**Russian Real Estate Valuation: Project 1**

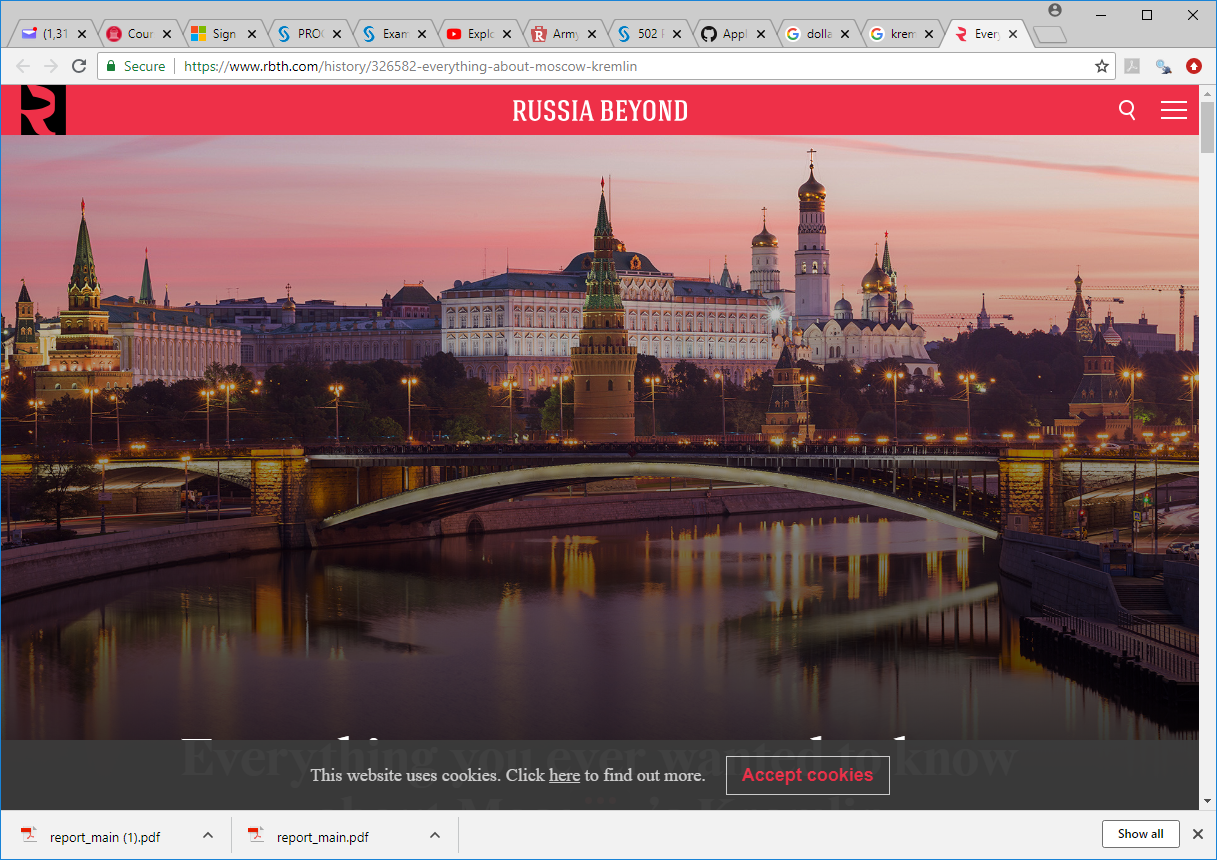
**MSDS 6371 Applied Statistics - Modeling and Inference**

**Dr. Anthony Tanaydin - Fall 2018**

**Allen Crane**

**Brock Friedrich**

**Matthew Trevathan**





**Introduction:**

For our project, we have been asked to assist the Russian bank Sberbank to predict real estate prices with more certainty, so they can provide more certainty to their customers and value to their shareholders.The goal of this analysis is to build 2 models for the prediction of Russian real estate property. One model will predict the individual property price given some or all of the variables, and another model to predict the mean price of the properties. Both models will use the TRAIN data set to build the model and the PREDICTION data set for the test model.

**Description of the data with a table or a reference to a table**

The TRAIN data set was used for this modeling effort. The original file consists of 30,471 observations with 292 variables each. Appendix A contains the complete data dictionary and explanation of variables.

