Projekt – Język R

Sławomir Karwowski

Studia podyplomowe – Data Scientist

Analiza, transformacja oraz zaprojektowanie modelu regresji liniowej przewidującego ceny domów w USA

Spis treści

1. Wstęp

Projekt ***Analiza, transformacja oraz zaprojektowanie modelu regresji liniowej przewidującego ceny domów w USA*** został wykonany przy użyciu Języka **R**, w środowisku **R Studio.**

Celem projektu było wykonanie modelu regresji liniowej w celu predykcji cen domów/mieszkań w USA. Zadanie zostało wykonane na zbiorze HousePrices dostępnego na stronie www.kaggle.com.

2. Analiza i zrozumienie danych

Zbiór, na którym wykonywany jest projekt to zbiór zawierający obserwacje – transakcji sprzedaży domów w USA. Zbiór treningowy zawiera 1460 obserwacji oraz 81 zmiennych. Zmienna SalePrice to zmienna, którą będziemy przewidywać. Reszta zmiennych to zmienne opisujące położenie, stan, cechy domu lub cechy transakcji.

Po wczytaniu danych została przeanalizowana statystyka każdej kolumny:

```
> summary(train_data)
       Id          MSSubClass      MSZoning         LotFrontage       LotArea          Street             Alley             LotShape
 Min.   :   1.0   Min.   : 20.0   Length:1460        Min.   : 21.00   Min.   :  1300   Length:1460        Length:1460        Length:1460
 1st Qu.: 365.8   1st Qu.: 20.0   Class :character   1st Qu.: 59.00   1st Qu.:  7554   Class :character   Class :character   Class :character
 Median : 730.5   Median : 50.0   Mode  :character   Median : 69.00   Median :  9478   Mode  :character   Mode  :character   Mode  :character
 Mean   : 730.5   Mean   : 56.9                      Mean   : 70.05   Mean   : 10517
 3rd Qu.:1095.2   3rd Qu.: 70.0                      3rd Qu.: 80.00   3rd Qu.: 11602
 Max.   :1460.0   Max.   :190.0                      Max.   :313.00   Max.   :215245
                                                     NA's   :259
 LandContour         Utilities          LotConfig          LandSlope         Neighborhood        Condition1         Condition2          BldgType
 Length:1460        Length:1460        Length:1460        Length:1460        Length:1460        Length:1460        Length:1460        Length:1460
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


  HouseStyle         OverallQual      OverallCond      YearBuilt     YearRemodAdd   RoofStyle           RoofMatl         Exterior1st
 Length:1460        Min.   : 1.000   Min.   :1.000   Min.   :1872   Min.   :1950   Length:1460        Length:1460        Length:1460
 Class :character   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Class :character   Class :character   Class :character
 Mode  :character   Median : 6.000   Median :5.000   Median :1973   Median :1994   Mode  :character   Mode  :character   Mode  :character
                    Mean   : 6.099   Mean   :5.575   Mean   :1971   Mean   :1985
                    3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004
                    Max.   :10.000   Max.   :9.000   Max.   :2010   Max.   :2010

 Exterior2nd        MasVnrType          MasVnrArea      ExterQual          ExterCond          Foundation          BsmtQual           BsmtCond
 Length:1460        Length:1460        Min.   :   0.0   Length:1460        Length:1460        Length:1460        Length:1460        Length:1460
 Class :character   Class :character   1st Qu.:   0.0   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Median :   0.0   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
                                       Mean   : 103.7
                                       3rd Qu.: 166.0
                                       Max.   :1600.0
                                       NA's   :8
 BsmtExposure       BsmtFinType1         BsmtFinSF1     BsmtFinType2         BsmtFinSF2       BsmtUnfSF       TotalBsmtSF       Heating
 Length:1460        Length:1460        Min.   :   0.0   Length:1460        Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Length:1460
 Class :character   Class :character   1st Qu.:   0.0   Class :character   1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   Class :character
 Mode  :character   Mode  :character   Median : 383.5   Mode  :character   Median :   0.00   Median : 477.5   Median : 991.5   Mode  :character
                                       Mean   : 443.6                      Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
                                       3rd Qu.: 712.2                      3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
                                       Max.   :5644.0                      Max.   :1474.00   Max.   :2336.0   Max.   :6110.0

  HeatingQC          CentralAir          Electrical          1stFlrSF       2ndFlrSF     LowQualFinSF       GrLivArea      BsmtFullBath
 Length:1460        Length:1460        Length:1460        Min.   : 334   Min.   :   0   Min.   :  0.000   Min.   : 334   Min.   :0.0000
 Class :character   Class :character   Class :character   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130   1st Qu.:0.0000
 Mode  :character   Mode  :character   Mode  :character   Median :1087   Median :   0   Median :  0.000   Median :1464   Median :0.0000
                                                          Mean   :1163   Mean   : 347   Mean   :  5.845   Mean   :1515   Mean   :0.4253
                                                          3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777   3rd Qu.:1.0000
                                                          Max.   :4692   Max.   :2065   Max.   :572.000   Max.   :5642   Max.   :3.0000
```

```
 BsmtHalfBath       FullBath        HalfBath       BedroomAbvGr    KitchenAbvGr   KitchenQual        TotRmsAbvGrd     Functional         Fireplaces
Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Length:1460        Min.   : 2.000   Length:1460        Min.   :0.000
1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Class :character   1st Qu.: 5.000   Class :character   1st Qu.:0.000
Median :0.00000   Median :2.000   Median :0.0000   Median :3.000   Median :1.000   Mode  :character   Median : 6.000   Mode  :character   Median :1.000
Mean   :0.05753   Mean   :1.565   Mean   :0.3829   Mean   :2.866   Mean   :1.047                      Mean   : 6.518                      Mean   :0.613
3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000                      3rd Qu.: 7.000                      3rd Qu.:1.000
Max.   :2.00000   Max.   :3.000   Max.   :2.0000   Max.   :8.000   Max.   :3.000                      Max.   :14.000                      Max.   :3.000

FireplaceQu        GarageType         GarageYrBlt   GarageFinish        GarageCars      GarageArea      GarageQual         GarageCond
Length:1460        Length:1460        Min.   :1900   Length:1460        Min.   :0.000   Min.   :   0.0   Length:1460        Length:1460
Class :character   Class :character   1st Qu.:1961   Class :character   1st Qu.:1.000   1st Qu.: 334.5   Class :character   Class :character
Mode  :character   Mode  :character   Median :1980   Mode  :character   Median :2.000   Median : 480.0   Mode  :character   Mode  :character
                                      Mean   :1979                      Mean   :1.767   Mean   : 473.0
                                      3rd Qu.:2002                      3rd Qu.:2.000   3rd Qu.: 576.0
                                      Max.   :2010                      Max.   :4.000   Max.   :1418.0
                                      NA's   :81
 PavedDrive          WoodDeckSF      OpenPorchSF     EnclosedPorch      3SsnPorch       ScreenPorch        PoolArea          PoolQC
Length:1460        Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Length:1460
Class :character   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   Class :character
Mode  :character   Median :  0.00   Median : 25.00   Median :  0.00   Median :  0.00   Median :  0.00   Median :  0.000   Mode  :character
                   Mean   : 94.24   Mean   : 46.66   Mean   : 21.95   Mean   :  3.41   Mean   : 15.06   Mean   :  2.759
                   3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
                   Max.   :857.00   Max.   :547.00   Max.   :552.00   Max.   :508.00   Max.   :480.00   Max.   :738.000

   Fence          MiscFeature          MiscVal             MoSold           YrSold       SaleType         SaleCondition        SalePrice
Length:1460        Length:1460        Min.   :    0.00   Min.   : 1.000   Min.   :2006   Length:1460        Length:1460        Min.   : 34900
Class :character   Class :character   1st Qu.:    0.00   1st Qu.: 5.000   1st Qu.:2007   Class :character   Class :character   1st Qu.:129975
Mode  :character   Mode  :character   Median :    0.00   Median : 6.000   Median :2008   Mode  :character   Mode  :character   Median :163000
                                      Mean   :   43.49   Mean   : 6.322   Mean   :2008                                          Mean   :180921
                                      3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009                                          3rd Qu.:214000
                                      Max.   :15500.00   Max.   :12.000   Max.   :2010                                          Max.   :755000
```

.

