Data622\_Final\_Project[[1]](#footnote-20)

’Alexis Mekueko

24 May 2022

To view all the files in this project, please see [My Github site](https://github.com/asmozo24/Machine_Learning_Final_Project).

# Problem Statement

How can we predict the sale price of a house?

# Data Collection

## Data source

We found some interesting dataset from data source: <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data>

This data was originally generated the Ames Housing, a division within the Ames city in Iowa state. Thanks to Kaggle for making the data available free of use. The data is about a sale record of houses with all the defined criteria. Below is the description of the variable present in the dataset.

For a quicker look, we made the data (no modification) available at [My Github site](https://github.com/asmozo24/Machine_Learning_Final_Project/blob/main/train.csv)

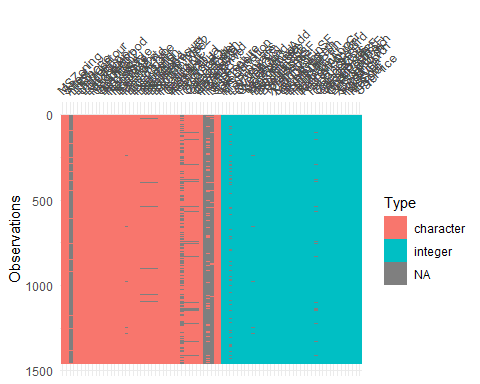
### Data Dictionary

## MSSubClass..Identifies.the.type.of.dwelling.involved.in.the.sale.  
## 1 20  
## 2 30  
## 3 40  
## 4 45  
## 5 50  
## 6 60  
## 7 70  
## 8 75  
## 9 80  
## 10 85  
## 11 90  
## 12 120  
## 13 150  
## 14 160  
## 15 180  
## 16 190  
## 17 MSZoning: Identifies the general zoning classification of the sale.  
## 18   
## 19 A  
## 20 C  
## 21 FV  
## 22 I  
## 23 RH  
## 24 RL  
## 25 RP  
## 26 RM  
## 27   
## 28 LotFrontage: Linear feet of street connected to property  
## 29 LotArea: Lot size in square feet  
## 30 Street: Type of road access to property  
## 31 Grvl  
## 32 Pave  
## 33   
## 34 Alley: Type of alley access to property  
## 35 Grvl  
## 36 Pave  
## 37 NA   
## 38   
## 39 LotShape: General shape of property  
## 40 Reg  
## 41 IR1  
## 42 IR2  
## 43 IR3  
## 44   
## 45 LandContour: Flatness of the property  
## 46 Lvl  
## 47 Bnk  
## 48 HLS  
## 49 Low  
## 50   
## 51 Utilities: Type of utilities available  
## 52   
## 53 AllPub  
## 54 NoSewr  
## 55 NoSeWa  
## 56 ELO  
## 57   
## 58 LotConfig: Lot configuration  
## 59 Inside  
## 60 Corner  
## 61 CulDSac  
## 62 FR2  
## 63 FR3  
## 64   
## 65 LandSlope: Slope of property  
## 66   
## 67 Gtl  
## 68 Mod  
## 69 Sev  
## 70   
## 71 Neighborhood: Physical locations within Ames city limits  
## 72 Blmngtn  
## 73 Blueste  
## 74 BrDale  
## 75 BrkSide  
## 76 ClearCr  
## 77 CollgCr  
## 78 Crawfor  
## 79 Edwards  
## 80 Gilbert  
## 81 IDOTRR  
## 82 MeadowV  
## 83 Mitchel  
## 84 Names  
## 85 NoRidge  
## 86 NPkVill  
## 87 NridgHt  
## 88 NWAmes  
## 89 OldTown  
## 90 SWISU  
## 91 Sawyer  
## 92 SawyerW  
## 93 Somerst  
## 94 StoneBr  
## 95 Timber  
## 96 Veenker  
## 97   
## 98   
## 99 Condition1: Proximity to various conditions  
## 100   
## 101 Artery  
## 102 Feedr  
## 103 Norm  
## 104 RRNn  
## 105 RRAn  
## 106 PosN  
## 107 PosA  
## 108 RRNe  
## 109 RRAe  
## 110   
## 111 Condition2: Proximity to various conditions (if more than one is present)  
## 112   
## 113 Artery  
## 114 Feedr  
## 115 Norm  
## 116 RRNn  
## 117 RRAn  
## 118 PosN  
## 119 PosA  
## 120 RRNe  
## 121 RRAe  
## 122   
## 123 BldgType: Type of dwelling  
## 124   
## 125 1Fam  
## 126 2FmCon  
## 127 Duplx  
## 128 TwnhsE  
## 129 TwnhsI  
## 130   
## 131 HouseStyle: Style of dwelling  
## 132   
## 133 1Story  
## 134 1.5Fin  
## 135 1.5Unf  
## 136 2Story  
## 137 2.5Fin  
## 138 2.5Unf  
## 139 SFoyer  
## 140 SLvl  
## 141   
## 142 OverallQual: Rates the overall material and finish of the house  
## 143 10  
## 144 9  
## 145 8  
## 146 7  
## 147 6  
## 148 5  
## 149 4  
## 150 3  
## 151 2  
## 152 1  
## 153   
## 154 OverallCond: Rates the overall condition of the house  
## 155 10  
## 156 9  
## 157 8  
## 158 7  
## 159 6  
## 160 5  
## 161 4  
## 162 3  
## 163 2  
## 164 1  
## 165   
## 166 YearBuilt: Original construction date  
## 167 YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)  
## 168 RoofStyle: Type of roof  
## 169 Flat  
## 170 Gable  
## 171 Gambrel  
## 172 Hip  
## 173 Mansard  
## 174 Shed  
## 175   
## 176 RoofMatl: Roof material  
## 177 ClyTile  
## 178 CompShg  
## 179 Membran  
## 180 Metal  
## 181 Roll  
## 182 Tar&Grv  
## 183 WdShake  
## 184 WdShngl  
## 185   
## 186 Exterior1st: Exterior covering on house  
## 187 AsbShng  
## 188 AsphShn  
## 189 BrkComm  
## 190 BrkFace  
## 191 CBlock  
## 192 CemntBd  
## 193 HdBoard  
## 194 ImStucc  
## 195 MetalSd  
## 196 Other  
## 197 Plywood  
## 198 PreCast  
## 199 Stone  
## 200 Stucco  
## 201 VinylSd  
## 202 Wd Sdng  
## 203 WdShing  
## 204   
## 205 Exterior2nd: Exterior covering on house (if more than one material)  
## 206 AsbShng  
## 207 AsphShn  
## 208 BrkComm  
## 209 BrkFace  
## 210 CBlock  
## 211 CemntBd  
## 212 HdBoard  
## 213 ImStucc  
## 214 MetalSd  
## 215 Other  
## 216 Plywood  
## 217 PreCast  
## 218 Stone  
## 219 Stucco  
## 220 VinylSd  
## 221 Wd Sdng  
## 222 WdShing  
## 223   
## 224 MasVnrType: Masonry veneer type  
## 225 BrkCmn  
## 226 BrkFace  
## 227 CBlock  
## 228 None  
## 229 Stone  
## 230   
## 231 MasVnrArea: Masonry veneer area in square feet  
## 232 ExterQual: Evaluates the quality of the material on the exterior   
## 233   
## 234 Ex  
## 235 Gd  
## 236 TA  
## 237 Fa  
## 238 Po  
## 239   
## 240 ExterCond: Evaluates the present condition of the material on the exterior  
## 241   
## 242 Ex  
## 243 Gd  
## 244 TA  
## 245 Fa  
## 246 Po  
## 247   
## 248 Foundation: Type of foundation  
## 249   
## 250 BrkTil  
## 251 CBlock  
## 252 PConc  
## 253 Slab  
## 254 Stone  
## 255 Wood  
## 256   
## 257 BsmtQual: Evaluates the height of the basement  
## 258 Ex  
## 259 Gd  
## 260 TA  
## 261 Fa  
## 262 Po  
## 263 NA  
## 264   
## 265 BsmtCond: Evaluates the general condition of the basement  
## 266 Ex  
## 267 Gd  
## 268 TA  
## 269 Fa  
## 270 Po  
## 271 NA  
## 272   
## 273 BsmtExposure: Refers to walkout or garden level walls  
## 274 Gd  
## 275 Av  
## 276 Mn  
## 277 No  
## 278 NA  
## 279   
## 280 BsmtFinType1: Rating of basement finished area  
## 281 GLQ  
## 282 ALQ  
## 283 BLQ  
## 284 Rec  
## 285 LwQ  
## 286 Unf  
## 287 NA  
## 288   
## 289 BsmtFinSF1: Type 1 finished square feet  
## 290 BsmtFinType2: Rating of basement finished area (if multiple types)  
## 291 GLQ  
## 292 ALQ  
## 293 BLQ  
## 294 Rec  
## 295 LwQ  
## 296 Unf  
## 297 NA  
## 298 BsmtFinSF2: Type 2 finished square feet  
## 299 BsmtUnfSF: Unfinished square feet of basement area  
## 300 TotalBsmtSF: Total square feet of basement area  
## 301 Heating: Type of heating  
## 302   
## 303 Floor  
## 304 GasA  
## 305 GasW  
## 306 Grav  
## 307 OthW  
## 308 Wall  
## 309   
## 310 HeatingQC: Heating quality and condition  
## 311 Ex  
## 312 Gd  
## 313 TA  
## 314 Fa  
## 315 Po  
## 316   
## 317 CentralAir: Central air conditioning  
## 318 N  
## 319 Y  
## 320   
## 321 Electrical: Electrical system  
## 322 SBrkr  
## 323 FuseA  
## 324 FuseF  
## 325 FuseP  
## 326 Mix  
## 327   
## 328 1stFlrSF: First Floor square feet  
## 329   
## 330 2ndFlrSF: Second floor square feet  
## 331 LowQualFinSF: Low quality finished square feet (all floors)  
## 332 GrLivArea: Above grade (ground) living area square feet  
## 333 BsmtFullBath: Basement full bathrooms  
## 334 BsmtHalfBath: Basement half bathrooms  
## 335 FullBath: Full bathrooms above grade  
## 336 HalfBath: Half baths above grade  
## 337 Bedroom: Bedrooms above grade (does NOT include basement bedrooms)  
## 338 Kitchen: Kitchens above grade  
## 339 KitchenQual: Kitchen quality  
## 340 Ex  
## 341 Gd  
## 342 TA  
## 343 Fa  
## 344 Po  
## 345   
## 346 TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)  
## 347 Functional: Home functionality (Assume typical unless deductions are warranted)  
## 348 Typ  
## 349 Min1  
## 350 Min2  
## 351 Mod  
## 352 Maj1  
## 353 Maj2  
## 354 Sev  
## 355 Sal  
## 356   
## 357 Fireplaces: Number of fireplaces  
## 358 FireplaceQu: Fireplace quality  
## 359 Ex  
## 360 Gd  
## 361 TA  
## 362 Fa  
## 363 Po  
## 364 NA  
## 365   
## 366 GarageType: Garage location  
## 367   
## 368 2Types  
## 369 Attchd  
## 370 Basment  
## 371 BuiltIn  
## 372 CarPort  
## 373 Detchd  
## 374 NA  
## 375   
## 376 GarageYrBlt: Year garage was built  
## 377   
## 378 GarageFinish: Interior finish of the garage  
## 379 Fin  
## 380 RFn  
## 381 Unf  
## 382 NA  
## 383   
## 384 GarageCars: Size of garage in car capacity  
## 385 GarageArea: Size of garage in square feet  
## 386 GarageQual: Garage quality  
## 387 Ex  
## 388 Gd  
## 389 TA  
## 390 Fa  
## 391 Po  
## 392 NA  
## 393   
## 394 GarageCond: Garage condition  
## 395 Ex  
## 396 Gd  
## 397 TA  
## 398 Fa  
## 399 Po  
## 400 NA  
## 401   
## 402 PavedDrive: Paved driveway  
## 403 Y  
## 404 P  
## 405 N  
## 406   
## 407 WoodDeckSF: Wood deck area in square feet  
## 408 OpenPorchSF: Open porch area in square feet  
## 409 EnclosedPorch: Enclosed porch area in square feet  
## 410 3SsnPorch: Three season porch area in square feet  
## 411 ScreenPorch: Screen porch area in square feet  
## 412 PoolArea: Pool area in square feet  
## 413 PoolQC: Pool quality  
## 414   
## 415 Ex  
## 416 Gd  
## 417 TA  
## 418 Fa  
## 419 NA  
## 420   
## 421 Fence: Fence quality  
## 422   
## 423 GdPrv  
## 424 MnPrv  
## 425 GdWo  
## 426 MnWw  
## 427 NA  
## 428   
## 429 MiscFeature: Miscellaneous feature not covered in other categories  
## 430   
## 431 Elev  
## 432 Gar2  
## 433 Othr  
## 434 Shed  
## 435 TenC  
## 436 NA  
## 437   
## 438 MiscVal: $Value of miscellaneous feature  
## 439 MoSold: Month Sold (MM)  
## 440 YrSold: Year Sold (YYYY)  
## 441 SaleType: Type of sale  
## 442   
## 443 WD   
## 444 CWD  
## 445 VWD  
## 446 New  
## 447 COD  
## 448 Con  
## 449 ConLw  
## 450 ConLI  
## 451 ConLD  
## 452 Oth  
## 453   
## 454 SaleCondition: Condition of sale  
## 455 Normal  
## 456 Abnorml  
## 457 AdjLand  
## 458 Alloca  
## 459 Family  
## 460 Partial  
## X  
## 1 1-STORY 