**Data Cleaning / Wrangling**

**(any renaming of variables or standardizing of values)**

**A.** The following variable names exceeded 32 characters, and therefore needed to be trimmed to shorter variable names:

**Before After .**

preschool\_education\_centers\_raion ps\_educ\_centers\_raion

school\_education\_centers\_top\_20\_raion s\_educ\_centers\_top20\_raion

raion\_build\_count\_with\_material\_info raion\_bld\_cnt\_material\_info

public\_transport\_station\_min\_walk pub\_trans\_stn\_min\_walk

**B.** Some variables exhibited significant curvature in the QQ plot, so the following variables were created using a log transformation on the original data:

**Before After .**

area\_m log\_area\_m = log(area\_m)

price\_doc log\_price\_doc=log(price\_doc)

**C.** In SAS, variables cannot begin with numeric values. As a result, for those of us coding in SAS, we needed rename the variables to begin with non-numeric values. Using “group\_” to proceed the variable names, we have the following:

**Before After .**

0\_6\_all group\_0\_6\_all

0\_6\_male group\_0\_6\_male

0\_6\_female group\_0\_6\_female

7\_14\_all group\_7\_14\_all

7\_14\_male group\_7\_14\_male

7\_14\_female group\_7\_14\_female

0\_17\_all group\_0\_17\_all

0\_17\_male group\_0\_17\_male

0\_17\_female group\_0\_17\_female

16\_29\_all group\_16\_29\_all

16\_29\_male group\_16\_29\_male

16\_29\_female group\_16\_29\_female

0\_13\_all group\_0\_13\_all

0\_13\_male group\_0\_13\_male

0\_13\_female group\_0\_13\_female

**D.** In SAS, variables cannot contain dashes/hyphens (“-”). As a result, for those of us coding in SAS, we needed rename the variables not include dashes/hyphens (“-”). Using an underscore (“\_”) to replace (“-”), we have the following:

**Before After .**

build\_count\_1921-1945 build\_count\_1921\_1945

'build\_count\_1946-1970 build\_count\_1946\_1970

'build\_count\_1971-1995 build\_count\_1971\_1995

**E.** Several numeric variables--assumed to be candidates for regression analysis--contained “NA” values. This caused them to be considered by SAS as character variables when the data was loaded into a SAS data set. There is probably a way to do the following in one step, but three steps were used to do this data processing step:

**1) The first step was to convert the “NA” values to missing (“”):**

**Before After .**

build\_year="NA" then build\_year=""

cafe\_avg\_price\_500="NA" then cafe\_avg\_price\_500=""

cafe\_avg\_price\_1000="NA" then cafe\_avg\_price\_1000=""

cafe\_sum\_1000\_max\_price\_avg="NA" then cafe\_sum\_1000\_max\_price\_avg=""

cafe\_sum\_1000\_min\_price\_avg="NA" then cafe\_sum\_1000\_min\_price\_avg=""

cafe\_sum\_500\_max\_price\_avg="NA" then cafe\_sum\_500\_max\_price\_avg=""

cafe\_sum\_500\_min\_price\_avg="NA" then cafe\_sum\_500\_min\_price\_avg=""

hospital\_beds\_raion="NA" then hospital\_beds\_raion=""

kitch\_sq="NA" then kitch\_sq=""

material="NA" then material=""

max\_floor="NA" then max\_floor=""

num\_room="NA" then num\_room=""

state="NA" then state=""

preschool\_quota="NA" then preschool\_quota=""

life\_sq="NA" then life\_sq=""

floor="NA" then floor=""

school\_quota="NA" then school\_quota=""

raion\_bld\_cnt\_material\_info="NA" then raion\_bld\_cnt\_material\_info=""

build\_count\_block="NA" then build\_count\_block=""

build\_count\_wood="NA" then build\_count\_wood=""

build\_count\_frame="NA" then build\_count\_frame=""

build\_count\_brick="NA" then build\_count\_brick=""

build\_count\_monolith="NA" then build\_count\_monolith=""

build\_count\_panel="NA" then build\_count\_panel=""

build\_count\_foam="NA" then build\_count\_foam=""

build\_count\_slag="NA" then build\_count\_slag=""

build\_count\_mix="NA" then build\_count\_mix=""

raion\_build\_count\_with\_builddate="NA" then raion\_build\_count\_with\_builddate=""

build\_count\_before\_1920="NA" then build\_count\_before\_1920=""

build\_count\_1921\_1945="NA" then build\_count\_1921\_1945=""

build\_count\_1946\_1970="NA" then build\_count\_1946\_1970=""

build\_count\_1971\_1995="NA" then build\_count\_1971\_1995=""

build\_count\_after\_1995="NA" then build\_count\_after\_1995=""

metro\_min\_walk="NA" then metro\_min\_walk=""

metro\_km\_walk="NA" then metro\_km\_walk=""

railroad\_station\_walk\_km="NA" then railroad\_station\_walk\_km=""

railroad\_station\_walk\_min="NA" then railroad\_station\_walk\_min=""