Wstępne obserwacje niektórych zmiennych:

- **ID** – Zmienna porządkowa – zbędna przy dalszej analizie
- **MSZoning** – typ okolicy (zmienna kategoryczna) – przypuszczenie o wpływie okolicy na cenę – ewentualna zamiana na zmienną numeryczną w celu użycia przy regresji liniowej
- **LotFrontage** – powierzchnia ulicy sąsiadująca z posiadłością – dużo NA (brak granicy); zmienna nieprzydatna przy dalszej analizie
- **LotArea** – powierzchnia działki (sprawdzenie wpływu na cenę)
- **Street** – rodzaj ulicy – zmienna kategoryczna
- **Alley** – typ chodnika (raczej brak wpływu na cenę)
- **Neighborhood** – okolica (dzielnica) przypuszczalny wpływ na cenę – zamiana kategorycznych wartości na numeryczne
- **OverallQual –** typ wykończenia i materiały – sprawdzenie wpływu na cenę
- **OverallCond –** stan budynku – sprawdzenie wpływu na cenę i korelacji ze zmienną OverallQual
- **YearBuilt –** rok budowy (zamiana na wiek mieszkania w latach)
- **YearRemod –** rok przebudowy (sprawdzenie korelacji z YearBuilt)
- **TotalBsmtSF –** całkowita powierzchnia piwnicy (zbadanie wpływu na cenę oraz korelacji z innymi zmiennymi opisującymi wielkość domu)
- **GrLivArea –** powierzchnia mieszkania/domu ponad powierzchnię terenu (zbadanie wpływu na cenę – potencjalnie silna korelacja)
- **TotRmsAbvGrd –** liczba pokoi (sprawdzenie zależności z wielkością mieszkania)
- **Fireplaces –** kominki (liczba) – sprawdzenie wpływu na cenę
- **GarageCars –** liczba miejsc samochodowych w garażu – sprawdzenie wpływu na cenę oraz zależności z innymi zmiennymi opisującymi garaż
- **SalePrices –** zmienna wyjściowa – cena domu/mieszkania – zmienna, którą będziemy przewidywać używając model regresji liniowej

- Zmienne kategoryczne, które po wstępnej analizie odrzuciłem z modelu takie jak: **HouseStyle, RoofMaterial, BsmtCond, BsmtFinType1, BsmtFinType2, Heating, Electrical, KitchenQual itp.**

Badając zmienne pod kątem użycia ich w modelu regresji liniowej zostały przyjęte założenia:

- Musi być liniowa zależność pomiędzy zmiennymi wejściowymi i wyjściową
- Rozkład błędów powinien być zbliżony do normalnego
- Brak autokorelacji błędów
- Zmienne wejściowe nie powinny być ze sobą silnie skorelowane

2.1. Badanie korelacji oraz wizualizacja danych w celu lepszego zrozumienia zbioru danych

- Na wstępie zostają sprawdzone liczności grup zmiennych kategorycznych. Na przykład zmienna charakteryzująca okolicę rozkłada się w grupach w następujący sposób:

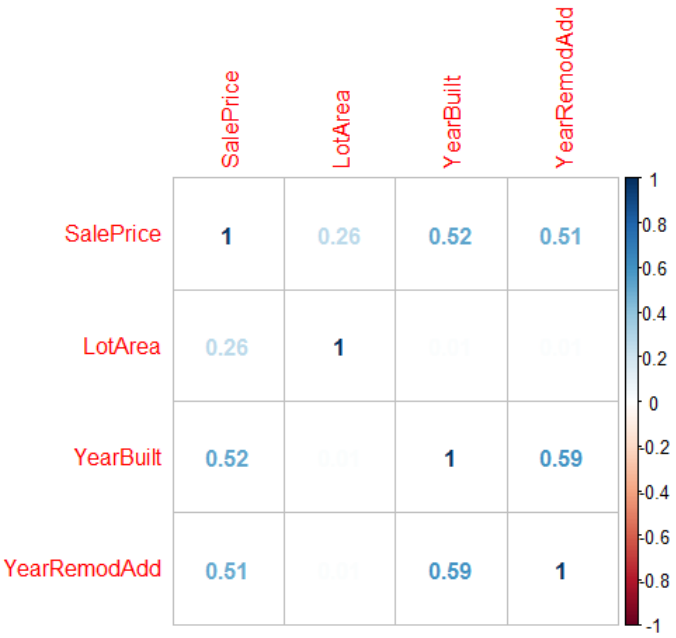
```
C (all)      FV      RH      RL      RM
     10      65      16    1151     218
```

- Sprawdzenie korelacji między wielkością parceli, a ceną sprzedaży:

```
> #check correlation between LotArea and SalePrice
> cor(log(t$LotArea), t$SalePrice)
[1] 0.3885203
> #check correlation between LotArea and SalePrice
> cor(t$LotArea, t$SalePrice)
[1] 0.2638434
```

  Z wartością zlogarytmowaną zauważamy większą korelację.

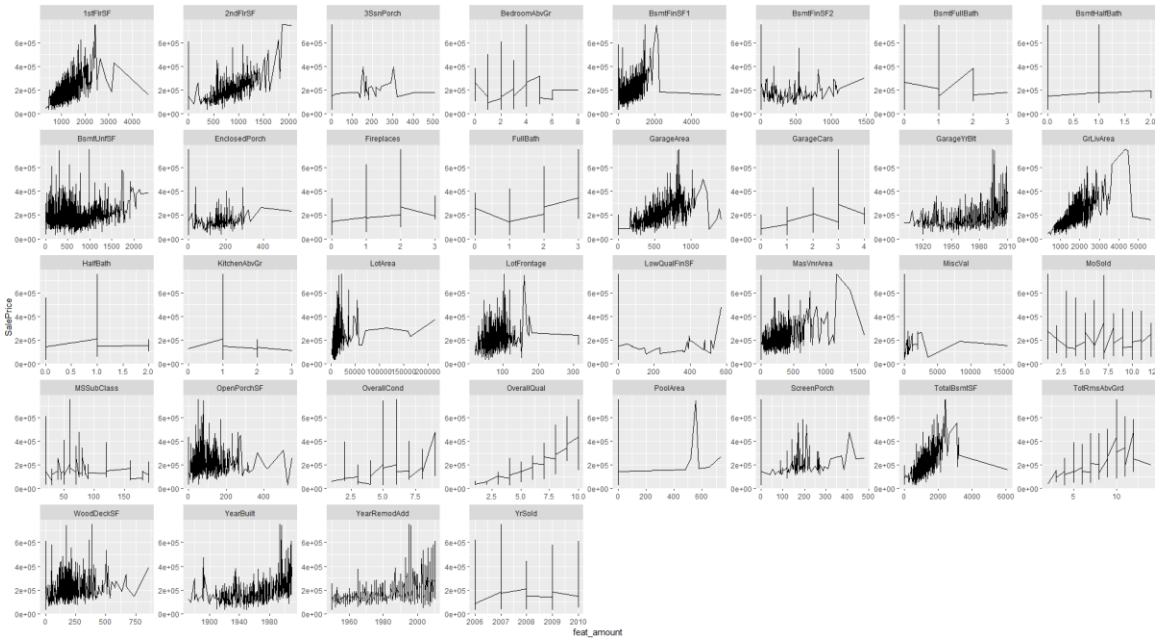
- Sprawdzenie korelacji pomiędzy zmiennymi opisującym wiek domu oraz czas remontu:



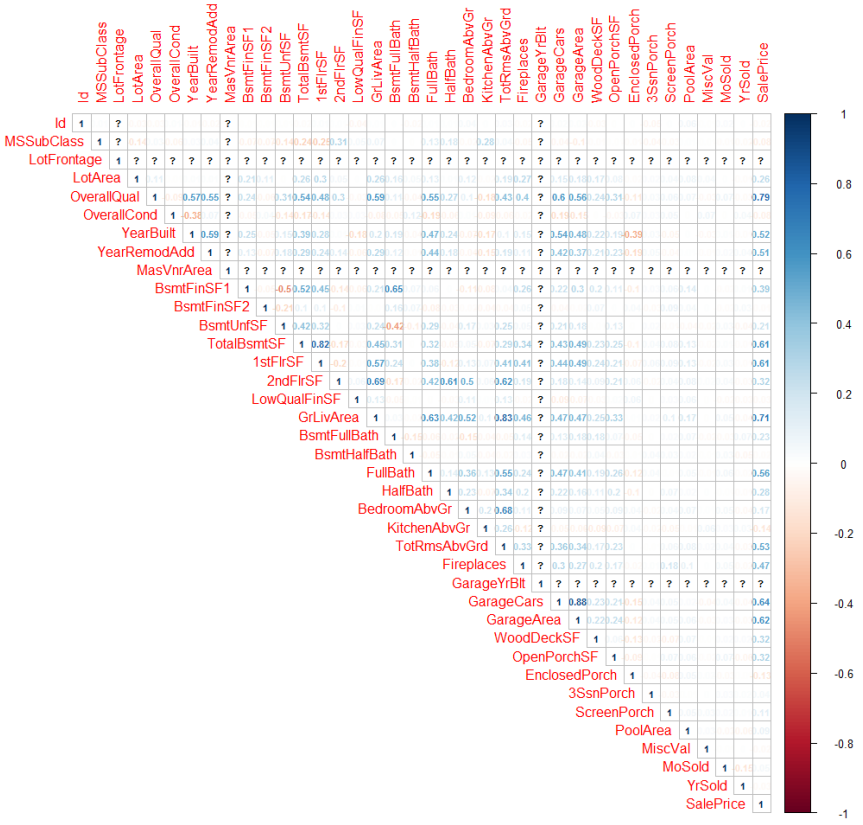
  Zauważalna znacząca korelacja między rokiem budowy, a rokiem przebudowy (remontu).

Po podzieleniu zbioru danych na dwa zbiory – jeden zawierający zmienne numeryczne, a drugi kategoryczne zostały sprawdzone kolejne korelacje.

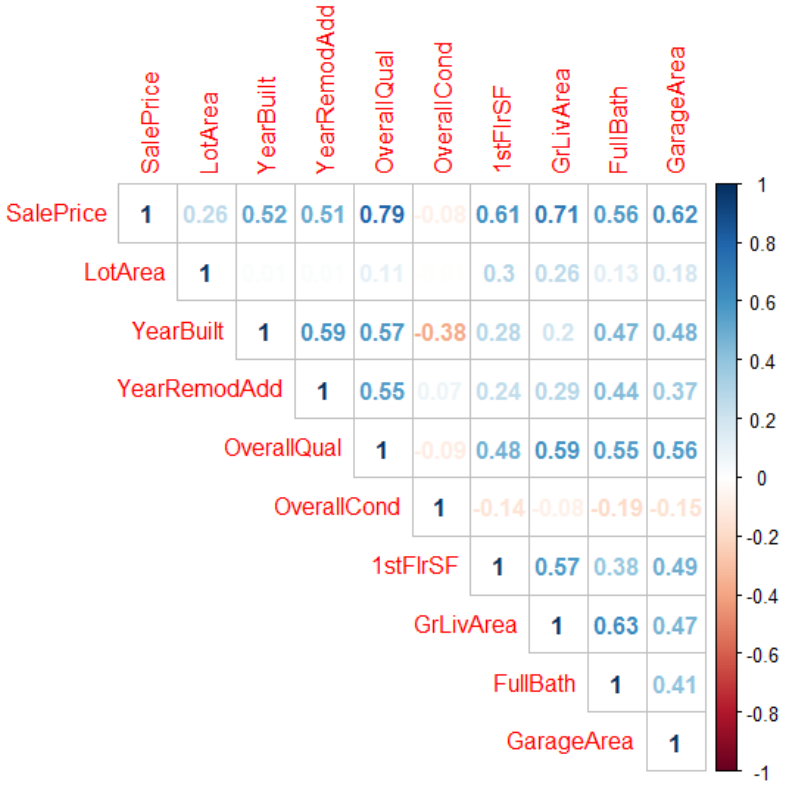
- Sprawdzenie zależności na wykresach oraz korelacji wszystkich zmiennych numerycznych (wykresy okazały się mało wyraźne ze względu na dużą liczbę zmiennych – w związku z czym przeprowadziłem tą samą analizę dzieląc zbiór danych numerycznych na podzbiory).



Zauważalne jest, które zmienne (pomimo wartości numerycznych) są kategoryczne.

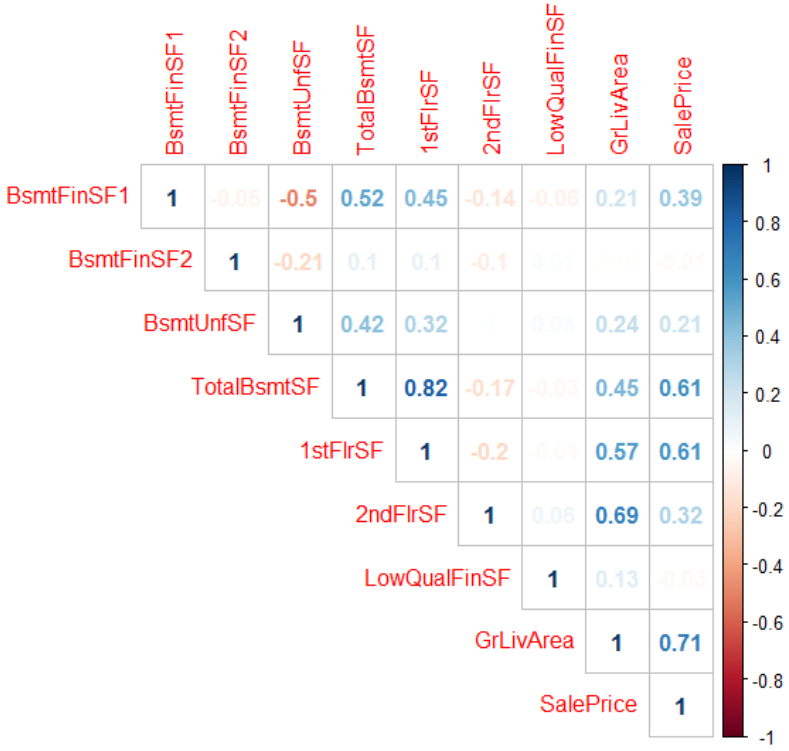


- Korelacja wybranych zmiennych (z potencjalnie dużym wpływem na cenę domu).



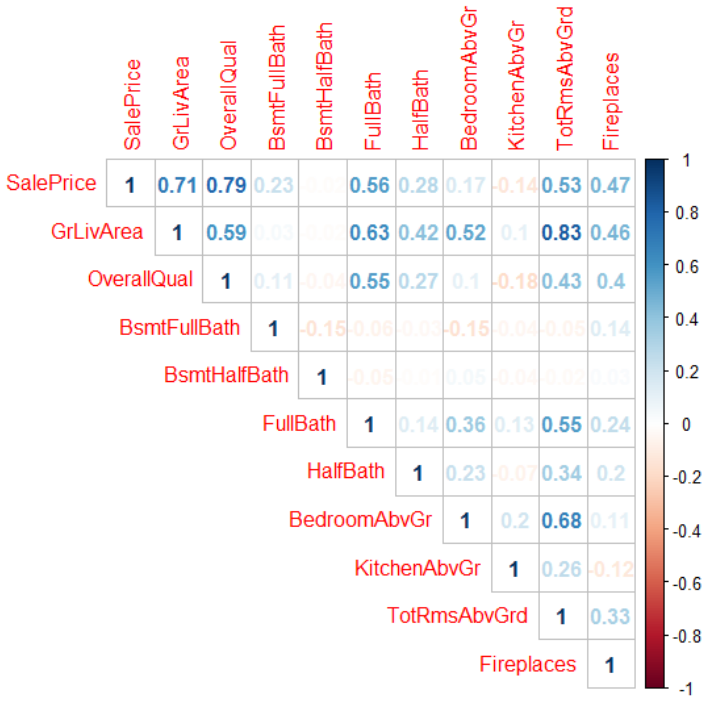
Zauważalna silna korelacja ze zmienna wyjściową takich zmiennych jak: GrLivArea czy OverallQual.

