# Sberbank Presentation

John, Jason, Fouad, Chris

# Outline

- Data Quality
    - From bad and missing data to imputation
- Exploratory data analysis
    - Feature Selection
- Modeling
    - Feature engineering
    - Multiple Linear Regression
    - Random Forest
    - Stacking
- Future Direction

# Correcting bad data

- Outliers either became NAs or more realistic values
  - floor = 77
  - state = 33 became 3, since state only ranges from 1 to 4
- Some values were entered for the wrong variable
  - kitch_sq = 2014 became the build_year value, since state min = 0, q1 = 1, median = 6, q3 = 9
- Some values were hard to interpret
  - These include life_sq, full_sq at 0 and 1 which became NAs
  - Life_sq  > full_sq or kitch_sq > full_sq
  - Num_room = 0
- Corrected unusable variables names
  - Demographic variables such as 0_6_all

# Types of missingness, imputation

| | sub_area | build_count_block | count |
|---|---|---|---|
| 1 | Ajeroport | 31 | 123 |
| 2 | Akademicheskoe | 81 | 211 |
| 3 | Alekseevskoe | 72 | 100 |
| 4 | Altuf'evskoe | 24 | 68 |
| 5 | Arbat | 3 | 15 |
| 6 | Babushkinskoe | 49 | 123 |

- Different types of missingness:
  - Missing Comp. at Random (MCAR): missingness is not dependent on another variable
  - *Missing at Random (MAR): missingness may be dependent on another variable
  - *Missing Not at Random (MNAR): missingness is dependent on another variable
- Missingness in Sberbank dataset
  - Missingness that could be imputed by sub_area
    - 18 of the 51 variables with missing values shared the same values within their sub_area
      - All of these begin with build_count_* (building material, span of years)
    - na.aggregate() in R used to impute missing values by sub_area
  - Missingness related to time
    - 8 building chars: build_year, floor, kitch_sq, life_sq, material, max_floor, num_room, state
    - 6 of the 8 were missing 100% of their data from Q3 2011 to Q1 2013
    - These values were imputed by their sub_area median or through KNN
      - KNN: more robust method as it considered numerous variables in imputation

# Missingness by quarter (2011 Q3 to 2015 Q2)

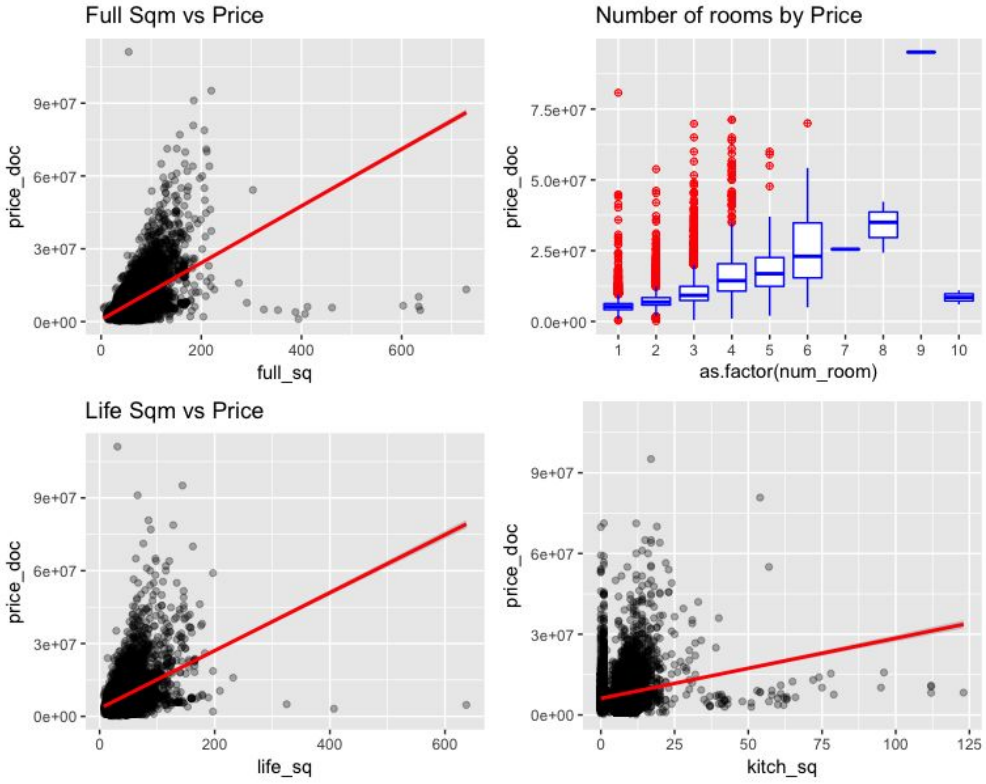
# Understanding Types of Features

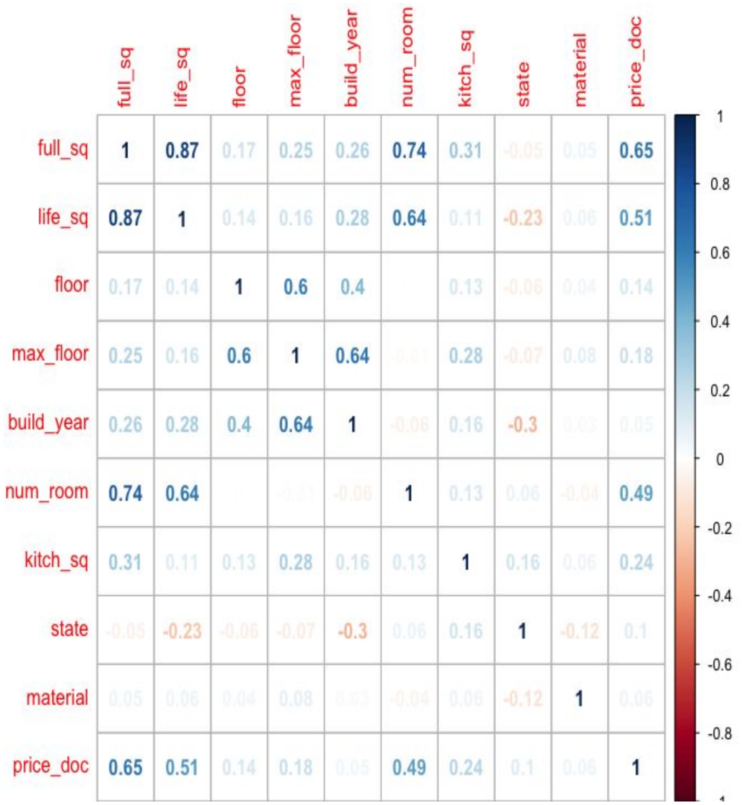
# Basic Variable Importance

- Evaluate Direct Correlation
- Remove obvious multicollinearity
    - i.e. Multiple cafe features
    - Which ones stay?
        - Variance
        - Hold more information
- Further Subgrouping
- Features that work within Linear Regression Model
- Noticeably missing features
    - Healthcare, education