1946 & NEWER ALL STYLES  
## 2 1-STORY 1945 & OLDER  
## 3 1-STORY W/FINISHED ATTIC ALL AGES  
## 4 1-1/2 STORY - UNFINISHED ALL AGES  
## 5 1-1/2 STORY FINISHED ALL AGES  
## 6 2-STORY 1946 & NEWER  
## 7 2-STORY 1945 & OLDER  
## 8 2-1/2 STORY ALL AGES  
## 9 SPLIT OR MULTI-LEVEL  
## 10 SPLIT FOYER  
## 11 DUPLEX - ALL STYLES AND AGES  
## 12 1-STORY PUD (Planned Unit Development) - 1946 & NEWER  
## 13 1-1/2 STORY PUD - ALL AGES  
## 14 2-STORY PUD - 1946 & NEWER  
## 15 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER  
## 16 2 FAMILY CONVERSION - ALL STYLES AND AGES  
## 17   
## 18   
## 19 Agriculture  
## 20 Commercial  
## 21 Floating Village Residential  
## 22 Industrial  
## 23 Residential High Density  
## 24 Residential Low Density  
## 25 Residential Low Density Park   
## 26 Residential Medium Density  
## 27   
## 28   
## 29   
## 30   
## 31 Gravel  
## 32 Paved  
## 33   
## 34   
## 35 Gravel  
## 36 Paved  
## 37 No alley access  
## 38   
## 39   
## 40 Regular  
## 41 Slightly irregular  
## 42 Moderately Irregular  
## 43 Irregular  
## 44   
## 45   
## 46 Near Flat/Level  
## 47 Banked - Quick and significant rise from street grade to building  
## 48 Hillside - Significant slope from side to side  
## 49 Depression  
## 50   
## 51   
## 52   
## 53 All public Utilities (E,G,W,& S)  
## 54 Electricity, Gas, and Water (Septic Tank)  
## 55 Electricity and Gas Only  
## 56 Electricity only  
## 57   
## 58   
## 59 Inside lot  
## 60 Corner lot  
## 61 Cul-de-sac  
## 62 Frontage on 2 sides of property  
## 63 Frontage on 3 sides of property  
## 64   
## 65   
## 66   
## 67 Gentle slope  
## 68 Moderate Slope  
## 69 Severe Slope  
## 70   
## 71   
## 72 Bloomington Heights  
## 73 Bluestem  
## 74 Briardale  
## 75 Brookside  
## 76 Clear Creek  
## 77 College Creek  
## 78 Crawford  
## 79 Edwards  
## 80 Gilbert  
## 81 Iowa DOT and Rail Road  
## 82 Meadow Village  
## 83 Mitchell  
## 84 North Ames  
## 85 Northridge  
## 86 Northpark Villa  
## 87 Northridge Heights  
## 88 Northwest Ames  
## 89 Old Town  
## 90 South & West of Iowa State University  
## 91 Sawyer  
## 92 Sawyer West  
## 93 Somerset  
## 94 Stone Brook  
## 95 Timberland  
## 96 Veenker  
## 97   
## 98   
## 99   
## 100   
## 101 Adjacent to arterial street  
## 102 Adjacent to feeder street  
## 103 Normal  
## 104 Within 200' of North-South Railroad  
## 105 Adjacent to North-South Railroad  
## 106 Near positive off-site feature--park, greenbelt, etc.  
## 107 Adjacent to postive off-site feature  
## 108 Within 200' of East-West Railroad  
## 109 Adjacent to East-West Railroad  
## 110   
## 111   
## 112   
## 113 Adjacent to arterial street  
## 114 Adjacent to feeder street  
## 115 Normal  
## 116 Within 200' of North-South Railroad  
## 117 Adjacent to North-South Railroad  
## 118 Near positive off-site feature--park, greenbelt, etc.  
## 119 Adjacent to postive off-site feature  
## 120 Within 200' of East-West Railroad  
## 121 Adjacent to East-West Railroad  
## 122   
## 123   
## 124   
## 125 Single-family Detached  
## 126 Two-family Conversion; originally built as one-family dwelling  
## 127 Duplex  
## 128 Townhouse End Unit  
## 129 Townhouse Inside Unit  
## 130   
## 131   
## 132   
## 133 One story  
## 134 One and one-half story: 2nd level finished  
## 135 One and one-half story: 2nd level unfinished  
## 136 Two story  
## 137 Two and one-half story: 2nd level finished  
## 138 Two and one-half story: 2nd level unfinished  
## 139 Split Foyer  
## 140 Split Level  
## 141   
## 142   
## 143 Very Excellent  
## 144 Excellent  
## 145 Very Good  
## 146 Good  
## 147 Above Average  
## 148 Average  
## 149 Below Average  
## 150 Fair  
## 151 Poor  
## 152 Very Poor  
## 153   
## 154   
## 155 Very Excellent  
## 156 Excellent  
## 157 Very Good  
## 158 Good  
## 159 Above Average  
## 160 Average  
## 161 Below Average  
## 162 Fair  
## 163 Poor  
## 164 Very Poor  
## 165   
## 166   
## 167   
## 168   
## 169 Flat  
## 170 Gable  
## 171 Gabrel (Barn)  
## 172 Hip  
## 173 Mansard  
## 174 Shed  
## 175   
## 176   
## 177 Clay or Tile  
## 178 Standard (Composite) Shingle  
## 179 Membrane  
## 180 Metal  
## 181 Roll  
## 182 Gravel & Tar  
## 183 Wood Shakes  
## 184 Wood Shingles  
## 185   
## 186   
## 187 Asbestos Shingles  
## 188 Asphalt Shingles  
## 189 Brick Common  
## 190 Brick Face  
## 191 Cinder Block  
## 192 Cement Board  
## 193 Hard Board  
## 194 Imitation Stucco  
## 195 Metal Siding  
## 196 Other  
## 197 Plywood  
## 198 PreCast  
## 199 Stone  
## 200 Stucco  
## 201 Vinyl Siding  
## 202 Wood Siding  
## 203 Wood Shingles  
## 204   
## 205   
## 206 Asbestos Shingles  
## 207 Asphalt Shingles  
## 208 Brick Common  
## 209 Brick Face  
## 210 Cinder Block  
## 211 Cement Board  
## 212 Hard Board  
## 213 Imitation Stucco  
## 214 Metal Siding  
## 215 Other  
## 216 Plywood  
## 217 PreCast  
## 218 Stone  
## 219 Stucco  
## 220 Vinyl Siding  
## 221 Wood Siding  
## 222 Wood Shingles  
## 223   
## 224   
## 225 Brick Common  
## 226 Brick Face  
## 227 Cinder Block  
## 228 None  
## 229 Stone  
## 230   
## 231   
## 232   
## 233   
## 234 Excellent  
## 235 Good  
## 236 Average/Typical  
## 237 Fair  
## 238 Poor  
## 239   
## 240   
## 241   
## 242 Excellent  
## 243 Good  
## 244 Average/Typical  
## 245 Fair  
## 246 Poor  
## 247   
## 248   
## 249   
## 250 Brick & Tile  
## 251 Cinder Block  
## 252 Poured Contrete  
## 253 Slab  
## 254 Stone  
## 255 Wood  
## 256   
## 257   
## 258 Excellent (100+ inches)  
## 259 Good (90-99 inches)  
## 260 Typical (80-89 inches)  
## 261 Fair (70-79 inches)  
## 262 Poor (<70 inches  
## 263 No Basement  
## 264   
## 265   
## 266 Excellent  
## 267 Good  
## 268 Typical - slight dampness allowed  
## 269 Fair - dampness or some cracking or settling  
## 270 Poor - Severe cracking, settling, or wetness  
## 271 No Basement  
## 272   
## 273   
## 274 Good Exposure  
## 275 Average Exposure (split levels or foyers typically score average or above)  
## 276 Mimimum Exposure  
## 277 No Exposure  
## 278 No Basement  
## 279   
## 280   
## 281 Good Living Quarters  
## 282 Average Living Quarters  
## 283 Below Average Living Quarters  
## 284 Average Rec Room  
## 285 Low Quality  
## 286 Unfinshed  
## 287 No Basement  
## 288   
## 289   
## 290   
## 291 Good Living Quarters  
## 292 Average Living Quarters  
## 293 Below Average Living Quarters  
## 294 Average Rec Room  
## 295 Low Quality  
## 296 Unfinshed  
## 297 No Basement  
## 298   
## 299   
## 300   
## 301   
## 302   
## 303 Floor Furnace  
## 304 Gas forced warm air furnace  
## 305 Gas hot water or steam heat  
## 306 Gravity furnace  
## 307 Hot water or steam heat other than gas  
## 308 Wall furnace  
## 309   
## 310   
## 311 Excellent  
## 312 Good  
## 313 Average/Typical  
## 314 Fair  
## 315 Poor  
## 316   
## 317   
## 318 No  
## 319 Yes  
## 320   
## 321   
## 322 Standard Circuit Breakers & Romex  
## 323 Fuse Box over 60 AMP and all Romex wiring (Average)  
## 324 60 AMP Fuse Box and mostly Romex wiring (Fair)  
## 325 60 AMP Fuse Box and mostly knob & tube wiring (poor)  
## 326 Mixed  
## 327   
## 328   
## 329   
## 330   
## 331   
## 332   
## 333   
## 334   
## 335   
## 336   
## 337   
## 338   
## 339   
## 340 Excellent  
## 341 Good  
## 342 Typical/Average  
## 343 Fair  
## 344 Poor  
## 345   
## 346   
## 347   
## 348 Typical Functionality  
## 349 Minor Deductions 1  
## 350 Minor Deductions 2  
## 351 Moderate Deductions  
## 352 Major Deductions 1  
## 353 Major Deductions 2  
## 354 Severely Damaged  
## 355 Salvage only  
## 356   
## 357   
## 358   
## 359 Excellent - Exceptional Masonry Fireplace  
## 360 Good - Masonry Fireplace in main level  
## 361 Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement  
## 362 Fair - Prefabricated Fireplace in basement  
## 363 Poor - Ben Franklin Stove  
## 364 No Fireplace  
## 365   
## 366   
## 367   
## 368 More than one type of garage  
## 369 Attached to home  
## 370 Basement Garage  
## 371 Built-In (Garage part of house - typically has room above garage)  
## 372 Car Port  
## 373 Detached from home  
## 374 No Garage  
## 375   
## 376   
## 377   
## 378   
## 379 Finished  
## 380 Rough Finished  
## 381 Unfinished  
## 382 No Garage  
## 383   
## 384   
## 385   
## 386   
## 387 Excellent  
## 388 Good  
## 389 Typical/Average  
## 390 Fair  
## 391 Poor  
## 392 No Garage  
## 393   
## 394   
## 395 Excellent  
## 396 Good  
## 397 Typical/Average  
## 398 Fair  
## 399 Poor  
## 400 No Garage  
## 401   
## 402   
## 403 Paved   
## 404 Partial Pavement  
## 405 Dirt/Gravel  
## 406   
## 407   
## 408   
## 409   
## 410   
## 411   
## 412   
## 413   
## 414   
## 415 Excellent  
## 416 Good  
## 417 Average/Typical  
## 418 Fair  
## 419 No Pool  
## 420   
## 421   
## 422   
## 423 Good Privacy  
## 424 Minimum Privacy  
## 425 Good Wood  
## 426 Minimum Wood/Wire  
## 427 No Fence  
## 428   
## 429   
## 430   
## 431 Elevator  
## 432 2nd Garage (if not described in garage section)  
## 433 Other  
## 434 Shed (over 100 SF)  
## 435 Tennis Court  
## 436 None  
## 437   
## 438   
## 439   
## 440   
## 441   
## 442   
## 443 Warranty Deed - Conventional  
## 444 Warranty Deed - Cash  
## 445 Warranty Deed - VA Loan  
## 446 Home just constructed and sold  
## 447 Court Officer Deed/Estate  
## 448 Contract 15% Down payment regular terms  
## 449 Contract Low Down payment and low interest  
## 450 Contract Low Interest  
## 451 Contract Low Down  
## 452 Other  
## 453   
## 454   
## 455 Normal Sale  
## 456 Abnormal Sale - trade, foreclosure, short sale  
## 457 Adjoining Land Purchase  
## 458 Allocation - two linked properties with separate deeds, typically condo with a garage unit  
## 459 Sale between family members  
## 460 Home was not completed when last assessed (associated with New Homes)