ID\_railroad\_station\_walk="NA" then ID\_railroad\_station\_walk=""

cafe\_sum\_1500\_min\_price\_avg="NA" then cafe\_sum\_1500\_min\_price\_avg=""

cafe\_sum\_1500\_max\_price\_avg="NA" then cafe\_sum\_1500\_max\_price\_avg=""

cafe\_avg\_price\_1500="NA" then cafe\_avg\_price\_1500=""

cafe\_sum\_2000\_min\_price\_avg="NA" then cafe\_sum\_2000\_min\_price\_avg=""

cafe\_sum\_2000\_max\_price\_avg="NA" then cafe\_sum\_2000\_max\_price\_avg=""

cafe\_avg\_price\_2000="NA" then cafe\_avg\_price\_2000=""

cafe\_sum\_3000\_min\_price\_avg="NA" then cafe\_sum\_3000\_min\_price\_avg=""

cafe\_sum\_3000\_max\_price\_avg="NA" then cafe\_sum\_3000\_max\_price\_avg=""

cafe\_avg\_price\_3000="NA" then cafe\_avg\_price\_3000=""

cafe\_sum\_5000\_min\_price\_avg="NA" then cafe\_sum\_5000\_min\_price\_avg=""

cafe\_sum\_5000\_max\_price\_avg="NA" then cafe\_sum\_5000\_max\_price\_avg=""

cafe\_avg\_price\_5000="NA" then cafe\_avg\_price\_5000=""

prom\_part\_5000="NA" then prom\_part\_5000=""

**2) The second step was to convert the missing variables (“”) to numeric:**

**Before After .**

build\_year = '' n\_build\_year = input(build\_year, 8.)

cafe\_avg\_price\_500 = '' n\_cafe\_avg\_price\_500 = input(cafe\_avg\_price\_500, 8.)

cafe\_avg\_price\_1000 = '' n\_cafe\_avg\_price\_1000 = input(cafe\_avg\_price\_1000, 8.)

cafe\_sum\_1000\_max\_price\_avg = '' n\_cafe\_sum\_1000\_max\_price\_avg = input(cafe\_sum\_1000\_max\_price\_avg, 8.)

cafe\_sum\_1000\_min\_price\_avg = '' \_cafe\_sum\_1000\_min\_price\_avg = input(cafe\_sum\_1000\_min\_price\_avg, 8.)

cafe\_sum\_500\_max\_price\_avg = '' \_cafe\_sum\_500\_max\_price\_avg = input(cafe\_sum\_500\_max\_price\_avg, 8.)

cafe\_sum\_500\_min\_price\_avg = '' \_cafe\_sum\_500\_min\_price\_avg = input(cafe\_sum\_500\_min\_price\_avg, 8.)

hospital\_beds\_raion = '' \_hospital\_beds\_raion = input(hospital\_beds\_raion, 8.)

kitch\_sq = '' n\_kitch\_sq = input(kitch\_sq, 8.)

material = '' n\_material = input(material, 8.)

max\_floor = '' n\_max\_floor = input(max\_floor, 8.)

num\_room = '' n\_num\_room = input(num\_room, 8.)

state = '' n\_state = input(state, 8.)

preschool\_quota = '' n\_preschool\_quota = input(preschool\_quota, 8.)

life\_sq ='' n\_life\_sq = input(life\_sq, 8.)

floor ='' n\_floor = input(floor, 8.)

school\_quota ='' n\_school\_quota = input(school\_quota, 8.)

raion\_bld\_cnt\_material\_info ='' n\_raion\_bld\_cnt\_material\_info = input(raion\_bld\_cnt\_material\_info, 8.)

build\_count\_block ='' n\_build\_count\_block = input(build\_count\_block, 8.)

build\_count\_wood ='' n\_build\_count\_wood = input(build\_count\_wood, 8.)

build\_count\_frame ='' n\_build\_count\_frame = input(build\_count\_frame, 8.)

build\_count\_brick ='' n\_build\_count\_brick = input(build\_count\_brick, 8.)

build\_count\_monolith ='' n\_build\_count\_monolith = input(build\_count\_monolith, 8.)

build\_count\_panel ='' n\_build\_count\_panel = input(build\_count\_panel, 8.)

build\_count\_foam ='' n\_build\_count\_foam = input(build\_count\_foam, 8.)

build\_count\_slag ='' n\_build\_count\_slag = input(build\_count\_slag, 8.)

build\_count\_mix ='' n\_build\_count\_mix = input(build\_count\_mix, 8.)

raion\_build\_count\_with\_builddate ='' n\_raion\_build\_count\_with\_date = input(raion\_build\_count\_with\_builddate, 8.)

build\_count\_before\_1920 ='' n\_build\_count\_before\_1920 = input(build\_count\_before\_1920, 8.)

build\_count\_1921\_1945 ='' n\_build\_count\_1921\_1945 = input(build\_count\_1921\_1945, 8.)

build\_count\_1946\_1970 ='' n\_build\_count\_1946\_1970 = input(build\_count\_1946\_1970, 8.)

