Data Exploration DATA621 Final

Modeling Housing Prices

Kai Lukowiak, Jaan Bernberg, Ilya Kats, Michael Muller

May 17, 2018

Abstract

This document provides an introduction to R Markdown, argues for its…

Contents

keywords: “pandoc, r markdown, knitr”

1. Introduction (Jaan)

Data Set Origin

Variables

Cases

Describe the background and motivation of your problem.

Number of complete cases original: 0
Number of complete cases after repairing known NA’s: 2,861 (≈ 98%)
Number of true NA’s: 58

2. Data Exploration (Kai)

3. Data Preparation (Kyle)

"The ggplot2 package is used to plot images in layers (Wickham 2009)."

Missing Value Imputation

Literature review

Methodology

Variable Transformations

Literature review

Methodology

4. Modeling (????)

Literature review

Methodology

5. Model Selection, Diagnostics & Conclusions (Ilya)

Literature review

Methodology

Appendix

Tables & Outputs

```
##     tidyverse         knitr      ggthemes          mice           VIM 
##          TRUE          TRUE          TRUE          TRUE          TRUE 
##         RCurl knitcitations 
##          TRUE          TRUE 
```

Variable	NA_qty	Pct_of_Tot
PoolQC	2909	99.66
MiscFeature	2814	96.40
Alley	2721	93.22
Fence	2348	80.44
FireplaceQu	1420	48.65
LotFrontage	486	16.65
GarageYrBlt	159	5.45
GarageFinish	159	5.45

Variable	NA_qty	Pct_of_Tot
GarageQual	159	5.45
GarageCond	159	5.45
Garage_Age_Yrs	159	5.45
GarageType	157	5.38
BsmtCond	82	2.81
BsmtExposure	82	2.81
BsmtQual	81	2.77
BsmtFinType2	80	2.74
BsmtFinType1	79	2.71
MasVnrType	24	0.82
MasVnrArea	23	0.79
MSZoning	4	0.14
Utilities	2	0.07
BsmtFullBath	2	0.07
BsmtHalfBath	2	0.07
Functional	2	0.07
Exterior1st	1	0.03
Exterior2nd	1	0.03
BsmtFinSF1	1	0.03
BsmtFinSF2	1	0.03
BsmtUnfSF	1	0.03
TotalBsmtSF	1	0.03
Electrical	1	0.03
KitchenQual	1	0.03
GarageCars	1	0.03
GarageArea	1	0.03
SaleType	1	0.03

Variable	NA_qty	Pct_of_Tot
PoolQC	2887	99.65
MiscFeature	2793	96.41
Alley	2700	93.20
Fence	2331	80.46
FireplaceQu	1414	48.81
LotFrontage	486	16.78
GarageYrBlt	158	5.45
GarageFinish	158	5.45
GarageQual	158	5.45
GarageCond	158	5.45
Garage_Age_Yrs	158	5.45
GarageType	157	5.42
BsmtQual	78	2.69
BsmtCond	78	2.69
BsmtExposure	78	2.69
BsmtFinType1	78	2.69
BsmtFinType2	78	2.69
MasVnrType	23	0.79
MasVnrArea	22	0.76
MSZoning	4	0.14
Utilities	2	0.07
BsmtFullBath	2	0.07