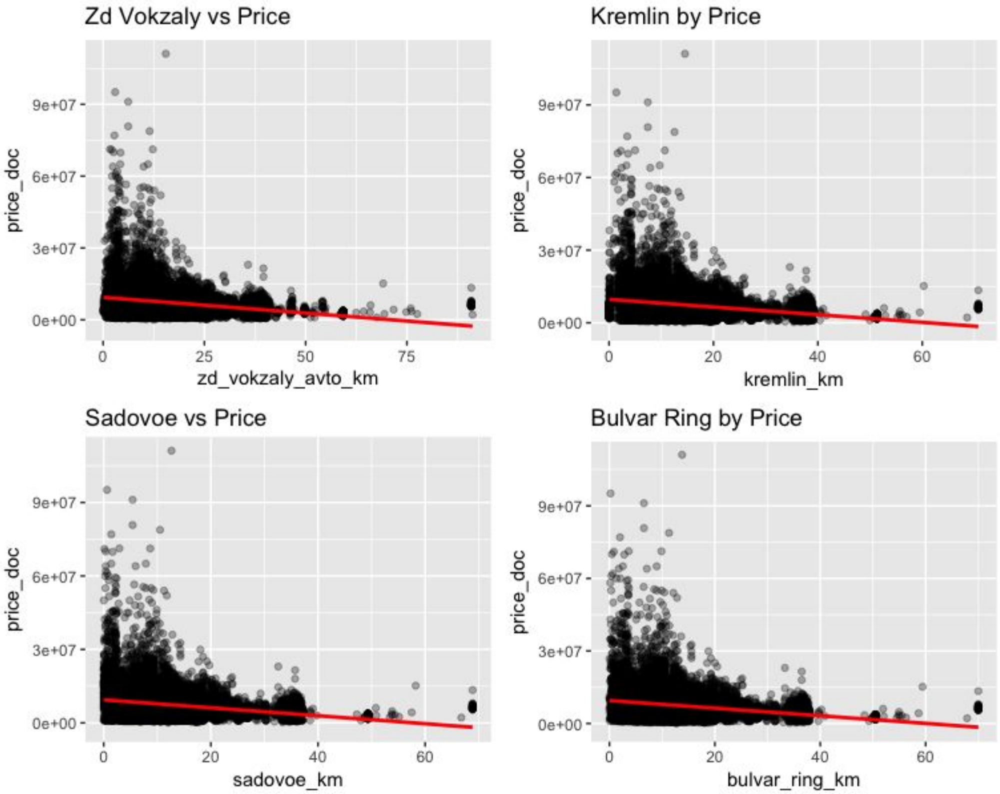
| | Var1 | Var2 | value |
|---|---|---|---|
| 1 | price_doc | price_doc | 1.000000000 |
| 2 | price_doc | full_sq | 0.564983378 |
| 3 | price_doc | life_sq | 0.455794263 |
| 4 | price_doc | num_room | 0.414936989 |
| 5 | price_doc | sport_count_5000 | 0.294967357 |
| 7 | price_doc | trc_count_5000 | 0.289433023 |
| 8 | price_doc | zd_vokzaly_avto_km | -0.284149955 |
| 9 | price_doc | sadovoe_km | -0.283710223 |
| 10 | price_doc | kremlin_km | -0.279332230 |
| 11 | price_doc | bulvar_ring_km | -0.279246594 |
| 13 | price_doc | ttk_km | -0.272680878 |
| 14 | price_doc | office_sqm_5000 | 0.270148536 |
| 17 | price_doc | nuclear_reactor_km | -0.257918995 |
| 18 | price_doc | sport_objects_raion | 0.252875408 |
| 20 | price_doc | cafe_count_5000_price_1000 | 0.240610222 |
| 21 | price_doc | stadium_km | -0.236996602 |
| 29 | price_doc | basketball_km | -0.223498087 |
| 32 | price_doc | kitch_sq | 0.221740810 |
| 34 | price_doc | university_km | -0.218596932 |
| 36 | price_doc | theater_km | -0.216094427 |
| 39 | price_doc | swim_pool_km | -0.211775178 |
| 40 | price_doc | catering_km | -0.210849909 |
| 42 | price_doc | thermal_power_plant_km | -0.210460129 |
| 43 | price_doc | workplaces_km | -0.209341663 |