build\_count\_1971\_1995 ='' n\_build\_count\_1971\_1995 = input(build\_count\_1971\_1995, 8.)

build\_count\_after\_1995 ='' n\_build\_count\_after\_1995 = input(build\_count\_after\_1995, 8.)

metro\_min\_walk ='' n\_metro\_min\_walk = input(metro\_min\_walk, 8.)

metro\_km\_walk ='' n\_metro\_km\_walk = input(metro\_km\_walk, 8.)

railroad\_station\_walk\_km ='' n\_railroad\_station\_walk\_km = input(railroad\_station\_walk\_km, 8.)

railroad\_station\_walk\_min ='' n\_railroad\_station\_walk\_min = input(railroad\_station\_walk\_min, 8.)

ID\_railroad\_station\_walk ='' n\_ID\_railroad\_station\_walk = input(ID\_railroad\_station\_walk, 8.)

cafe\_sum\_1500\_min\_price\_avg ='' n\_cafe\_sum\_1500\_min\_price\_avg = input(cafe\_sum\_1500\_min\_price\_avg, 8.)

cafe\_sum\_1500\_max\_price\_avg ='' n\_cafe\_sum\_1500\_max\_price\_avg = input(cafe\_sum\_1500\_max\_price\_avg, 8.)

cafe\_avg\_price\_1500 ='' n\_cafe\_avg\_price\_1500 = input(cafe\_avg\_price\_1500, 8.)

cafe\_sum\_2000\_min\_price\_avg ='' n\_cafe\_sum\_2000\_min\_price\_avg = input(cafe\_sum\_2000\_min\_price\_avg, 8.)

cafe\_sum\_2000\_max\_price\_avg ='' n\_cafe\_sum\_2000\_max\_price\_avg = input(cafe\_sum\_2000\_max\_price\_avg, 8.)

cafe\_avg\_price\_2000 ='' n\_cafe\_avg\_price\_2000 = input(cafe\_avg\_price\_2000, 8.)

cafe\_sum\_3000\_min\_price\_avg ='' n\_cafe\_sum\_3000\_min\_price\_avg = input(cafe\_sum\_3000\_min\_price\_avg, 8.)

cafe\_sum\_3000\_max\_price\_avg ='' n\_cafe\_sum\_3000\_max\_price\_avg = input(cafe\_sum\_3000\_max\_price\_avg, 8.)

cafe\_avg\_price\_3000 ='' n\_cafe\_avg\_price\_3000 = input(cafe\_avg\_price\_3000, 8.)

cafe\_sum\_5000\_min\_price\_avg ='' n\_cafe\_sum\_5000\_min\_price\_avg = input(cafe\_sum\_5000\_min\_price\_avg, 8.)

cafe\_sum\_5000\_max\_price\_avg ='' n\_cafe\_sum\_5000\_max\_price\_avg = input(cafe\_sum\_5000\_max\_price\_avg, 8.)

cafe\_avg\_price\_5000 ='' n\_cafe\_avg\_price\_5000 = input(cafe\_avg\_price\_5000, 8.)

prom\_part\_5000 ='' n\_prom\_part\_5000 = input(prom\_part\_5000, 8.)