Variable	NA_qty	Pct_of_Tot
BsmtHalfBath	2	0.07
Functional	2	0.07
Exterior1st	1	0.03
Exterior2nd	1	0.03
BsmtFinSF1	1	0.03
BsmtFinSF2	1	0.03
BsmtUnfSF	1	0.03
TotalBsmtSF	1	0.03
Electrical	1	0.03
KitchenQual	1	0.03
GarageCars	1	0.03
GarageArea	1	0.03
SaleType	1	0.03

```
## [1] 2887
```

```
##        Id           MSSubClass        MSZoning     LotFrontage
##  Min.   :   1.0   Min.   : 20.00   C (all):  25   Min.   :  0.00
##  1st Qu.: 730.5   1st Qu.: 20.00   FV     : 139   1st Qu.: 43.00
##  Median :1460.0   Median : 50.00   RH     :  26   Median : 63.00
##  Mean   :1460.0   Mean   : 57.14   RL     :2265   Mean   : 57.77
##  3rd Qu.:2189.5   3rd Qu.: 70.00   RM     : 460   3rd Qu.: 78.00
##  Max.   :2919.0   Max.   :190.00   NA's   :   4   Max.   :313.00
##
##     LotArea        Street      Alley      LotShape   LandContour
##  Min.   :  1300   Grvl:  12   Grvl: 120   IR1: 968   Bnk: 117
##  1st Qu.:  7478   Pave:2907   Pave:  78   IR2:  76   HLS: 120
##  Median :  9453               NoA :2700   IR3:  16   Low:  60
##  Mean   : 10168               NA's:  21   Reg:1859   Lvl:2622
##  3rd Qu.: 11570
##  Max.   :215245
##
##   Utilities      LotConfig    LandSlope   Neighborhood    Condition1
##  AllPub:2916   Corner : 511   Gtl:2778   NAmes  : 443   Norm   :2511
##  NoSeWa:   1   CulDSac: 176   Mod: 125   CollgCr: 267   Feedr  : 164
##  NA's  :   2   FR2    :  85   Sev:  16   OldTown: 239   Artery :  92
##                FR3    :  14              Edwards: 194   RRAn   :  50
##                Inside :2133              Somerst: 182   PosN   :  39
##                                          NridgHt: 166   RRAe   :  28
##                                          (Other):1428   (Other):  35
##    Condition2     BldgType      HouseStyle    OverallQual
##  Norm   :2889   1Fam  :2425   1Story :1471   Min.   : 1.000
##  Feedr  :  13   2fmCon:  62   2Story : 872   1st Qu.: 5.000
##  Artery :   5   Duplex: 109   1.5Fin : 314   Median : 6.000
##  PosA   :   4   Twnhs :  96   SLvl   : 128   Mean   : 6.089
##  PosN   :   4   TwnhsE: 227   SFoyer :  83   3rd Qu.: 7.000
##  RRNn   :   2                 2.5Unf :  24   Max.   :10.000
##  (Other):   2                 (Other):  27
##   OverallCond      YearBuilt     YearRemodAdd    RoofStyle
##  Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  20
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1965   Gable  :2310
##  Median :5.000   Median :1973   Median :1993   Gambrel:  22
##  Mean   :5.565   Mean   :1971   Mean   :1984   Hip    : 551
##  3rd Qu.:6.000   3rd Qu.:2001   3rd Qu.:2004   Mansard:  11
##  Max.   :9.000   Max.   :2010   Max.   :2010   Shed   :   5
##
##     RoofMatl     Exterior1st    Exterior2nd     MasVnrType
##  CompShg:2876   VinylSd:1025   VinylSd:1014   BrkCmn :  25
##  Tar&Grv:  23   MetalSd: 450   MetalSd: 447   BrkFace: 879
##  WdShake:   9   HdBoard: 442   HdBoard: 406   None   :1742
##  WdShngl:   7   Wd Sdng: 411   Wd Sdng: 391   Stone  : 249
##  ClyTile:   1   Plywood: 221   Plywood: 270   NA's   :  24
##  Membran:   1   (Other): 369   (Other): 390
##  (Other):   2   NA's   :   1   NA's   :   1
##    MasVnrArea     ExterQual ExterCond  Foundation   BsmtQual
##  Min.   :   0.0   Ex: 107   Ex:  12   BrkTil: 311   Ex  : 258
##  1st Qu.:   0.0   Fa:  35   Fa:  67   CBlock:1235   Fa  :  88
##  Median :   0.0   Gd: 979   Gd: 299   PConc :1308   Gd  :1209
##  Mean   : 102.2   TA:1798   Po:   3   Slab  :  49   TA  :1283
##  3rd Qu.: 164.0             TA:2538   Stone :  11   NoB :  78
##  Max.   :1600.0                       Wood  :   5   NA's:   3
##  NA's   :23
##  BsmtCond    BsmtExposure  BsmtFinType1   BsmtFinSF1      BsmtFinType2
##  Fa  : 104   Av  : 418    Unf    :851   Min.   :   0.0   Unf    :2493
##  Gd  : 122   Gd  : 276    GLQ    :849   1st Qu.:   0.0   Rec    : 105
##  Po  :   5   Mn  : 239    ALQ    :429   Median : 368.5   LwQ    :  87
##  TA  :2606   No  :1904    Rec    :288   Mean   : 441.4   NoB    :  78
##  NoB :  78   NoB :  78    BLQ    :269   3rd Qu.: 733.0   BLQ    :  68
##  NA's:   4   NA's:   4    (Other):232   Max.   :5644.0   (Other):  86
##                           NA's   :  1   NA's   :1        NA's   :   2
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF      Heating
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Floor:   1
