**MSDS-6372 Applied Statistics - Team Project #1 Report**

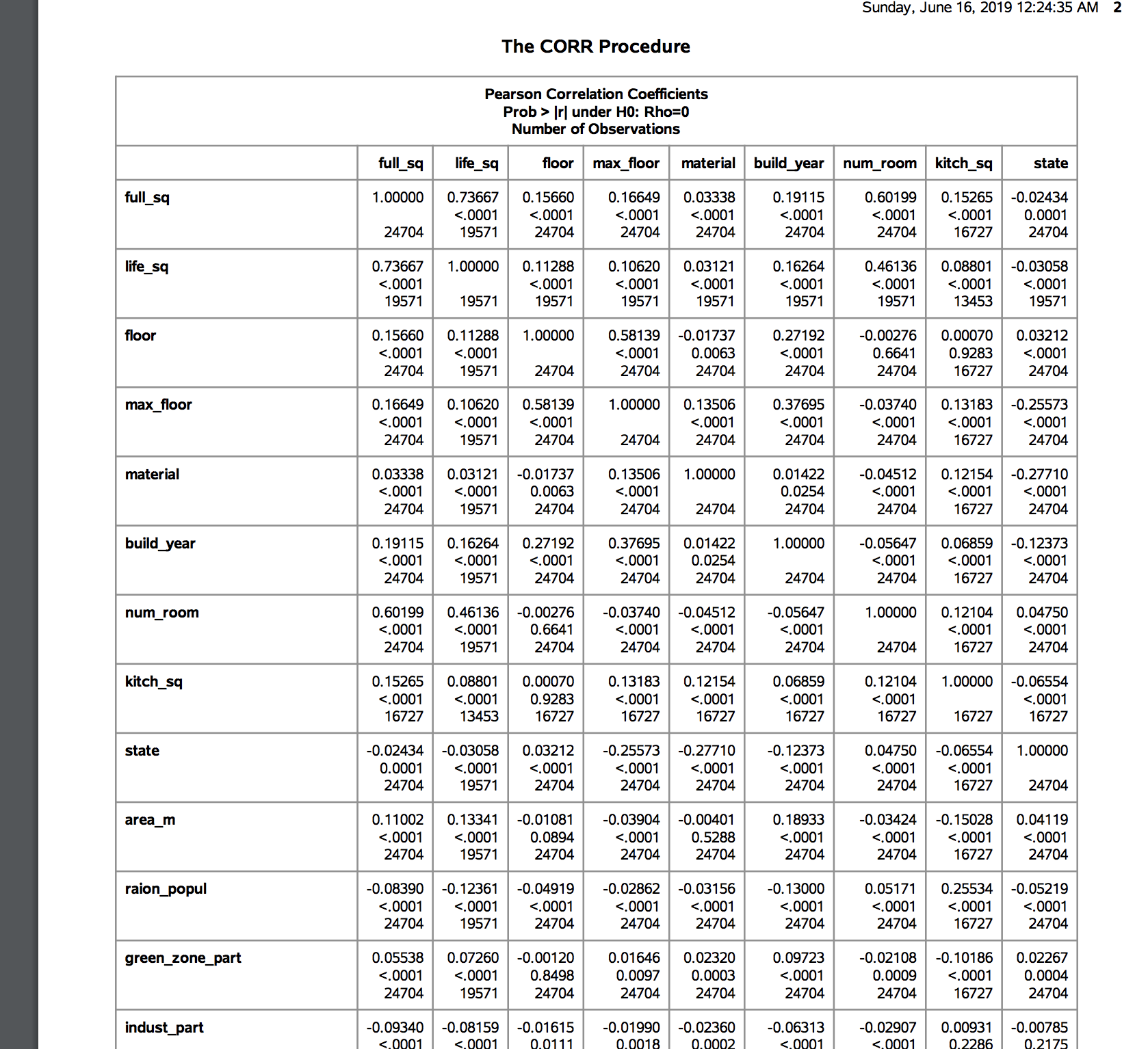
**GOAL1 - Prediction of Individual Property Values**

* Introduction

We were tasked with modeling and making projections of real estate prices using data provided by Sberbank of Russia. The data provided were in two files, modelingData.csv and projectionData.csv.

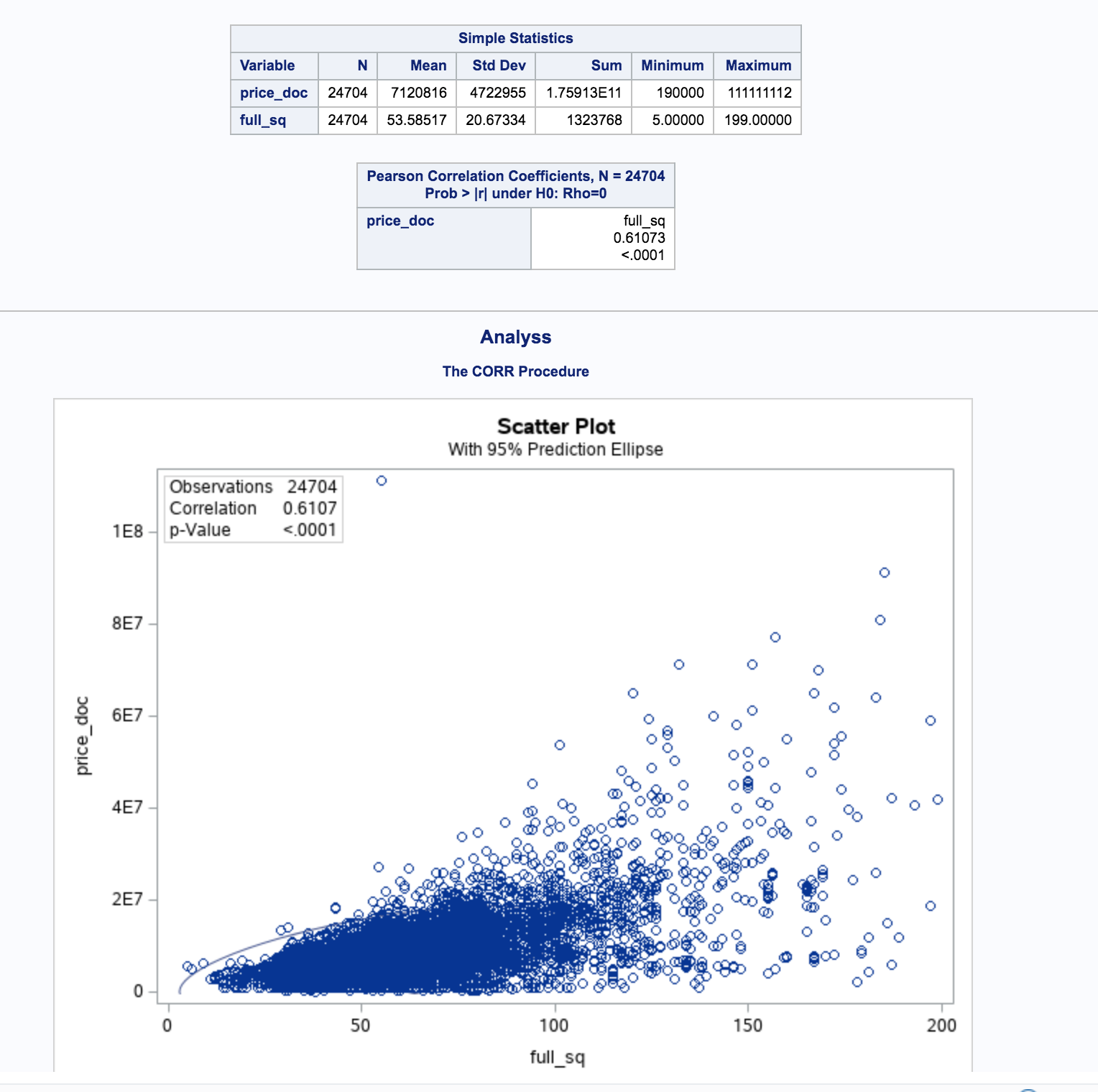
Using the data we were asked to answer two questions. The first item to address was the development of a model that minimized the Root Mean Squared Logarithmic Error (RMSLE). The second item to address was to make projections of the mean price of properties from July 2015 to July 2016. For second topic, please refer to the section [GOAL2]

* Data Description
  1. The modelingData.csv file contains properties sold between August 20, 2011 and June 30, 2015
     + 25471 observations / rows
     + 292 variables / columns
  2. The prediction target file, predictionData.CSV contains predictions on properties sold between July 2015 and May 2016
     + 5000 observations / rows
     + 291 variables / columns
  3. For further details on data description, please refer to the Appendix 1
* Data Cleaning / Wrangling
  1. The data was adjusted using primarily the means of the variable column. For instance, we revised the observation with build\_year 4965 as 1965 since it seems that there were typos
  2. Detailed data cleaning works are explained in Appendix 1
* EDA
  1. Outliers
     + We found some outliers in our analysis and either we changed it or dropped the values if they looked too unreasonable. The code details are in Data wrangling part
  2. Multicollinearity
     + We can see some of the independent variables are highly correlated to each other and they might impact coefficients for linear regressions
     + As we can see in the CORR output, the correlation between life\_sq and full\_sq and num\_room and full\_sq are sufficiently high to produce effects due to multi collinearity



[Display: Multicollinearity Analysis]

* 1. Assumptions
     + Independence: For our analysis we assumed that the collected data are independently collected. There is no indication of data being dependent on other.
     + Normal Distribution: The response variable is assumed to be normally distributed across the measurements taken for independent variables.
     + Linearity: We checked area parameters and they seem to be linearly correlated to our dependent variable: price doc. For the variables we considered we didn’t find any variable that needed transformation.
     + As we can see in the diagram below, the price increases linearly as full\_sq values increase. We can assume this data to be linearly related to Price doc and add this to the model without transformation