1. **The third step was to drop the original variable names:**

**Dropped:**

build\_year cafe\_avg\_price\_500

cafe\_avg\_price\_1000 cafe\_sum\_1000\_max\_price\_avg

cafe\_sum\_1000\_min\_price\_avg cafe\_sum\_500\_max\_price\_avg

cafe\_sum\_500\_min\_price\_avg hospital\_beds\_raion

kitch\_sq material

max\_floor num\_room

State preschool\_quota

life\_sq floor

school\_quota raion\_bld\_cnt\_material\_info

build\_count\_block build\_count\_wood

build\_count\_frame build\_count\_brick

build\_count\_monolith build\_count\_panel

build\_count\_foam build\_count\_slag

build\_count\_mix raion\_build\_count\_with\_builddate

build\_count\_before\_1920 build\_count\_1921\_1945

build\_count\_1946\_1970 build\_count\_1971\_1995

build\_count\_after\_1995 metro\_min\_walk

metro\_km\_walk railroad\_station\_walk\_km

railroad\_station\_walk\_min ID\_railroad\_station\_walk

cafe\_sum\_1500\_min\_price\_avg cafe\_sum\_1500\_max\_price\_avg

cafe\_avg\_price\_1500 cafe\_sum\_2000\_min\_price\_avg

cafe\_sum\_2000\_max\_price\_avg cafe\_avg\_price\_2000

cafe\_sum\_3000\_min\_price\_avg cafe\_sum\_3000\_max\_price\_avg

cafe\_avg\_price\_3000 cafe\_sum\_5000\_min\_price\_avg

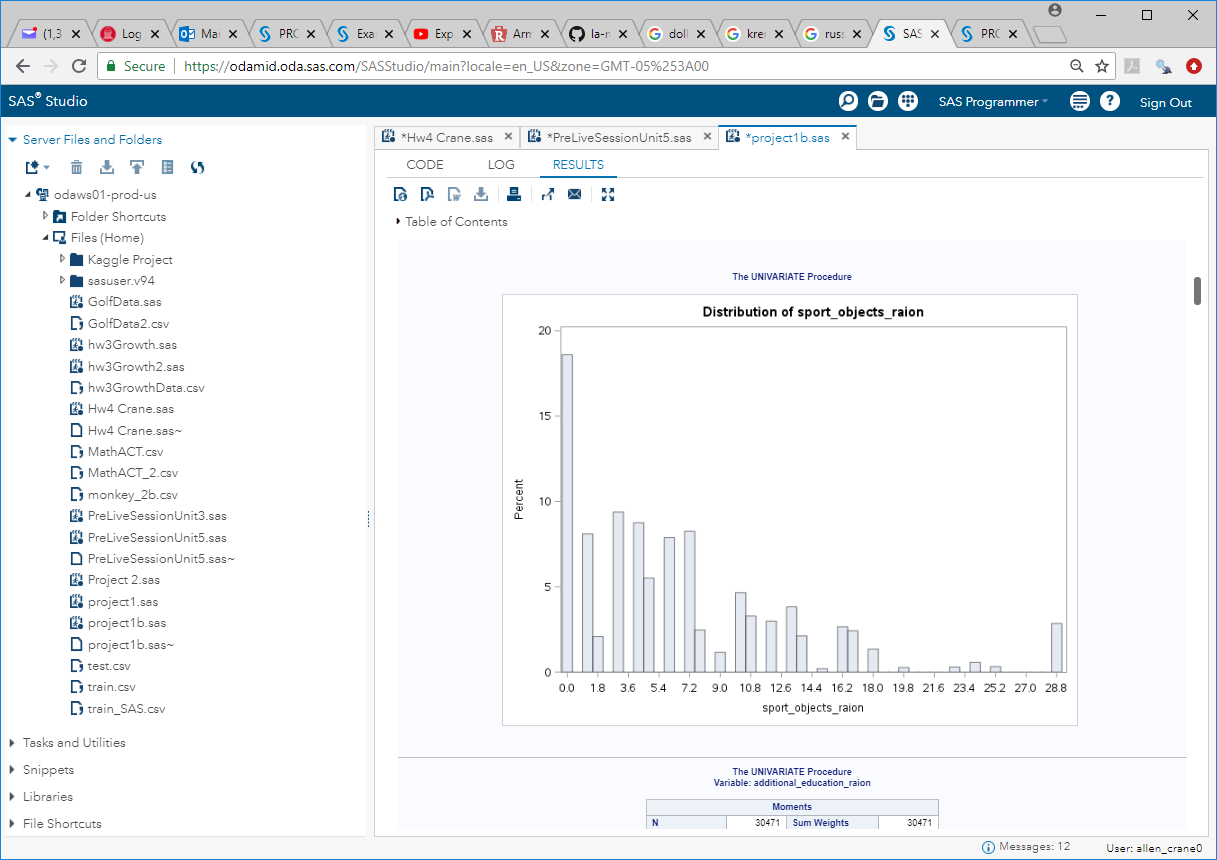
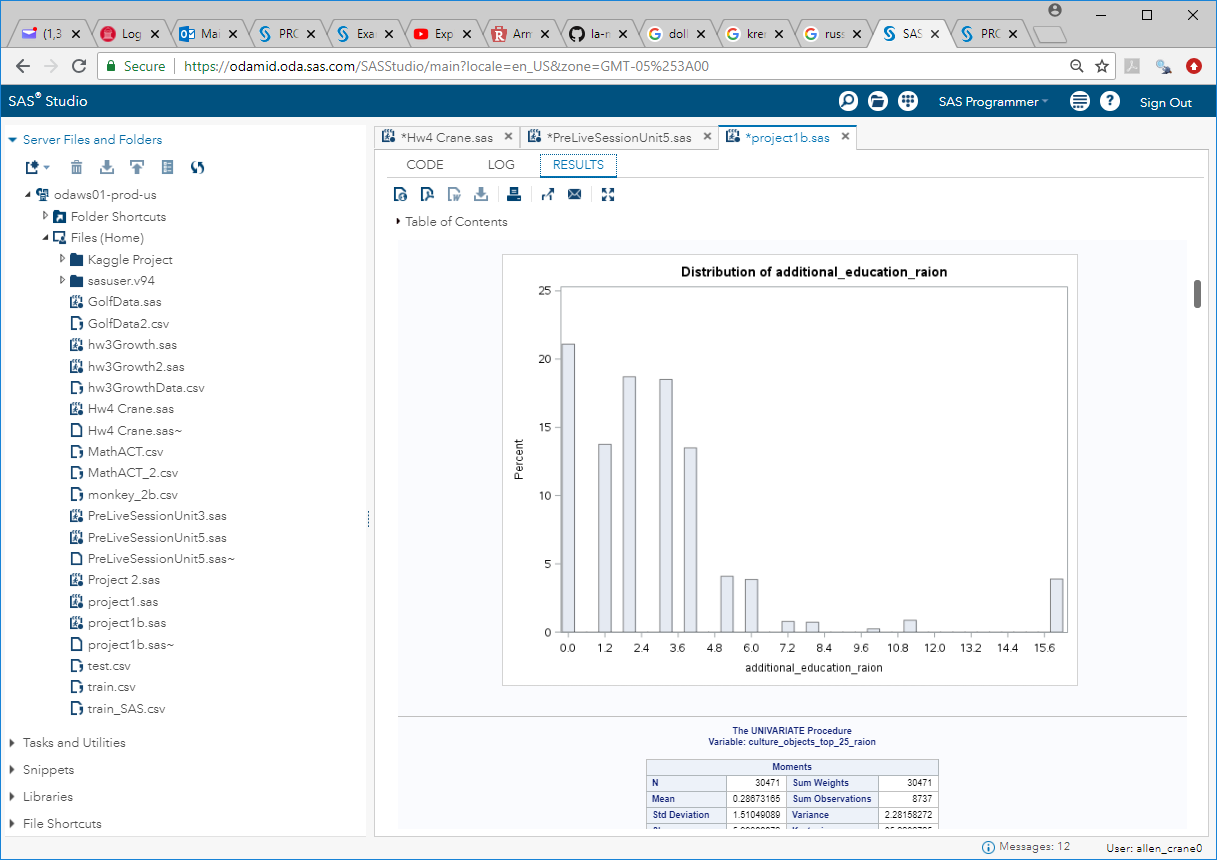
cafe\_sum\_5000\_max\_price\_avg cafe\_avg\_price\_5000