## Data Overview/Charateristic

## 'data.frame': 1460 obs. of 81 variables:  
## $ Id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ MSSubClass : int 60 20 60 70 60 50 20 60 50 190 ...  
## $ MSZoning : chr "RL" "RL" "RL" "RL" ...  
## $ LotFrontage : int 65 80 68 60 84 85 75 NA 51 50 ...  
## $ LotArea : int 8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...  
## $ Street : chr "Pave" "Pave" "Pave" "Pave" ...  
## $ Alley : chr NA NA NA NA ...  
## $ LotShape : chr "Reg" "Reg" "IR1" "IR1" ...  
## $ LandContour : chr "Lvl" "Lvl" "Lvl" "Lvl" ...  
## $ Utilities : chr "AllPub" "AllPub" "AllPub" "AllPub" ...  
## $ LotConfig : chr "Inside" "FR2" "Inside" "Corner" ...  
## $ LandSlope : chr "Gtl" "Gtl" "Gtl" "Gtl" ...  
## $ Neighborhood : chr "CollgCr" "Veenker" "CollgCr" "Crawfor" ...  
## $ Condition1 : chr "Norm" "Feedr" "Norm" "Norm" ...  
## $ Condition2 : chr "Norm" "Norm" "Norm" "Norm" ...  
## $ BldgType : chr "1Fam" "1Fam" "1Fam" "1Fam" ...  
## $ HouseStyle : chr "2Story" "1Story" "2Story" "2Story" ...  
## $ OverallQual : int 7 6 7 7 8 5 8 7 7 5 ...  
## $ OverallCond : int 5 8 5 5 5 5 5 6 5 6 ...  
## $ YearBuilt : int 2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...  
## $ YearRemodAdd : int 2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...  
## $ RoofStyle : chr "Gable" "Gable" "Gable" "Gable" ...  
## $ RoofMatl : chr "CompShg" "CompShg" "CompShg" "CompShg" ...  
## $ Exterior1st : chr "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...  
## $ Exterior2nd : chr "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...  
## $ MasVnrType : chr "BrkFace" "None" "BrkFace" "None" ...  
## $ MasVnrArea : int 196 0 162 0 350 0 186 240 0 0 ...  
## $ ExterQual : chr "Gd" "TA" "Gd" "TA" ...  
## $ ExterCond : chr "TA" "TA" "TA" "TA" ...  
## $ Foundation : chr "PConc" "CBlock" "PConc" "BrkTil" ...  
## $ BsmtQual : chr "Gd" "Gd" "Gd" "TA" ...  
## $ BsmtCond : chr "TA" "TA" "TA" "Gd" ...  
## $ BsmtExposure : chr "No" "Gd" "Mn" "No" ...  
## $ BsmtFinType1 : chr "GLQ" "ALQ" "GLQ" "ALQ" ...  
## $ BsmtFinSF1 : int 706 978 486 216 655 732 1369 859 0 851 ...  
## $ BsmtFinType2 : chr "Unf" "Unf" "Unf" "Unf" ...  
## $ BsmtFinSF2 : int 0 0 0 0 0 0 0 32 0 0 ...  
## $ BsmtUnfSF : int 150 284 434 540 490 64 317 216 952 140 ...  
## $ TotalBsmtSF : int 856 1262 920 756 1145 796 1686 1107 952 991 ...  
## $ Heating : chr "GasA" "GasA" "GasA" "GasA" ...  
## $ HeatingQC : chr "Ex" "Ex" "Ex" "Gd" ...  
## $ CentralAir : chr "Y" "Y" "Y" "Y" ...  
## $ Electrical : chr "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...  
## $ X1stFlrSF : int 856 1262 920 961 1145 796 1694 1107 1022 1077 ...  
## $ X2ndFlrSF : int 854 0 866 756 1053 566 0 983 752 0 ...  
## $ LowQualFinSF : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ GrLivArea : int 1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...  
## $ BsmtFullBath : int 1 0 1 1 1 1 1 1 0 1 ...  
## $ BsmtHalfBath : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ FullBath : int 2 2 2 1 2 1 2 2 2 1 ...  
## $ HalfBath : int 1 0 1 0 1 1 0 1 0 0 ...  
## $ BedroomAbvGr : int 3 3 3 3 4 1 3 3 2 2 ...  
## $ KitchenAbvGr : int 1 1 1 1 1 1 1 1 2 2 ...  
## $ KitchenQual : chr "Gd" "TA" "Gd" "Gd" ...  
## $ TotRmsAbvGrd : int 8 6 6 7 9 5 7 7 8 5 ...  
## $ Functional : chr "Typ" "Typ" "Typ" "Typ" ...  
## $ Fireplaces : int 0 1 1 1 1 0 1 2 2 2 ...  
## $ FireplaceQu : chr NA "TA" "TA" "Gd" ...  
## $ GarageType : chr "Attchd" "Attchd" "Attchd" "Detchd" ...  
## $ GarageYrBlt : int 2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...  
## $ GarageFinish : chr "RFn" "RFn" "RFn" "Unf" ...  
## $ GarageCars : int 2 2 2 3 3 2 2 2 2 1 ...  
## $ GarageArea : int 548 460 608 642 836 480 636 484 468 205 ...  
## $ GarageQual : chr "TA" "TA" "TA" "TA" ...  
## $ GarageCond : chr "TA" "TA" "TA" "TA" ...  
## $ PavedDrive : chr "Y" "Y" "Y" "Y" ...  
## $ WoodDeckSF : int 0 298 0 0 192 40 255 235 90 0 ...  
## $ OpenPorchSF : int 61 0 42 35 84 30 57 204 0 4 ...  
## $ EnclosedPorch: int 0 0 0 272 0 0 0 228 205 0 ...  
## $ X3SsnPorch : int 0 0 0 0 0 320 0 0 0 0 ...  
## $ ScreenPorch : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ PoolArea : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ PoolQC : chr NA NA NA NA ...  
## $ Fence : chr NA NA NA NA ...  
## $ MiscFeature : chr NA NA NA NA ...  
## $ MiscVal : int 0 0 0 0 0 700 0 350 0 0 ...  
## $ MoSold : int 2 5 9 2 12 10 8 11 4 1 ...  
## $ YrSold : int 2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...  
## $ SaleType : chr "WD" "WD" "WD" "WD" ...  
## $ SaleCondition: chr "Normal" "Normal" "Normal" "Abnorml" ...  
## $ SalePrice : int 208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...