##  1st Qu.:   0.00   1st Qu.: 220.0   1st Qu.: 793.0   GasA :2874
##  Median :   0.00   Median : 467.0   Median : 989.5   GasW :  27
##  Mean   :  49.58   Mean   : 560.8   Mean   :1051.8   Grav :   9
##  3rd Qu.:   0.00   3rd Qu.: 805.5   3rd Qu.:1302.0   OthW :   2
##  Max.   :1526.00   Max.   :2336.0   Max.   :6110.0   Wall :   6
##  NA's   :1         NA's   :1        NA's   :1
##  HeatingQC CentralAir Electrical     X1stFlrSF      X2ndFlrSF
##  Ex:1493   N: 196     FuseA: 188   Min.   : 334   Min.   :   0.0
##  Fa:  92   Y:2723     FuseF:  50   1st Qu.: 876   1st Qu.:   0.0
##  Gd: 474              FuseP:   8   Median :1082   Median :   0.0
##  Po:   3              Mix  :   1   Mean   :1160   Mean   : 336.5
##  TA: 857              SBrkr:2671   3rd Qu.:1388   3rd Qu.: 704.0
##                       NA's :   1   Max.   :5095   Max.   :2065.0
##
##   LowQualFinSF        GrLivArea     BsmtFullBath     BsmtHalfBath
##  Min.   :   0.000   Min.   : 334   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:   0.000   1st Qu.:1126   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :   0.000   Median :1444   Median :0.0000   Median :0.00000
##  Mean   :   4.694   Mean   :1501   Mean   :0.4299   Mean   :0.06136
##  3rd Qu.:   0.000   3rd Qu.:1744   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1064.000   Max.   :5642   Max.   :3.0000   Max.   :2.00000
##                                    NA's   :2        NA's   :2
##     FullBath        HalfBath       BedroomAbvGr   KitchenAbvGr
##  Min.   :0.000   Min.   :0.0000   Min.   :0.00   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.00   1st Qu.:1.000
##  Median :2.000   Median :0.0000   Median :3.00   Median :1.000
##  Mean   :1.568   Mean   :0.3803   Mean   :2.86   Mean   :1.045
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.00   3rd Qu.:1.000
##  Max.   :4.000   Max.   :2.0000   Max.   :8.00   Max.   :3.000
##
##  KitchenQual  TotRmsAbvGrd      Functional     Fireplaces     FireplaceQu
##  Ex  : 205   Min.   : 2.000   Typ    :2717   Min.   :0.0000   Ex  :  43
##  Fa  :  70   1st Qu.: 5.000   Min2   :  70   1st Qu.:0.0000   Fa  :  74
##  Gd  :1151   Median : 6.000   Min1   :  65   Median :1.0000   Gd  : 744
##  TA  :1492   Mean   : 6.452   Mod    :  35   Mean   :0.5971   Po  :  46
##  NA's:   1   3rd Qu.: 7.000   Maj1   :  19   3rd Qu.:1.0000   TA  : 592
##              Max.   :15.000   (Other):  11   Max.   :4.0000   NoFp:1414
##                               NA's   :   2                    NA's:   6
##    GarageType    GarageYrBlt    GarageFinish   GarageCars
##  2Types :  23   Min.   :1895   Fin : 719    Min.   :0.000
##  Attchd :1723   1st Qu.:1961   RFn : 811    1st Qu.:1.000
##  Basment:  36   Median :1984   Unf :1230    Median :2.000
##  BuiltIn: 186   Mean   :2412   NoG : 158    Mean   :1.767
##  CarPort:  15   3rd Qu.:2003   NA's:   1    3rd Qu.:2.000
##  Detchd : 779   Max.   :9999                Max.   :5.000
##  NoG    : 157   NA's   :1                   NA's   :1
##    GarageArea     GarageQual  GarageCond  PavedDrive   WoodDeckSF
##  Min.   :   0.0   Ex  :   3   Ex  :   3   N: 216     Min.   :   0.00
##  1st Qu.: 320.0   Fa  : 124   Fa  :  74   P:  62     1st Qu.:   0.00
##  Median : 480.0   Gd  :  24   Gd  :  15   Y:2641     Median :   0.00
##  Mean   : 472.9   Po  :   5   Po  :  14              Mean   :  93.71
##  3rd Qu.: 576.0   TA  :2604   TA  :2654              3rd Qu.: 168.00
##  Max.   :1488.0   NoG : 158   NoG : 158              Max.   :1424.00
##  NA's   :1        NA's:   1   NA's:   1
##   OpenPorchSF     EnclosedPorch      X3SsnPorch       ScreenPorch
##  Min.   :  0.00   Min.   :   0.0   Min.   :  0.000   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.:   0.0   1st Qu.:  0.000   1st Qu.:  0.00
##  Median : 26.00   Median :   0.0   Median :  0.000   Median :  0.00
##  Mean   : 47.49   Mean   :  23.1   Mean   :  2.602   Mean   : 16.06
##  3rd Qu.: 70.00   3rd Qu.:   0.0   3rd Qu.:  0.000   3rd Qu.:  0.00
##  Max.   :742.00   Max.   :1012.0   Max.   :508.000   Max.   :576.00
##
##     PoolArea        PoolQC       Fence      MiscFeature    MiscVal
##  Min.   :  0.000   Ex  :   4   GdPrv: 118   Gar2:   5   Min.   :    0.00
##  1st Qu.:  0.000   Fa  :   2   GdWo : 112   Othr:   4   1st Qu.:    0.00
##  Median :  0.000   Gd  :   4   MnPrv: 329   Shed:  95   Median :    0.00
##  Mean   :  2.252   NoP :2887   MnWw :  12   TenC:   1   Mean   :   50.83
##  3rd Qu.:  0.000   NA's:  22   NoF  :2331   NoM :2793   3rd Qu.:    0.00
##  Max.   :800.000               NA's :  17   NA's:  21   Max.   :17000.00
##
##      MoSold           YrSold        SaleType    SaleCondition
##  Min.   : 1.000   Min.   :2006   WD     :2525   Abnorml: 190
##  1st Qu.: 4.000   1st Qu.:2007   New    : 239   AdjLand:  12
##  Median : 6.000   Median :2008   COD    :  87   Alloca :  24
##  Mean   : 6.213   Mean   :2008   ConLD  :  26   Family :  46
##  3rd Qu.: 8.000   3rd Qu.:2009   CWD    :  12   Normal :2402
##  Max.   :12.000   Max.   :2010   (Other):  29   Partial: 245
##                                  NA's   :   1
##    SalePrice         d_name          House_Age_Yrs    RemodAdd_Age_Yrs
##  Min.   : 34900   Length:2919        Min.   : -1.00   Min.   :-2.00
##  1st Qu.:129975   Class :character   1st Qu.:  7.00   1st Qu.: 4.00
##  Median :163000   Mode  :character   Median : 35.00   Median :15.00
##  Mean   :180921                      Mean   : 36.48   Mean   :23.53
##  3rd Qu.:214000                      3rd Qu.: 54.50   3rd Qu.:43.00
##  Max.   :755000                      Max.   :136.00   Max.   :60.00
##  NA's   :1459
##  Garage_Age_Yrs
##  Min.   :-200.00
##  1st Qu.:   5.00
##  Median :  25.00
##  Mean   :  28.07
##  3rd Qu.:  46.00
##  Max.   : 114.00
##  NA's   :1

## [1] 176

## [1] 0

## [1] 2861

## [1] 58
```