****

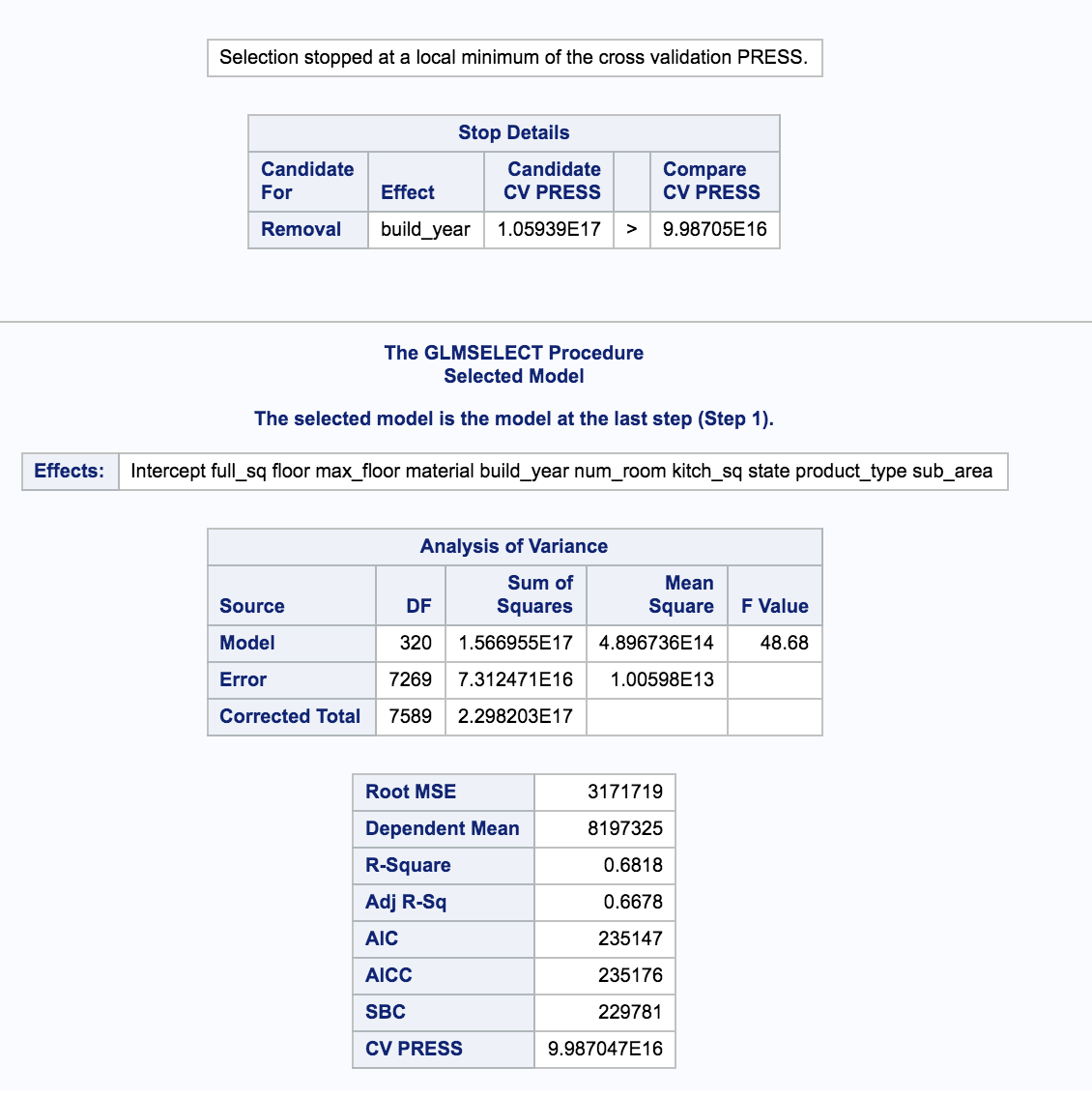
[Display: Scatter Plot for full\_sq and price\_doc]

* Variable Selection
  1. Based on the forward selection, backward selection, stepwise selection and Lasso, we choose 25 relevant variables with statistical significance
     + full\_sq life\_sq
     + floor
     + max\_floor
     + material
     + build\_year
     + num\_room
     + kitch\_sq
     + state
     + product\_type
     + sub\_area
     + area\_m
     + raion\_popul
     + green\_zone\_part
     + indust\_part children\_preschool preschool\_quota
     + children\_school school\_quota
     + school\_education\_centers\_raion
     + hospital\_beds\_raion healthcare\_centers\_raion
     + university\_top\_20\_raion
     + sport\_objects\_raion additional\_education\_raion
     + culture\_objects\_top\_25
     + culture\_objects\_top\_25\_raion
     + shopping\_centers\_raion office\_raion
     + thermal\_power\_plant\_raion incineration\_raion
     + oil\_chemistry\_raion
     + radiation\_raion railroad\_terminal\_raion big\_market\_raion
     + nuclear\_reactor\_raion
     + detention\_facility\_raion
* Model Selection
  1. After comparing the forward selection, backward selection, stepwise selection and Lasso techniques, we concluded that the backward selection is giving us the best result
  2. Adjusted R, AIC and AICC favored this method
  3. Summary Comparison

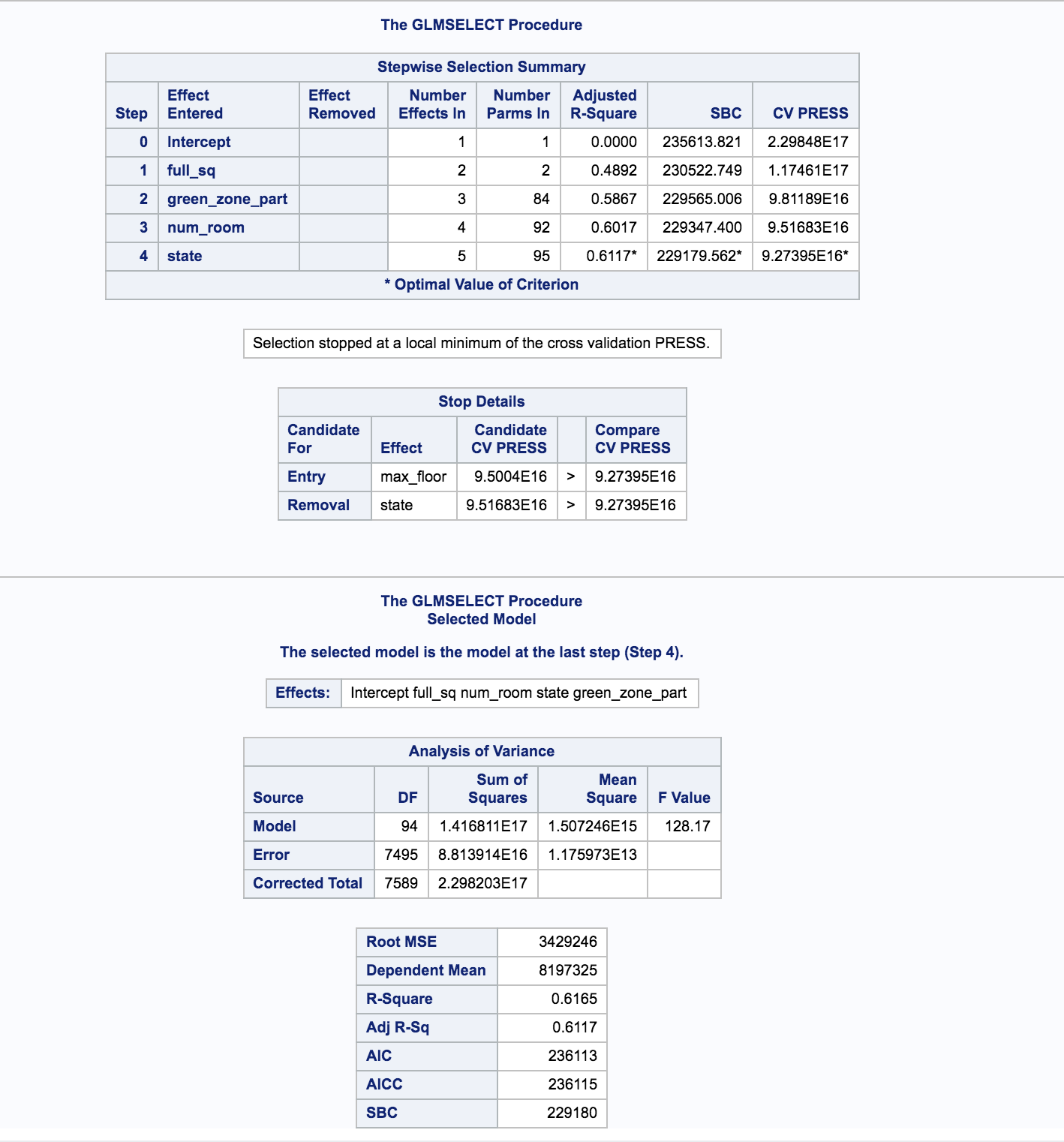
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Backward**  **Selection** | Stepwise  Selection | Forward  Selection | Lasso |
| Adjusted R-SQ | **0.67** | 0.61 | 0.61 | 0.48 |
| AIC | **235147** | 236113 | 236113 | 238259 |
| AICC | **235176** | 236115 | 236115 | 238259 |

* 1. Backward Selection Results





* 1. Stepwise Selection Results



* 1. Forward Selection Results and Lasso Modeling Results



* Final Prediction
  1. We created different predictions based on selection models. Our final prediction is based on the backward selection model since that has the lowest Adjusted R squared Value
  2. But some of the predictions have missing values due to the limitation of the chosen model. So we have used some reasonable number to set values for these numbers such as averages

**GOAL2**

* Introduction

Sberbank, one of Russia’s largest bank, would like to predict average house prices based on historic trend. We used the time series analysis technic on aggregated monthly data to solve this problem

* Data Wrangling
  1. The timestamp if each house purchase record are written in the number of days since 12/30/1899. So we transform the data into the familiar YYYY-MM-DD format with R