prom\_part\_5000

**Exploratory Data Analysis (EDA) (Complete with summary statistics, descriptions, tables and/or plots etc.)**

1. **Outlier Identification and Handling**

After some initial univariate testing, we began to see some patterns emerge. For example, several variables that end in “raion” (Russian for “neighborhood”) had the same pattern. These are all countable variables. Note how there are many zero values, followed by an approximate normal distribution, with a notable increase at the far right side of these graphs:



The same is true for many of the variables ending in “\_raion” (“neighborhood, in Russian). Further investigation shows that these neighborhoods (a.k.a. “sub\_area” in our data) correspond to affluent and government sections of greater Moscow, Presneskoe (Presnensky) and Tversekoe (Tverskoy). These data show that there are some sub\_areas that are culturally and administratively dense enough to legitimize these data findings. Therefore, we will not omit these variables due to simply being outliers, else we may risk losing material information about their predictive power for property values in certain areas.

***Presnensky District*** *is an upscale area known for the State Museum of Oriental Art and quiet parks like Patriarch’s Ponds. Child-friendly attractions include the Moscow Zoo, and hands-on exhibits at the nearby Moscow Planetarium. A variety of restaurants serving European and fusion fare line Bolshaya Nikitskaya Street, which is also home to classic repertory at the Mayakovsky Theater.*

***Tverskoy District*** *is a district of Central Administrative Okrug of the federal city of Moscow, Russia. Population: 75,378; 75,955. The district extends from Kitai-gorod northwest to Belorussky and Savyolovsky Rail Terminals.*

*(source: Wikipedia)*

***Presnensk******y District Tverskoy District***

1. **Missing value identification, summary and possible imputation (mean, median, regression.) This may also be considered “Data Wrangling”.**

For our analysis, several numeric variables contained non-trivial percentages of missing data (percent of the data with “NA” listed as the value). Since the variable with the highest missing value rate (hospital\_beds\_raion) is 47%, throwing out the observations with missing values would limit us to just over half of the data set. Similarly, removing the entire variable(s) with missing values could cause us to miss significant coefficients for our predictive models. The following is a list of all variables in the TRAIN data set with missing values, and the missing value percentage of all observations for each variable)

