MSDS-6372 - Project 1

Akbar Thobani

Michael Weatherford

Tai Chowdhury

**Goal 1: Prediction of Individual Property Values**

Introduction

Home purchasing can be one of the largest purchases of an individual’s life. The price you pay for your home will affect your finances, budget, lifestyle and future. The sizeable loans, and long-term lengths involved with mortgages, also makes it a high-risk lending scenario for the bank. Being able to accurately predict a properties value based on the parameters of the home can be extremely beneficial to both the homeowner and lender.

With this study we have two main goals; first, we would like to predict the value of the property based off descriptive variables of the property second, we would like to forecast the mean property price for Jul 2015- July 2016. We are analyzing these goals for properties in Russia and were supplied data from a Russian financial institution named Sberbank.

Data Description

Test Test

Data Cleaning/Wrangling

Initial exploration of the data sets shows that we are missing approximately 5.6% of the data for both model data and project data sets (Figure 2). We did not want to remove the data rows and lose possibly important data in that row, so we imputed the missing data using Multivariate Imputation via Chained Equations (MICE). The problem using MICE for this data set is that there are many unbalanced factor variables. If you were to dummy code these variables there is a high chance that one column a linear combination of another. To correct for this, we changed the imputation method, so it was not a stochastic method. So, we utilized classification and regression trees to complete the imputation of missing data.

Figure 2 “Missing Variable Plot”

