Examining Russian Realty Prices

Springboard Capstone Project

Dylan Distasio

May 24, 2017

Introduction

This capstone project is based on a Kaggle competition hosted by Sberbank. The full details from Kaggle are as follows:

Housing costs demand a significant investment from both consumers and developers. And when it comes to planning a budget-whether personal or corporate-the last thing anyone needs is uncertainty about one of their biggets expenses. Sberbank, Russia's oldest and largest bank, helps their customers by making predictions about realty prices so renters, developers, and lenders are more confident when they sign a lease or purchase a building.

Although the housing market is relatively stable in Russia, the country's volatile economy makes forecasting prices as a function of apartment characteristics a unique challenge. Complex interactions between housing features such as number of bedrooms and location are enough to make pricing predictions complicated. Adding an unstable economy to the mix means Sberbank and their customers need more than simple regression models in their arsenal.

In this competition, Sberbank is challenging Kagglers to develop algorithms which use a broad spectrum of features to predict realty prices. Competitors will rely on a rich dataset that includes housing data and macroeconomic patterns. An accurate forecasting model will allow Sberbank to provide more certainty to their customers in an uncertain economy.

Overall approach

Exploratory Data Analysis, feature engineering, and model building was performed to construct a linear regression model to attempt to predict realty price fluctuations in Russia based on the provided data.

Data information

The client, Sberbank, provided a number of files related to Russian realty prices:

- train.csv - a training dataset containing 30,471 observations with 292 potential features related to the Russian realty market including sales price data provided.
- test.csv - a dataset to test the predictive model on containing 7,662 observations with 291 (sales price removed) of the same features of the training dataset.
- macro.csv - a dataset consisting of 2,484 historical records with 100 macroeconomic features of potential interest.
- data_dictionary.txt - a text file containing a data dictionary for features in the training, test, and macroeconomic datasets.

Exploratory Data Analysis

The training, test, and macroeconomic datasets were read into dataframes in R.

The training and macroeconomic dataframes were joined together based on the timestamp field common to both for further analyis.

Initial findings

- An erronenous outlier in the feature related to property condition was removed as part of data cleanup.
- A histogram of historical sales prices is positively skewed with a long tail.
- Initial correlation analysis of the features selected regarding the property shows the strongest correlation between the square footage of the property, and the sales price.
- Initial correlation analysis of the features selected related to macroeconomic factors shows weak correlations with the sales price.
- Initial review of average sales price based on subregion of Russia revealed substantial differences in sales prices between regions. This pointed towards the important of this feature for incorporation into models.
- There may be a time lag involved with the response of the sales price to macroeconomic factors which may be difficult to capture in many basic models.
- There are a number of potential features with a large % of missing values that may need to be massaged if used for building a model. These were visualized, and counts of missing values were also generated.
- Average sales prices showed a rising trend throughout the time period captured in the training dataset which may make any model more brittle at predicting future prices.
- There appeared to be seasonality in prices based on time of year. This may be a worthwhile avenue of investigation in building later, more sophisticated predictive models.
- Property condition was also examined as a feature. The majority of properties were in moderately good condition (2-3) in this dataset with a small number of properties in the best condition on a relative basis. Visually plotting property condition versus sales price suggested higher average sales prices as property condition improved.

Model Construction

Linear regression was selected as an initial model for a number of reasons including: * It is a relatively simple model to understand and explain * It can be relatively successful at predicting a dependent continuous variable depending on how linear the relationship is with the independent ones.

Initial linear regression model (dwelling features only)

The initial linear regression model constructed was based on a subset of dwelling related features originally found in the training dataset.