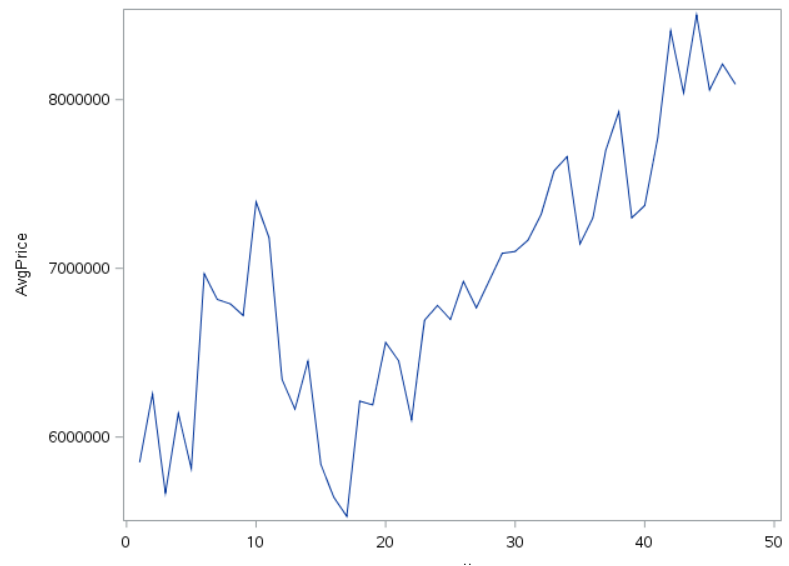
|  |
| --- |
| # Select ID, timestamp and price, and transform to YYYY-MM-DD  goal2\_data <- modelingData[c("id", "timestamp", "price\_doc")]  goal2\_data$timestamp <- as.Date(goal2\_data$timestamp, origin = "1899-12-30") |

* 1. With Aggregate function in R, we could get the average monthly price. And we also define sequential month number to make the time series analysis easier

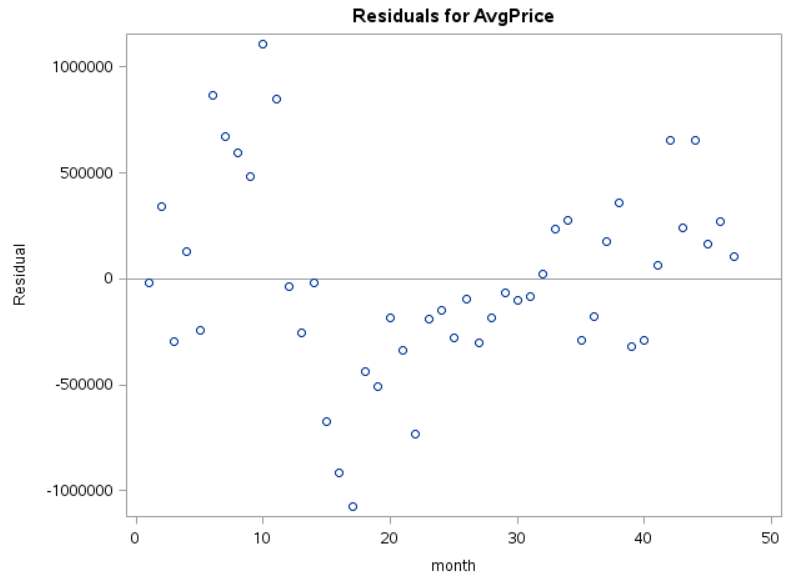
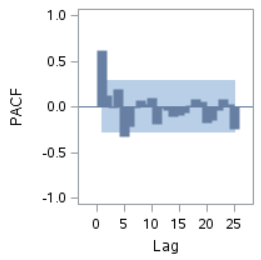
|  |
| --- |
| # Aggregate by month-year  goal2\_agg <- aggregate(goal2\_data, by=list(goal2\_data$monthYear), FUN = mean)  goal2\_agg <- goal2\_agg[c("Group.1", "price\_doc")]  names(goal2\_agg) <- c("monthYear", "AvgPrice")  goal2\_agg$monthNumber <- seq.int(nrow(goal2\_agg)) |

* Plotting Time Series

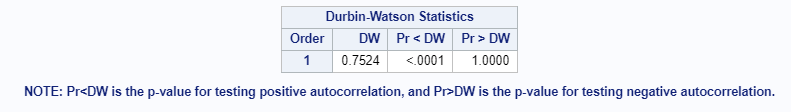
The plotting and subsequent analysis is conducted with the SAS tool. We can see the overall trend of the mean price increase accompanied with oscillating ups and downs.



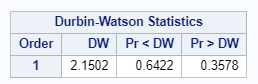
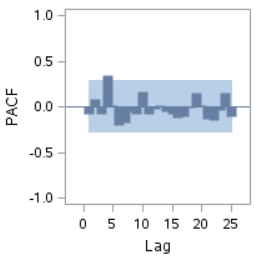
* Residual Series Modeling
  1. When fitting the data to OLS, we see the residuals are clustered along recent past. Positive residuals follow another positive residuals from previous observation, and so forth for negative residuals – hunch for autocorrelation
  2. The PACF result also suggest the autocorrelation. So we will Durbin-Watson statistics

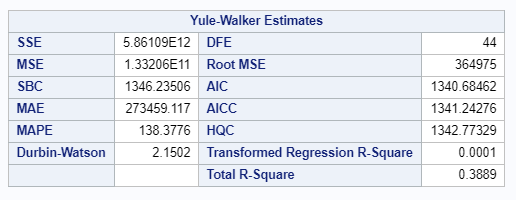
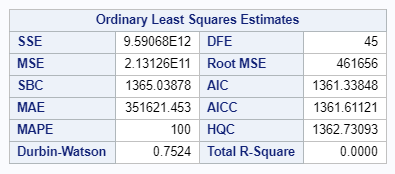
* 1. DW statistics value and Pr suggest there is a positive autocorrelation



* 1. Once corrected with AR(1), we see the residual autocorrelation is corrected based on DW value near to 2.0

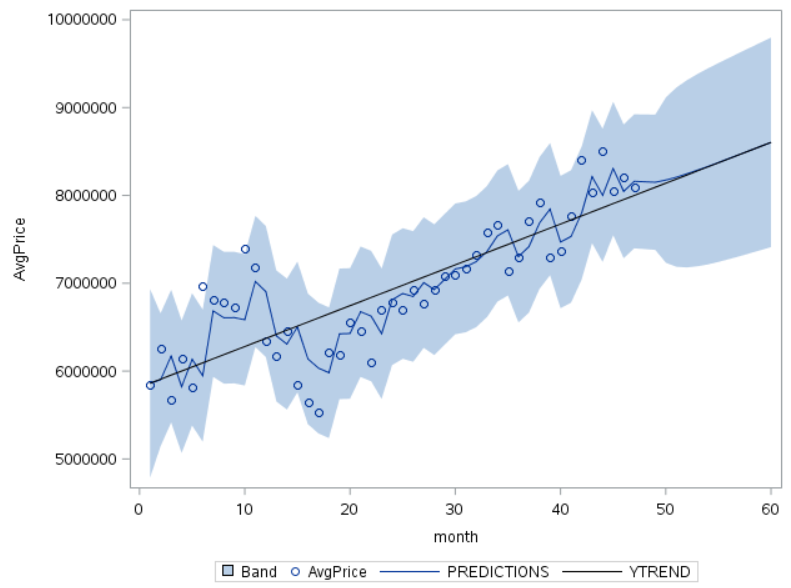
* 1. When comparing the AIC and SBC, we see corrected models are favored compared to the OLS. The figures are lower with Y-W estimates



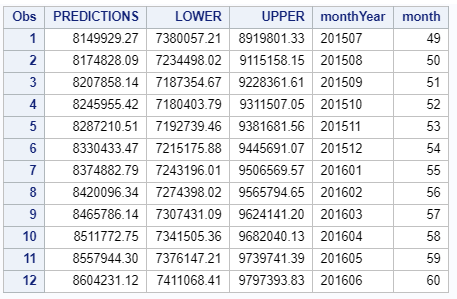
* Average House Price Prediction
  1. The monthly average price from July 2015 to June 2016 are predicted with the corrected estimates from SAS

|  |
| --- |
| PROC AUTOREG DATA=GOAL2\_PRED ;  MODEL AVGPRICE = MONTH / NLAG = (1) DWPROB;  OUTPUT OUT = PREDS\_PRICE P = PREDICTIONS LCL = LOWER UCL = UPPER PM = YTREND;  RUN; |

* 1. The SAS package help combine the trend estimate with the forecast residual analyzed in previous steps. The plot with 95% confidence interval, predictions and trend values are as following



* 1. Predictions of the housing price with 95% confidence interval



**Appendix 1: Data Dictionary**

**Major Variables**

|  |  |
| --- | --- |
| Variables | Description |
| price\_doc | sale price (this is the target variable) |
| id | transaction id |
| timestamp | date of transaction |
| full\_sq | total area in square meters, including loggias, balconies and other non-residential areas |
| life\_sq | living area in square meters, excluding loggias, balconies and other non-residential areas |
| floor | for apartments, floor of the building |
| max\_floor | number of floors in the building |
| material | wall material |
| build\_year | year built |
| num\_room | number of living rooms |
| kitch\_sq | kitchen area |
| state | apartment condition |
| product\_type | owner-occupier purchase or investment |
| sub\_area | name of the district |