## Warning: `gather\_()` was deprecated in tidyr 1.2.0.  
## Please use `gather()` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was generated.



This data is composed of 80 features for 1460 observations. The datatypes are mixed of integer and character. The target variable or the variable of interest is ‘SalePrice’. According to the dictionary, ‘SalePrice’ is the price at which a house was sold. Other variables are criteria used to derive the price of the house. We believe 81 variables for a price determination is excessive. Not all these variable are relevant/pertinent to the target variable. In addition, some variables have missing values. Therefore, we will need to reduce the number of feature by pertinence to SalePrice and cleaning the missing values.

# Data Cleaning

Let’s see the missing values.

##   
## Attaching package: 'skimr'

## The following object is masked from 'package:naniar':  
##   
## n\_complete

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:arules':  
##   
## intersect, setdiff, union

## The following object is masked from 'package:tsibble':  
##   
## interval

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

Data summary

|  |  |
| --- | --- |
| Name | df1 |
| Number of rows | 1460 |
| Number of columns | 81 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 43 |
| numeric | 38 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

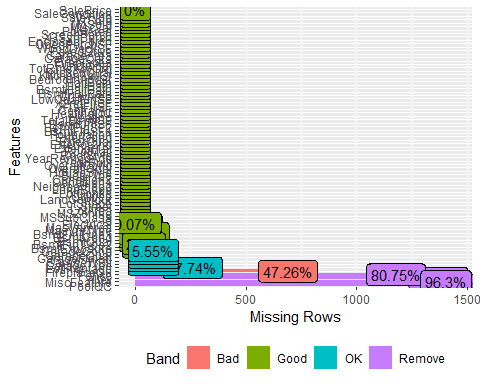
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| MSZoning | 0 | 1.00 | 2 | 7 | 0 | 5 | 0 |
| Street | 0 | 1.00 | 4 | 4 | 0 | 2 | 0 |
| Alley | 1369 | 0.06 | 4 | 4 | 0 | 2 | 0 |
| LotShape | 0 | 1.00 | 3 | 3 | 0 | 4 | 0 |
| LandContour | 0 | 1.00 | 3 | 3 | 0 | 4 | 0 |
| Utilities | 0 | 1.00 | 6 | 6 | 0 | 2 | 0 |
| LotConfig | 0 | 1.00 | 3 | 7 | 0 | 5 | 0 |
| LandSlope | 0 | 1.00 | 3 | 3 | 0 | 3 | 0 |
| Neighborhood | 0 | 1.00 | 5 | 7 | 0 | 25 | 0 |
| Condition1 | 0 | 1.00 | 4 | 6 | 0 | 9 | 0 |
| Condition2 | 0 | 1.00 | 4 | 6 | 0 | 8 | 0 |
| BldgType | 0 | 1.00 | 4 | 6 | 0 | 5 | 0 |
| HouseStyle | 0 | 1.00 | 4 | 6 | 0 | 8 | 0 |
| RoofStyle | 0 | 1.00 | 3 | 7 | 0 | 6 | 0 |
| RoofMatl | 0 | 1.00 | 4 | 7 | 0 | 8 | 0 |
| Exterior1st | 0 | 1.00 | 5 | 7 | 0 | 15 | 0 |
| Exterior2nd | 0 | 1.00 | 5 | 7 | 0 | 16 | 0 |
| MasVnrType | 8 | 0.99 | 4 | 7 | 0 | 4 | 0 |
| ExterQual | 0 | 1.00 | 2 | 2 | 0 | 4 | 0 |
| ExterCond | 0 | 1.00 | 2 | 2 | 0 | 5 | 0 |
| Foundation | 0 | 1.00 | 4 | 6 | 0 | 6 | 0 |
| BsmtQual | 37 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtCond | 37 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtExposure | 38 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtFinType1 | 37 | 0.97 | 3 | 3 | 0 | 6 | 0 |
| BsmtFinType2 | 38 | 0.97 | 3 | 3 | 0 | 6 | 0 |
| Heating | 0 | 1.00 | 4 | 5 | 0 | 6 | 0 |
| HeatingQC | 0 | 1.00 | 2 | 2 | 0 | 5 | 0 |
| CentralAir | 0 | 1.00 | 1 | 1 | 0 | 2 | 0 |
| Electrical | 1 | 1.00 | 3 | 5 | 0 | 5 | 0 |
| KitchenQual | 0 | 1.00 | 2 | 2 | 0 | 4 | 0 |
| Functional | 0 | 1.00 | 3 | 4 | 0 | 7 | 0 |
| FireplaceQu | 690 | 0.53 | 2 | 2 | 0 | 5 | 0 |
| GarageType | 81 | 0.94 | 6 | 7 | 0 | 6 | 0 |
| GarageFinish | 81 | 0.94 | 3 | 3 | 0 | 3 | 0 |
| GarageQual | 81 | 0.94 | 2 | 2 | 0 | 5 | 0 |
| GarageCond | 81 | 0.94 | 2 | 2 | 0 | 5 | 0 |
| PavedDrive | 0 | 1.00 | 1 | 1 | 0 | 3 | 0 |
| PoolQC | 1453 | 0.00 | 2 | 2 | 0 | 3 | 0 |
| Fence | 1179 | 0.19 | 4 | 5 | 0 | 4 | 0 |
| MiscFeature | 1406 | 0.04 | 4 | 4 | 0 | 4 | 0 |
| SaleType | 0 | 1.00 | 2 | 5 | 0 | 9 | 0 |
| SaleCondition | 0 | 1.00 | 6 | 7 | 0 | 6 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1.00 | 730.50 | 421.61 | 1 | 365.75 | 730.5 | 1095.25 | 1460 | ▇▇▇▇▇ |
| MSSubClass | 0 | 1.00 | 56.90 | 42.30 | 20 | 20.00 | 50.0 | 70.00 | 190 | ▇▅▂▁▁ |
| LotFrontage | 259 | 0.82 | 70.05 | 24.28 | 21 | 59.00 | 69.0 | 80.00 | 313 | ▇▃▁▁▁ |
| LotArea | 0 | 1.00 | 10516.83 | 9981.26 | 1300 | 7553.50 | 9478.5 | 11601.50 | 215245 | ▇▁▁▁▁ |
| OverallQual | 0 | 1.00 | 6.10 | 1.38 | 1 | 5.00 | 6.0 | 7.00 | 10 | ▁▂▇▅▁ |
| OverallCond | 0 | 1.00 | 5.58 | 1.11 | 1 | 5.00 | 5.0 | 6.00 | 9 | ▁▁▇▅▁ |
| YearBuilt | 0 | 1.00 | 1971.27 | 30.20 | 1872 | 1954.00 | 1973.0 | 2000.00 | 2010 | ▁▂▃▆▇ |
| YearRemodAdd | 0 | 1.00 | 1984.87 | 20.65 | 1950 | 1967.00 | 1994.0 | 2004.00 | 2010 | ▅▂▂▃▇ |
| MasVnrArea | 8 | 0.99 | 103.69 | 181.07 | 0 | 0.00 | 0.0 | 166.00 | 1600 | ▇▁▁▁▁ |
| BsmtFinSF1 | 0 | 1.00 | 443.64 | 456.10 | 0 | 0.00 | 383.5 | 712.25 | 5644 | ▇▁▁▁▁ |
| BsmtFinSF2 | 0 | 1.00 | 46.55 | 161.32 | 0 | 0.00 | 0.0 | 0.00 | 1474 | ▇▁▁▁▁ |
| BsmtUnfSF | 0 | 1.00 | 567.24 | 441.87 | 0 | 223.00 | 477.5 | 808.00 | 2336 | ▇▅▂▁▁ |
| TotalBsmtSF | 0 | 1.00 | 1057.43 | 438.71 | 0 | 795.75 | 991.5 | 1298.25 | 6110 | ▇▃▁▁▁ |
| X1stFlrSF | 0 | 1.00 | 1162.63 | 386.59 | 334 | 882.00 | 1087.0 | 1391.25 | 4692 | ▇▅▁▁▁ |
| X2ndFlrSF | 0 | 1.00 | 346.99 | 436.53 | 0 | 0.00 | 0.0 | 728.00 | 2065 | ▇▃▂▁▁ |
| LowQualFinSF | 0 | 1.00 | 5.84 | 48.62 | 0 | 0.00 | 0.0 | 0.00 | 572 | ▇▁▁▁▁ |
| GrLivArea | 0 | 1.00 | 1515.46 | 525.48 | 334 | 1129.50 | 1464.0 | 1776.75 | 5642 | ▇▇▁▁▁ |
| BsmtFullBath | 0 | 1.00 | 0.43 | 0.52 | 0 | 0.00 | 0.0 | 1.00 | 3 | ▇▆▁▁▁ |
| BsmtHalfBath | 0 | 1.00 | 0.06 | 0.24 | 0 | 0.00 | 0.0 | 0.00 | 2 | ▇▁▁▁▁ |
| FullBath | 0 | 1.00 | 1.57 | 0.55 | 0 | 1.00 | 2.0 | 2.00 | 3 | ▁▇▁▇▁ |
| HalfBath | 0 | 1.00 | 0.38 | 0.50 | 0 | 0.00 | 0.0 | 1.00 | 2 | ▇▁▅▁▁ |
| BedroomAbvGr | 0 | 1.00 | 2.87 | 0.82 | 0 | 2.00 | 3.0 | 3.00 | 8 | ▁▇▂▁▁ |
| KitchenAbvGr | 0 | 1.00 | 1.05 | 0.22 | 0 | 1.00 | 1.0 | 1.00 | 3 | ▁▇▁▁▁ |
| TotRmsAbvGrd | 0 | 1.00 | 6.52 | 1.63 | 2 | 5.00 | 6.0 | 7.00 | 14 | ▂▇▇▁▁ |
| Fireplaces | 0 | 1.00 | 0.61 | 0.64 | 0 | 0.00 | 1.0 | 1.00 | 3 | ▇▇▁▁▁ |
| GarageYrBlt | 81 | 0.94 | 1978.51 | 24.69 | 1900 | 1961.00 | 1980.0 | 2002.00 | 2010 | ▁▁▅▅▇ |
| GarageCars | 0 | 1.00 | 1.77 | 0.75 | 0 | 1.00 | 2.0 | 2.00 | 4 | ▁▃▇▂▁ |
| GarageArea | 0 | 1.00 | 472.98 | 213.80 | 0 | 334.50 | 480.0 | 576.00 | 1418 | ▂▇▃▁▁ |
| WoodDeckSF | 0 | 1.00 | 94.24 | 125.34 | 0 | 0.00 | 0.0 | 168.00 | 857 | ▇▂▁▁▁ |
| OpenPorchSF | 0 | 1.00 | 46.66 | 66.26 | 0 | 0.00 | 25.0 | 68.00 | 547 | ▇▁▁▁▁ |
| EnclosedPorch | 0 | 1.00 | 21.95 | 61.12 | 0 | 0.00 | 0.0 | 0.00 | 552 | ▇▁▁▁▁ |
| X3SsnPorch | 0 | 1.00 | 3.41 | 29.32 | 0 | 0.00 | 0.0 | 0.00 | 508 | ▇▁▁▁▁ |
| ScreenPorch | 0 | 1.00 | 15.06 | 55.76 | 0 | 0.00 | 0.0 | 0.00 | 480 | ▇▁▁▁▁ |
| PoolArea | 0 | 1.00 | 2.76 | 40.18 | 0 | 0.00 | 0.0 | 0.00 | 738 | ▇▁▁▁▁ |
| MiscVal | 0 | 1.00 | 43.49 | 496.12 | 0 | 0.00 | 0.0 | 0.00 | 15500 | ▇▁▁▁▁ |
| MoSold | 0 | 1.00 | 6.32 | 2.70 | 1 | 5.00 | 6.0 | 8.00 | 12 | ▃▆▇▃▃ |
| YrSold | 0 | 1.00 | 2007.82 | 1.33 | 2006 | 2007.00 | 2008.0 | 2009.00 | 2010 | ▇▇▇▇▅ |
| SalePrice | 0 | 1.00 | 180921.20 | 79442.50 | 34900 | 129975.00 | 163000.0 | 214000.00 | 755000 | ▇▅▁▁▁ |