The features incorporated along with their impact on the model are below:

```
Call:
lm(formula = price_doc ~ ., data = comb[home])
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-87894791  -1331623    193956   1407975  61674251

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.601e+06  1.893e+05 -19.020  < 2e-16 ***
full_sq      1.474e+05  2.236e+03  65.919  < 2e-16 ***
life_sq     -9.756e+02  5.038e+02  -1.937   0.0528 .
floor        4.356e+04  7.383e+03   5.900 3.72e-09 ***
max_floor    6.514e+03  6.277e+03   1.038   0.2994
build_year   4.421e+02  8.372e+01   5.280 1.31e-07 ***
num_room     1.048e+05  5.318e+04   1.971   0.0487 *
kitch_sq    -7.187e+02  1.075e+03  -0.668   0.5040
state        9.372e+05  4.164e+04  22.508  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3812000 on 14806 degrees of freedom
  (2096 observations deleted due to missingness)
Multiple R-squared:  0.4306,    Adjusted R-squared:  0.4303
F-statistic:  1400 on 8 and 14806 DF,  p-value: < 2.2e-16
```

Second iteration of linear regression model (dwelling+macroeconomic features)

A second linear regression model was constructed incorporating a subset of macroeconomic features added to the existing dwelling based model.

This only had a slight effect on adjusted-R square, which was not surprising based on earlier correlation matrix analysis.

```
Call:
lm(formula = price_doc ~ ., data = comb[modelfeatures])

Residuals:
      Min        1Q    Median        3Q       Max
-89222498  -1322586    224859   1381799  51723569

Coefficients: (3 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.523e+05  6.052e+06   0.058   0.9536
full_sq            1.498e+05  2.391e+03  62.632  < 2e-16 ***
life_sq           -8.048e+02  4.881e+02  -1.649   0.0992 .
floor              4.105e+04  7.774e+03   5.281 1.31e-07 ***
max_floor         -9.880e+02  6.606e+03  -0.150   0.8811
build_year        -1.338e+00  1.952e-01  -6.857 7.36e-12 ***
num_room          -6.506e+03  5.600e+04  -0.116   0.9075
kitch_sq           2.713e+01  1.252e+03   0.022   0.9827
state              1.005e+06  4.318e+04  23.278  < 2e-16 ***
gdp_quart         -1.505e+01  1.120e+02  -0.134   0.8930
gdp_quart_growth  -4.488e+04  2.901e+05  -0.155   0.8770
cpi               -1.557e+03  1.716e+04  -0.091   0.9277
ppi               -1.921e+03  8.419e+03  -0.228   0.8195
usdrub             6.667e+04  8.175e+04   0.815   0.4148
```

```
eurrub            -4.993e+04  6.013e+04  -0.830   0.4063
brent             -2.760e+03  1.364e+04  -0.202   0.8397
gdp_annual         6.582e+01  6.078e+01   1.083   0.2788
gdp_annual_growth         NA         NA      NA       NA
micex_rgbi_tr     -1.830e+04  1.957e+04  -0.935   0.3497
mortgage_value    -2.637e-03  1.550e-01  -0.017   0.9864
mortgage_rate     -2.651e+05  2.783e+05  -0.953   0.3408
grp                       NA         NA      NA       NA
grp_growth                NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3680000 on 12482 degrees of freedom
  (17969 observations deleted due to missingness)
Multiple R-squared:  0.4368,    Adjusted R-squared:  0.436
F-statistic: 509.6 on 19 and 12482 DF,  p-value: < 2.2e-16
```

Final iteration of linear regression model (dwelling+macroeconomic+region features)

The final iteration of the linear regression model combined dwelling, macroeconomic, and sub region related information.

Adding in one region related feature significantly increased the adjusted R-square to 0.5595.

```
Call:
lm(formula = price_doc ~ ., data = comb[modelfeatures])

Residuals:
      Min        1Q    Median        3Q       Max
-88295064   -642791    482513   1231095  48461879

Coefficients: (147 not defined because of singularities)
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -1.736e+06  5.392e+06  -0.322 0.747472
full_sq                           1.465e+05  2.222e+03  65.915  < 2e-16 ***
life_sq                          -5.357e+02  4.330e+02  -1.237 0.216005
floor                             4.412e+04  6.994e+03   6.308 2.92e-10 ***
max_floor                         2.891e+04  6.280e+03   4.603 4.21e-06 ***
build_year                       -2.670e+02  8.725e+01  -3.061 0.002213 **
num_room                         -1.768e+05  5.092e+04  -3.473 0.000516 ***
kitch_sq                          6.191e+02  1.111e+03   0.557 0.577471
state                             5.478e+05  4.449e+04  12.313  < 2e-16 ***
gdp_quart                         4.904e+01  9.954e+01   0.493 0.622229
gdp_quart_growth                 -1.494e+05  2.579e+05  -0.579 0.562490
cpi                               4.018e+03  1.526e+04   0.263 0.792313
ppi                               1.593e+02  7.485e+03   0.021 0.983025
usdrub                           -7.633e+02  7.271e+04  -0.010 0.991624
eurrub                            1.437e+04  5.348e+04   0.269 0.788101
brent                            -2.546e+03  1.213e+04  -0.210 0.833721
gdp_annual                        4.446e+01  5.402e+01   0.823 0.410492
gdp_annual_growth                        NA         NA      NA       NA
micex_rgbi_tr                    -6.666e+03  1.737e+04  -0.384 0.701163
mortgage_value                   -3.746e-02  1.379e-01  -0.272 0.785929
```