	n	mean	sd	median	min	max	kurtosis
LotFrontage	2919	5.776670e+01	33.4816355	63.0	0	313	2.1693381
LotArea	2919	1.016811e+04	7886.9963591	9453.0	1300	215245	264.3133838
OverallQual	2919	6.089072e+00	1.4099472	6.0	1	10	0.0629498
OverallCond	2919	5.564577e+00	1.1131307	5.0	1	9	1.4717941
YearBuilt	2919	1.971313e+03	30.2914415	1973.0	1872	2010	-0.5142007
YearRemodAdd	2919	1.984264e+03	20.8943442	1993.0	1950	2010	-1.3473139
MasVnrArea	2896	1.022013e+02	179.3342530	0.0	0	1600	9.2278531
BsmtFinSF1	2918	4.414232e+02	455.6108259	368.5	0	5644	6.8841727
BsmtFinSF2	2918	4.958225e+01	169.2056111	0.0	0	1526	18.7872826
BsmtUnfSF	2918	5.607721e+02	439.5436594	467.0	0	2336	0.3985396
TotalBsmtSF	2918	1.051778e+03	440.7662581	989.5	0	6110	9.1250560
X1stFlrSF	2919	1.159582e+03	392.3620787	1082.0	334	5095	6.9357030
X2ndFlrSF	2919	3.364837e+02	428.7014555	0.0	0	2065	-0.4253575
LowQualFinSF	2919	4.694416e+00	46.3968245	0.0	0	1064	174.5095701
GrLivArea	2919	1.500760e+03	506.0510451	1444.0	334	5642	4.1076200
BsmtFullBath	2917	4.298937e-01	0.5247356	0.0	0	3	-0.7380409
BsmtHalfBath	2917	6.136440e-02	0.2456869	0.0	0	2	14.8083680
FullBath	2919	1.568003e+00	0.5529693	2.0	0	4	-0.5409486
HalfBath	2919	3.802672e-01	0.5028716	0.0	0	2	-1.0350789
BedroomAbvGr	2919	2.860226e+00	0.8226931	3.0	0	8	1.9326437
KitchenAbvGr	2919	1.044536e+00	0.2144620	1.0	0	3	19.7264407
TotRmsAbvGrd	2919	6.451524e+00	1.5693791	6.0	2	15	1.1621540
Fireplaces	2919	5.971223e-01	0.6461294	1.0	0	4	0.0721322
GarageYrBlt	2918	2.412418e+03	1815.6634616	1984.0	1895	9999	13.5081444
GarageCars	2918	1.766621e+00	0.7616243	2.0	0	5	0.2335170
GarageArea	2918	4.728746e+02	215.3948150	480.0	0	1488	0.9334205
WoodDeckSF	2919	9.370983e+01	126.5265893	0.0	0	1424	6.7212891
OpenPorchSF	2919	4.748681e+01	67.5754934	26.0	0	742	10.9070384
EnclosedPorch	2919	2.309832e+01	64.2442456	0.0	0	1012	28.3058078
X3SsnPorch	2919	2.602261e+00	25.1881693	0.0	0	508	149.0477443
ScreenPorch	2919	1.606235e+01	56.1843651	0.0	0	576	17.7300026