**Neighborhood Features**

|  |  |
| --- | --- |
| **Variables** | **Description** |
| area\_m | Area, sq.m. |
| raion\_popul | Number of municipality population. district |
| green\_zone\_part | Proportion of area of greenery in the total area |
| indust\_part | Share of industrial zones in area of the total area |
| children\_preschool | Number of pre-school age population |
| preschool\_quota | Number of seats in pre-school organizations |
| preschool\_education\_centers\_raion | Number of pre-school institutions |
| children\_school | Population of school-age children |
| school\_quota | Number of high school seats in area |
| school\_education\_centers\_raion | Number of high school institutions |
| school\_education\_centers\_top\_20\_raion | Number of high schools of the top 20 best schools in Moscow |
| hospital\_beds\_raion | Number of hospital beds for the district |
| healthcare\_centers\_raion | Number of healthcare centers in district |
| university\_top\_20\_raion | Number of higher education institutions in the top ten ranking of the Federal rank |
| sport\_objects\_raion | Number of higher education institutions |
| additional\_education\_raion | Number of additional education organizations |
| culture\_objects\_top\_25 | Presence of the key objects of cultural heritage (significant objects for the level of the RF constituent entities, city) |
| culture\_objects\_top\_25\_raion | Number of objects of cultural heritage |
| shopping\_centers\_raion | Number of malls and shopping centers in district |
| office\_raion | Number of malls and shopping centers in district |
| thermal\_power\_plant\_raion | Presence of thermal power station in district |
| incineration\_raion | Presence of incinerators |
| oil\_chemistry\_raion | Presence of dirty industries |
| radiation\_raion | Presence of radioactive waste disposal |
| railroad\_terminal\_raion | Presence of the railroad terminal in district |
| big\_market\_raion | Presence of large grocery / wholesale markets |
| nuclear\_reactor\_raion | Presence of existing nuclear reactors |
| detention\_facility\_raion | Presence of detention centers, prisons |
| full\_all | Total number of population in the municipality |
| male\_f | Male population |
| female\_f | Female population |
| young\_all | Population younger than working age |
| young\_male | Male population younger than working age |
| young\_female | Feale population younger than working age |
| work\_all | Working-age population |
| work\_male | Male working-age population |
| work\_female | Female working-age population |
| ekder\_all | Population older than working age |
| ekder\_male | Male population older than working age |
| ekder\_female | Female population older than working age |
| 0\_6\_all | Population aged 0-6 |
| 0\_6\_male | Male population aged 0-7 |
| 0\_6\_female | Female population aged 0-8 |
| 7\_14\_all | Population aged 7-14 |
| 7\_14\_male | Male population aged 7-14 |
| 7\_14\_female | Female population aged 7-14 |
| 0\_17\_all | Population aged 0-17 |
| 0\_17\_male | Male population aged 0-17 |
| 0\_17\_female | Female population aged 0-17 |
| 16\_29\_all | Population aged 16-19 |
| 16\_29\_male | Male population aged 16-19 |
| 16\_29\_female | Female population aged 16-19 |
| 0\_13\_all | Population aged 0-13 |
| 0\_13\_male | Male population aged 0-13 |
| 0\_13\_female | Female population aged 0-13 |
| raion\_build\_count\_with\_material\_info | Number of building with material info in district |
| build\_count\_block | Share of block buildings |
| build\_count\_wood | Share of wood buildings |
| build\_count\_frame | Share of frame buildings |
| build\_count\_brick | Share of brick buildings |
| build\_count\_monolith | Share of monolith buildings |
| build\_count\_panel | Share of panel buildings |
| build\_count\_foam | Share of foam buildings |
| build\_count\_slag | Share of slag buildings |
| build\_count\_mix | Share of mixed buildings |
| raion\_build\_count\_with\_builddate\_info | Number of building with build year info in district |
| build\_count\_before\_1920 | Share of before\_1920 buildings |
| build\_count\_1921-1945 | Share of 1921-1945 buildings |
| build\_count\_1946-1970 | Share of 1946-1970 buildings |
| build\_count\_1971-1995 | Share of 1971-1995 buildings |
| build\_count\_after\_1995 | Share of after\_1995 buildings |
| 7\_14\_male | Male population aged 7-14 |
| 7\_14\_female | Female population aged 7-14 |
| 0\_17\_all | Population aged 0-17 |
| 0\_17\_male | Male population aged 0-17 |
| 0\_17\_female | Female population aged 0-17 |
| 16\_29\_all | Population aged 16-19 |
| 16\_29\_male | Male population aged 16-19 |
| 16\_29\_female | Female population aged 16-19 |
| 0\_13\_all | Population aged 0-13 |
| 0\_13\_male | Male population aged 0-13 |
| 0\_13\_female | Female population aged 0-13 |
| metro\_min\_avto | Time to subway by car, min. |
| metro\_km\_avto | Distance to subway by car, km |
| metro\_min\_walk | Time to metro by foot |
| metro\_km\_walk | Distance to the metro, km |
| kindergarten\_km | Distance to kindergarten |
| school\_km | Distance to high school |
| park\_km | Distance to park |
| green\_zone\_km | Distance to green zone |
| industrial\_zone\_km | Distance to industrial zone |
| water\_treatment\_km | Distance to water treatment |
| cemetery\_km | Distance to the cemetery |
| incineration\_km | Distance to the incineration |
| railroad\_station\_walk\_km | Distance to the railroad station (walk) |
| railroad\_station\_walk\_min | Time to the railroad station (walk) |
| ID\_railroad\_station\_walk | Nearest railroad station id (walk) |
| railroad\_station\_avto\_km | Distance to the railroad station (avto) |
| railroad\_station\_avto\_min | Time to the railroad station (avto) |
| ID\_railroad\_station\_avto | Nearest railroad station id (avto) |
| public\_transport\_station\_km | Distance to the public transport station (walk) |
| public\_transport\_station\_min\_walk | Time to the public transport station (walk) |
| water\_km | Distance to the water reservoir / river |
| water\_1line | First line to the river (150 m) |
| mkad\_km | Distance to MKAD (Moscow Circle Auto Road) |
| ttk\_km | Distance to the TTC (Third Transport Ring) |
| sadovoe\_km | Distance to the Garden Ring |
| bulvar\_ring\_km | The distance to the Boulevard Ring |
| kremlin\_km | Distance to the city center (Kremlin) |
| big\_road1\_km | Distance to Nearest major road |
| ID\_big\_road1 | Nearest big road id |
| big\_road1\_1line | First line to the road (100 m for highways, 250 m to MKAD) |
| big\_road2\_km | The distance to next distant major road |
| ID\_big\_road2 | 2nd nearest big road id |
| railroad\_km | Distance to the railway / Moscow Central Ring / open areas Underground |
| railroad\_1line | First line to the railway (100 m) |
| zd\_vokzaly\_avto\_km | Distance to train station |
| ID\_railroad\_terminal | Nearest railroad terminal id |
| bus\_terminal\_avto\_km | Distance to bus terminal (avto) |
| ID\_bus\_terminal | Nearest bus terminal id |
| oil\_chemistry\_km | Distance to dirty industries |
| nuclear\_reactor\_km | Distance to nuclear reactor |
| radiation\_km | Distance to burial of radioactive waste |
| power\_transmission\_line\_km | Distance to power transmission line |
| thermal\_power\_plant\_km | Distance to thermal power plant |
| ts\_km | Distance to power station |
| big\_market\_km | Distance to grocery / wholesale markets |
| market\_shop\_km | Distance to markets and department stores |
| fitness\_km | Distance to fitness |
| swim\_pool\_km | Distance to swimming pool |
| ice\_rink\_km | Distance to ice palace |
| stadium\_km | Distance to stadium |
| basketball\_km | Distance to the basketball courts |
| hospice\_morgue\_km | Distance to hospice/morgue |
| detention\_facility\_km | Distance to detention facility |
| public\_healthcare\_km | Distance to public healthcare |
| university\_km | Distance to universities |
| workplaces\_km | Distance to workplaces |
| shopping\_centers\_km | Distance to shopping centers |
| office\_km | Distance to business centers/ offices |
| additional\_education\_km | Distance to additional education |
| preschool\_km | Distance to preschool education organizations |
| big\_church\_km | Distance to large church |
| church\_synagogue\_km | Distance to Christian churches and Synagogues |
| mosque\_km | Distance to mosques |
| theater\_km | Distance to theater |
| museum\_km | Distance to museums |
| exhibition\_km | Distance to exhibition |
| catering\_km | Distance to catering |
| ecology | Ecological zone where the house is located |
| green\_part\_500 | The share of green zones in 500 meters zone |
| prom\_part\_500 | The share of industrial zones in 500 meters zone |
| office\_count\_500 | The number of office space in 500 meters zone |
| office\_sqm\_500 | The square of office space in 500 meters zone |
| trc\_count\_500 | The number of shopping malls in 500 meters zone |
| trc\_sqm\_500 | The square of shopping malls in 500 meters zone |
| cafe\_count\_500 | The number of cafes or restaurants in 500 meters zone |
| cafe\_sum\_500\_min\_price\_avg | Cafes and restaurant min average bill in 500 meters zone |
| cafe\_sum\_500\_max\_price\_avg | Cafes and restaurant max average bill in 500 meters zone |
| cafe\_avg\_price\_500 | Cafes and restaurant average bill in 500 meters zone |
| cafe\_count\_500\_na\_price | Cafes and restaurant bill N/A in 500 meters zone |
| cafe\_count\_500\_price\_500 | Cafes and restaurant bill, average under 500 in 500 meters zone |
| fcafe\_count\_500\_price\_1000 | Cafes and restaurant bill, average 500-1000 in 500 meters zone |
| cafe\_count\_500\_price\_1500 | Cafes and restaurant bill, average 1000-1500 in 500 meters zone |
| cafe\_count\_500\_price\_2500 | Cafes and restaurant bill, average 1500-2500 in 500 meters zone |
| cafe\_count\_500\_price\_4000 | Cafes and restaurant bill, average 2500-4000 in 500 meters zone |
| cafe\_count\_500\_price\_high | Cafes and restaurant bill, average over 4000 in 500 meters zone |