mortgage_rate	-2.216e+05	2.473e+05	-0.896	0.370261	
grp	NA	NA	NA	NA	
grp_growth	NA	NA	NA	NA	
sub_areaAkademicheskoe	8.681e+05	5.078e+05	1.709	0.087390	.
sub_areaAlekseevskoe	2.273e+04	6.269e+05	0.036	0.971075	
sub_areaAltuf'evskoe	-2.721e+06	6.959e+05	-3.910	9.30e-05	***
sub_areaArbat	8.040e+06	1.510e+06	5.325	1.03e-07	***
sub_areaBabushkinskoe	-1.340e+06	6.158e+05	-2.176	0.029574	*
sub_areaBasmannoe	1.332e+06	6.124e+05	2.175	0.029667	*
sub_areaBegovoe	1.222e+06	7.185e+05	1.700	0.089136	.
sub_areaBeskudnikovskoe	-2.590e+06	5.299e+05	-4.887	1.04e-06	***
sub_areaBibirevo	-2.278e+06	5.071e+05	-4.491	7.14e-06	***
sub_areaBirjulevo Vostochnoe	-3.137e+06	4.826e+05	-6.500	8.33e-11	***
sub_areaBirjulevo Zapadnoe	-3.347e+06	5.630e+05	-5.945	2.84e-09	***
sub_areaBogorodskoe	-2.037e+06	4.749e+05	-4.290	1.80e-05	***
sub_areaBrateevo	-2.956e+06	5.125e+05	-5.767	8.27e-09	***
sub_areaButyrskoe	-1.043e+06	6.024e+05	-1.731	0.083530	.
sub_areaCaricyno	-1.927e+06	4.922e+05	-3.914	9.11e-05	***
sub_areaCheremushki	4.639e+05	5.451e+05	0.851	0.394760	
sub_areaChertanovo Central'noe	-1.510e+06	5.266e+05	-2.867	0.004149	**
sub_areaChertanovo Juzhnoe	-2.291e+06	5.017e+05	-4.566	5.02e-06	***
sub_areaChertanovo Severnoe	-1.550e+06	5.288e+05	-2.930	0.003394	**
sub_areaDanilovskoe	-5.817e+05	5.153e+05	-1.129	0.258929	
sub_areaDmitrovskoe	-2.629e+06	5.209e+05	-5.046	4.58e-07	***
sub_areaDonskoe	-6.517e+05	5.910e+05	-1.103	0.270223	
sub_areaDorogomilovo	3.188e+06	7.028e+05	4.536	5.79e-06	***
sub_areaFilevskij Park	-6.565e+05	5.803e+05	-1.131	0.257917	
sub_areaFili Davydkovo	9.436e+04	5.666e+05	0.167	0.867741	
sub_areaGagarinskoe	2.497e+06	6.443e+05	3.876	0.000107	***
sub_areaGol'janovo	-2.548e+06	4.726e+05	-5.392	7.10e-08	***
sub_areaGolovinskoe	-2.275e+06	4.972e+05	-4.575	4.81e-06	***
sub_areaHamovniki	8.175e+06	6.355e+05	12.863	< 2e-16	***
sub_areaHoroshevo-Mnevniki	-1.487e+06	4.873e+05	-3.051	0.002284	**
sub_areaHoroshevskoe	1.494e+06	5.765e+05	2.591	0.009578	**
sub_areaHovrino	-1.481e+06	5.259e+05	-2.817	0.004859	**
sub_areaIvanovskoe	-2.566e+06	4.928e+05	-5.207	1.95e-07	***
sub_areaIzmajlovo	-1.469e+06	4.847e+05	-3.032	0.002437	**
sub_areaJakimanka	6.141e+05	6.611e+05	0.929	0.352934	
sub_areaJaroslavskoe	-2.117e+06	5.754e+05	-3.680	0.000235	***
sub_areaJasenevo	-1.694e+06	5.042e+05	-3.360	0.000781	***
sub_areaJuzhnoe Butovo	-2.880e+06	4.537e+05	-6.347	2.27e-10	***
sub_areaJuzhnoe Medvedkovo	-2.692e+06	5.861e+05	-4.593	4.41e-06	***
sub_areaJuzhnoe Tushino	-1.630e+06	5.177e+05	-3.148	0.001647	**
sub_areaJuzhnoportovoe	-5.068e+05	5.533e+05	-0.916	0.359735	
sub_areaKapotnja	-3.344e+06	7.898e+05	-4.235	2.31e-05	***
sub_areaKon'kovo	7.520e+03	5.052e+05	0.015	0.988125	
sub_areaKoptevo	-1.626e+06	4.957e+05	-3.280	0.001039	**
sub_areaKosino-Uhtomskoe	-3.328e+06	4.865e+05	-6.841	8.26e-12	***
sub_areaKotlovka	-1.108e+06	5.418e+05	-2.044	0.040959	*
sub_areaKrasnosel'skoe	1.070e+06	8.669e+05	1.234	0.217276	
sub_areaKrjukovo	-4.156e+06	4.655e+05	-8.928	< 2e-16	***
sub_areaKrylatskoe	1.070e+06	6.172e+05	1.734	0.082911	.
sub_areaKuncevo	-1.007e+06	5.132e+05	-1.961	0.049884	*
sub_areaKurkino	-1.325e+06	8.061e+05	-1.644	0.100128	

```
sub_areaKuz'minki                          -1.787e+06  4.976e+05  -3.592 0.000329 ***
sub_areaLefortovo                          -1.060e+06  5.964e+05  -1.778 0.075423 .
sub_areaLevoberezhnoe                      -1.698e+06  5.719e+05  -2.970 0.002989 **
sub_areaLianozovo                          -2.635e+06  6.451e+05  -4.084 4.45e-05 ***