- Korelacja zmiennych związanych z piwnicą domu.



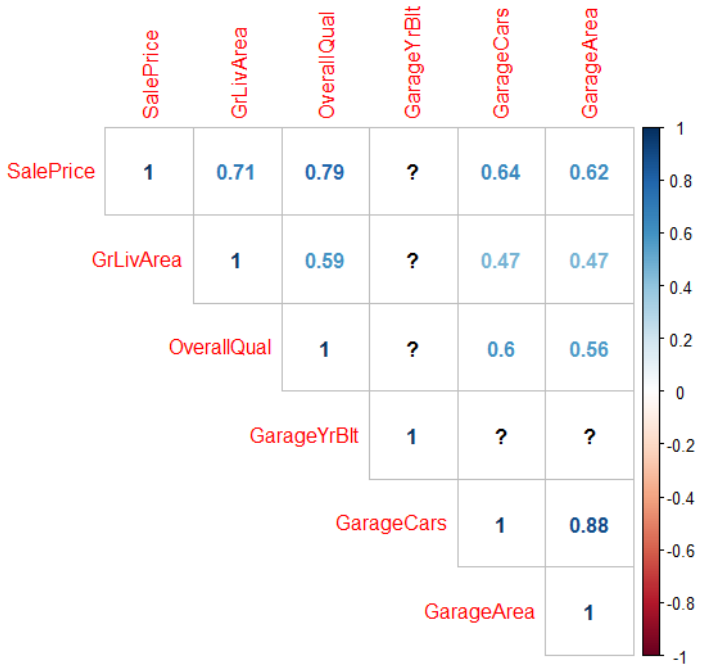
Zauważalna korelacja między zmienną TotalBsmtSF a ceną domu (przy okazji nie ma silnej korelacji z inną istotną zmienną GrLivArea).

- Korelacja zmiennych związanych z łazienkami i kominkami.



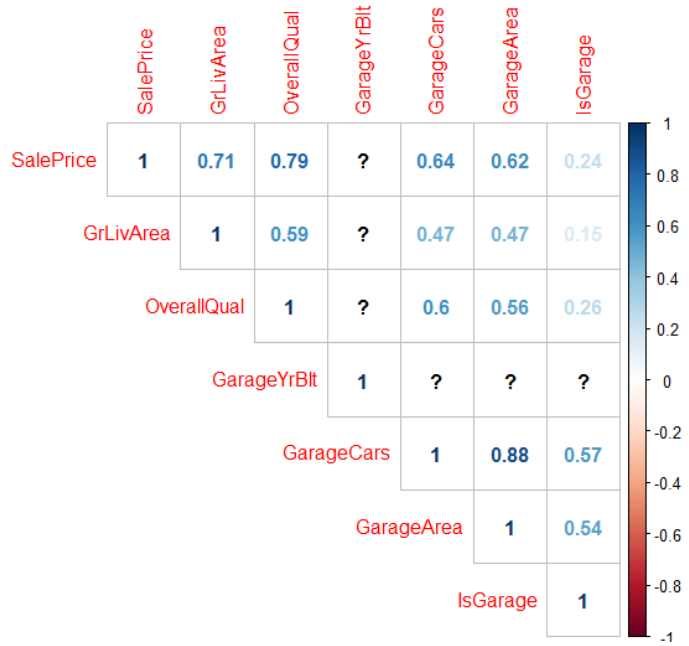
Zauważalna silna korelacja między ilością pokojów, a wielkością mieszkania. Zmienna określająca liczbę kominków jest skorelowana z ceną – może być wartościową zmienną przy predykcji ceny, gdyż nie koniecznie musi zależeć od wielkości mieszkania, a może świadczyć o jego standardzie.

- Korelacje zmiennych charakteryzujących garaże w domu.



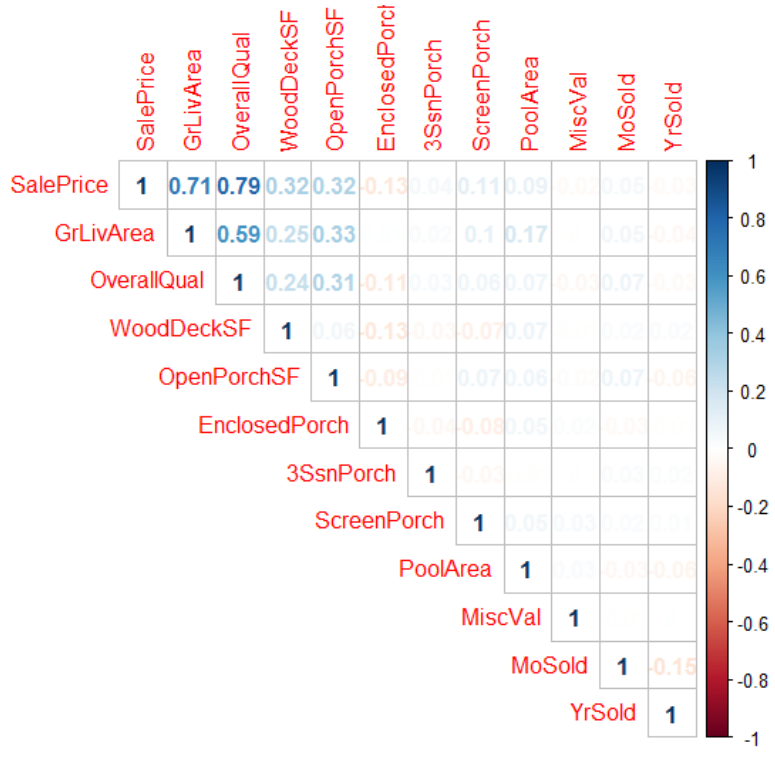
Silna korelacja między wielkością garażu i liczbą miejsc samochodowych – zależność intuicyjna (w związku z tym jedna z tych zmiennych zostanie wzięta pod uwagę do modelu).

- Sprawdzenie czy zmienna mówiąca, czy w domu jest garaż może być lepszą zmienną do modelu. W tym celu została stworzona zmienna IsGarage przyjmująca wartość 1 , gdy w domu jest garaż oraz wartość 0, gdy tego garażu nie ma.

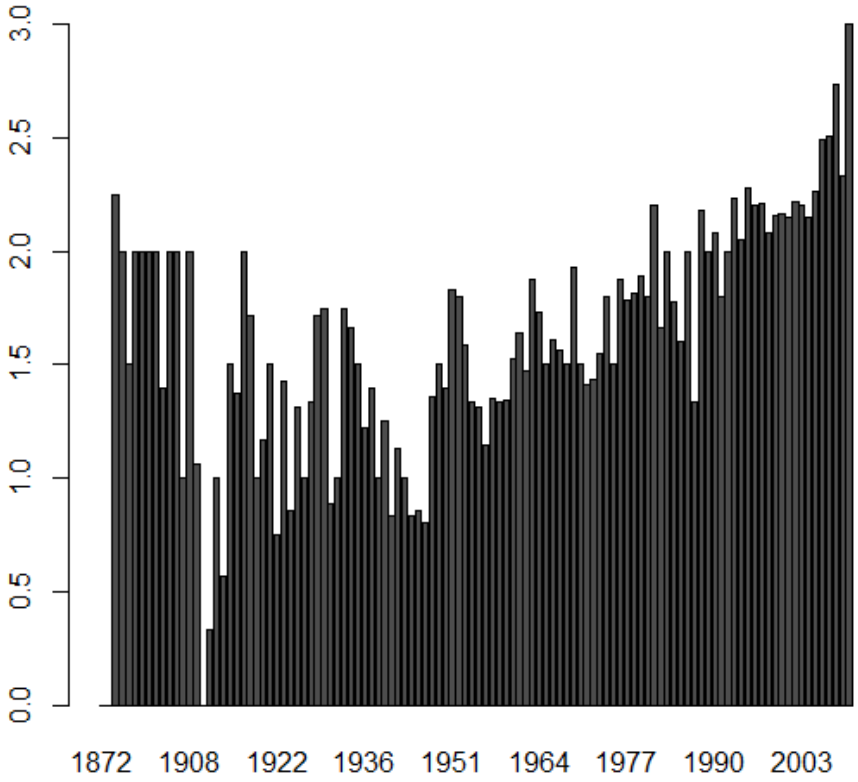


Jako wniosek z tej części – do modelu zostanie wykorzystana zmienna GarageCars mówiąca na ile samochodów jest garaż.