	n	mean	sd	median	min	max	kurtosis
PoolArea	2919	2.251799e+00	35.6639460	0.0	0	800	297.9135190
MiscVal	2919	5.082597e+01	567.4022106	0.0	0	17000	562.7189675
MoSold	2919	6.213087e+00	2.7147618	6.0	1	12	-0.4573565
YrSold	2919	2.007793e+03	1.3149645	2008.0	2006	2010	-1.1564874
House_Age_Yrs	2919	3.647996e+01	30.3361823	35.0	-1	136	-0.5057903
RemodAdd_Age_Yrs	2919	2.352826e+01	20.8920609	15.0	-2	60	-1.3388441
Garage_Age_Yrs	2918	2.806923e+01	25.8003331	25.0	-200	114	1.6139352

```
##     MSZoning     Street      Alley      LotShape   LandContour
##  C (all):  25   Grvl:  12   Grvl: 120   IR1: 968   Bnk: 117
##  FV     : 139   Pave:2907   Pave:  78   IR2:  76   HLS: 120
##  RH     :  26               NoA :2700   IR3:  16   Low:  60
##  RL     :2265               NA's:  21   Reg:1859   Lvl:2622
##  RM     : 460
##  NA's   :   4
##
##   Utilities      LotConfig    LandSlope   Neighborhood    Condition1
##  AllPub:2916   Corner : 511   Gtl:2778   NAmes  : 443   Norm   :2511
##  NoSeWa:   1   CulDSac: 176   Mod: 125   CollgCr: 267   Feedr  : 164
##  NA's  :   2   FR2    :  85   Sev:  16   OldTown: 239   Artery :  92
##                FR3    :  14              Edwards: 194   RRAn   :  50
##                Inside :2133              Somerst: 182   PosN   :  39
##                                          NridgHt: 166   RRAe   :  28
##                                          (Other):1428   (Other):  35
##    Condition2     BldgType      HouseStyle     RoofStyle       RoofMatl
##  Norm   :2889   1Fam  :2425   1Story :1471   Flat   :  20   CompShg:2876
##  Feedr  :  13   2fmCon:  62   2Story : 872   Gable  :2310   Tar&Grv:  23
##  Artery :   5   Duplex: 109   1.5Fin : 314   Gambrel:  22   WdShake:   9
##  PosA   :   4   Twnhs :  96   SLvl   : 128   Hip    : 551   WdShngl:   7
##  PosN   :   4   TwnhsE: 227   SFoyer :  83   Mansard:  11   ClyTile:   1
##  RRNn   :   2                 2.5Unf :  24   Shed   :   5   Membran:   1
##  (Other):   2                 (Other):  27                  (Other):   2
##   Exterior1st     Exterior2nd     MasVnrType     ExterQual ExterCond
##  VinylSd:1025   VinylSd:1014   BrkCmn :  25   Ex: 107   Ex:  12
##  MetalSd: 450   MetalSd: 447   BrkFace: 879   Fa:  35   Fa:  67
##  HdBoard: 442   HdBoard: 406   None   :1742   Gd: 979   Gd: 299
##  Wd Sdng: 411   Wd Sdng: 391   Stone  : 249   TA:1798   Po:   3
##  Plywood: 221   Plywood: 270   NA's   :  24             TA:2538
##  (Other): 369   (Other): 390
##  NA's   :   1   NA's   :   1
##   Foundation    BsmtQual     BsmtCond     BsmtExposure  BsmtFinType1
##  BrkTil: 311   Ex  : 258   Fa  : 104   Av  : 418    Unf    :851
##  CBlock:1235   Fa  :  88   Gd  : 122   Gd  : 276    GLQ    :849
##  PConc :1308   Gd  :1209   Po  :   5   Mn  : 239    ALQ    :429
##  Slab  :  49   TA  :1283   TA  :2606   No  :1904    Rec    :288
##  Stone :  11   NoB :  78   NoB :  78   NoB :  78    BLQ    :269
##  Wood  :   5   NA's:   3   NA's:   4   NA's:   4    (Other):232
##                                                     NA's   :  1
##   BsmtFinType2   Heating     HeatingQC CentralAir Electrical   KitchenQual
##  Unf    :2493   Floor:   1   Ex:1493   N: 196     FuseA: 188   Ex  : 205
##  Rec    : 105   GasA :2874   Fa:  92   Y:2723     FuseF:  50   Fa  :  70
##  LwQ    :  87   GasW :  27   Gd: 474              FuseP:   8   Gd  :1151
##  NoB    :  78   Grav :   9   Po:   3             Mix  :   1   TA  :1492
##  BLQ    :  68   OthW :   2   TA: 857             SBrkr:2671   NA's:   1
##  (Other):  86   Wall :   6                       NA's :   1
##  NA's   :   2
##    Functional   FireplaceQu   GarageType   GarageFinish GarageQual
##  Typ    :2717   Ex  :  43   2Types :  23   Fin : 719    Ex  :   3
##  Min2   :  70   Fa  :  74   Attchd :1723   RFn : 811    Fa  : 124
##  Min1   :  65   Gd  : 744   Basment:  36   Unf :1230    Gd  :  24
##  Mod    :  35   Po  :  46   BuiltIn: 186   NoG : 158    Po  :   5
##  Maj1   :  19   TA  : 592   CarPort:  15   NA's:   1    TA  :2604
##  (Other):  11   NoFp:1414   Detchd : 779                NoG : 158
##  NA's   :   2   NA's:   6   NoG    : 157                NA's:   1
##  GarageCond  PavedDrive  PoolQC       Fence      MiscFeature
##  Ex  :   3   N: 216     Ex  :   4   GdPrv: 118   Gar2:   5
##  Fa  :  74   P:  62     Fa  :   2   GdWo : 112   Othr:   4
##  Gd  :  15   Y:2641     Gd  :   4   MnPrv: 329   Shed:  95
##  Po  :  14              NoP :2887   MnWw :  12   TenC:   1
##  TA  :2654              NA's:  22   NoF  :2331   NoM :2793
##  NoG : 158                          NA's :  17   NA's:  21
##  NA's:   1
##     SaleType    SaleCondition
##  WD     :2525   Abnorml: 190
##  New    : 239   AdjLand:  12
##  COD    :  87   Alloca :  24
##  ConLD  :  26   Family :  46
##  CWD    :  12   Normal :2402
##  (Other):  29   Partial: 245
##  NA's   :   1
```