sub_areaLjublino                           -2.302e+06  4.731e+05  -4.866 1.15e-06 ***
sub_areaLomonosovskoe                       3.239e+06  5.593e+05   5.791 7.16e-09 ***
sub_areaLosinoostrovskoe                   -1.629e+06  5.450e+05  -2.989 0.002804 **
sub_areaMar'ina Roshha                     -4.410e+05  6.055e+05  -0.728 0.466390
sub_areaMar'ino                            -2.513e+06  4.460e+05  -5.636 1.78e-08 ***
sub_areaMarfino                            -9.106e+05  6.491e+05  -1.403 0.160727
sub_areaMatushkino                         -3.714e+06  5.647e+05  -6.577 5.00e-11 ***
sub_areaMeshhanskoe                         2.288e+06  6.398e+05   3.576 0.000350 ***
sub_areaMetrogorodok                       -1.953e+06  6.315e+05  -3.092 0.001994 **
sub_areaMitino                             -1.656e+06  4.534e+05  -3.652 0.000261 ***
sub_areaMolzhaninovskoe                    -1.071e+07  2.337e+06  -4.583 4.62e-06 ***
sub_areaMoskvorech'e-Saburovo              -2.098e+06  5.921e+05  -3.544 0.000396 ***
sub_areaMozhajskoe                         -1.841e+06  5.043e+05  -3.650 0.000263 ***
sub_areaNagatino-Sadovniki                 -1.403e+06  5.377e+05  -2.608 0.009111 **
sub_areaNagatinskij Zaton                  -7.875e+05  4.999e+05  -1.575 0.115259
sub_areaNagornoe                           -3.269e+05  5.027e+05  -0.650 0.515461
sub_areaNekrasovka                         -4.065e+06  4.469e+05  -9.094  < 2e-16 ***
sub_areaNizhegorodskoe                     -2.088e+06  6.351e+05  -3.288 0.001013 **
sub_areaNovo-Peredelkino                   -3.462e+06  5.682e+05  -6.094 1.13e-09 ***
sub_areaNovogireevo                        -1.933e+06  5.157e+05  -3.749 0.000178 ***
sub_areaNovokosino                         -2.824e+06  5.297e+05  -5.331 9.95e-08 ***
sub_areaObruchevskoe                        1.012e+06  5.624e+05   1.798 0.072126 .
sub_areaOchakovo-Matveevskoe               -2.212e+06  5.324e+05  -4.154 3.28e-05 ***
sub_areaOrehovo-Borisovo Juzhnoe           -2.125e+06  5.160e+05  -4.118 3.85e-05 ***
sub_areaOrehovo-Borisovo Severnoe          -2.065e+06  5.232e+05  -3.947 7.97e-05 ***
sub_areaOstankinskoe                       -8.324e+04  6.890e+05  -0.121 0.903850
sub_areaOtradnoe                           -1.958e+06  4.804e+05  -4.075 4.63e-05 ***
sub_areaPechatniki                         -2.294e+06  5.141e+05  -4.462 8.19e-06 ***
sub_areaPerovo                             -2.223e+06  4.822e+05  -4.609 4.09e-06 ***
sub_areaPokrovskoe Streshnevo              -5.937e+05  5.648e+05  -1.051 0.293245
sub_areaPoselenie Desjonovskoe             -4.085e+06  5.245e+05  -7.788 7.34e-15 ***
sub_areaPoselenie Filimonkovskoe           -4.263e+06  4.906e+05  -8.690  < 2e-16 ***
sub_areaPoselenie Kievskij                 -6.853e+06  3.280e+06  -2.089 0.036695 *
sub_areaPoselenie Kokoshkino               -7.643e+06  1.922e+06  -3.978 7.00e-05 ***
sub_areaPoselenie Krasnopahorskoe          -5.235e+06  1.024e+06  -5.110 3.26e-07 ***
sub_areaPoselenie Marushkinskoe            -4.929e+06  1.676e+06  -2.941 0.003280 **
sub_areaPoselenie Mihajlovo-Jarcevskoe     -5.303e+06  3.278e+06  -1.618 0.105723
sub_areaPoselenie Moskovskij               -5.131e+06  4.996e+05 -10.271  < 2e-16 ***
sub_areaPoselenie Mosrentgen               -3.979e+06  1.157e+06  -3.438 0.000589 ***
sub_areaPoselenie Novofedorovskoe          -5.667e+06  6.866e+05  -8.254  < 2e-16 ***