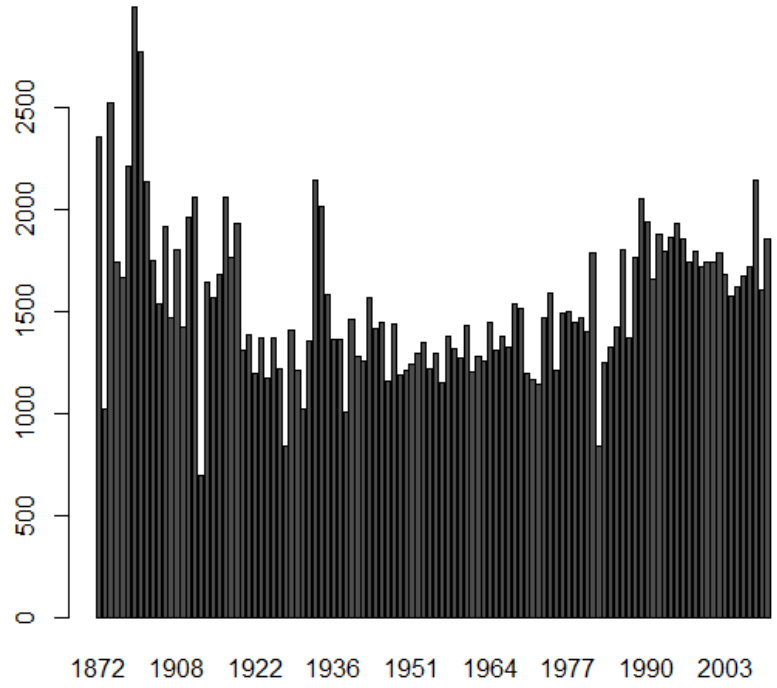
- Sprawdzenie korelacji innych (mniej znaczących zmiennych) oraz wpływu na te wybrane do modelu.



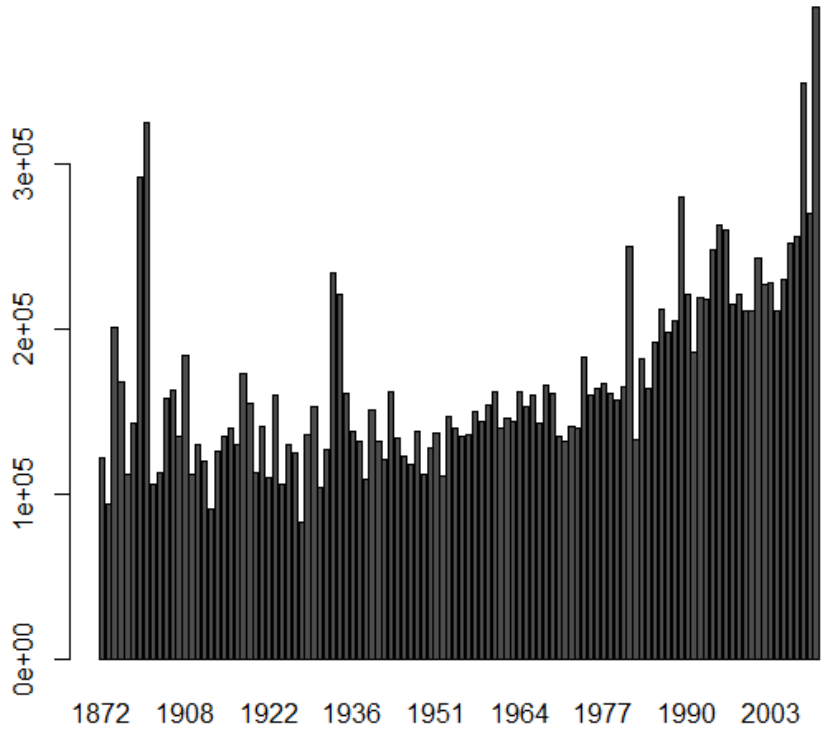
- Sprawdzenie zależności pomiędzy wybranymi zmiennymi, a ceną mieszkania/domu na wykresie słupkowym.



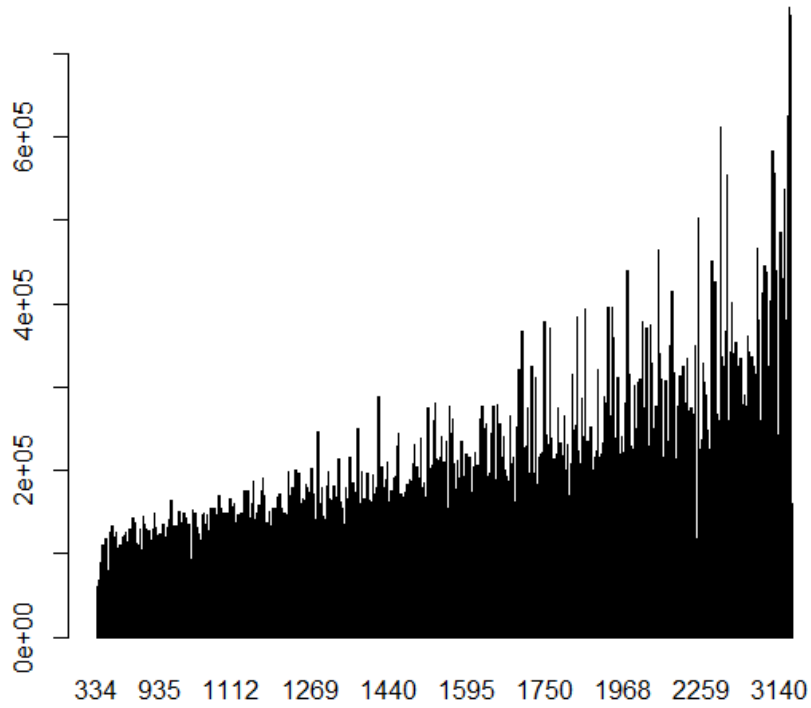
Nieznacznie więcej miejsc parkingowych jest w nowszych domach.



Brak widocznego wpływu wielkości mieszkania powyżej powierzchnią terenu, a rokiem wybudowania.