| big\_church\_count\_500 | The number of big churches in 500 meters zone |
| church\_count\_500 | The number of churches in 500 meters zone |
| mosque\_count\_500 | The number of mosques in 500 meters zone |
| leisure\_count\_500 | The number of leisure facilities in 500 meters zone |
| sport\_count\_500 | The number of sport facilities in 500 meters zone |
| market\_count\_500 | The number of markets in 500 meters zone |
| green\_part\_1000 | The share of green zones in 1000 meters zone |
| prom\_part\_1000 | The share of industrial zones in 1000 meters zone |
| office\_count\_1000 | The number of office space in 1000 meters zone |
| office\_sqm\_1000 | The square of office space in 1000 meters zone |
| trc\_count\_1000 | The number of shopping malls in 1000 meters zone |
| trc\_sqm\_1000 | The square of shopping malls in 1000 meters zone |
| cafe\_count\_1000 | The number of cafes or restaurants in 1000 meters zone |
| cafe\_sum\_1000\_min\_price\_avg | Cafes and restaurant min average bill in 1000 meters zone |
| cafe\_sum\_1000\_max\_price\_avg | Cafes and restaurant max average bill in 1000 meters zone |
| cafe\_avg\_price\_1000 | Cafes and restaurant average bill in 1000 meters zone |
| cafe\_count\_1000\_na\_price | Cafes and restaurant bill N/A in 1000 meters zone |
| cafe\_count\_1000\_price\_500 | Cafes and restaurant bill, average under 500 in 1000 meters zone |
| cafe\_count\_1000\_price\_1000 | Cafes and restaurant bill, average 500-1000 in 1000 meters zone |
| cafe\_count\_1000\_price\_1500 | Cafes and restaurant bill, average 1000-1500 in 1000 meters zone |
| cafe\_count\_1000\_price\_2500 | Cafes and restaurant bill, average 1500-2500 in 1000 meters zone |
| cafe\_count\_1000\_price\_4000 | Cafes and restaurant bill, average 2500-4000 in 1000 meters zone |
| cafe\_count\_1000\_price\_high | Cafes and restaurant bill, average over 4000 in 1000 meters zone |
| big\_church\_count\_1000 | The number of big churches in 1000 meters zone |
| church\_count\_1000 | The number of churches in 1000 meters zone |
| mosque\_count\_1000 | The number of mosques in 1000 meters zone |
| leisure\_count\_1000 | The number of leisure facilities in 1000 meters zone |
| sport\_count\_1000 | The number of sport facilities in 1000 meters zone |
| market\_count\_1000 | The number of markets in 1000 meters zone |
| green\_part\_1500 | The share of green zones in 1500 meters zone |
| prom\_part\_1500 | The share of industrial zones in 1500 meters zone |
| office\_count\_1500 | The number of office space in 1500 meters zone |
| office\_sqm\_1500 | The square of office space in 1500 meters zone |
| trc\_count\_1500 | The number of shopping malls in 1500 meters zone |
| trc\_sqm\_1500 | The square of shopping malls in 1500 meters zone |
| cafe\_count\_1500 | The number of cafes or restaurants in 1500 meters zone |
| cafe\_sum\_1500\_min\_price\_avg | Cafes and restaurant min average bill in 1500 meters zone |
| cafe\_sum\_1500\_max\_price\_avg | Cafes and restaurant max average bill in 1500 meters zone |
| cafe\_avg\_price\_1500 | Cafes and restaurant average bill in 1500 meters zone |
| cafe\_count\_1500\_na\_price | Cafes and restaurant bill N/A in 1500 meters zone |
| cafe\_count\_1500\_price\_500 | Cafes and restaurant bill, average under 500 in 1500 meters zone |
| cafe\_count\_1500\_price\_1000 | Cafes and restaurant bill, average 500-1000 in 1500 meters zone |
| cafe\_count\_1500\_price\_1500 | Cafes and restaurant bill, average 1000-1500 in 1500 meters zone |
| cafe\_count\_1500\_price\_2500 | Cafes and restaurant bill, average 1500-2500 in 1500 meters zone |
| cafe\_count\_1500\_price\_4000 | Cafes and restaurant bill, average 2500-4000 in 1500 meters zone |
| cafe\_count\_1500\_price\_high | Cafes and restaurant bill, average over 4000 in 1500 meters zone |
| big\_church\_count\_1500 | The number of big churches in 1500 meters zone |
| church\_count\_1500 | The number of churches in 1500 meters zone |
| mosque\_count\_1500 | The number of mosques in 1500 meters zone |
| leisure\_count\_1500 | The number of leisure facilities in 1500 meters zone |
| sport\_count\_1500 | The number of sport facilities in 1500 meters zone |
| market\_count\_1500 | The number of markets in 1500 meters zone |
| green\_part\_2000 | The share of green zones in 2000 meters zone |
| prom\_part\_2000 | The share of industrial zones in 2000 meters zone |
| office\_count\_2000 | The number of office space in 2000 meters zone |
| office\_sqm\_2000 | The square of office space in 2000 meters zone |
| trc\_count\_2000 | The number of shopping malls in 2000 meters zone |
| trc\_sqm\_2000 | The square of shopping malls in 2000 meters zone |
| cafe\_count\_2000 | The number of cafes or restaurants in 1500 meters zone |
| cafe\_sum\_2000\_min\_price\_avg | Cafes and restaurant min average bill in 2000 meters zone |
| cafe\_sum\_2000\_max\_price\_avg | Cafes and restaurant max average bill in 2000 meters zone |
| cafe\_avg\_price\_2000 | Cafes and restaurant average bill in 2000 meters zone |
| cafe\_count\_2000\_na\_price | Cafes and restaurant bill N/A in 2000 meters zone |
| cafe\_count\_2000\_price\_500 | Cafes and restaurant bill, average under 500 in 2000 meters zone |
| cafe\_count\_2000\_price\_1000 | Cafes and restaurant bill, average 500-1000 in 2000 meters zone |
| cafe\_count\_2000\_price\_1500 | Cafes and restaurant bill, average 1000-1500 in 2000 meters zone |
| cafe\_count\_2000\_price\_2500 | Cafes and restaurant bill, average 1500-2500 in 2000 meters zone |
| cafe\_count\_2000\_price\_4000 | Cafes and restaurant bill, average 2500-4000 in 2000 meters zone |
| cafe\_count\_2000\_price\_high | Cafes and restaurant bill, average over 4000 in 2000 meters zone |
| big\_church\_count\_2000 | The number of big churches in 2000 meters zone |
| church\_count\_2000 | The number of churches in 2000 meters zone |
| mosque\_count\_2000 | The number of mosques in 2000 meters zone |
| leisure\_count\_2000 | The number of leisure facilities in 2000 meters zone |
| sport\_count\_2000 | The number of sport facilities in 2000 meters zone |
| market\_count\_2000 | The number of markets in 2000 meters zone |
| green\_part\_3000 | The share of green zones in 3000 meters zone |
| prom\_part\_3000 | The share of industrial zones in 3000 meters zone |
| office\_count\_3000 | The number of office space in 3000 meters zone |
| office\_sqm\_3000 | The square of office space in 3000 meters zone |
| trc\_count\_3000 | The number of shopping malls in 3000 meters zone |
| trc\_sqm\_3000 | The square of shopping malls in 3000 meters zone |
| cafe\_count\_3000 | The number of cafes or restaurants in 1500 meters zone |
| cafe\_sum\_3000\_min\_price\_avg | Cafes and restaurant min average bill in 3000 meters zone |
| cafe\_sum\_3000\_max\_price\_avg | Cafes and restaurant max average bill in 3000 meters zone |
| cafe\_avg\_price\_3000 | Cafes and restaurant average bill in 3000 meters zone |
| cafe\_count\_3000\_na\_price | Cafes and restaurant bill N/A in 3000 meters zone |
| cafe\_count\_3000\_price\_500 | Cafes and restaurant bill, average under 500 in 3000 meters zone |
| cafe\_count\_3000\_price\_1000 | Cafes and restaurant bill, average 500-1000 in 3000 meters zone |
| cafe\_count\_3000\_price\_1500 | Cafes and restaurant bill, average 1000-1500 in 3000 meters zone |
| cafe\_count\_3000\_price\_2500 | Cafes and restaurant bill, average 1500-2500 in 3000 meters zone |
| cafe\_count\_3000\_price\_4000 | Cafes and restaurant bill, average 2500-4000 in 3000 meters zone |
| cafe\_count\_3000\_price\_high | Cafes and restaurant bill, average over 4000 in 3000 meters zone |
| big\_church\_count\_3000 | The number of big churches in 3000 meters zone |
| church\_count\_3000 | The number of churches in 3000 meters zone |
| mosque\_count\_3000 | The number of mosques in 3000 meters zone |
| leisure\_count\_3000 | The number of leisure facilities in 3000 meters zone |
| sport\_count\_3000 | The number of sport facilities in 3000 meters zone |
| market\_count\_3000 | The number of markets in 3000 meters zone |
| green\_part\_5000 | The share of green zones in 5000 meters zone |
| prom\_part\_5000 | The share of industrial zones in 5000 meters zone |
| office\_count\_5000 | The number of office space in 5000 meters zone |
| office\_sqm\_5000 | The square of office space in 5000 meters zone |
| trc\_count\_5000 | The number of shopping malls in 5000 meters zone |
| trc\_sqm\_5000 | The square of shopping malls in 5000 meters zone |
| cafe\_count\_5000 | The number of cafes or restaurants in 1500 meters zone |
| cafe\_sum\_5000\_min\_price\_avg | Cafes and restaurant min average bill in 5000 meters zone |
| cafe\_sum\_5000\_max\_price\_avg | Cafes and restaurant max average bill in 5000 meters zone |
| cafe\_avg\_price\_5000 | Cafes and restaurant average bill in 5000 meters zone |
| cafe\_count\_5000\_na\_price | Cafes and restaurant bill N/A in 5000 meters zone |
| cafe\_count\_5000\_price\_500 | Cafes and restaurant bill, average under 500 in 5000 meters zone |
| cafe\_count\_5000\_price\_1000 | Cafes and restaurant bill, average 500-1000 in 5000 meters zone |
| cafe\_count\_5000\_price\_1500 | Cafes and restaurant bill, average 1000-1500 in 5000 meters zone |
| cafe\_count\_5000\_price\_2500 | Cafes and restaurant bill, average 1500-2500 in 5000 meters zone |
| cafe\_count\_5000\_price\_4000 | Cafes and restaurant bill, average 2500-4000 in 5000 meters zone |
| cafe\_count\_5000\_price\_high | Cafes and restaurant bill, average over 4000 in 5000 meters zone |
| big\_church\_count\_5000 | The number of big churches in 5000 meters zone |
| church\_count\_5000 | The number of churches in 5000 meters zone |
| mosque\_count\_5000 | The number of mosques in 5000 meters zone |
| leisure\_count\_5000 | The number of leisure facilities in 5000 meters zone |
| sport\_count\_5000 | The number of sport facilities in 5000 meters zone |
| market\_count\_5000 | The number of markets in 5000 meters zone |