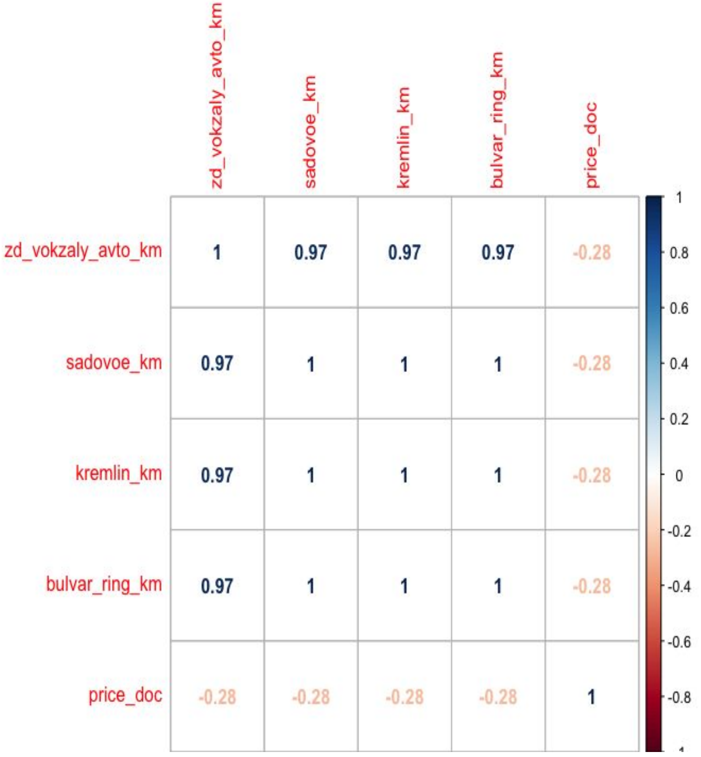
# EDA of Building Characteristics



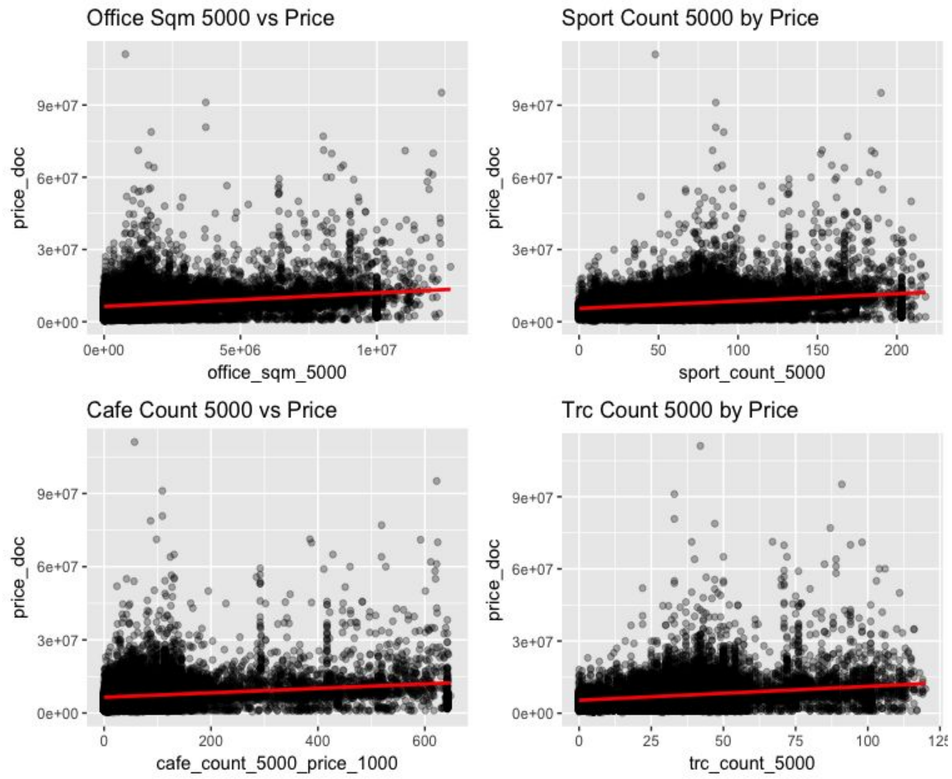



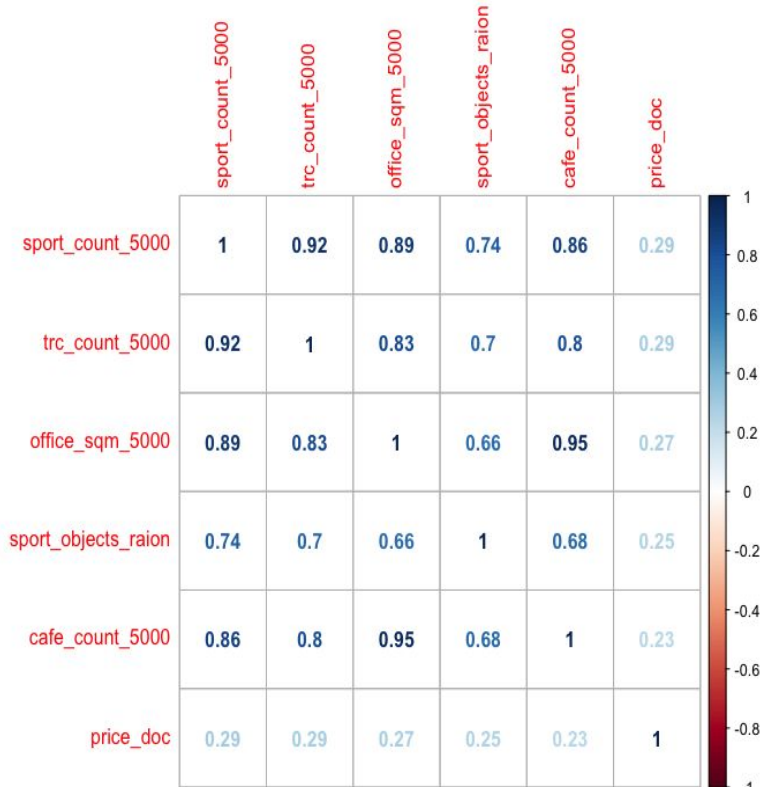
# EDA of Location Characteristics