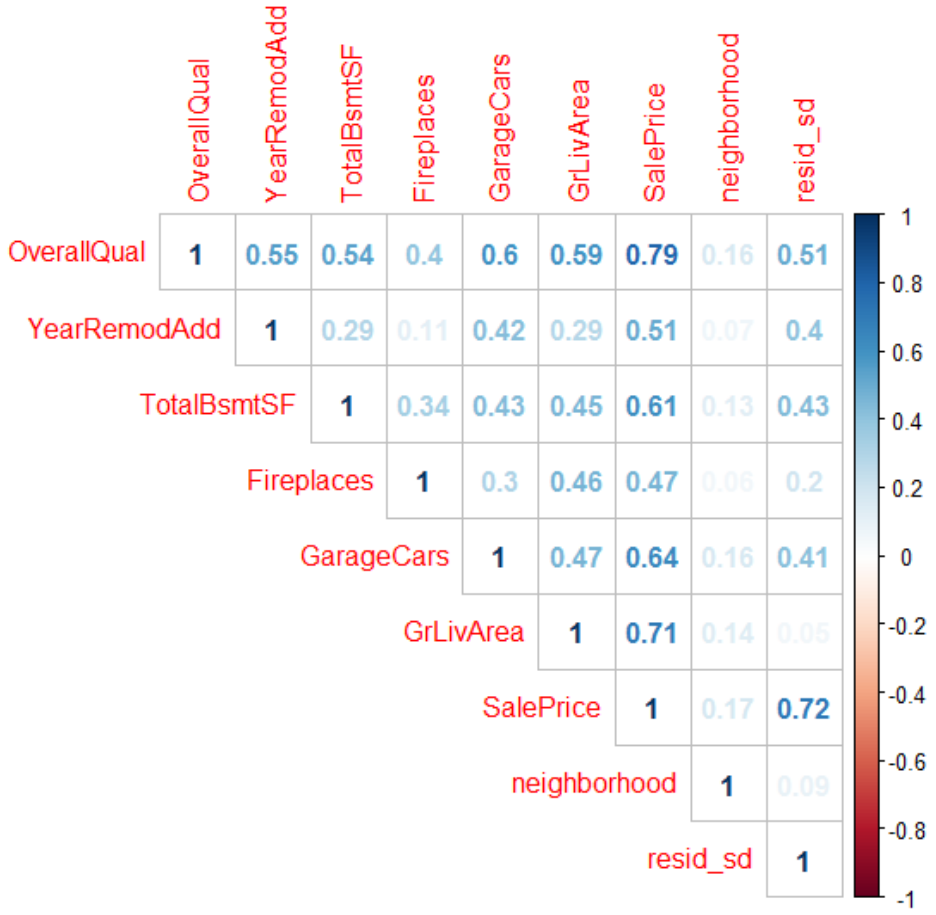


Zauważalnie droższe są nowe mieszkania (brak wpływu wielkości co może być mylące).



Widoczna mocna korelacja mówiąca o wpływie wielkości mieszkania na cenę sprzedaży.

- Ze zmiennych kategorycznych wybrano zmienną Neighborhood, która mów o okolicy (dzielnicy) w jakiej został sprzedany dom. W tym celu zamieniono wartości kategoryczne na wartości numeryczne porządkowe – w celu użycia zmiennej w modelu regresji liniowej.
- Sprawdzono wpływ wybranych zmiennych na zmienną wyjściową przy jednoczesnym wykluczeniu wpływu potencjalnie najsilniej skorelowanej zmiennej GrLivArea – korelacja zmiennych z resztami z modelu regresji liniowej zmiennej GrLivArea oraz zmiennej wyjściowej SalePrice – resid_sd.



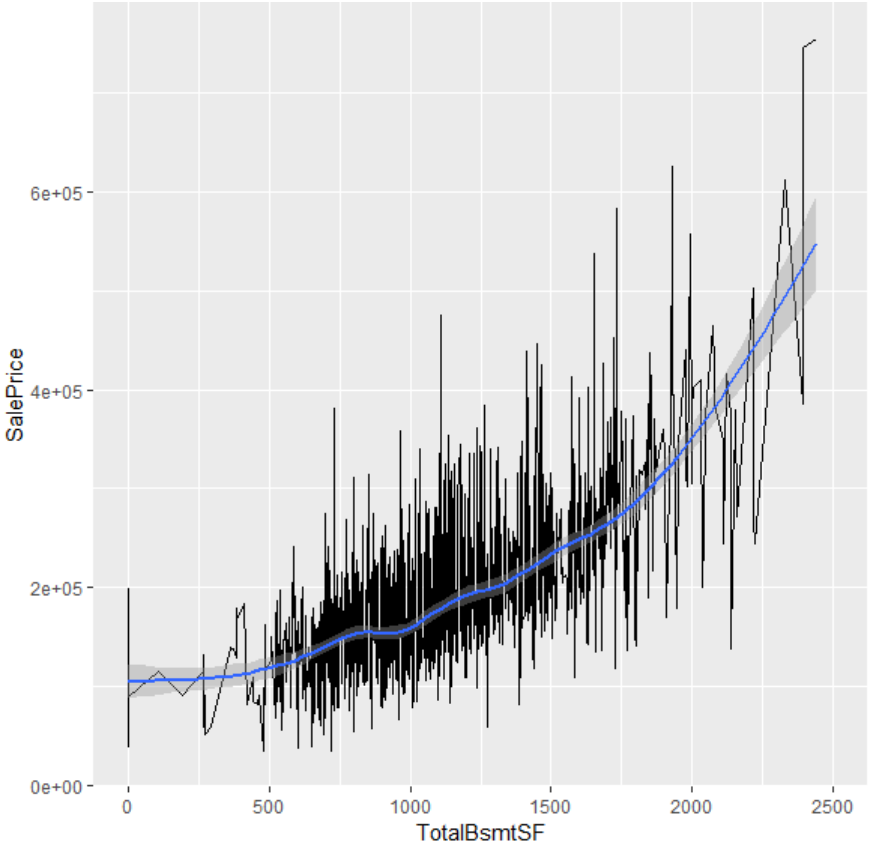
Zarówno zmienna Fireplaces jak i zmienna Neighborhood nie wykazuje dużej zależności, natomiast możliwe że jest to związane z mocnym wpływem, którejś ze zmiennych.

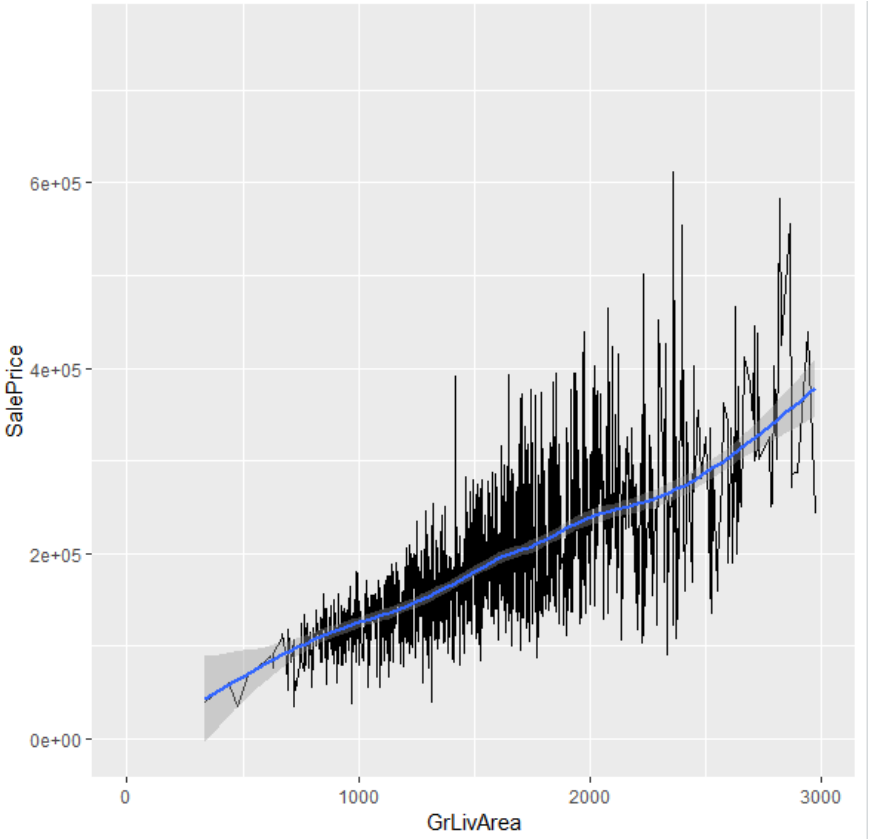
3. Przekształcenia danych

Ostatecznie do modelu oraz dalszych przekształceń zostały użyte zmienne:

- **OverallQual** – jakość domu (użyte materiały itd.).
- **YearRemodAdd** – przebudowa – remont domu.
- **TotalBsmtSF** – powierzchnia piwnicy.
- **Fireplaces** – liczba kominków.
- **GarageCars** – liczba miejsc parkingowych.
- **GrLivArea** – powierzchnia domu powyżej powierzchni terenu.
- **SalePrice** – zmienna wyjściowa.