sub_areaPoselenie Pervomajskoe             -7.184e+06  6.700e+05 -10.722  < 2e-16 ***
sub_areaPoselenie Rjazanovskoe             -5.370e+06  1.023e+06  -5.251 1.54e-07 ***
sub_areaPoselenie Rogovskoe                -1.205e+07  1.399e+06  -8.614  < 2e-16 ***
sub_areaPoselenie Shhapovskoe              -7.879e+06  3.279e+06  -2.403 0.016288 *
sub_areaPoselenie Shherbinka               -4.747e+06  5.196e+05  -9.134  < 2e-16 ***
sub_areaPoselenie Sosenskoe                -4.445e+06  4.545e+05  -9.780  < 2e-16 ***
sub_areaPoselenie Vnukovskoe               -3.873e+06  4.763e+05  -8.133 4.60e-16 ***
sub_areaPoselenie Voronovskoe              -4.616e+06  1.510e+06  -3.057 0.002239 **
sub_areaPoselenie Voskresenskoe            -3.539e+06  4.976e+05  -7.113 1.20e-12 ***
sub_areaPreobrazhenskoe                    -1.499e+06  5.392e+05  -2.780 0.005447 **
```

```
sub_areaPresnenskoe                    3.642e+06  5.198e+05   7.006 2.58e-12 ***
sub_areaProspekt Vernadskogo           9.205e+05  6.162e+05   1.494 0.135214
sub_areaRamenki                        1.383e+06  5.053e+05   2.737 0.006218 **
sub_areaRjazanskij                    -2.008e+06  5.078e+05  -3.955 7.71e-05 ***
sub_areaRostokino                     -1.050e+06  7.028e+05  -1.494 0.135246
sub_areaSavelki                       -3.975e+06  5.756e+05  -6.905 5.25e-12 ***
sub_areaSavelovskoe                   -2.344e+05  6.230e+05  -0.376 0.706714
sub_areaSevernoe                      -3.995e+06  9.603e+05  -4.160 3.20e-05 ***
sub_areaSevernoe Butovo               -2.201e+06  5.220e+05  -4.217 2.49e-05 ***
sub_areaSevernoe Izmajlovo            -1.873e+06  5.299e+05  -3.534 0.000410 ***
sub_areaSevernoe Medvedkovo           -1.731e+06  5.471e+05  -3.164 0.001557 **
sub_areaSevernoe Tushino              -1.063e+06  4.977e+05  -2.136 0.032692 *
sub_areaShhukino                      -3.225e+05  5.197e+05  -0.621 0.534927
sub_areaSilino                        -4.527e+06  5.822e+05  -7.775 8.12e-15 ***
sub_areaSokol                         -6.535e+05  6.450e+05  -1.013 0.310976
sub_areaSokol'niki                     1.877e+06  6.644e+05   2.826 0.004725 **
sub_areaSokolinaja Gora               -1.253e+06  4.970e+05  -2.520 0.011734 *
sub_areaSolncevo                      -4.006e+06  5.169e+05  -7.749 9.97e-15 ***
sub_areaStaroe Krjukovo               -4.399e+06  6.197e+05  -7.098 1.33e-12 ***
sub_areaStrogino                      -3.096e+05  4.833e+05  -0.641 0.521750
sub_areaSviblovo                      -1.309e+06  6.125e+05  -2.137 0.032605 *
sub_areaTaganskoe                      1.424e+06  5.232e+05   2.721 0.006522 **
sub_areaTekstil'shhiki                -1.732e+06  5.052e+05  -3.429 0.000607 ***
sub_areaTeplyj Stan                   -1.155e+06  5.590e+05  -2.066 0.038888 *
sub_areaTimirjazevskoe                -8.302e+05  5.322e+05  -1.560 0.118780
sub_areaTroickij okrug                -5.882e+06  6.101e+05  -9.641  < 2e-16 ***
sub_areaTroparevo-Nikulino            -9.051e+05  6.171e+05  -1.467 0.142492
sub_areaTverskoe                      -1.838e+06  4.860e+05  -3.782 0.000156 ***
sub_areaVeshnjaki                     -2.461e+06  4.972e+05  -4.950 7.53e-07 ***
sub_areaVnukovo                       -4.012e+06  8.861e+05  -4.527 6.03e-06 ***
sub_areaVojkovskoe                    -9.119e+05  5.852e+05  -1.558 0.119172
sub_areaVostochnoe                    -2.681e+06  1.510e+06  -1.775 0.075907 .
sub_areaVostochnoe Degunino           -2.202e+06  5.651e+05  -3.897 9.79e-05 ***