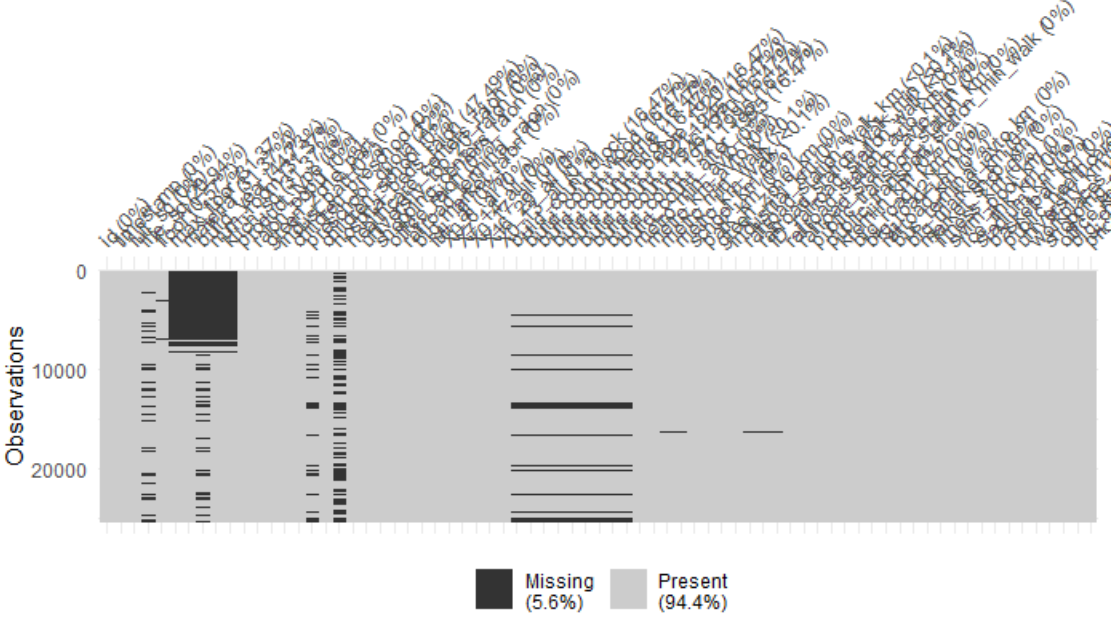
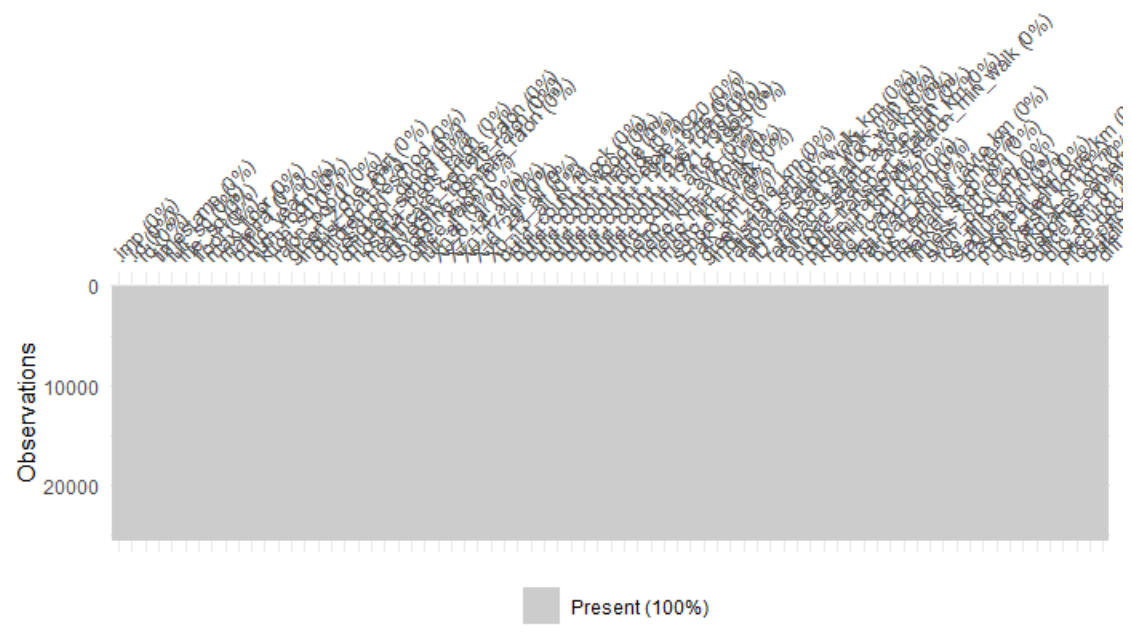
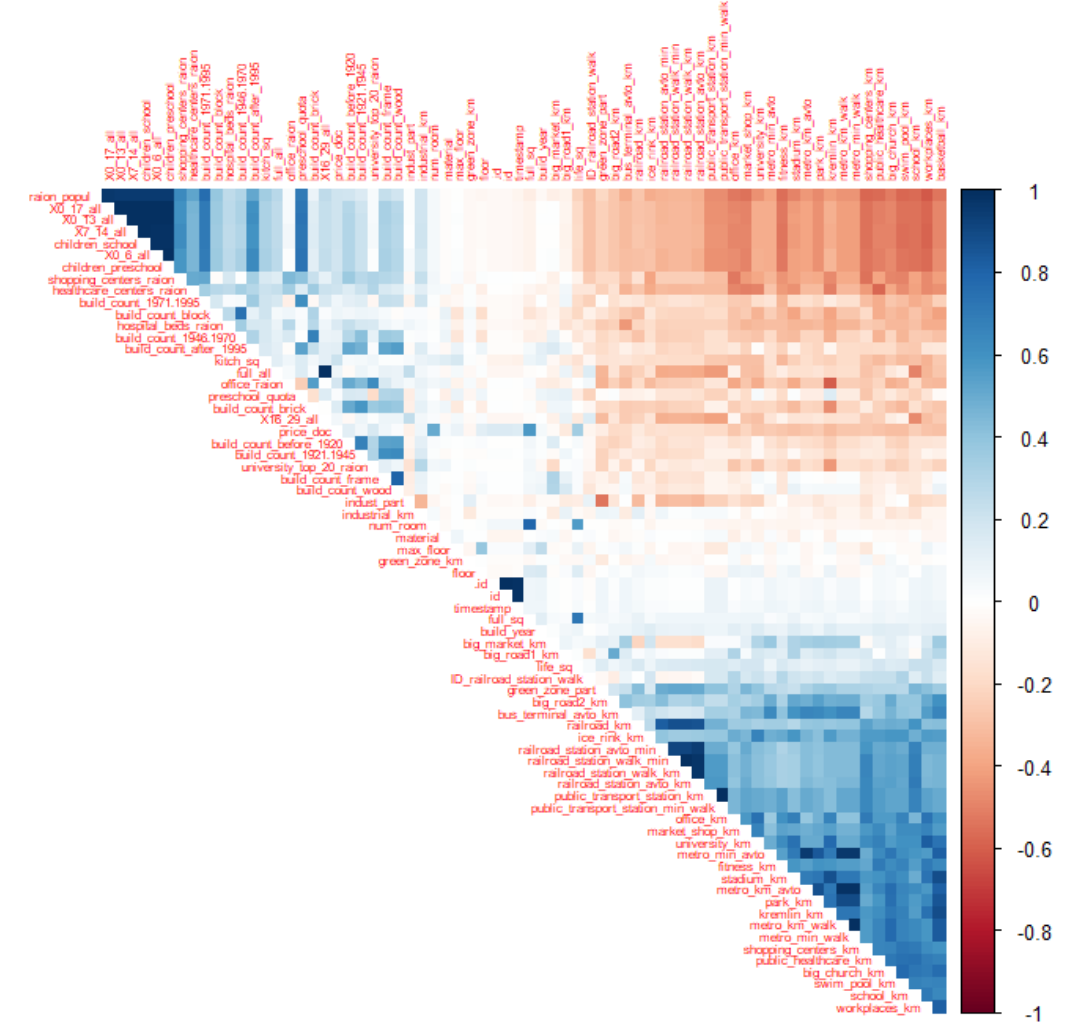


Figure 3 “Missing Variable Plot After Imputation”



EDA

1. Outlier identification was conducted on all continuous, numerical values. We used the Inter-Quartile-Range (IQR) method. Each column was evaluated and both the top and bottom 25% values were extracted, and any values beyond 1.5\*IQR was considered an outlier and was replaced with either 5% or 95%th observation value.
2. Please see “Data Cleaning/Wrangling” section for a description for missing value identification.
3. We checked for collinearity using a correlation plot, which plots all variables against each other. The closer a value is to 1 or -1 the higher probability that those variables are collinearly related.



1. Checking assumptions
2. Variable Selection
3. Anything else

Modeling

Test Test

Prediction

Test Test

Conclusion

Test Test

Goal 2: Prediction of Mean Property Value by Year

Introduction

Test Test

Data Cleaning/Wrangling

Test Test

Plot Time Series

Test Test

Model the Residual Series

Test Test

Forecast

Test Test

Conclusion