Wykresy przedstawiające zależności między wybranymi zmiennymi a zmienną wyjściową.





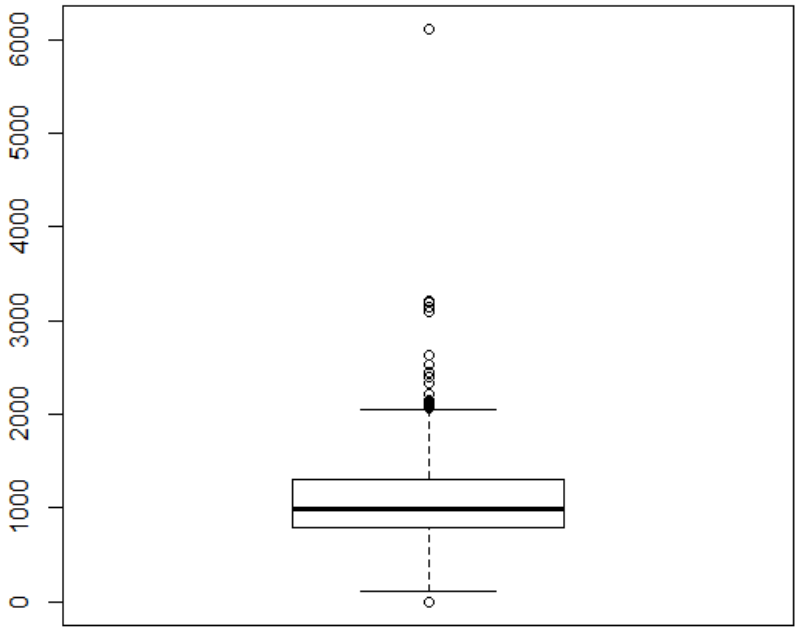
Na etapie przekształcania i transformacji zmiennych zostały wykonane następujące kroki.

- Zmiana wartości daty ostatniego remontu na liczbę lat od ostatniego remontu. Po przekształceniu statystyka zmiennej YearRemodAdd prezentuje się następująco:

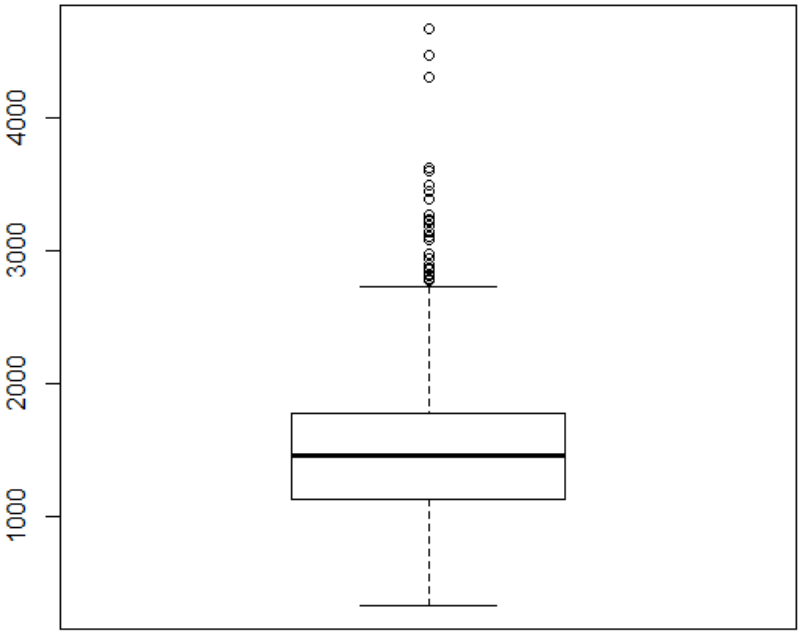
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 9.00   15.00   25.00   34.13   52.00   69.00
```

  Zauważam, że najwięcej wartości jest z roku 1950 co może świadczyć, że zawarte są tam wszystkie transakcje również z lat wcześniejszych.

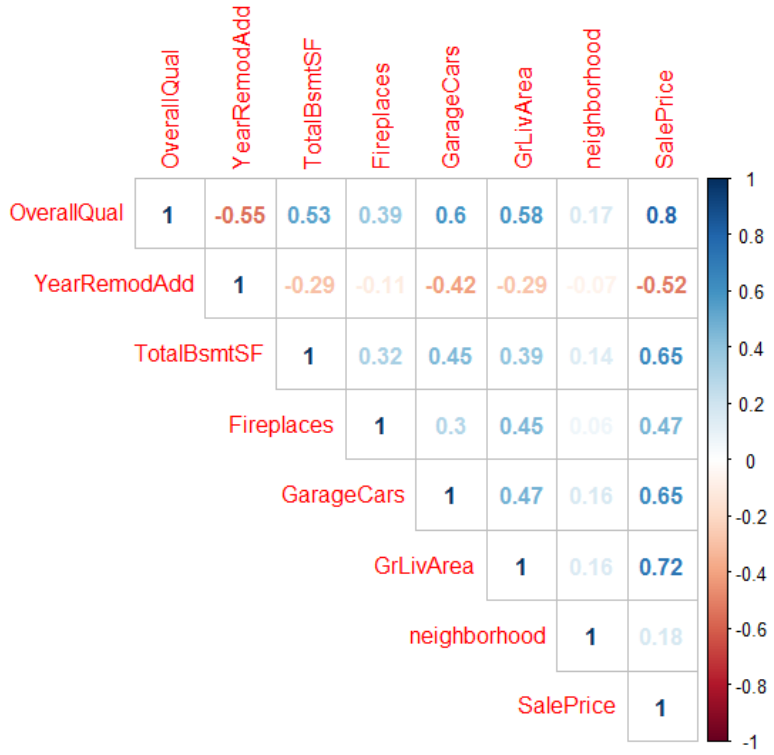
- Usunięcie wartości odstającej ze zmiennej TotalBsmtSF (wartość powyżej 5000).



- Usunięcie wartości odstających ze zmiennej GrLivArea (wartości powyżej 4000).



- Dodanie kolumny mówiącej o cenie za 1 stopę kwadratową domu – usunięcie pozycji, dla których cena jednostkowa jest niższa niż 60 i jakość wykonania (OverallQual) jest wyższa niż 8. Te wartości zostały uznane jako anomalie i mogą niekorzystnie wpływać na wyniki predykcji.
- Ostateczna korelacja między zmiennymi wziętymi do modelu



4. Model regresji liniowej

Na wstępie zbiór treningowy został podzielony w stosunku 0,85 i 0,15. Mniejszy podzbiór zostanie użyty do przetestowania modelu oraz ewentualnej optymalizacji przed sprawdzeniem modelu na danych testowych.

Model regresji liniowej prezentuje się następująco:

```
lm(formula = SalePrice ~ OverallQual + YearRemodAdd + TotalBsmtSF +
    Fireplaces + GarageCars + GrLivArea + neighborhood, data = train)