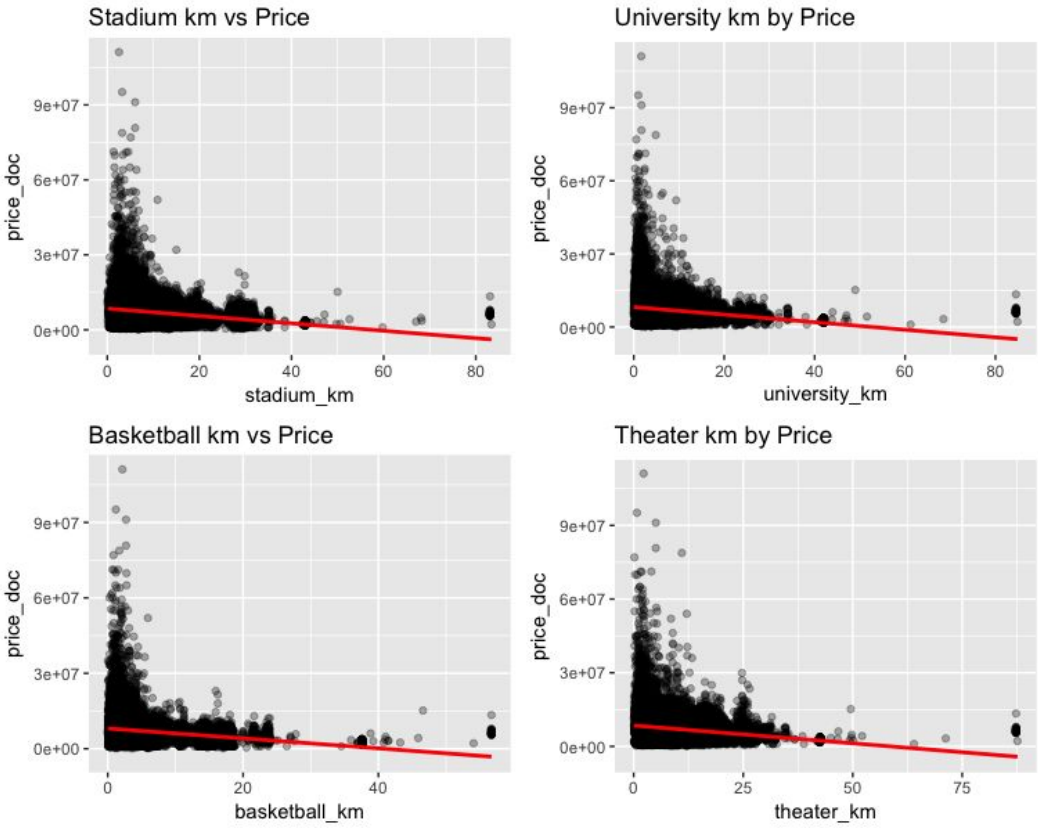



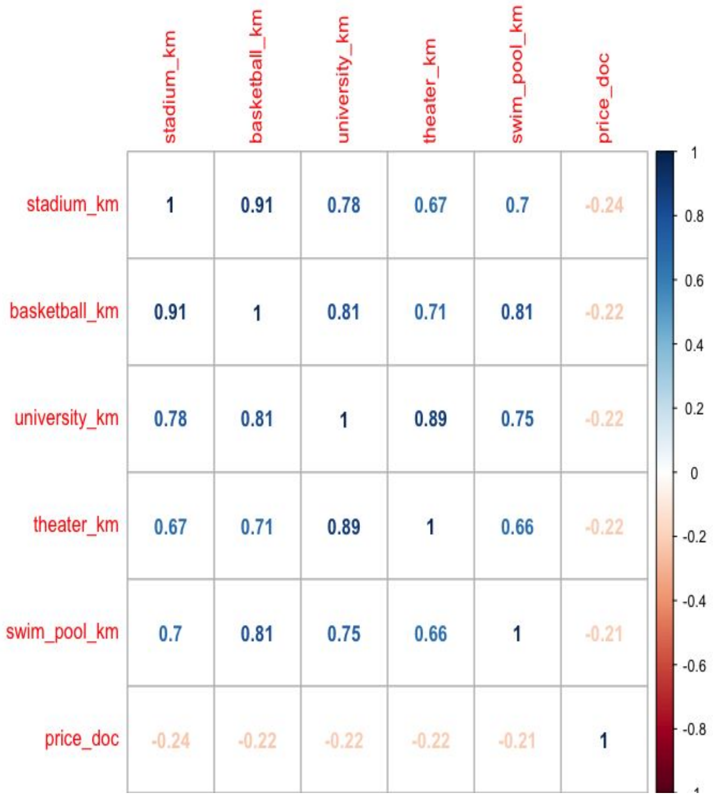
# EDA of Lifestyle Characteristics



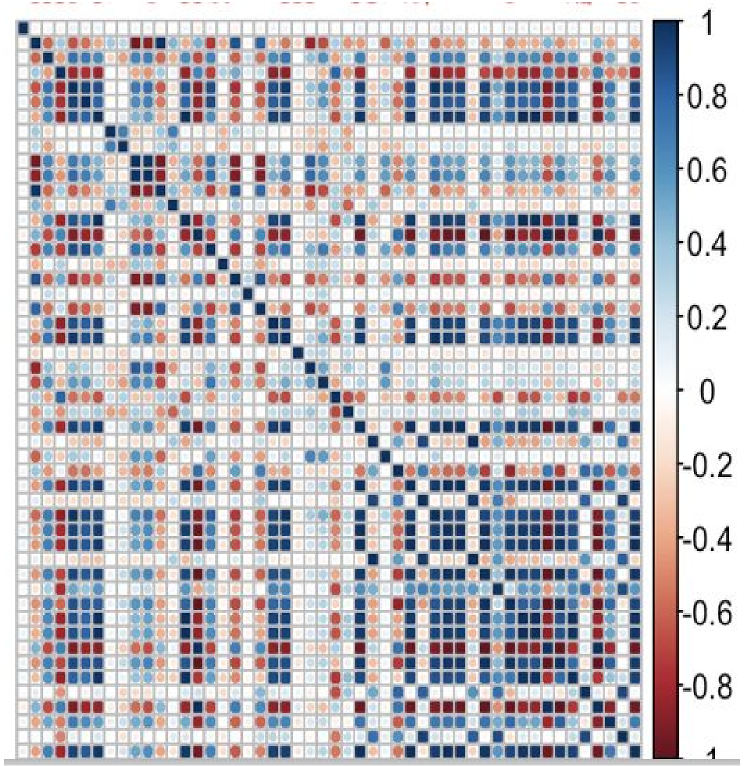
# EDA of Education / Cultural Characteristics