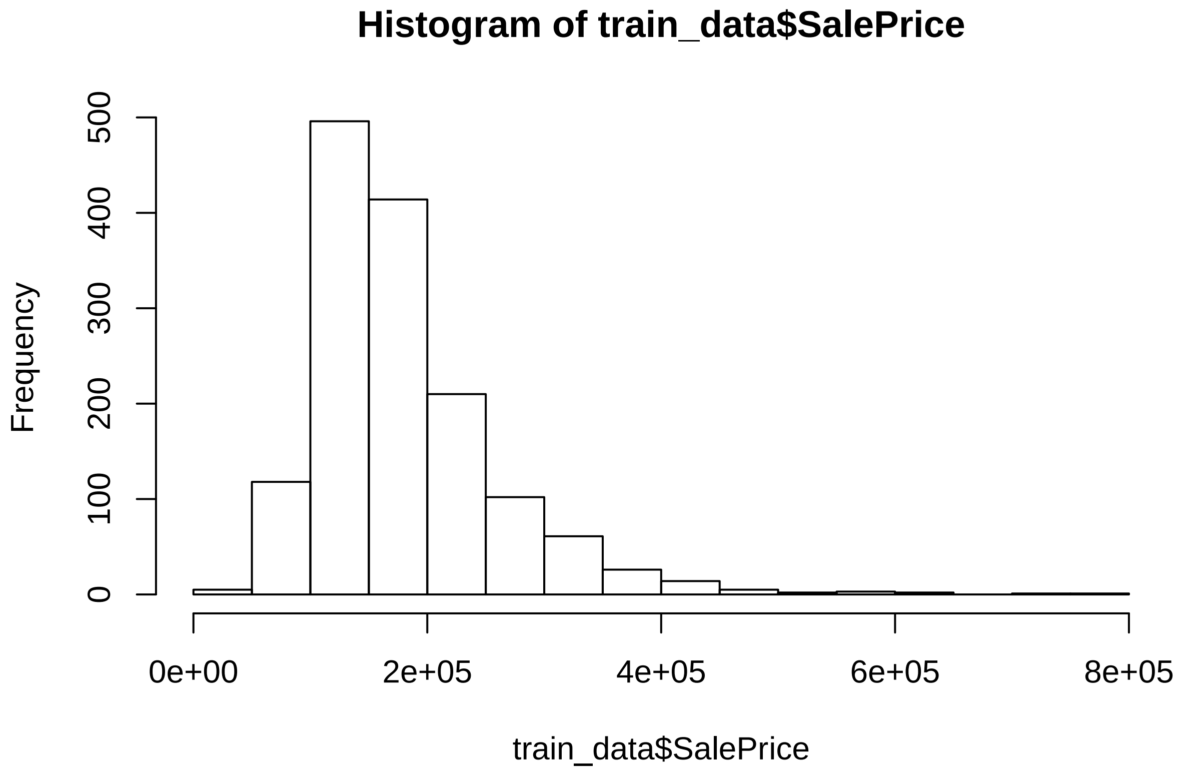
```
## [1] 1460   84
```

```
## [1] 1459   84
```