There are 43 character(categorical) variables and 38 numerical variables. We know there are 1460 observations, so seeing the number (or percent) missing values per variable can give us an idea of the cleaning approach.

## Id MSSubClass MSZoning LotFrontage LotArea   
## 0 0 0 259 0   
## Street Alley LotShape LandContour Utilities   
## 0 1369 0 0 0   
## LotConfig LandSlope Neighborhood Condition1 Condition2   
## 0 0 0 0 0   
## BldgType HouseStyle OverallQual OverallCond YearBuilt   
## 0 0 0 0 0   
## YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd   
## 0 0 0 0 0   
## MasVnrType MasVnrArea ExterQual ExterCond Foundation   
## 8 8 0 0 0   
## BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1   
## 37 37 38 37 0   
## BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating   
## 38 0 0 0 0   
## HeatingQC CentralAir Electrical X1stFlrSF X2ndFlrSF   
## 0 0 1 0 0   
## LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath   
## 0 0 0 0 0   
## HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd   
## 0 0 0 0 0   
## Functional Fireplaces FireplaceQu GarageType GarageYrBlt   
## 0 0 690 81 81   
## GarageFinish GarageCars GarageArea GarageQual GarageCond   
## 81 0 0 81 81   
## PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch   
## 0 0 0 0 0   
## ScreenPorch PoolArea PoolQC Fence MiscFeature   
## 0 0 1453 1179 1406   
## MiscVal MoSold YrSold SaleType SaleCondition   
## 0 0 0 0 0   
## SalePrice   
## 0



By visual inspection, we can remove these variables: PoolQC, Fence, Alley, MiscFeature, FireplaceQu, LotFrontage. because missing too many value…it probably means these variables are not important … houses that have these variables with values are probably extra..extra…or too special…

LotFrontage = LotFrontage: Linear feet of street connected to property. We want to remove this variable due to the 17.4% missing values and the fact that every house is built under the city regulation.

We will remove variable ‘Id’. There is no need for it or because it is just an indication for record purpose.

## LotFrontage Alley MasVnrType MasVnrArea BsmtQual BsmtCond   
## 259 1369 8 8 37 37   
## BsmtExposure BsmtFinType1 BsmtFinType2 Electrical FireplaceQu GarageType   
## 38 37 38 1 690 81   
## GarageYrBlt GarageFinish GarageQual GarageCond PoolQC Fence   
## 81 81 81 81 1453 1179   
## MiscFeature   
## 1406

Above are variables with missing values greater than 0. I like this already!

This is where the benefit of having friends who are in housing for sale or construction count.

Next, we will evaluate (does it really matter? can it be done easily? does it cost a lot ?) all other variables with missing values

GarageFinish (81 missing values): Interior finish of the garage…average homebuyers don’t care about it… the condition of the garage is sufficient … remove

GarageQual (81 missing values): Garage quality, this might mean the quality of the materials used to build the garage. Remove…

GarageCond (81 missing values): Garage condition, this might mean the condition(look) at the time the house is sold.

GarageType(81 missing values): garage location……garage condition is enough ….remove

GarageYrBlt(81 missing values): garage year built…same with yearbuilt…remove

MasVnrType(07 missing values): Masonry veneer type…the type of material used to construct the house(bricks, stone…)…we can keep it for now

MasVnrArea(07 missing values): Masonry veneer area in square feet…this is the kind of work homebuyers won’t do. Because it involve doing the math to find the ratio of the veneer area covered Vs. not covered….remove

BsmtQual( 037 missing values): Evaluates the height of the basement…will be removed because another variable described the same thing.

BsmtCond (37 missing values): Evaluates the general condition of the basement….we can keep…

BsmtExposure(38 missing values): Refers to walkout or garden level walls..average homebuyers don’t care about it. remove

BsmtFinType1 (37 missing values): Rating of basement finished area…described already…remove

BsmtFinType2 ( 38 missing values): Type 1 finished square feet…no tricky math for buyer… remove

Electrical…we will only remove the 01 missing values…no hurts

## [1] "MSSubClass" "MSZoning" "LotArea" "Street"   
## [5] "LotShape" "LandContour" "Utilities" "LotConfig"   
## [9] "LandSlope" "Neighborhood" "Condition1" "Condition2"   
## [13] "BldgType" "HouseStyle" "OverallQual" "OverallCond"   
## [17] "YearBuilt" "YearRemodAdd" "RoofStyle" "RoofMatl"   
## [21] "Exterior1st" "Exterior2nd" "MasVnrType" "ExterQual"   
## [25] "ExterCond" "Foundation" "BsmtCond" "BsmtFinSF1"   
## [29] "BsmtFinSF2" "BsmtUnfSF" "TotalBsmtSF" "Heating"   
## [33] "HeatingQC" "CentralAir" "Electrical" "X1stFlrSF"   
## [37] "X2ndFlrSF" "LowQualFinSF" "GrLivArea" "BsmtFullBath"   
## [41] "BsmtHalfBath" "FullBath" "HalfBath" "BedroomAbvGr"   
## [45] "KitchenAbvGr" "KitchenQual" "TotRmsAbvGrd" "Functional"   
## [49] "Fireplaces" "GarageCars" "GarageArea" "GarageCond"   
## [53] "PavedDrive" "WoodDeckSF" "OpenPorchSF" "EnclosedPorch"  
## [57] "X3SsnPorch" "ScreenPorch" "PoolArea" "MiscVal"   
## [61] "MoSold" "YrSold" "SaleType" "SaleCondition"  
## [65] "SalePrice"