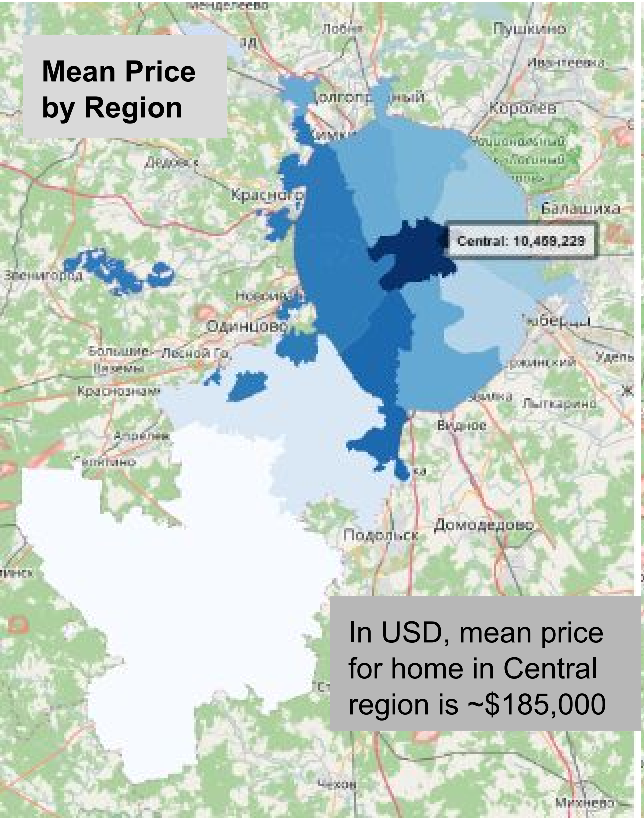



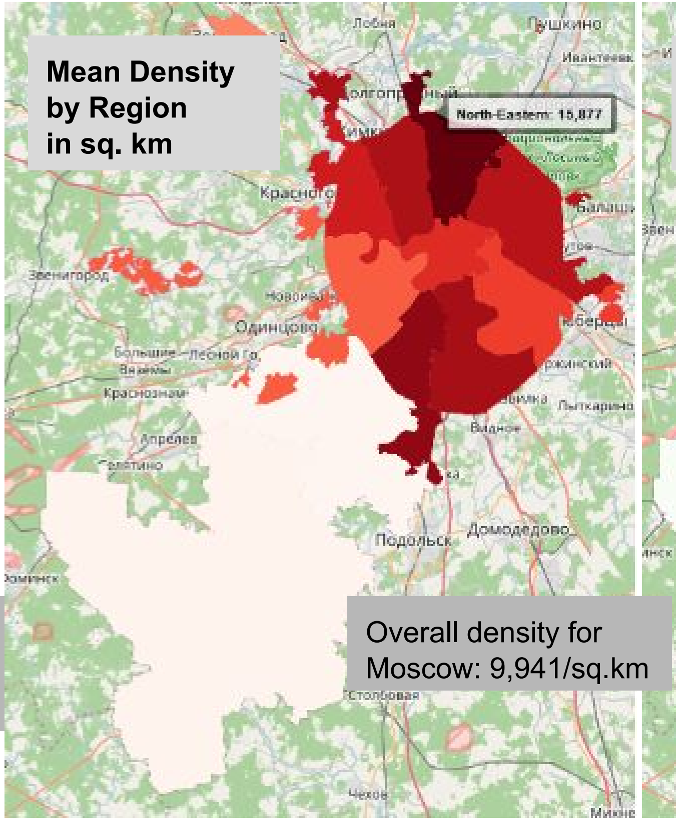
# EDA of Macro Dataset

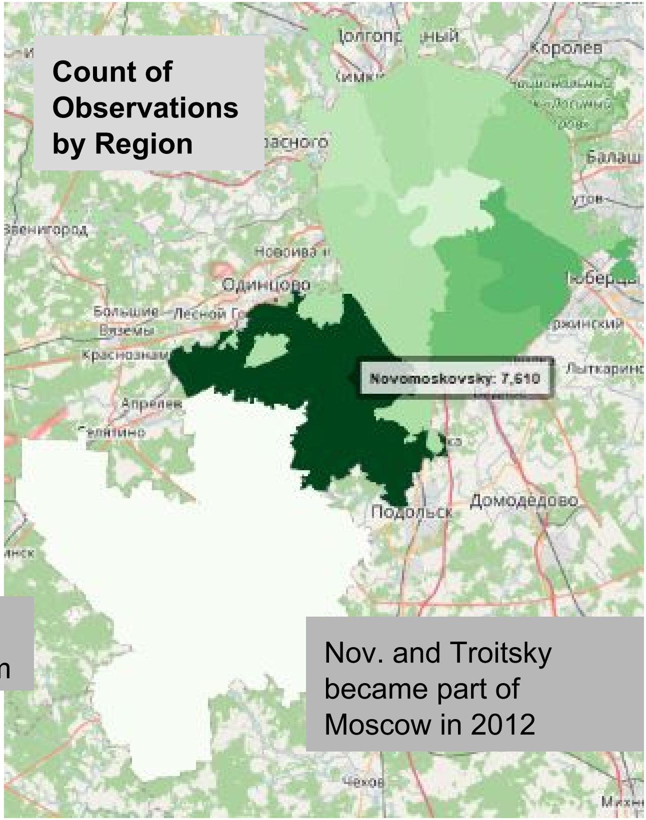
| | Var1 | Var2 | value |
|---|---|---|---|
| 1 | price_doc | price_doc | 1.0000000000 |
| 2 | price_doc | labor_force | 0.1031398753 |
| 3 | price_doc | unprofitable_enterpr_share | 0.1027078295 |
| 4 | price_doc | profitable_enterpr_share | -0.1027078295 |
| 5 | price_doc | retail_trade_turnover_per_cap | 0.1022785834 |
| 6 | price_doc | gdp_annual_growth | -0.1022061047 |
| 7 | price_doc | retail_trade_turnover | 0.1021741366 |
| 8 | price_doc | fin_res_per_cap | -0.1013108991 |
| 9 | price_doc | construction_value | 0.1011695026 |
| 10 | price_doc | grp | 0.1010008240 |
| 11 | price_doc | salary | 0.1009771114 |
| 12 | price_doc | employment | 0.1009067241 |
| 13 | price_doc | cpi | 0.1001319260 |
| 14 | price_doc | fixed_basket | 0.0997143889 |
| 15 | price_doc | deposits_value | 0.0963939738 |
| 16 | price_doc | gdp_deflator | 0.0957839678 |
| 17 | price_doc | invest_fixed_assets | 0.0953011545 |



# Price, density, and count by Moscow administrative region

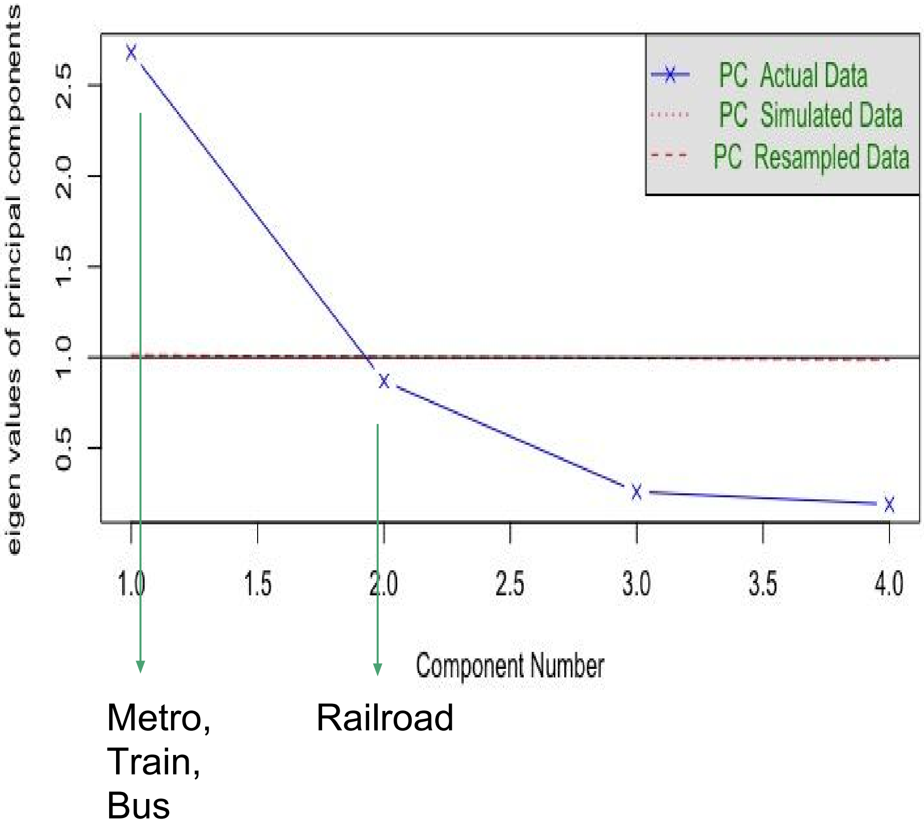


# Feature Selection/Engineering

- Subgroup
  - Grouping variables (e.g., cafe or schools)
- Feature Engineering
  - High correlation
    - PCA (distance to transportation)
  - Large number of features
    - Clustering