sub_areaVostochnoe Izmajlovo          -1.837e+06  5.262e+05  -3.492 0.000482 ***
sub_areaVyhino-Zhulebino              -2.337e+06  4.765e+05  -4.905 9.45e-07 ***
sub_areaZamoskvorech'e                 4.719e+06  7.658e+05   6.162 7.42e-10 ***
sub_areaZapadnoe Degunino             -3.694e+06  5.018e+05  -7.361 1.94e-13 ***
sub_areaZjablikovo                    -2.375e+06  5.833e+05  -4.072 4.68e-05 ***
sub_areaZjuzino                       -8.894e+05  4.932e+05  -1.804 0.071331 .
sub_area.1Akademicheskoe                      NA         NA      NA       NA
sub_area.1Alekseevskoe                        NA         NA      NA       NA
sub_area.1Altuf'evskoe                        NA         NA      NA       NA
sub_area.1Arbat                               NA         NA      NA       NA
sub_area.1Babushkinskoe                       NA         NA      NA       NA
sub_area.1Basmannoe                           NA         NA      NA       NA
sub_area.1Begovoe                             NA         NA      NA       NA
sub_area.1Beskudnikovskoe                     NA         NA      NA       NA
sub_area.1Bibirevo                            NA         NA      NA       NA
sub_area.1Birjulevo Vostochnoe                NA         NA      NA       NA
sub_area.1Birjulevo Zapadnoe                  NA         NA      NA       NA
sub_area.1Bogorodskoe                         NA         NA      NA       NA
sub_area.1Brateevo                            NA         NA      NA       NA
sub_area.1Butyrskoe                           NA         NA      NA       NA
sub_area.1Caricyno                            NA         NA      NA       NA
```

```
sub_area.1Cheremushki                              NA         NA      NA       NA
sub_area.1Chertanovo Central'noe                   NA         NA      NA       NA
sub_area.1Chertanovo Juzhnoe                       NA         NA      NA       NA
sub_area.1Chertanovo Severnoe                      NA         NA      NA       NA
sub_area.1Danilovskoe                              NA         NA      NA       NA
sub_area.1Dmitrovskoe                              NA         NA      NA       NA
sub_area.1Donskoe                                  NA         NA      NA       NA
sub_area.1Dorogomilovo                             NA         NA      NA       NA
sub_area.1Filevskij Park                           NA         NA      NA       NA
sub_area.1Fili Davydkovo                           NA         NA      NA       NA
sub_area.1Gagarinskoe                              NA         NA      NA       NA
sub_area.1Gol'janovo                               NA         NA      NA       NA
sub_area.1Golovinskoe                              NA         NA      NA       NA
sub_area.1Hamovniki                                NA         NA      NA       NA
sub_area.1Horoshevo-Mnevniki                       NA         NA      NA       NA
sub_area.1Horoshevskoe                             NA         NA      NA       NA
sub_area.1Hovrino                                  NA         NA      NA       NA
sub_area.1Ivanovskoe                               NA         NA      NA       NA
 [ reached getOption("max.print") -- omitted 111 rows ]
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3252000 on 12337 degrees of freedom
  (4410 observations deleted due to missingness)
Multiple R-squared:  0.5652,    Adjusted R-squared:  0.5595
F-statistic: 98.39 on 163 and 12337 DF,  p-value: < 2.2e-16
```