**Variable Missing Value % (Percent“NA”) .**

hospital\_beds\_raion 47%

build\_year 45%

state 44%

cafe\_sum\_500\_min\_price\_avg 44%

cafe\_sum\_500\_max\_price\_avg 44%

cafe\_avg\_price\_500 44%

max\_floor 31%

material 31%

num\_room 31%

kitch\_sq 31%

preschool\_quota 22%

school\_quota 22%

cafe\_sum\_1000\_min\_price\_avg 21%

cafe\_sum\_1000\_max\_price\_avg 21%

cafe\_avg\_price\_1000 21%

life\_sq 21%

raion\_build\_count\_with\_material\_info 16%

build\_count\_block 16%

build\_count\_wood 16%

build\_count\_frame 16%

build\_count\_brick 16%

build\_count\_monolith 16%

build\_count\_panel 16%

build\_count\_foam 16%

build\_count\_slag 16%

build\_count\_mix 16%

raion\_build\_count\_with\_builddate\_info 16%

build\_count\_before\_1920 16%

build\_count\_1921-1945 16%

build\_count\_1946-1970 16%

build\_count\_1971-1995 16%

build\_count\_after\_1995 16%

cafe\_sum\_2000\_min\_price\_avg 6%

cafe\_sum\_2000\_max\_price\_avg 6%

cafe\_avg\_price\_2000 6%

cafe\_sum\_3000\_min\_price\_avg 3%

cafe\_sum\_3000\_max\_price\_avg 3%

cafe\_avg\_price\_3000 3%

cafe\_sum\_5000\_min\_price\_avg 1%

cafe\_sum\_5000\_max\_price\_avg 1%

cafe\_avg\_price\_5000 1%

prom\_part\_5000 1%

floor 1%

**Expectation maximization (EM) method for imputation of missing values**

According to Schlomer, Bauman, and Card, “This method is one of several maximum likelihood (ML) approaches. In all ML strategies, observed data are used to estimate parameters, which are then used to estimate the missing scores. These ML strategies have demonstrated superiority to deletion, nonstochastic imputation, and stochastic regression imputation methods (Roth, 1994) for multivariate normal distributions.The EM strategy is based on a recursive process: The missing data have information that is useful in estimating various parameters, and the estimated parameter has information that is useful in finding the most likely value of the missing data (Bennett, 2001). Thus, the EM method is an iterative procedure with two steps in each iteration: In the expectation step, the process is similar to the regression-based imputation. First, starting values for the parameters (e.g., means, covariances) are obtained with available data. Regression methods are used to impute, on the basis of these initial values, the values for the missing data. When this step is completed, in the maximization step new values for the parameters are calculated with the newly imputed data along with the original observed data. Then the process starts over with the expectation step and continues until the estimates change very little from one iteration to the next (i.e., until the estimates converge; Allison, 2001).” (*Source: Gabriel L. Schlomer, Sheri Bauman, and Noel A. Card, “Best Practices for Missing Data Management in Counseling Psychology”, Journal of Counseling Psychology 2010, Vol. 57, No. 1, 1–10)*

Below is the generic code for the PROC MI command: (see SAS code in Appendix for detail)

proc mi data=train2 seed=501213

mu0=0 out=train4;