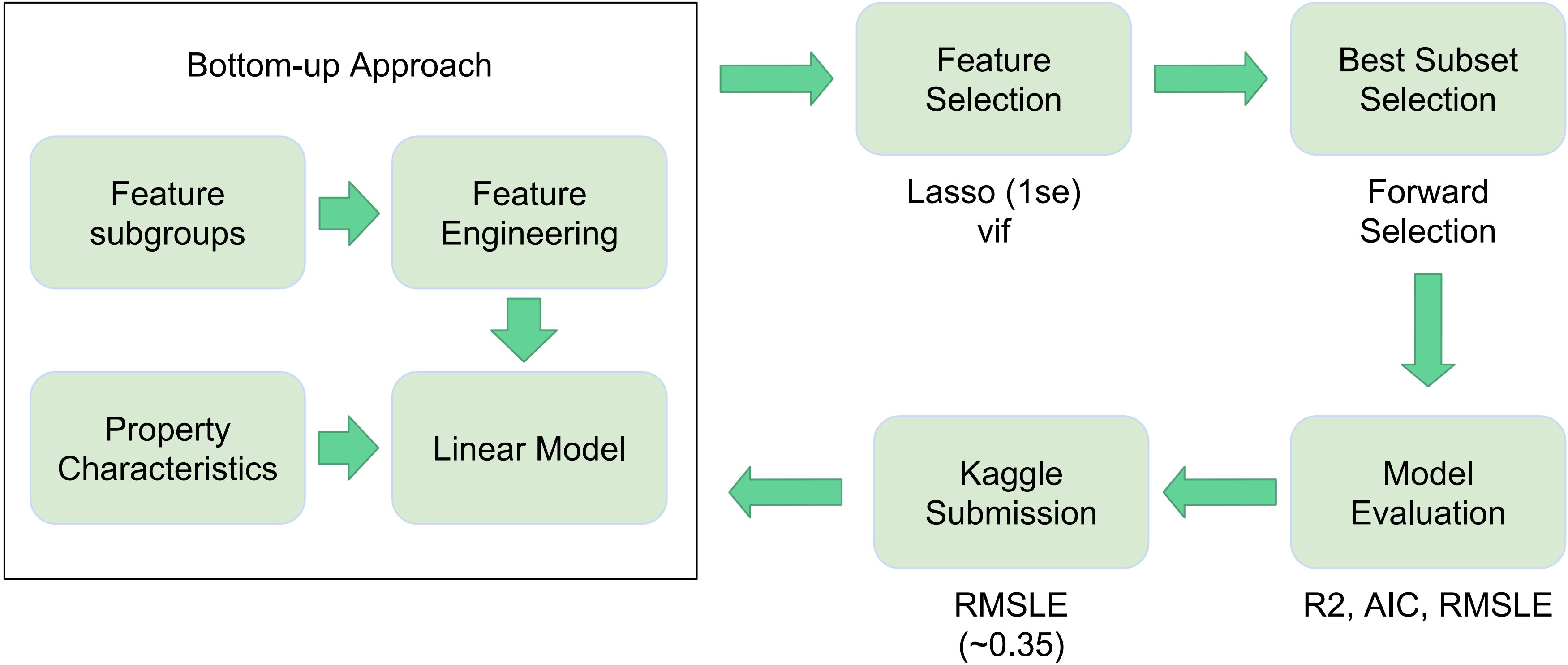
# Clustering - Negative Exposure

## Hierarchical Clustering

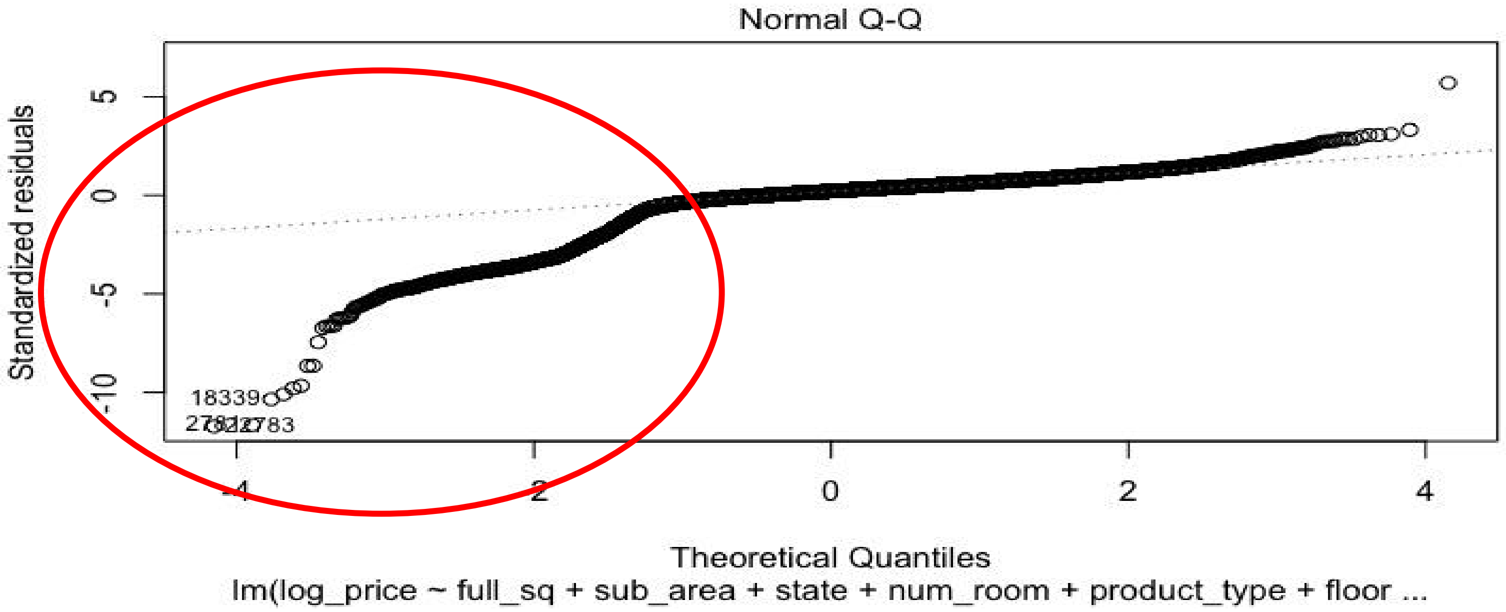
- Underlying patterns with presence of negative risk factors in the neighborhood
- Euclidean Distance
- Average Linkage

| Cluster | Name | Oil_Chem (dirty industry) | Radioactive Waste | Nuclear Reactors | Thermal Power | Incinerators |
|---|---|---|---|---|---|---|
| 1 | Safe Environment | 0 | 0 | 0 | 0 | 0 |
| 2 | Nuclear reactor/ radioactive waste | 0 | 1 | 1 | 0 | 0 |
| 3 | Dangerous Industrial Area | 1 | 1 | 0 | 1 | 0 |