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotCon
524	60	RL	130	40094	Pave	NA	IR1	Bnk	AllPub	Inside
2296	60	RL	134	16659	Pave	NA	IR1	Lvl	AllPub	Corner
2550	20	RL	128	39290	Pave	NA	IR1	Bnk	AllPub	Inside
2593	20	RL	68	8298	Pave	NA	IR1	HLS	AllPub	Inside

```
## Warning in do_read_bib(file, encoding = .Encoding, srcfile):
## bibliography.bib:9:0
##  syntax error, unexpected TOKEN_COMMA, expecting TOKEN_ABBREV or TOKEN_VALUE
##  Dropping the entry `wickham_ggplot2_2009` (starting at line 1)
```

Plots



```
## Warning in do_read_bib(file, encoding = .Encoding, srcfile):
## bibliography.bib:9:0
##  syntax error, unexpected TOKEN_COMMA, expecting TOKEN_ABBREV or TOKEN_VALUE
##  Dropping the entry `wickham_ggplot2_2009` (starting at line 1)
```

Bibliography

(???)

Code

```
##Use PDF for Final Paper
  # html_document:
  #   theme: yeti
  #   code_folding: hide
  #   toc: true
  #   toc_float:
  #     collapsed: true
  #     smooth_scroll: false

knitr::opts_chunk$set(
                      error = F
                      , message = F
                      #,tidy = T
                      , cache = T
                      , warning = T
                      , results = 'hide' #suppress code output
                      , echo = F #suppress code
                      , fig.show = 'hide' #suppress plots
                      )

install_load <- function(pkg){
  # Load packages & Install them if needed.
  # CODE SOURCE: https://gist.github.com/stevenworthington/3178163
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg)) install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE, quietly = TRUE, warn.conflicts = FALSE)
}

# required packages
packages <- c("tidyverse","knitr", "ggthemes", "mice", "VIM", "RCurl", "knitcitations")

install_load(packages)

##Read data
url_train <- "https://raw.githubusercontent.com/kaiserxc/DATA621FinalProject/master/house-prices-advance
url_test <-  "https://raw.githubusercontent.com/kaiserxc/DATA621FinalProject/master/house-prices-advance

stand_read <- function(url){
  return(read.csv(text = getURL(url)))
}

o_train <-
  stand_read(url_train) %>%
  mutate(d_name = 'train')
o_test <- stand_read(url_test) %>%
  mutate(SalePrice = NA, d_name = 'test')

full_set <- rbind(o_train, o_test)
# x <- plot_missing(full_set)
na_review <- function(df){
  # returns df of vars w/ NA qty desc.
  na_qty <- colSums(is.na(df)) %>% as.data.frame(stringsAsFactors=F)
  colnames(na_qty) <- c("NA_qty")
  na_qty <- cbind('Variable' = rownames(na_qty), na_qty) %>%
    select(Variable, NA_qty)
  rownames(na_qty) <- NULL

  na_qty <- na_qty %>%
    arrange(desc(NA_qty)) %>% filter(NA_qty > 0) %>%
    mutate(Variable = as.character(Variable)) %>%
    mutate(Pct_of_Tot =  round(NA_qty/nrow(df), 4) * 100)

  return(na_qty)
}
```

```r
first_pass <- full_set %>%
  # first_pass is train.csv and test.csv combined for NA reviews
  # and imputation planning and calculated columns
  mutate(House_Age_Yrs = YrSold - YearBuilt,
         RemodAdd_Age_Yrs = YrSold - YearRemodAdd,
         Garage_Age_Yrs = YrSold - GarageYrBlt)
naVars <- na_review(first_pass %>% select(-SalePrice))
naVars


set_aside <- c(2600, 2504, 2421, 2127, 2041, 2186, 2525, 1488, 949, 2349, 2218, 2219, 333)
#View(first_pass[is.na(first_pass$PoolQC), ]) # 2600, 2504, 2421
#View(first_pass[is.na(first_pass$GarageFinish), ]) # 2127
#View(first_pass[is.na(first_pass$GarageQual), ]) # 2127
#View(first_pass[is.na(first_pass$GarageCond), ]) # 2127
#View(first_pass[is.na(first_pass$BsmtCond), ]) # 2041, 2186, 2525
#View(first_pass[is.na(first_pass$BsmtExposure), ]) # 1488, 949, 2349
#View(first_pass[is.na(first_pass$BsmtQual), ]) # 2218, 2219
#View(first_pass[is.na(first_pass$BsmtFinType2), ]) # 333
#View(first_pass[is.na(first_pass$MasVnrType), ]) #

#qty
# first_pass[first_pass$PoolArea == 0, ]     # 2,906
# first_pass[is.na(first_pass$PoolQC), ]
# first_pass[is.na(first_pass$Alley), ]      # 2,721
# first_pass[is.na(first_pass$Fence), ]      # 2,348
# first_pass[first_pass$Fireplaces == 0, ]   # 1,420
# first_pass[is.na(first_pass$GarageType),]   # 157
# first_pass[is.na(first_pass$GarageArea),]   # 1
# first_pass[is.na(first_pass$GarageFinish),] # 159
# first_pass[first_pass$GarageArea == 0, ]   # 158
# first_pass[first_pass$TotalBsmtSF == 0, ]  # 79
# first_pass[is.na(first_pass$Electrical),]   # 1
set_asideA <- '2600|2504|2421|2127|2041|2186|2525|1488|949|2349|2218|2219|333' # 13
set_asideB <- '|2550|524|2296|2593' # negative values in '_Age' columns

x <- first_pass %>%
  # exclude set_aside observations to fill in known NA's
  filter(!grepl(paste0(set_asideA, set_asideB), Id))

naVarsx <- na_review(x %>% select(-SalePrice))
naVarsx


nrow(x[x$PoolArea==0, ])   # 2,887
# x[is.na(x$MiscFeature),]   # 2,793
# x[is.na(x$Alley),]         # 2,700
# x[is.na(x$Fence),]         # 2,331
# x[is.na(x$FireplaceQu),]   # 1,414
# nrow(x[x$LotFrontage==0, ])# 486
# x[is.na(x$GarageArea),]    # 158
# x[x$TotalBsmtSF == 0, ]    # 78
obtain_data <- function(df){
  # like first_pass but with imputation that addresses
  # observations that have known NA's
  df %>%