We continue to remove by pertinence (if it seems like extra or one can live without it in the area…) to the sale price. LandContour,LotShape, Street, LotConfig, LandSlope, Condition2, RoofStyle, Exterior2nd, OverallQual, Foundation, TotalBsmtSF,BsmtUnfSF, Heating, Electrical, X1stFlrSF, X2ndFlrSF, LowQualFinSF, BsmtFullBath, BsmtHalfBath, TotRmsAbvGrd, PavedDrive, EnclosedPorch, 3SsnPorch, ScreenPorch, MoSold,MSSubClass,Condition1

## [1] "MSZoning" "LotArea" "Utilities" "Neighborhood"   
## [5] "BldgType" "HouseStyle" "OverallCond" "YearBuilt"   
## [9] "YearRemodAdd" "RoofMatl" "Exterior1st" "MasVnrType"   
## [13] "ExterQual" "ExterCond" "BsmtCond" "BsmtFinSF1"   
## [17] "BsmtFinSF2" "HeatingQC" "CentralAir" "X1stFlrSF"   
## [21] "GrLivArea" "FullBath" "HalfBath" "BedroomAbvGr"   
## [25] "KitchenAbvGr" "KitchenQual" "Functional" "Fireplaces"   
## [29] "GarageCars" "GarageArea" "GarageCond" "WoodDeckSF"   
## [33] "OpenPorchSF" "PoolArea" "MiscVal" "YrSold"   
## [37] "SaleType" "SaleCondition" "SalePrice"

## MasVnrType BsmtCond GarageCond   
## 8 37 81

No need to do imputation by mean or other numerical imputations. We just need to replace the value according to the definition of the variable and remove the row where ‘NA’ has no meaning and it is very low missing values (ex: 10 missing values will no infere on 1460 observations)

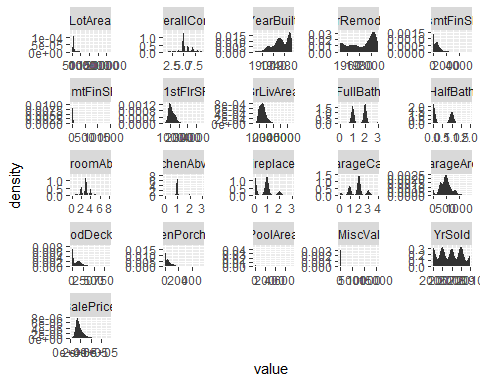
## [1] "MSZoning" "LotArea" "Utilities" "Neighborhood"   
## [5] "BldgType" "HouseStyle" "OverallCond" "YearBuilt"   
## [9] "YearRemodAdd" "RoofMatl" "Exterior1st" "MasVnrType"   
## [13] "ExterQual" "ExterCond" "BsmtCond" "BsmtFinSF1"   
## [17] "BsmtFinSF2" "HeatingQC" "CentralAir" "X1stFlrSF"   
## [21] "GrLivArea" "FullBath" "HalfBath" "BedroomAbvGr"   
## [25] "KitchenAbvGr" "KitchenQual" "Functional" "Fireplaces"   
## [29] "GarageCars" "GarageArea" "GarageCond" "WoodDeckSF"   
## [33] "OpenPorchSF" "PoolArea" "MiscVal" "YrSold"   
## [37] "SaleType" "SaleCondition" "SalePrice"

Remained 36 variables after cleaning… let’s convert character variables to categorical ones.

# Exploratory Data Analysis (EDA)

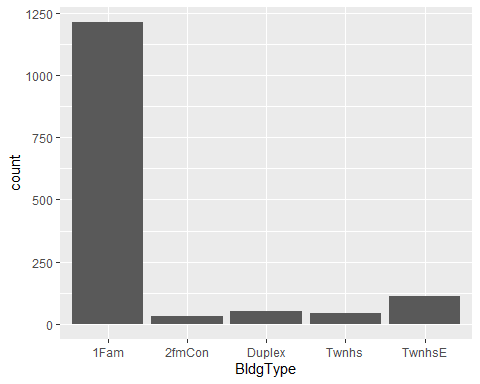
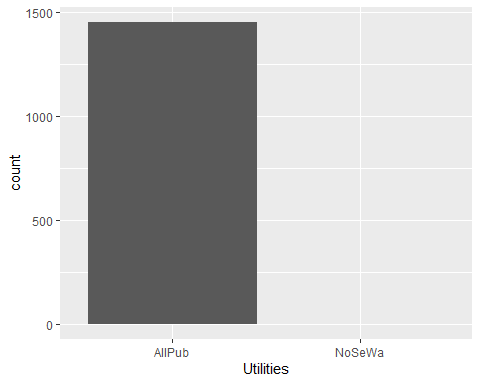
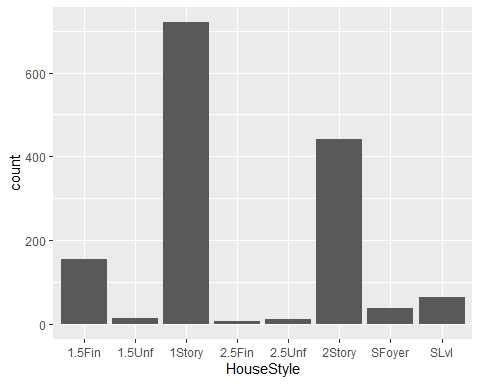
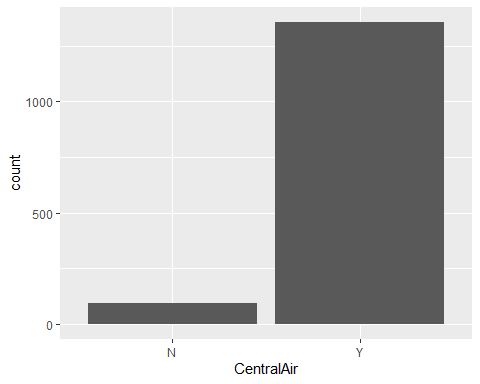
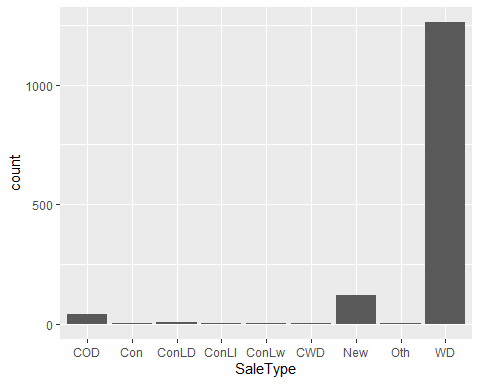
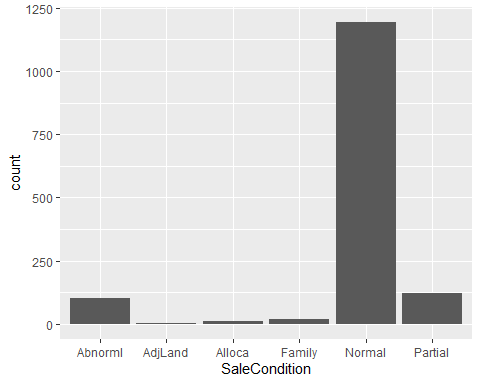
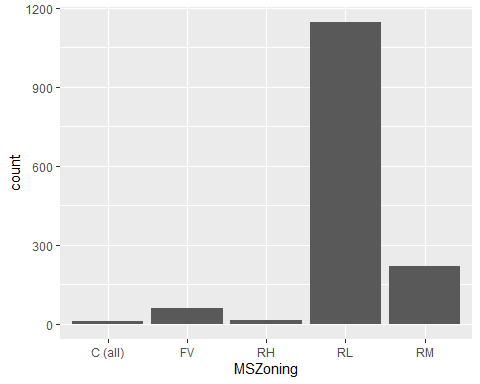
## MSZoning LotArea Utilities Neighborhood   
## Length:1452 Min. : 1300 Length:1452 Length:1452   
## Class :character 1st Qu.: 7539 Class :character Class :character   
## Mode :character Median : 9478 Mode :character Mode :character   
## Mean : 10507   
## 3rd Qu.: 11600   
## Max. :215245   
## BldgType HouseStyle OverallCond YearBuilt   
## Length:1452 Length:1452 Min. :1.000 Min. :1872   
## Class :character Class :character 1st Qu.:5.000 1st Qu.:1954   
## Mode :character Mode :character Median :5.000 Median :1972   
## Mean :5.579 Mean :1971   
## 3rd Qu.:6.000 3rd Qu.:2000   
## Max. :9.000 Max. :2010   
## YearRemodAdd RoofMatl Exterior1st MasVnrType   
## Min. :1950 Length:1452 Length:1452 Length:1452   
## 1st Qu.:1966 Class :character Class :character Class :character   
## Median :1993 Mode :character Mode :character Mode :character   
## Mean :1985   
## 3rd Qu.:2004   
## Max. :2010   
## ExterQual ExterCond BsmtCond BsmtFinSF1   
## Length:1452 Length:1452 Length:1452 Min. : 0.0   
## Class :character Class :character Class :character 1st Qu.: 0.0   
## Mode :character Mode :character Mode :character Median : 381.0   
## Mean : 442.0   
## 3rd Qu.: 706.5   
## Max. :5644.0   
## BsmtFinSF2 HeatingQC CentralAir X1stFlrSF   
## Min. : 0.00 Length:1452 Length:1452 Min. : 334   
## 1st Qu.: 0.00 Class :character Class :character 1st Qu.: 882   
## Median : 0.00 Mode :character Mode :character Median :1086   
## Mean : 46.81 Mean :1161   
## 3rd Qu.: 0.00 3rd Qu.:1391   
## Max. :1474.00 Max. :4692   
## GrLivArea FullBath HalfBath BedroomAbvGr   
## Min. : 334 Min. :0.000 Min. :0.0000 Min. :0.000   
## 1st Qu.:1128 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:2.000   
## Median :1462 Median :2.000 Median :0.0000 Median :3.000   
## Mean :1514 Mean :1.563 Mean :0.3815 Mean :2.867   
## 3rd Qu.:1776 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :5642 Max. :3.000 Max. :2.0000 Max. :8.000   
## KitchenAbvGr KitchenQual Functional Fireplaces   
## Min. :0.000 Length:1452 Length:1452 Min. :0.0000   
## 1st Qu.:1.000 Class :character Class :character 1st Qu.:0.0000   
## Median :1.000 Mode :character Mode :character Median :1.0000   
## Mean :1.046 Mean :0.6123   
## 3rd Qu.:1.000 3rd Qu.:1.0000   
## Max. :3.000 Max. :3.0000   
## GarageCars GarageArea GarageCond WoodDeckSF   
## Min. :0.000 Min. : 0.0 Length:1452 Min. : 0.00   
## 1st Qu.:1.000 1st Qu.: 327.8 Class :character 1st Qu.: 0.00   
## Median :2.000 Median : 478.0 Mode :character Median : 0.00   
## Mean :1.765 Mean : 472.5 Mean : 94.42   
## 3rd Qu.:2.000 3rd Qu.: 576.0 3rd Qu.:168.00   
## Max. :4.000 Max. :1418.0 Max. :857.00   
## OpenPorchSF PoolArea MiscVal YrSold   
## Min. : 0.00 Min. : 0.000 Min. : 0.00 Min. :2006   
## 1st Qu.: 0.00 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.:2007   
## Median : 24.00 Median : 0.000 Median : 0.00 Median :2008   
## Mean : 46.39 Mean : 2.774 Mean : 43.73 Mean :2008   
## 3rd Qu.: 68.00 3rd Qu.: 0.000 3rd Qu.: 0.00 3rd Qu.:2009   
## Max. :547.00 Max. :738.000 Max. :15500.00 Max. :2010   
## SaleType SaleCondition SalePrice   
## Length:1452 Length:1452 Min. : 34900   
## Class :character Class :character 1st Qu.:129900   
## Mode :character Mode :character Median :162700   
## Mean :180615   
## 3rd Qu.:214000   
## Max. :755000