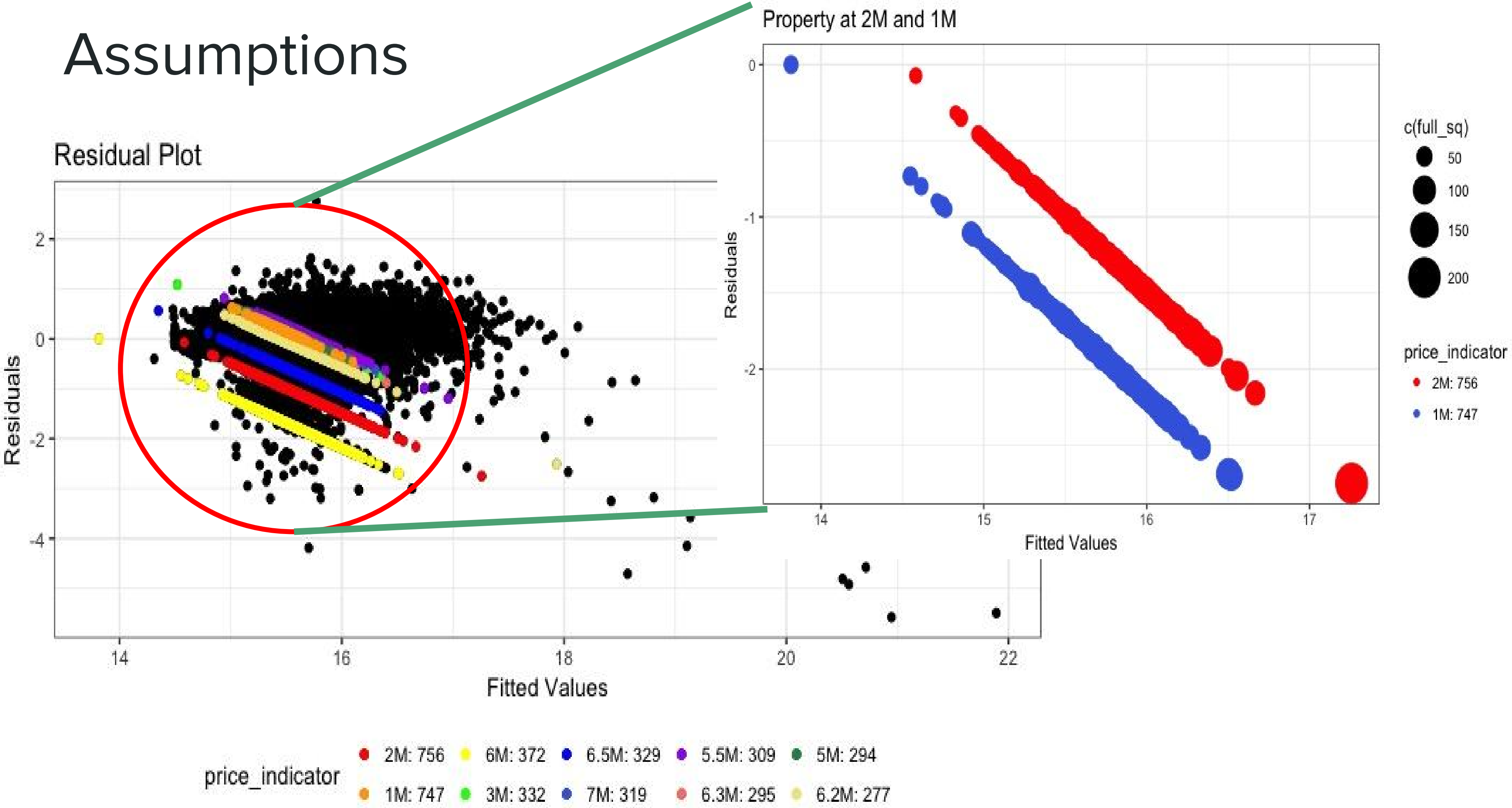
# Multiple Linear Regression



# Assumptions - Normality



# Assumptions



# Tree Model - Random Forest

Why we use random Forest?

1. Identify moderately strong predictor

2. Investigate nonlinear pattern from the data set.

3. Feature Selection

# Random Forest - R vs Python

Variables: full_sq, life_sq, floor ,num_room,

product_type, sub area

10 fold Cross-Validated with 500 trees

Processing time > 6 hr

RMSLE :0.34587

# Random Forest Model: Sub_Area

- Sub_Area plays an important role in the linear model.
- Is sub_area also important in Random Forest?

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 15.1280582 | 0.0465972 | 324.656 | < 2e-16 | *** |
| full_sq | 0.0089520 | 0.0001613 | 55.491 | < 2e-16 | *** |
| sub_areaAkademicheskoe | -0.0769393 | 0.0551126 | -1.396 | 0.162713 | |
| sub_areaAlekseevskoe | -0.1358809 | 0.0654612 | -2.076 | 0.037926 | * |
| sub_areaAltuf'evskoe | -0.4746849 | 0.0734004 | -6.467 | 1.01e-10 | *** |
| sub_areaArbat | 0.4317588 | 0.1378805 | 3.131 | 0.001741 | ** |
| sub_areaBabushkinskoe | -0.2320874 | 0.0620333 | -3.741 | 0.000183 | *** |
| sub_areaBasmannoe | 0.0706176 | 0.0671457 | 1.052 | 0.292942 | |
| sub_areaBegovoe | 0.0432474 | 0.0764597 | 0.566 | 0.571654 | |
| sub_areaBeskudnikovskoe | -0.3907625 | 0.0580224 | -6.735 | 1.67e-11 | *** |
| sub_areaBibirevo | -0.2036914 | 0.0546441 | -3.728 | 0.000194 | *** |
| sub_areaBirjulevo Vostochnoe | -0.3765963 | 0.0533200 | -7.063 | 1.67e-12 | *** |
| sub_areaBirjulevo Zapadnoe | -0.4093570 | 0.0630870 | -6.489 | 8.79e-11 | *** |
| sub_areaBogorodskoe | -0.3208415 | 0.0518880 | -6.183 | 6.36e-10 | *** |
| sub_areaBrateevo | -0.2597758 | 0.0569880 | -4.558 | 5.17e-06 | *** |
| sub_areaButyrskoe | -0.2529606 | 0.0652785 | -3.875 | 0.000107 | *** |
| sub_areaCaricyno | -0.3117460 | 0.0548726 | -5.681 | 1.35e-08 | *** |
| sub_areaCheremushki | 0.0242696 | 0.0587077 | 0.413 | 0.679319 | |
| sub_areaChertanovo Central'noe | -0.2523700 | 0.0560234 | -4.505 | 6.67e-06 | *** |
| sub_areaChertanovo Juzhnoe | -0.2947973 | 0.0530845 | -5.553 | 2.83e-08 | *** |
| sub_areaChertanovo Severnoe | -0.1401899 | 0.0558157 | -2.512 | 0.012022 | * |
| sub_areaDanilovskoe | -0.0850491 | 0.0556942 | -1.527 | 0.126753 | |
| sub_areaDmitrovskoe | -0.3329230 | 0.0572245 | -5.818 | 6.02e-09 | *** |

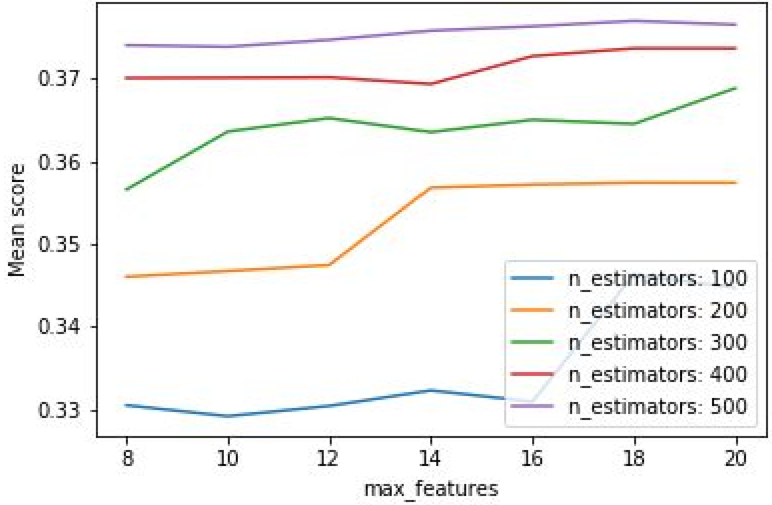
# Random Forest: Sub_Area

Modeling:

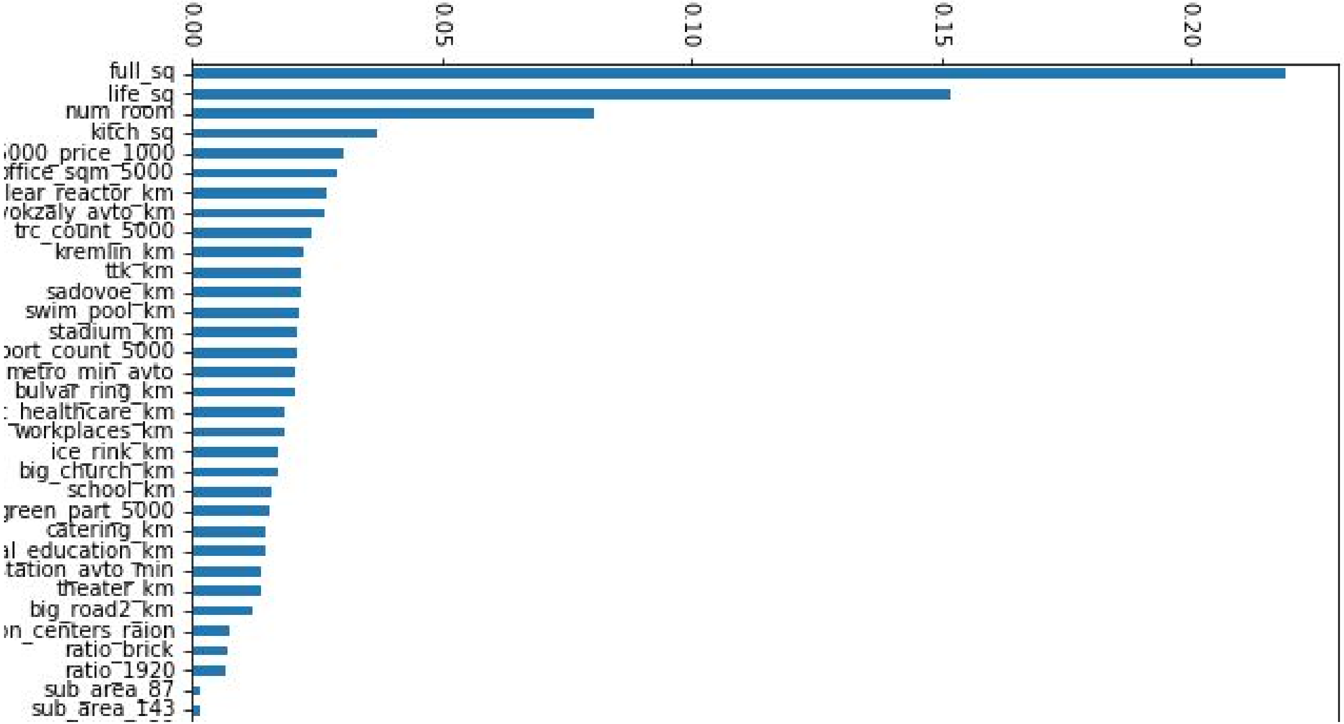
1. Using grid search and 10 fold cross validation to find the best parameter for max features and number of trees.
2. Number of variables : 177 (dummy coded 146 for sub area)

Result:

1. Best parameter: 400 trees, 20 variables.
2. Testing Score: 0.3769
3. Training Score: 0.4901
4. RMSLE 0.3432
5. Sub_Area is not as important as in linear



# Random Forest: Sub_area Feature Importance



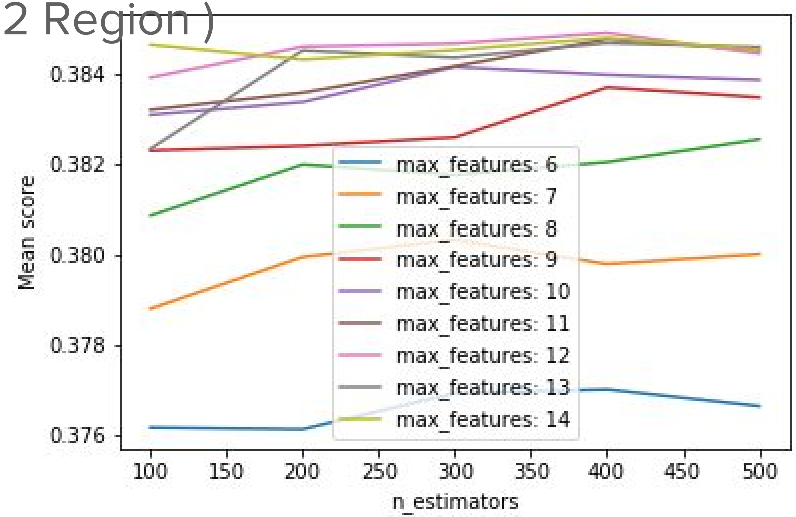
# Random Forest: OKRUG (Admin. Region)

Modeling:

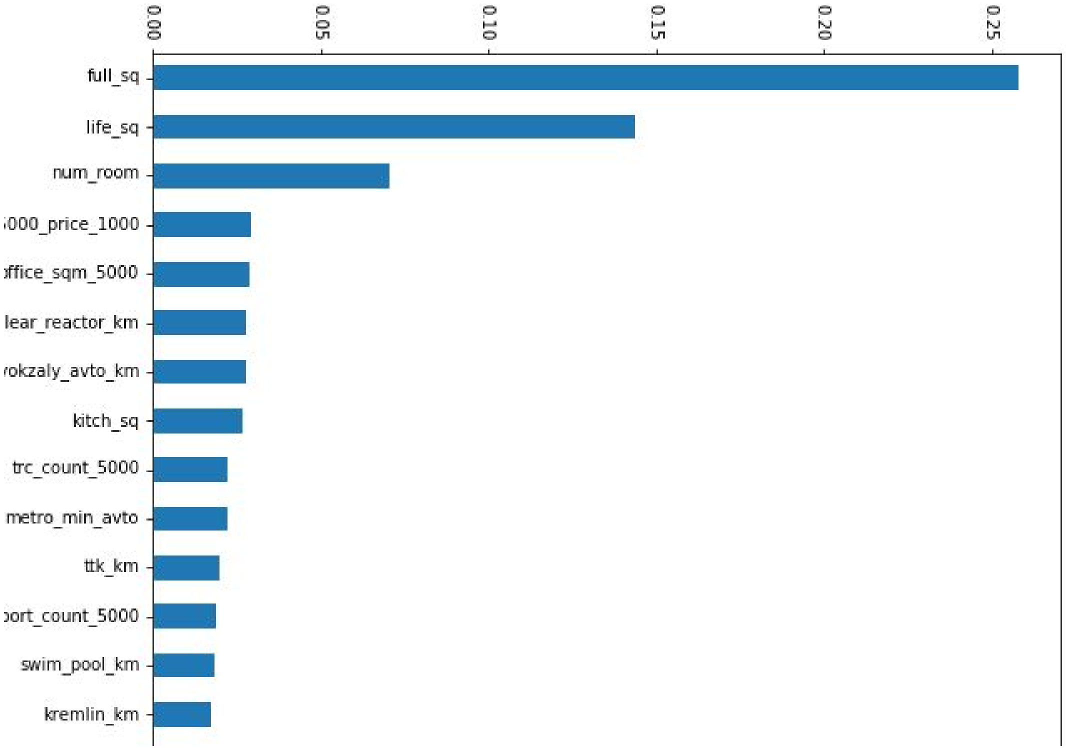
1. Using grid search and cross validation to find the best parameter for max features and number of trees.
2. Number of variables : 57 (dummy coded 12 Region )

Result:

1. Best esitmator : 400 trees & 12 features.
2. Test Score: 0.3849
3. Training Score: 0.5272
4. Slightly improve RMSLE to 0.34206



# Random Forest: Feature Importance OKRUG (Admin. Region)



# Stacking


Kaggle Score

~0.35

Kaggle Score

0.34326

Kaggle Score

0.33972

# Conclusion and Future Direction

- Conclusion (based on Random Forest):
  - Apartment Characteristics (e.g., full_sq, life_sq, no_rooms, and kitch_sq)
  - Neighborhood Characteristics (e.g., cafes, railroad, nuclear_reactor*, metro_dist, and office)
- Future Direction
  - Time Dependency
  - Further analysis of subareas
  - More Industry Knowledge

# Thank You!

Any Questions?

# RF and MLR Residual Comparison (optional)