Findings

Based on a linear regression model utilizing a subset of dwelling, macroeconomic, and region related features, it was difficult to obtain a highly predictive model of the training data as measured by adjusted R-squared.

However, a relatively large number of potential features had a significant amount of missing data that may need to be addressed to improve model performance. Additional data cleaning and feature incorporation may lead to better adjusted R-Squared results.

Due to a rank deficiency issue with the model, predictive capability on the test data set was unable to be tested. This could be due to a number of factors including an inadequate sample size for the model generated, or that multiple features were not linearly independent.

Future Work

- Continue to explore the data for additional cleaning, and address missing values. One possibility that may result in improved model performance would be removing additional obvious outliers, and populating missing values with either the mean or some other interpolated value(s).
- Examine additional features for inclusion into a new model. There are a large number of remaining features in the dataset that may prove useful in prediction, but due to the limited scope of this exercise, they were excluded from the feature engineering process.
- There may also be the opportunity to do additional feature engineering to remove features that are highly correlated with each other. Doing so may alleviate the issue with the previously built linear regression models failing on the test dataset due to a rank deficiency error.

- Investigate potential lag between macroeconomic indicators and sales prices to engineer features that may be more predictive. Using time series data as a forecasting tool can be difficult, especially when there may be different lag between various macroeconomic features and the sales price.
- Examine additional model types that may prove a better solution than linear regression in predicting sales prices. Logistical regression is one related possibility. There are a whole host of other available models to explore also. Ultimately, an ensemble approach may yield the best results.