Coefficients:
 (Intercept)   OverallQual  YearRemodAdd   TotalBsmtSF    Fireplaces    GarageCars     GrLivArea  neighborhood
   -61117.34      17748.09       -462.55         42.93       9031.85      13495.56         48.76         80.38
```

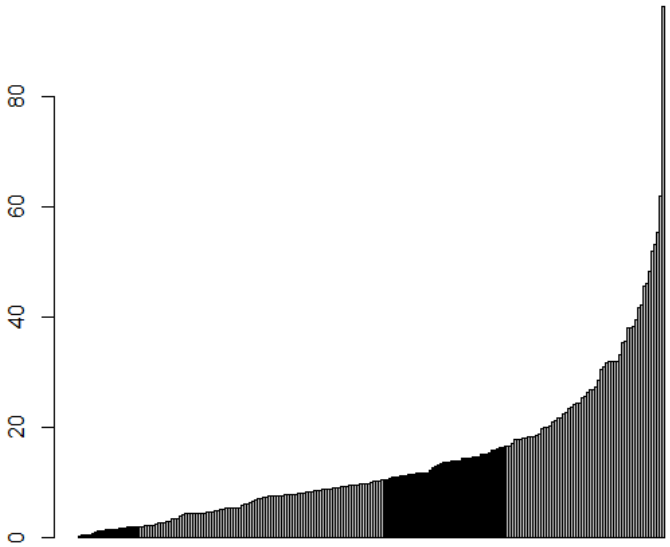
Przy sprawdzaniu modelu podzbiorem zbioru treningowego otrzymano następujące wyniki.

Średni błąd: 21842,36.

Średni błąd procentowy: 13.65944.

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.09649  5.16243  9.82609 13.65944 17.74433 96.39737
```

Procentowy rozkład błędów:



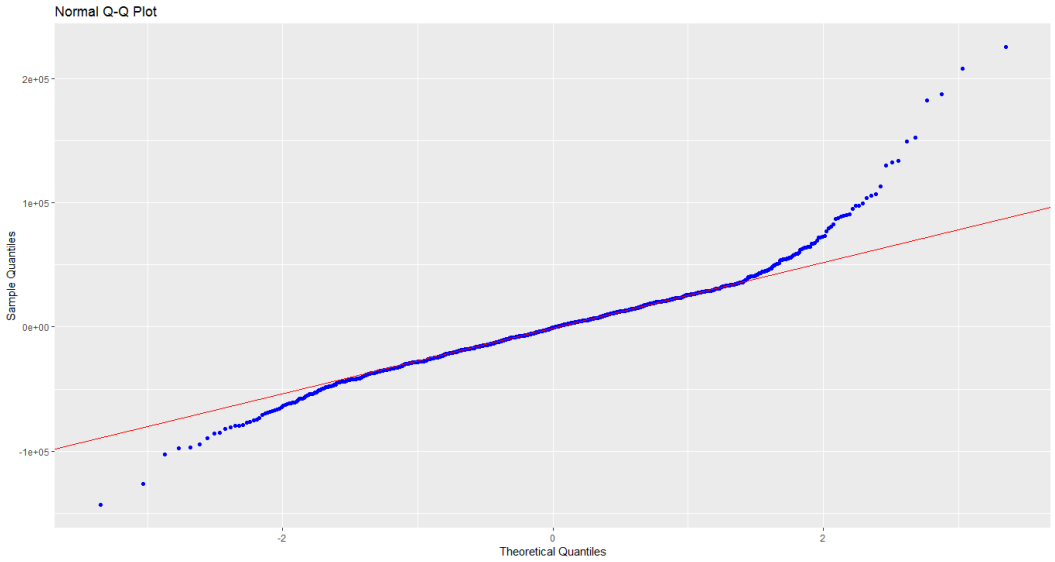
Następnie wykonano predykcje cen domów na zbiorze testowym.

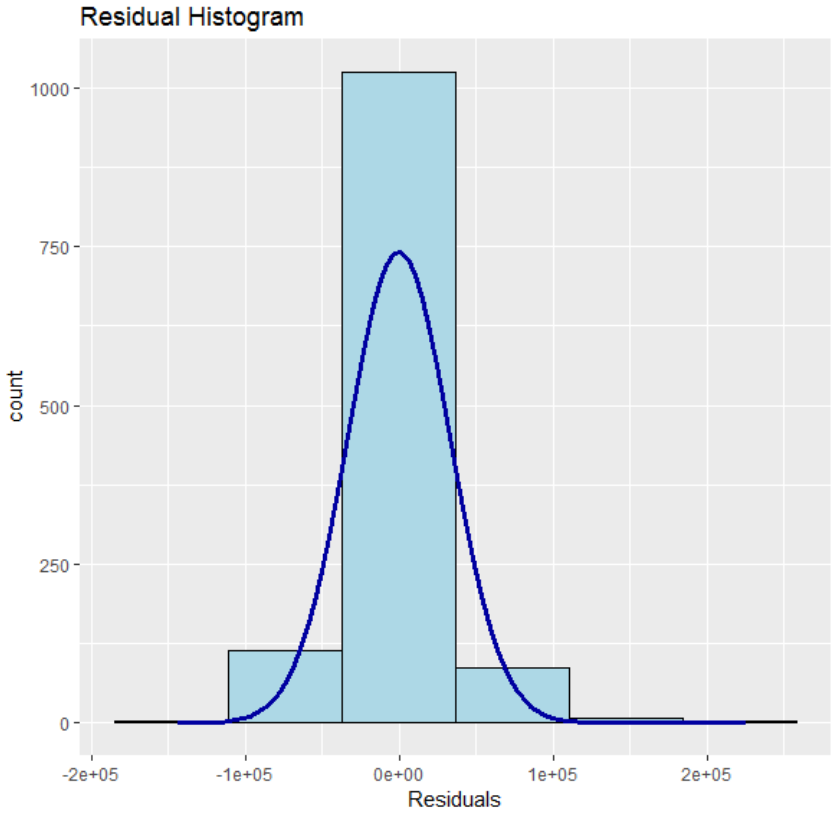
	OverallQual	YearRemodAdd	TotalBsmtSF	Fireplaces	GarageCars	GrLivArea	neighborhood	Price_Predicted
1	5	58	882	0	1	896	13	97464.596
2	6	61	1329	0	1	1329	13	154069.187
3	5	21	928	1	2	1629	9	173674.993
4	6	21	926	1	2	1604	9	190241.052
5	8	27	1280	0	2	1280	22	215098.844
6	6	25	763	1	2	1655	9	183731.385
7	6	12	1168	0	2	1187	9	175991.090
8	6	21	789	1	2	1465	9	177639.919
9	7	29	1300	1	2	1341	9	208178.530
10	4	49	882	0	2	882	13	96610.610
11	7	20	1405	1	2	1337	13	216997.741
12	6	48	483	0	1	987	3	106087.802
13	5	48	525	0	1	1092	3	95080.540
14	6	44	855	1	2	1456	15	169987.074
15	7	44	836	0	1	836	15	134864.487
16	9	9	1590	1	3	2334	16	327101.380
17	8	9	1544	0	3	1544	16	260533.088
18	9	14	1698	1	3	1698	16	299186.347
19	8	13	1822	1	3	1822	16	293118.895
20	9	15	2846	2	3	2696	16	405411.915

Po sprawdzeniu wyników dla modelu regresji liniowej bez zmiennej neighborhood średni błąd jest nieznacznie wyższy i wynosi: 21860,5.

5. Wnioski

- Projekt Analiza, transformacja oraz zaprojektowanie modelu regresji liniowej przewidującego ceny domów w USA wykonano przy użyciu języka R w środowisku R Studio.
- Wykonano analizę oraz wizualizacje danych, transformacje danych oraz model predykcyjny regresji liniowej.
- Średni błąd predykcji wyniósł mniej niż 14%. Mediana błędów wyniosła mniej niż 10%.
- Wykres Q-Q oraz rozkład błędów wygląda następująco
-





- Reszty z modelu nie pozostają w autokorelacji – co zostało sprawdzone testem Durbina – Watsona z wykorzystaniem biblioteki car.
- Usunięcie zmiennej kategorycznej Neighborhood z modelu nieznacznie obniża precyzję predykcji.
- Uproszczenie modelu przez wybranie tylko trzech zmiennych (potencjalnie najsilniej skorelowanych ze zmienną wyjściową) OverallQual, TotalBsmtSF, GrLivArea zwiększa błąd predykcji o około 2000.
- W celu poprawy modelu można zastosować metodę krzyżowej walidacji.