    mutate(PoolQC = fct_explicit_na(PoolQC, na_level='NoP'),
           MiscFeature = fct_explicit_na(MiscFeature, na_level='NoM'),
           Alley = fct_explicit_na(Alley, na_level='NoA'),
           Fence = fct_explicit_na(Fence, na_level = 'NoF'),
           FireplaceQu = fct_explicit_na(FireplaceQu, na_level = 'NoFp'),
           LotFrontage = ifelse(is.na(LotFrontage), 0, LotFrontage),

           # Note GarageYrBlt set to 9999 may be a problem
           GarageYrBlt = ifelse(is.na(GarageYrBlt), 9999, GarageYrBlt),
           GarageFinish = fct_explicit_na(GarageFinish, na_level = 'NoG'),
           GarageQual = fct_explicit_na(GarageQual, na_level = 'NoG'),
           GarageCond = fct_explicit_na(GarageCond, na_level = 'NoG'),
           # NOTE: Garage_Age_Yrs: 0 doesn't seem appropriate...
           Garage_Age_Yrs = ifelse(is.na(Garage_Age_Yrs), 0, Garage_Age_Yrs),
           GarageType = fct_explicit_na(GarageType, na_level = 'NoG'),

           BsmtQual = fct_explicit_na(BsmtQual, na_level = 'NoB'),
           BsmtCond = fct_explicit_na(BsmtCond, na_level = 'NoB'),
           BsmtExposure = fct_explicit_na(BsmtExposure, na_level = 'NoB'),
           BsmtFinType1 = fct_explicit_na(BsmtFinType1, na_level = 'NoB'),
           BsmtFinType2 = fct_explicit_na(BsmtFinType2, na_level = 'NoB')
           )
}
probl_obs <- full_set %>%
  mutate(House_Age_Yrs = YrSold - YearBuilt,
         RemodAdd_Age_Yrs = YrSold - YearRemodAdd,
         Garage_Age_Yrs = YrSold - GarageYrBlt) %>%
  filter(grepl(paste0(set_asideA, set_asideB), Id))

known_obs <- full_set %>%
  filter(!grepl(paste0(set_asideA, set_asideB), Id)) %>%
  mutate(House_Age_Yrs = YrSold - YearBuilt,
         RemodAdd_Age_Yrs = YrSold - YearRemodAdd,
         Garage_Age_Yrs = YrSold - GarageYrBlt)

full_set_clean <- rbind(obtain_data(known_obs), probl_obs) %>% arrange(Id)
#View(full_set_clean)
summary(full_set_clean)
naVarsy <- na_review(full_set_clean %>% select(-SalePrice))
sum(naVarsy$NA_qty) # 176
# unique(full_set_clean$Alley) # NoA  Grvl Pave <NA>, levels: Grvl Pave NoA
# unique(full_set_clean$PoolQC) # NoP  Ex   <NA> Fa   Gd, levels: Ex Fa Gd NoP
# unique(full_set_clean$GarageYrBlt) # character!
var_types <- function(df){
  # returns df of Variable name and Type from df
  var_df <- sapply(df, class) %>% as.data.frame()
  colnames(var_df) <- c("Var_Type")
  var_df <- cbind(var_df, 'Variable' = rownames(var_df)) %>%
    select(Variable, Var_Type) %>%
    mutate(Variable = as.character(Variable),Var_Type = as.character(Var_Type))
  return(var_df)
}

var_review <- var_types(full_set_clean %>% select(-c(Id,SalePrice,d_name)))

fac_vars <- var_review %>% filter(Var_Type == 'factor') %>%
  select(Variable) %>% t() %>% as.character() # 43 total length(fac_vars)
num_vars <- var_review %>% filter(grepl('character|integer|numeric', Var_Type)) %>%
  select(Variable) %>% t() %>% as.character() # 39 total but see GarageYrBlt #length(num_vars)
sum(complete.cases(full_set %>% select(-SalePrice)))       # 0
sum(complete.cases(full_set_clean %>% select(-SalePrice))) # 2,861 ~ 98%
nrow(full_set_clean) - 2861 # 58 NA
stat_info <- psych::describe(full_set_clean %>% select(num_vars, -Id, -d_name))
stat_info[c(2:nrow(stat_info)),c(2:5,8:9,13:ncol(stat_info)-1)]
summary(full_set_clean %>% select(fac_vars, -Id, -SalePrice, -d_name))

train_data <- full_set_clean %>% filter(d_name == 'train') %>% select(-d_name)
test_data <- full_set_clean %>% filter(d_name == 'test') %>% select(-d_name)
##View(train_data)
dim(train_data)
dim(test_data)
# Data Exploration Plots
#plot_boxplot()
full_set_clean %>%
  filter(Garage_Age_Yrs < 0 | RemodAdd_Age_Yrs < 0 | Garage_Age_Yrs < 0) # Ids c(524, 2296, 2550, 2593)
hist(train_data$SalePrice)
# init = mice(first_pass, maxit=0)
# meth = init$method
# predM = init$predictorMatrix
#
# # The code below will remove the variable as a
# # predictor but still will be imputed. Just for
# # illustration purposes, I select the BMI
# # variable to not be included as predictor during
# # imputation.
# predM[, c('SalePrice')] = 0
#
# # If you want to skip a variable from imputation
# # use the code below. This variable will be
# # used for prediction.
# meth[] = ""
#
# # Now let specify the methods for imputing the
# # missing values. There are specific methods
# # for continues, binary and ordinal variables.
# # I set different methods for each variable.
# # You can add more than one variable in each method.
#
# meth[c("BsmtExposure", "BsmtFinType2", "MasVnrType",
#        "MasVnrArea", "Electrical")]="norm"
#
#
# imputed = mice(clinsurf, method=meth,
#                predictorMatrix=predM, m=5,
#                printFlag = F)
# #Create a dataset after imputation.
# imputed <- complete(imputed)
# sapply(imputed, FUN = function(x) sum(is.na(x)))
# NA Qtys
# CAR_AGE 510
# HOME_VAL 464
# YOJ 454
# INCOME 445
# AGE 6



#cleanbib()
#options("citation_format" = "pandoc")
read.bibtex(file = "bibliography.bib")
```

Wickham, Hadley. 2009. *Ggplot2*. New York, NY. http://ggplot2.org.