## No id variables; using all as measure variables



Above is the distribution of numerical variables…yearsold looks like many years counted. in fact, those variables with multiple mode, mean the variable have a set of of values that repeat. Hard to see off values (outliers).

below is the distribution of categorical variables, not all them will be plot because some of them have too many levels to be plotted all at once…maybe there is another technique



The way we visualize these categorical variables is that some of them can be redefined numerical since the categorical level in some of them are null or close to null. Potentially use some rating… 0 means nothing 1 means poor and so on with …2, 3, 4, 5..

Of course this is not the only way to transform them…

# Feature Engineering

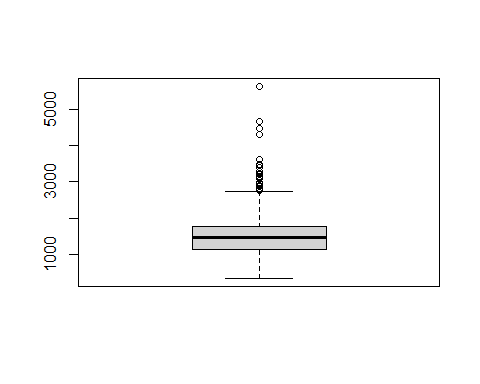
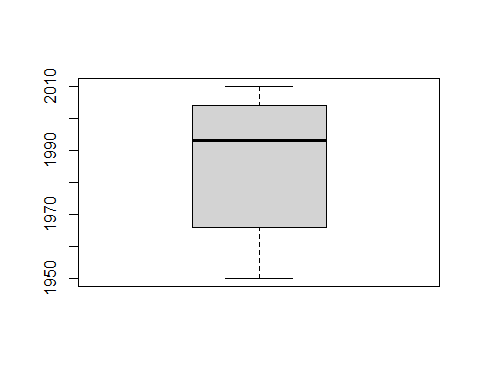
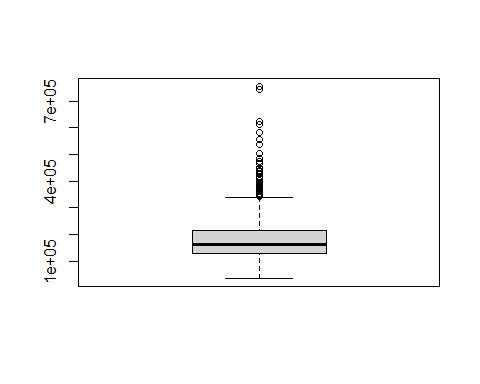
| Var1 | Freq |
| --- | --- |
| 1950 | 178 |
| 1951 | 4 |
| 1952 | 5 |
| 1953 | 10 |
| 1954 | 14 |
| 1955 | 9 |
| 1956 | 10 |
| 1957 | 9 |
| 1958 | 15 |
| 1959 | 18 |
| 1960 | 12 |
| 1961 | 8 |
| 1962 | 14 |
| 1963 | 13 |
| 1964 | 11 |
| 1965 | 19 |
| 1966 | 15 |
| 1967 | 12 |
| 1968 | 17 |
| 1969 | 14 |
| 1970 | 26 |
| 1971 | 18 |
| 1972 | 20 |
| 1973 | 11 |
| 1974 | 7 |
| 1975 | 9 |
| 1976 | 30 |
| 1977 | 25 |
| 1978 | 16 |
| 1979 | 10 |
| 1980 | 12 |
| 1981 | 8 |
| 1982 | 7 |
| 1983 | 5 |
| 1984 | 7 |
| 1985 | 9 |
| 1986 | 5 |
| 1987 | 10 |
| 1988 | 9 |
| 1989 | 11 |
| 1990 | 15 |
| 1991 | 14 |
| 1992 | 17 |
| 1993 | 19 |
| 1994 | 22 |
| 1995 | 31 |
| 1996 | 36 |
| 1997 | 25 |
| 1998 | 36 |
| 1999 | 30 |
| 2000 | 55 |
| 2001 | 21 |
| 2002 | 46 |
| 2003 | 50 |
| 2004 | 62 |
| 2005 | 73 |
| 2006 | 96 |
| 2007 | 74 |
| 2008 | 39 |
| 2009 | 23 |
| 2010 | 6 |

| Var1 | Freq |
| --- | --- |
| 2006 | 313 |
| 2007 | 327 |
| 2008 | 301 |
| 2009 | 337 |
| 2010 | 174 |

Not sure what to think of variables with years…does it affect the target variable? Let’s transform categorical to factor.

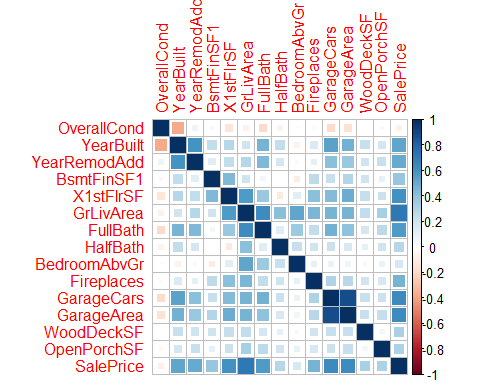
## Rows: 1,452  
## Columns: 39  
## $ MSZoning <fct> RL, RL, RL, RL, RL, RL, RL, RL, RM, RL, RL, RL, RL, RL, ~  
## $ LotArea <int> 8450, 9600, 11250, 9550, 14260, 14115, 10084, 10382, 612~  
## $ Utilities <fct> AllPub, AllPub, AllPub, AllPub, AllPub, AllPub, AllPub, ~  
## $ Neighborhood <fct> CollgCr, Veenker, CollgCr, Crawfor, NoRidge, Mitchel, So~  
## $ BldgType <fct> 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 1Fam, 2f~  
## $ HouseStyle <fct> 2Story, 1Story, 2Story, 2Story, 2Story, 1.5Fin, 1Story, ~  
## $ OverallCond <int> 5, 8, 5, 5, 5, 5, 5, 6, 5, 6, 5, 5, 6, 5, 5, 8, 7, 5, 5,~  
## $ YearBuilt <int> 2003, 1976, 2001, 1915, 2000, 1993, 2004, 1973, 1931, 19~  
## $ YearRemodAdd <int> 2003, 1976, 2002, 1970, 2000, 1995, 2005, 1973, 1950, 19~  
## $ RoofMatl <fct> CompShg, CompShg, CompShg, CompShg, CompShg, CompShg, Co~  
## $ Exterior1st <fct> VinylSd, MetalSd, VinylSd, Wd Sdng, VinylSd, VinylSd, Vi~  
## $ MasVnrType <fct> BrkFace, None, BrkFace, None, BrkFace, None, Stone, Ston~  
## $ ExterQual <fct> Gd, TA, Gd, TA, Gd, TA, Gd, TA, TA, TA, TA, Ex, TA, Gd, ~  
## $ ExterCond <fct> TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, ~  
## $ BsmtCond <fct> TA, TA, TA, Gd, TA, TA, TA, TA, TA, TA, TA, TA, TA, TA, ~  
## $ BsmtFinSF1 <int> 706, 978, 486, 216, 655, 732, 1369, 859, 0, 851, 906, 99~  
## $ BsmtFinSF2 <int> 0, 0, 0, 0, 0, 0, 0, 32, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~  
## $ HeatingQC <fct> Ex, Ex, Ex, Gd, Ex, Ex, Ex, Ex, Gd, Ex, Ex, Ex, TA, Ex, ~  
## $ CentralAir <fct> Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y,~  
## $ X1stFlrSF <int> 856, 1262, 920, 961, 1145, 796, 1694, 1107, 1022, 1077, ~  
## $ GrLivArea <int> 1710, 1262, 1786, 1717, 2198, 1362, 1694, 2090, 1774, 10~  
## $ FullBath <int> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 3, 1, 2, 1, 1, 1, 2, 1,~  
## $ HalfBath <int> 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,~  
## $ BedroomAbvGr <int> 3, 3, 3, 3, 4, 1, 3, 3, 2, 2, 3, 4, 2, 3, 2, 2, 2, 2, 3,~  
## $ KitchenAbvGr <int> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1,~  
## $ KitchenQual <fct> Gd, TA, Gd, Gd, Gd, TA, Gd, TA, TA, TA, TA, Ex, TA, Gd, ~  
## $ Functional <fct> Typ, Typ, Typ, Typ, Typ, Typ, Typ, Typ, Min1, Typ, Typ, ~  
## $ Fireplaces <int> 0, 1, 1, 1, 1, 0, 1, 2, 2, 2, 0, 2, 0, 1, 1, 0, 1, 0, 0,~  
## $ GarageCars <int> 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 3, 1, 3, 1, 2, 2, 2, 2,~  
## $ GarageArea <int> 548, 460, 608, 642, 836, 480, 636, 484, 468, 205, 384, 7~  
## $ GarageCond <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "T~  
## $ WoodDeckSF <int> 0, 298, 0, 0, 192, 40, 255, 235, 90, 0, 0, 147, 140, 160~  
## $ OpenPorchSF <int> 61, 0, 42, 35, 84, 30, 57, 204, 0, 4, 0, 21, 0, 33, 213,~  
## $ PoolArea <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~  
## $ MiscVal <int> 0, 0, 0, 0, 0, 700, 0, 350, 0, 0, 0, 0, 0, 0, 0, 0, 700,~  
## $ YrSold <int> 2008, 2007, 2008, 2006, 2008, 2009, 2007, 2009, 2008, 20~  
## $ SaleType <fct> WD, WD, WD, WD, WD, WD, WD, WD, WD, WD, WD, New, WD, New~  
## $ SaleCondition <fct> Normal, Normal, Normal, Abnorml, Normal, Normal, Normal,~  
## $ SalePrice <int> 208500, 181500, 223500, 140000, 250000, 143000, 307000, ~

