                    0
##    5:                 0                           0                    0
##   ---
## 2226:                 0                           0                    0
## 2227:                 0                           1                    0
## 2228:                 0                           1                    0
## 2229:                 0                           0                    0
## 2230:                 0                           1                    0
##       sq_footage_miss zip_code_miss median_income_miss
##    1:               1             0                  0
##    2:               0             0                  0
##    3:               0             0                  0
##    4:               1             0                  0
##    5:               0             0                  0
##   ---
## 2226:               1             0                  0
## 2227:               1             0                  0
## 2228:               0             0                  0
## 2229:               0             0                  0
## 2230:               0             0                  0
```