**Data Wrangling Detail**

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Original  Value | Transformed Value | Method / Explanation |
| timestamp | Integer | Date | Dates are the common way of presenting time values |
| full\_sq | NA | 54.3 | Used means of the category for values greater than or equal to 2 |
| full\_sq | 0 | 54.3 | Used means of the category for values greater than or equal to 2 |
| full\_sq | 1 | 54.3 | Used means of the category for values greater than or equal to 2 |
| life\_sq | NA | 34.6 | Used means of the category for values greater than or equal to 2 |
| life\_sq | 0 | 34.6 | Used means of the category for values greater than or equal to 2 |
| life\_sq | 1 | 34.6 | Used means of the category for values greater than or equal to 2 |
| floor | NA | 8 | Used means of the category |
| floor | 0 | 1 | All structures must have at least one floor. |
| max\_floor | NA | Set to “floor” category value | Maximum floor value is unknowable.  Set to most reliable data about the number of floors for the structure. |
| max\_floor | 0 | Set to “floor” category value | Maximum floor value is unknowable.  Set to most reliable data about the number of floors for the structure. |
| material | NA | 1 | Most common material type |
| build\_year | 4965 | 1965 | Appears to be a typo |
| build\_year | 1691 | 1991 | Appears to be a typo |
| build\_year | 20052009 | 2007 | Original value appears to be a range of 2005-2009.  Used the means of the range. |
| build\_year | 71 | 1971 | Appears to be missing the 1900 in front of the year |
| build\_year | NA | 1985 | Used means of the category for values greater than 215. |
| build\_year | 0 | 1985 | Used means of the category for values greater than 215. |
| build\_year | 1 | 1985 | Used means of the category for values greater than 215. |
| build\_year | 3 | 1985 | Used means of the category for values greater than 215. |
| build\_year | 20 | 1985 | Used means of the category for values greater than 215. |
| build\_year | 215 | 1985 | Used means of the category for values greater than 215. |
| num\_rooms | NA | 2 | Used means of the category for values greater than 2. |
| num\_rooms | 0 | 1 | Unable to have a zero rooms |
| kitchen\_sq | NA | 6.0 | Used means of the category |
| kitchen\_sq | 1974 | 1 | The value was the same as the “build\_year” category |
| kitchen\_sq | 2013 | 1 | The value was the same as the “build\_year” category |
| kitchen\_sq | 2014 | 1 | The value was the same as the “build\_year” category |
| kitchen\_sq | 620 | 1 | The value was larger than the entire structure. |
| kitchen\_sq | 1970 | 1 | The value was larger than the entire structure. |
| state | NA | 2 | Used the means of the category. |
| state | 33 | 3 | Appears to be a typo |
| preschool\_quota | NA | 3273.8 | Used the means of the category. |
| school\_quota | NA | 1194.8 | Used the means of the category. |
| raion\_build\_count\_with\_material\_info | NA | 328.2 | Used the means of the category. |
| build\_cont\_wood | NA | 50.2 | Used the means of the category. |
| build\_count\_frame | NA | 40.3 | Used the means of the category. |
| build\_count\_brick | NA | 108.1 | Used the means of the category. |
| build\_count\_  monolith | NA | 12.0 | Used the means of the category. |
| build\_count\_panel | NA | 107.5 | Used the means of the category. |
| build\_count\_foam | NA | 0.2 | Used the means of the category. |
| build\_count\_slag | NA | 4.5 | Used the means of the category. |
| build\_count\_mix | NA | 0.6 | Used the means of the category. |
| raion\_build\_count\_with\_builddate\_info | NA | 327.9 | Used the means of the category. |
| build\_count\_before\_1920 | NA | 18.8 | Used the means of the category. |
| build\_count\_1921-  1945 | NA | 26.6 | Used the means of the category. |
| build\_count\_1946-  1970 | NA | 141.0 | Used the means of the category. |
| build\_count\_1971-  1995 | NA | 80.4 | Used the means of the category. |
| build\_count\_after\_  1995 | NA | 61.1 | Used the means of the category. |
| metro\_min\_walk | NA | 42.8 | Used the means of the category. |
| metro\_km\_walk | NA | 5.5 | Used the means of the category. |
| railroad\_station\_  walk\_km | NA | 4.4 | Used the means of the category. |
| railroad\_station\_  walk\_min | NA | 52.7 | Used the means of the category. |
| ID\_railroad\_station\_walk | NA | 38.9 | Used the means of the category. |
| cafe\_sum\_500\_min\_price\_avg | NA | 740.3 | Used the means of the category. |
| cafe\_sum\_500\_max\_price\_avg | NA | 1245.6 | Used the means of the category. |
| cafe\_avg\_price\_500 | NA | 993.0 | Used the means of the category. |
| cafe\_sum\_1000\_  min\_price\_avg | NA | 710.1 | Used the means of the category. |
| cafe\_sum\_1000\_  max\_price\_avg | NA | 1205.4 | Used the means of the category. |
| cafe\_avg\_price\_  1000 | NA | 957.7 | Used the means of the category. |
| cafe\_sum\_1500\_  min\_price\_avg | NA | 713.2 | Used the means of the category. |
| cafe\_sum\_1500\_  max\_price\_avg | NA | 1204.7 | Used the means of the category. |
| cafe\_avg\_price\_  1500 | NA | 958.9 | Used the means of the category. |
| cafe\_sum\_2000\_  min\_price\_avg | NA | 719.8 | Used the means of the category. |
| cafe\_sum\_2000\_  max\_price\_avg | NA | 1210.8 | Used the means of the category. |
| cafe\_avg\_price\_  2000 | NA | 965.3 | Used the means of the category. |
| cafe\_sum\_3000\_  min\_price\_avg | NA | 765.7 | Used the means of the category. |
| cafe\_sum\_3000\_  max\_price\_avg | NA | 1283.1 | Used the means of the category. |
| cafe\_avg\_price\_  3000 | NA | 1024.4 | Used the means of the category. |
| prom\_part\_5000 | NA | 10.3 | Used the means of the category. |
| cafe\_sum\_5000\_  min\_price\_avg | NA | 765.5 | Used the means of the category. |
| cafe\_sum\_5000\_  max\_price\_avg | NA | 1278.9 | Used the means of the category. |
| cafe\_avg\_price\_  5000 | NA | 1022.2 | Used the means of the category. |

**Appendix 2: Code**

1. Goal 1- Data importing with SAS

/\* MODELING DATA \*/

%web\_drop\_table(WORK.MODELING);

FILENAME REFFILE '/folders/myfolders/DataFiles/modelingData\_Adj\_06-13.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT= WORK.MODELING;

GETNAMES=YES;

guessingrows=12200;

RUN;

PROC CONTENTS DATA=WORK.MODELING; RUN;

%web\_open\_table(WORK.MODELING);

/\* PROJECTION DATA \*/

%web\_drop\_table(WORK.PROJECTION);

FILENAME REFFILE '/folders/myfolders/DataFiles/projectionData\_Adj\_06-13.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT= WORK.PROJECTION;

GETNAMES=YES;

guessingrows= max;

RUN;

PROC CONTENTS DATA=WORK.PROJECTION; RUN;

%web\_open\_table(WORK.PROJECTION);

1. Goal 1 – SAS code for data wrangling

/\* MODELING DATA CLEANING \*/

data WORK.CLEAN\_MODELING ;

set WORK.MODELING;

region=compress(sub\_area,"'");

if full\_sq = 5326 then delete;

if full\_sq >200 then delete;

if full\_sq <2 then full\_sq = 54.3;

if missing(full\_sq) then full\_sq = 54.3;

if life\_sq = 7478 then delete;

/\* if full\_sq >200 then delete; \*/

if full\_sq <2 then full\_sq = 34.6;

if missing(full\_sq) then full\_sq = 34.6;

if floor = 77 then delete;

if floor = 0 then floor =1;

if missing(floor) then floor = 8;

if max\_floor = 0 then max\_floor=floor;

if max\_floor > 98 then delete;

if missing(max\_floor) then max\_floor=floor;

if missing(material) then material = 1;

if build\_year = 4965 then build\_year = 1965;

if build\_year = 1691 then build\_year = 1991;

if build\_year = 20052009 then build\_year = 2007;

if build\_year = 71 then build\_year = 1971;

if build\_year < 220 then delete;

if missing(build\_year) then build\_year = 1985;

if num\_rooms = 0 then num\_rooms = 1;

if missing(num\_rooms) then num\_rooms = 2;

if kitch\_sq = 1974 then kitch\_sq = 1;

if kitch\_sq = 2013 then kitch\_sq = 1;

if kitch\_sq = 2014 then kitch\_sq = 1;

if kitch\_sq > 500 then delete;

if missing(kitch\_sq) then kitch\_sq = 6;

if state = 33 then state = 3;

if missing(state) then state = 2;

if missing(preschool\_quota) then preschool\_quota = 3273.8;

if missing(school\_quota) then school\_quota = 1194.8;

if missing(raion\_build\_count\_with\_material\_) then raion\_build\_count\_with\_material\_ = 328.2;

if missing(build\_cont\_wood) then build\_cont\_wood = 50.2;

if missing(build\_count\_frame) then build\_count\_frame = 40.3;

if missing(build\_count\_brick) then build\_count\_brick = 108.1;

if missing(build\_count\_monolith) then build\_count\_monolith = 12.0;

if missing(build\_count\_panel) then build\_count\_panel = 107.5;

if missing(build\_count\_foam) then build\_count\_foam = 0.2;

if missing(build\_count\_slag) then build\_count\_slag = 4.5;

if missing(build\_count\_mix) then build\_count\_mix = 0.6;

if missing(raion\_build\_count\_with\_builddate) then raion\_build\_count\_with\_builddate = 327.9;