Let’s see some boxplot



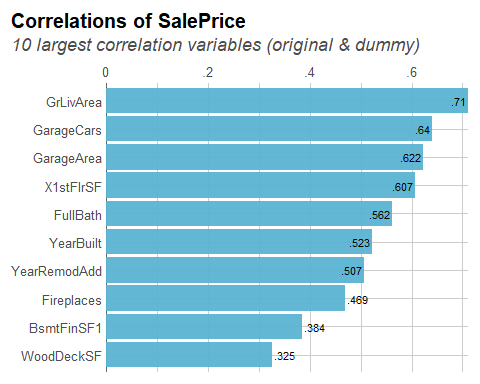
As we can see, SalePrice, GrLivArea have some outliers… For the sake that we want to run regression analysis, we limit the variables to numerical moving forward.

Let’s see some correlations



Based on the correlation plot, if we want to select those with high correlation with the target variable. Let’s run another correlation function to help us select these variables rather than relying on visual (nothing wrong with visual)…

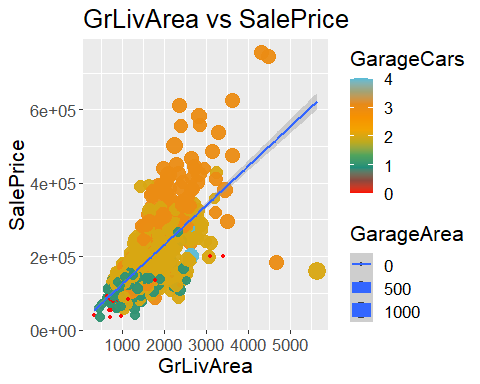
## Warning in .font\_global(font, quiet = FALSE): Font 'Arial Narrow' is not  
## installed, has other name, or can't be found



We will keep the variables with 0.5 above… there are really 08 variables ( )

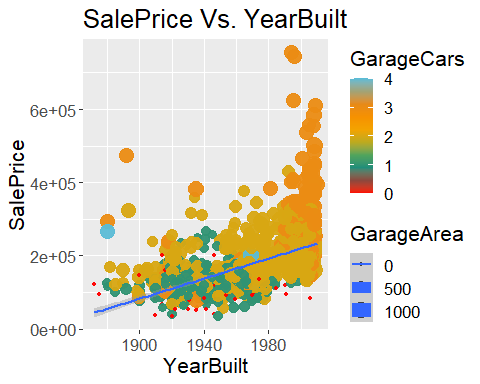
Let’s see SalesPrice Vs. GrLivArea

## `geom\_smooth()` using formula 'y ~ x'



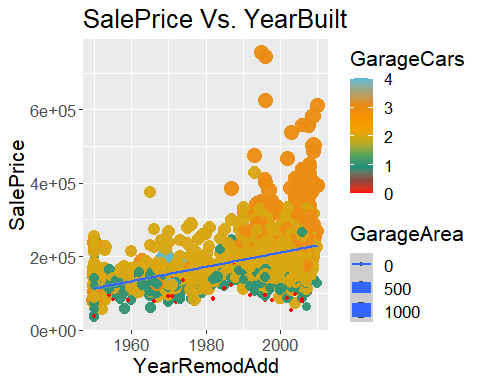
Let’s see SalesPrice Vs. YearBuilt

## `geom\_smooth()` using formula 'y ~ x'



Let’s see SalesPrice Vs. YearRemodAdd

## `geom\_smooth()` using formula 'y ~ x'



Definitely high correlation and linear relationship….

# Modelling

We will run multiple linear regression and Neutral Network. There are definitely regression variables in this dataset. the neutral network is to see how other algorithm will perform, specially those on the unsupervised category. Neutral network seem to have perfomance in neuron analysis (according to neuroscientists who performed them…)

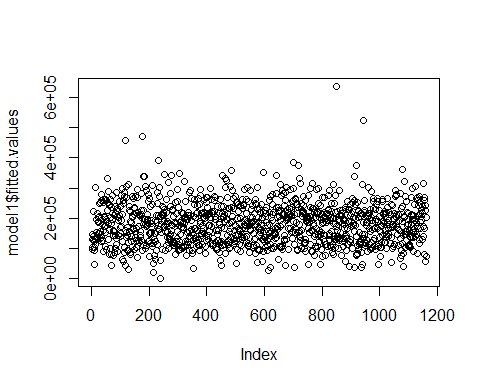
### Splitting the data

Split will be 0.8 train and 0.2 test

## Multiple Linear Regression (model1)

We want to use glm() function due to more than one independent/predictors.

##   
## Call:  
## glm(formula = SalePrice ~ ., data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -476681 -20824 -3747 17360 296387   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.114e+06 1.469e+05 -14.391 < 2e-16 \*\*\*  
## YearBuilt 4.827e+02 6.049e+01 7.980 3.52e-15 \*\*\*  
## YearRemodAdd 5.875e+02 7.809e+01 7.524 1.07e-13 \*\*\*  
## X1stFlrSF 3.499e+01 4.143e+00 8.445 < 2e-16 \*\*\*  
## GrLivArea 6.928e+01 3.665e+00 18.902 < 2e-16 \*\*\*  
## FullBath -4.560e+03 3.309e+03 -1.378 0.169   
## GarageCars 1.886e+04 3.732e+03 5.053 5.04e-07 \*\*\*  
## GarageArea 1.271e+01 1.294e+01 0.982 0.326   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 1844051805)  
##   
## Null deviance: 7.4845e+12 on 1161 degrees of freedom  
## Residual deviance: 2.1280e+12 on 1154 degrees of freedom  
## AIC: 28099  
##   
## Number of Fisher Scoring iterations: 2



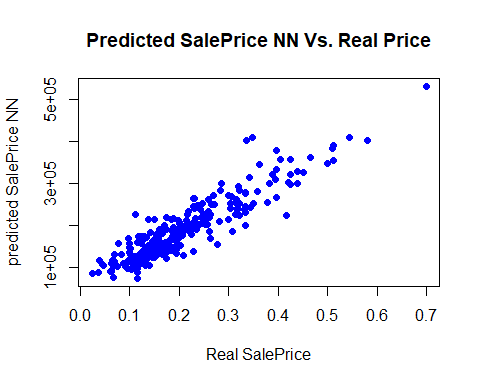
Wow! All the variables show good significance to the SalePrice.

## Neutral Network (model2)

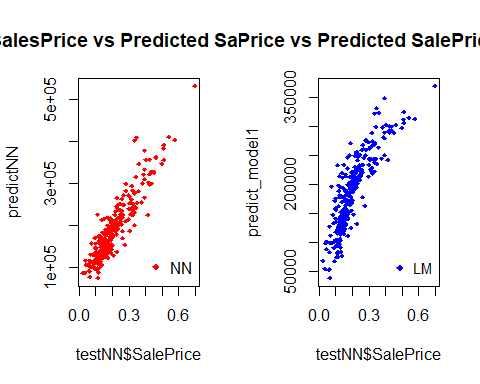
##   
## Attaching package: 'neuralnet'

## The following object is masked from 'package:dplyr':  
##   
## compute

Well done!

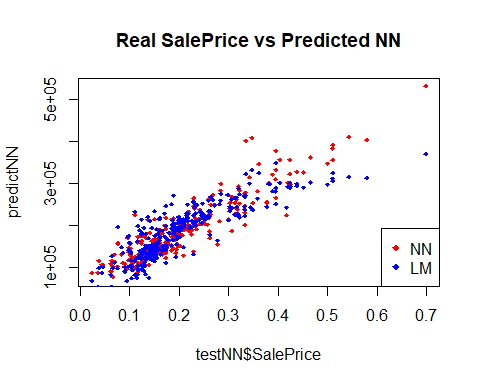
predicting NN 

## MSE\_model2  
## MSE\_model1 922874054.912377  
## 1477004184.67592 1



Neutral Network seems about equal performance with multiple linear regression.

A closer look



It is a hard call here…but we want to go with Multiple linear regression since it has a better MSE and p-value.

# References

1 - <file:///C:/Users/owner/Downloads/622-Article%20Text-961-2-10-20220308.pdf>

2 - <http://neuralnetworksanddeeplearning.com/chap1.html>

3- <https://cran.r-project.org/web/packages/dlookr/vignettes/EDA.html>

4- <https://stats.oarc.ucla.edu/stat/data/rmarkdown/rmarkdown_seminar_flat.html#links-internal-and-external>

5- <https://plotly.com/r/line-charts/>

6- <https://www.r-bloggers.com/2015/09/fitting-a-neural-network-in-r-neuralnet-package/>

1. Machine\_learning. [↑](#footnote-ref-20)