if missing(build\_count\_before\_1920) then build\_count\_before\_1920 = 18.8;

if missing(build\_count\_1921\_1945) then build\_count\_1921\_1945 = 26.6;

if missing(build\_count\_1946\_1970) then build\_count\_1946\_1970 = 141.0;

if missing(build\_count\_1971\_1995) then build\_count\_1971\_1995 = 80.4;

if missing(build\_count\_after\_1995) then build\_count\_after\_1995 = 61.1;

if missing(metro\_min\_walk) then metro\_min\_walk = 42.8;

if missing(metro\_km\_walk) then metro\_km\_walk = 5.5;

if missing(railroad\_station\_walk\_km) then railroad\_station\_walk\_km = 4.4;

if missing(railroad\_station\_walk\_min) then railroad\_station\_walk\_min = 52.7;

if missing(ID\_railroad\_station\_walk) then ID\_railroad\_station\_walk = 38.9;

if missing(cafe\_sum\_500\_min\_price\_avg) then cafe\_sum\_500\_min\_price\_avg = 740.3;

if missing(cafe\_sum\_500\_max\_price\_avg) then cafe\_sum\_500\_max\_price\_avg = 1245.6;

if missing(cafe\_avg\_price\_500) then cafe\_avg\_price\_500 = 993.0;

if missing(cafe\_sum\_1000\_min\_price\_avg) then cafe\_sum\_1000\_min\_price\_avg = 710.1;

if missing(cafe\_sum\_1000\_max\_price\_avg) then cafe\_sum\_1000\_max\_price\_avg = 1205.4;

if missing(cafe\_avg\_price\_1000) then cafe\_avg\_price\_1000 = 957.7;

if missing(cafe\_sum\_1500\_min\_price\_avg) then cafe\_sum\_1500\_min\_price\_avg = 713.2;

if missing(cafe\_sum\_1500\_max\_price\_avg) then cafe\_sum\_1500\_max\_price\_avg = 1204.7;

if missing(cafe\_avg\_price\_1500) then cafe\_avg\_price\_1500 = 958.9;

if missing(cafe\_sum\_2000\_min\_price\_avg) then cafe\_sum\_2000\_min\_price\_avg = 719.8;

if missing(cafe\_sum\_2000\_max\_price\_avg) then cafe\_sum\_2000\_max\_price\_avg = 1210.8;

if missing(cafe\_avg\_price\_2000) then cafe\_avg\_price\_2000 = 965.3;

if missing(cafe\_sum\_3000\_min\_price\_avg) then cafe\_sum\_3000\_min\_price\_avg = 765.7;

if missing(cafe\_sum\_3000\_max\_price\_avg) then cafe\_sum\_3000\_max\_price\_avg = 1283.1;

if missing(cafe\_avg\_price\_3000) then cafe\_avg\_price\_3000 = 1024.4;

if missing(prom\_part\_5000) then prom\_part\_5000 = 10.3;

if missing(cafe\_sum\_5000\_min\_price\_avg) then cafe\_sum\_5000\_min\_price\_avg = 765.5;

if missing(cafe\_sum\_5000\_max\_price\_avg) then cafe\_sum\_5000\_max\_price\_avg = 1278.9;

if missing(cafe\_avg\_price\_5000) then cafe\_avg\_price\_5000 = 1022.2;

idd = input(VAR1, 8.);

run;

/\* proc print data=WORK.CLEAN\_MODELING (obs=10); run; \*/

/\* PROJECTION DATA CLEANING \*/

data WORK.CLEAN\_PROJECTION ;

set WORK.PROJECTION;

region=compress(sub\_area,"'");

if full\_sq = 5326 then delete;

if full\_sq >200 then delete;

if full\_sq <2 then full\_sq = 54.3;

if missing(full\_sq) then full\_sq = 54.3;

if life\_sq = 7478 then delete;

/\* if full\_sq >200 then delete; \*/

if full\_sq <2 then full\_sq = 34.6;

if missing(full\_sq) then full\_sq = 34.6;

if floor = 77 then delete;

if floor = 0 then floor =1;

if missing(floor) then floor = 8;

if max\_floor = 0 then max\_floor =floor;

if max\_floor > 98 then delete;

if missing(max\_floor) then max\_floor=floor;

if missing(material) then material = 1;

if build\_year = 4965 then build\_year = 1965;

if build\_year = 1691 then build\_year = 1991;

if build\_year = 20052009 then build\_year = 2007;

if build\_year = 71 then build\_year = 1971;

if build\_year < 220 then delete;

if missing(build\_year) then build\_year = 1985;

if num\_rooms = 0 then num\_rooms = 1;

if missing(num\_rooms) then num\_rooms = 2;

if kitch\_sq = 1974 then kitch\_sq = 1;

if kitch\_sq = 2013 then kitch\_sq = 1;

if kitch\_sq = 2014 then kitch\_sq = 1;

if kitch\_sq > 500 then delete;

if missing(kitch\_sq) then kitch\_sq = 6;

if state = 33 then state = 3;

if missing(state) then state = 2;

if missing(preschool\_quota) then preschool\_quota = 3273.8;

if missing(school\_quota) then school\_quota = 1194.8;

if missing(raion\_build\_count\_with\_material\_) then raion\_build\_count\_with\_material\_ = 328.2;

if missing(build\_cont\_wood) then build\_cont\_wood = 50.2;

if missing(build\_count\_frame) then build\_count\_frame = 40.3;

if missing(build\_count\_brick) then build\_count\_brick = 108.1;

if missing(build\_count\_monolith) then build\_count\_monolith = 12.0;

if missing(build\_count\_panel) then build\_count\_panel = 107.5;

if missing(build\_count\_foam) then build\_count\_foam = 0.2;

if missing(build\_count\_slag) then build\_count\_slag = 4.5;

if missing(build\_count\_mix) then build\_count\_mix = 0.6;

if missing(raion\_build\_count\_with\_builddate) then raion\_build\_count\_with\_builddate = 327.9;

if missing(build\_count\_before\_1920) then build\_count\_before\_1920 = 18.8;

if missing(build\_count\_1921\_1945) then build\_count\_1921\_1945 = 26.6;

if missing(build\_count\_1946\_1970) then build\_count\_1946\_1970 = 141.0;

if missing(build\_count\_1971\_1995) then build\_count\_1971\_1995 = 80.4;

if missing(build\_count\_after\_1995) then build\_count\_after\_1995 = 61.1;

if missing(metro\_min\_walk) then metro\_min\_walk = 42.8;

if missing(metro\_km\_walk) then metro\_km\_walk = 5.5;

if missing(railroad\_station\_walk\_km) then railroad\_station\_walk\_km = 4.4;

if missing(railroad\_station\_walk\_min) then railroad\_station\_walk\_min = 52.7;

if missing(ID\_railroad\_station\_walk) then ID\_railroad\_station\_walk = 38.9;

if missing(cafe\_sum\_500\_min\_price\_avg) then cafe\_sum\_500\_min\_price\_avg = 740.3;

if missing(cafe\_sum\_500\_max\_price\_avg) then cafe\_sum\_500\_max\_price\_avg = 1245.6;

if missing(cafe\_avg\_price\_500) then cafe\_avg\_price\_500 = 993.0;

if missing(cafe\_sum\_1000\_min\_price\_avg) then cafe\_sum\_1000\_min\_price\_avg = 710.1;

if missing(cafe\_sum\_1000\_max\_price\_avg) then cafe\_sum\_1000\_max\_price\_avg = 1205.4;

if missing(cafe\_avg\_price\_1000) then cafe\_avg\_price\_1000 = 957.7;

if missing(cafe\_sum\_1500\_min\_price\_avg) then cafe\_sum\_1500\_min\_price\_avg = 713.2;

if missing(cafe\_sum\_1500\_max\_price\_avg) then cafe\_sum\_1500\_max\_price\_avg = 1204.7;

if missing(cafe\_avg\_price\_1500) then cafe\_avg\_price\_1500 = 958.9;

if missing(cafe\_sum\_2000\_min\_price\_avg) then cafe\_sum\_2000\_min\_price\_avg = 719.8;

if missing(cafe\_sum\_2000\_max\_price\_avg) then cafe\_sum\_2000\_max\_price\_avg = 1210.8;

if missing(cafe\_avg\_price\_2000) then cafe\_avg\_price\_2000 = 965.3;

if missing(cafe\_sum\_3000\_min\_price\_avg) then cafe\_sum\_3000\_min\_price\_avg = 765.7;

if missing(cafe\_sum\_3000\_max\_price\_avg) then cafe\_sum\_3000\_max\_price\_avg = 1283.1;

if missing(cafe\_avg\_price\_3000) then cafe\_avg\_price\_3000 = 1024.4;

if missing(prom\_part\_5000) then prom\_part\_5000 = 10.3;

if missing(cafe\_sum\_5000\_min\_price\_avg) then cafe\_sum\_5000\_min\_price\_avg = 765.5;

if missing(cafe\_sum\_5000\_max\_price\_avg) then cafe\_sum\_5000\_max\_price\_avg = 1278.9;

if missing(cafe\_avg\_price\_5000) then cafe\_avg\_price\_5000 = 1022.2;

log\_full\_sq= log(full\_sq);

idd= idd+28000;

run;

1. Goal 1 – SAS code to check assumptions

**proc** **corr** data=housing.clean\_modeling rank PEARSON

plots (maxpoints = **40000**)= scatter(alpha=**.05** );

var full\_sq life\_sq floor max\_floor material;

with price\_doc ;

title Analyss ;

**run**;

**proc** **corr** data=housing.clean\_modeling rank PEARSON

plots (maxpoints = **40000**)= all;

var full\_sq;

with price\_doc ;

title Analyss ;

**run**;

**proc** **corr** data= housing.clean\_modeling rank PEARSON

plots (maxpoints = **40000**)= scatter(alpha=**.05** );

var build\_year num\_room

kitch\_sq state area\_m ;

with price\_doc ;

title Analyss ;

**run**;

**proc** **corr** data=work.refined\_data rank PEARSON

plots (maxpoints = **40000**)= scatter(alpha=**.05** );

var raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school

;

with price\_doc ;

title Analyss ;

**run**;

**proc** **corr** data=work.refined\_data rank PEARSON

plots (maxpoints = **40000**)= scatter(alpha=**.05** );

var children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion

1. Goal 1 – SAS code to check multicollinearity

**proc** **corr** data= housing.clean\_modeling;

var full\_sq life\_sq floor max\_floor material build\_year num\_room

kitch\_sq state area\_m raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion university\_top\_20\_raion

sport\_objects\_raion additional\_education\_raion

culture\_objects\_top\_25\_raion shopping\_centers\_raion ;

**run**;

**proc** **glm** data=housing.clean\_modeling;

class preschool\_quota children\_school school\_quota;

model price\_doc = full\_sq raion\_popul life\_sq preschool\_quota children\_school school\_quota;

**run**;

1. Goal 1 – SAS code for model comparison

/\* Backward selection. \*/

**proc** **glmselect** data=housing.merged\_data;

class life\_sq floor max\_floor material build\_year num\_room

kitch\_sq state product\_type sub\_area area\_m raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion university\_top\_20\_raion

sport\_objects\_raion additional\_education\_raion culture\_objects\_top\_25

culture\_objects\_top\_25\_raion shopping\_centers\_raion office\_raion

thermal\_power\_plant\_raion incineration\_raion oil\_chemistry\_raion

radiation\_raion railroad\_terminal\_raion big\_market\_raion nuclear\_reactor\_raion

detention\_facility\_raion;

model price\_doc = full\_sq life\_sq floor max\_floor material build\_year num\_room

kitch\_sq state product\_type sub\_area area\_m raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion university\_top\_20\_raion

sport\_objects\_raion additional\_education\_raion culture\_objects\_top\_25

culture\_objects\_top\_25\_raion shopping\_centers\_raion office\_raion

thermal\_power\_plant\_raion incineration\_raion oil\_chemistry\_raion

radiation\_raion railroad\_terminal\_raion big\_market\_raion nuclear\_reactor\_raion

detention\_facility\_raion

/ selection=Backward(stop=CV) cvmethod=random(**5**) stats=adjrsq;

output out = results p = predict ;

**run**;

**data** result\_backward\_selection ;

set results;

price\_doc = predict;

if missing(predict) then price\_doc = **140000**;

keep id price\_doc;

where idd > **28000**;

**proc** **export** data=result\_backward\_selection dbms=csv

outfile="/folders/myfolders/stats2\_hw/term\_proj1/output/result\_backward\_selection.csv"

replace;

**run**;

**proc** **print** data = result\_Backward\_selection;

**run**;

/\* Stepwise. \*/

**proc** **glmselect** data=housing.merged\_data;

class life\_sq floor max\_floor material build\_year num\_room

kitch\_sq state product\_type sub\_area area\_m raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion university\_top\_20\_raion

sport\_objects\_raion additional\_education\_raion culture\_objects\_top\_25

culture\_objects\_top\_25\_raion shopping\_centers\_raion office\_raion

thermal\_power\_plant\_raion incineration\_raion oil\_chemistry\_raion

radiation\_raion railroad\_terminal\_raion big\_market\_raion nuclear\_reactor\_raion

detention\_facility\_raion;

model price\_doc = full\_sq life\_sq floor max\_floor material build\_year num\_room

kitch\_sq state product\_type sub\_area area\_m raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion university\_top\_20\_raion

sport\_objects\_raion additional\_education\_raion culture\_objects\_top\_25

culture\_objects\_top\_25\_raion shopping\_centers\_raion office\_raion

thermal\_power\_plant\_raion incineration\_raion oil\_chemistry\_raion

radiation\_raion railroad\_terminal\_raion big\_market\_raion nuclear\_reactor\_raion

detention\_facility\_raion

/ selection=Stepwise(stop=CV) cvmethod=random(**5**) stats=adjrsq;

output out = results p = predict ;

**run**;

/\* LASSO \*/

proc glmselect data= WORK.MERGED\_DATA;

class full\_sq life\_sq floor max\_floor material build\_year num\_room

kitch\_sq state product\_type sub\_area area\_m raion\_popul green\_zone\_part

indust\_part children\_preschool preschool\_quota

children\_school school\_quota school\_education\_centers\_raion

hospital\_beds\_raion healthcare\_centers\_raion university\_top\_20\_raion

sport\_objects\_raion additional\_education\_raion culture\_objects\_top\_25

culture\_objects\_top\_25\_raion shopping\_centers\_raion office\_raion

thermal\_power\_plant\_raion incineration\_raion oil\_chemistry\_raion

radiation\_raion railroad\_terminal\_raion big\_market\_raion nuclear\_reactor\_raion

detention\_facility\_raion;

model price\_doc = full\_sq -- detention\_facility\_raion

/ selection=LASSO(choose=CV stop=CV) CVdetails;

output out = results p = predict ;

run;

data result\_LASSO\_selection ;

set results;

price\_doc = predict;

if missing(predict) then price\_doc = 140000;

keep id price\_doc;

where idd > 28000;

proc print data = result\_LASSO\_selection; run;

1. Goal 2 – R code for data wrangling

# Project 01, Goal 02 - Pankaj, Karl, Shawn

# ---------------------------------------------------------

# MSDS 6372 - Project#1 - Goal 2

#

# Part A - Data Wrangling for Time Series Analysis

# Pankaj Kumar, Shawn Jung, Karl Jurek

# ---------------------------------------------------------

# reading the original modeling data

modelingData <- read.csv("C:/Users/shawn/OneDrive/SMU/Applied Stat/Project01/modelingData.csv")

# Select ID, timestamp and price, and transform to YYYY-MM-DD

goal2\_data <- modelingData[c("id", "timestamp", "price\_doc")]

goal2\_data$timestamp <- as.Date(goal2\_data$timestamp, origin = "1899-12-30")

# Parsing day, month, year and month-year as separate columns

goal2\_data$day <- format.Date(goal2\_data$timestamp, "%d")

goal2\_data$month <- format.Date(goal2\_data$timestamp, "%m")

goal2\_data$year <- format.Date(goal2\_data$timestamp, "%Y")

goal2\_data$monthYear <- paste(goal2\_data$year, goal2\_data$month, sep="")

# Aggregate by month-year

goal2\_agg <- aggregate(goal2\_data, by=list(goal2\_data$monthYear), FUN = mean)

goal2\_agg <- goal2\_agg[c("Group.1", "price\_doc")]

names(goal2\_agg) <- c("monthYear", "AvgPrice")

goal2\_agg$monthNumber <- seq.int(nrow(goal2\_agg))

# write to CSV to load from SAS

write.csv(goal2\_agg, file="goal2data.csv")

1. Goal 2 – SAS code for Time Series Analysis

/\* ---------------------------------------------------------

\* MSDS 6372 - Project#1 - Goal 2

\* Part B - Time Seriese Analysis of Average House Price

\* Pankaj Kumar, Shawn Jung, Karl Jurek

\* --------------------------------------------------------- \*/

/\* set data file location as a separate path \*/

FILENAME REFILL '/home/shawnj0/sasuser.v94/AppliedStat/goal2data.csv';

/\* load the wrangled data as GOAL2 \*/

PROC IMPORT DATAFILE=REFILL

DBMS=CSV

OUT= GOAL2;

GETNAMES=YES;

RUN;

/\* inspect if the data is ok \*/

PROC CONTENTS DATA=GOAL2; RUN;

PROC PRINT DATA=GOAL2;RUN;

/\* visually inspect the data \*/

PROC SGPLOT DATA=GOAL2;

SERIESE X = MONTH Y= AVGPRICE;

RUN;

/\* Fit a simple lineare regression \*/

PROC GLM DATA=GOAL2 PLOT=ALL ;

MODEL AVGPRICE = MONTH /SOLUTION;

RUN;

/\* obtain residuals and plot them \*/

PROC REG DATA=GOAL2;

MODEL AVGPRICE = MONTH / R ;

OUTPUT OUT=GOAL2RESID R=RESID ;

RUN;

PROC SGPLOT DATA=GOAL2RESID;

SERIESE X = MONTH Y= RESID;

RUN;

/\* Check autocorrelation on residuals \*/

PROC AUTOREG DATA=GOAL2RESID; /\* without any lag \*/

MODEL RESID = MONTH / DWPROB;

RUN;

PROC AUTOREG DATA=GOAL2RESID; /\* try lag 1 \*/

MODEL RESID = MONTH / NLAG = (1) DWPROB;

RUN;

/\* Prediction of residuals - prepare data \*/

DATA FORPRED;

INPUT monthYear $ month @@ ; cards;

201507 49

201508 50

201509 51

201510 52

201511 53

201512 54

201601 55

201602 56

201603 57

201604 58

201605 59

201606 60

;

RUN;

DATA GOAL2\_PRED; /\*combine two sets \*/

SET GOAL2 FORPRED;

RUN;

/\* running prediction \*/

PROC AUTOREG DATA=GOAL2\_PRED ;

MODEL AVGPRICE = MONTH / NLAG = (1) DWPROB;

OUTPUT OUT = PREDS\_PRICE P = PREDICTIONS LCL = LOWER UCL = UPPER PM = YTREND;

RUN;

PROC SGPLOT DATA = PREDS\_PRICE;

BAND X = MONTH UPPER = UPPER LOWER = LOWER;

SCATTER X = MONTH Y = AVGPRICE;

SERIES X = MONTH Y = PREDICTIONS;

SERIES X = MONTH Y = YTREND / LINEATTRS= (COLOR=BLACK);

RUN;

PROC PRINT DATA=PREDS\_PRICE;RUN;

/\* forecating July 2015 to June 2016 \*/

DATA FORCASTED\_PRICE;

SET PREDS\_PRICE;

KEEP MONTH MONTHYEAR PREDICTIONS LOWER UPPER;

WHERE MONTH > 47;

;

PROC PRINT DATA=FORCASTED\_PRICE;RUN;