var n\_build\_year

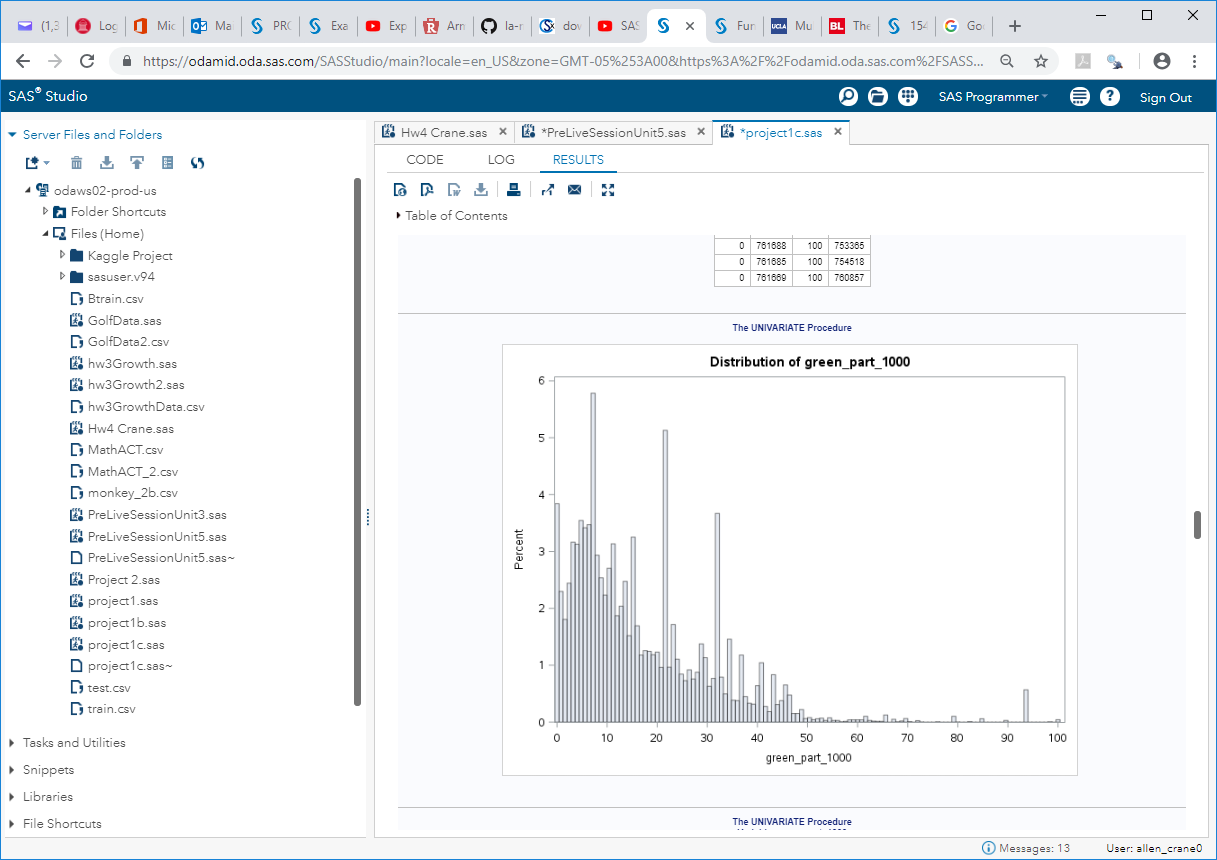
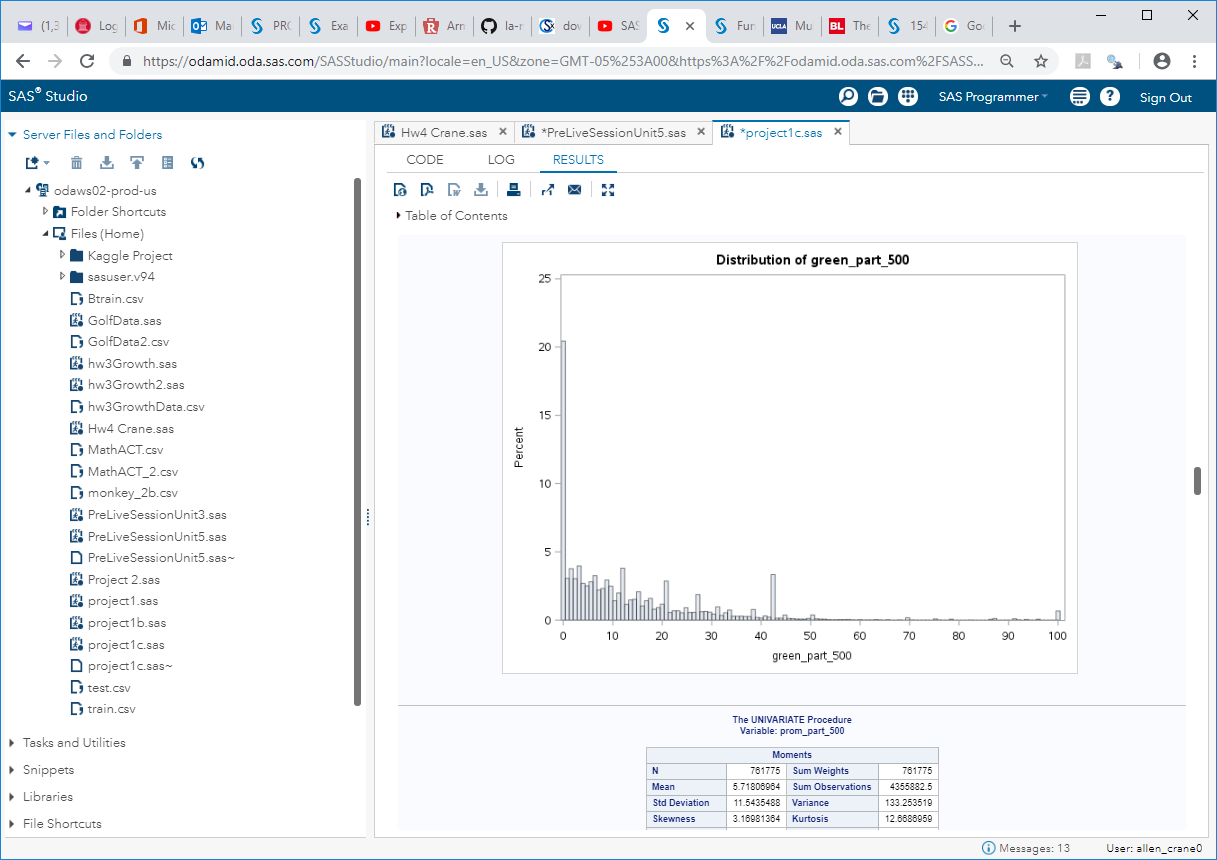
…

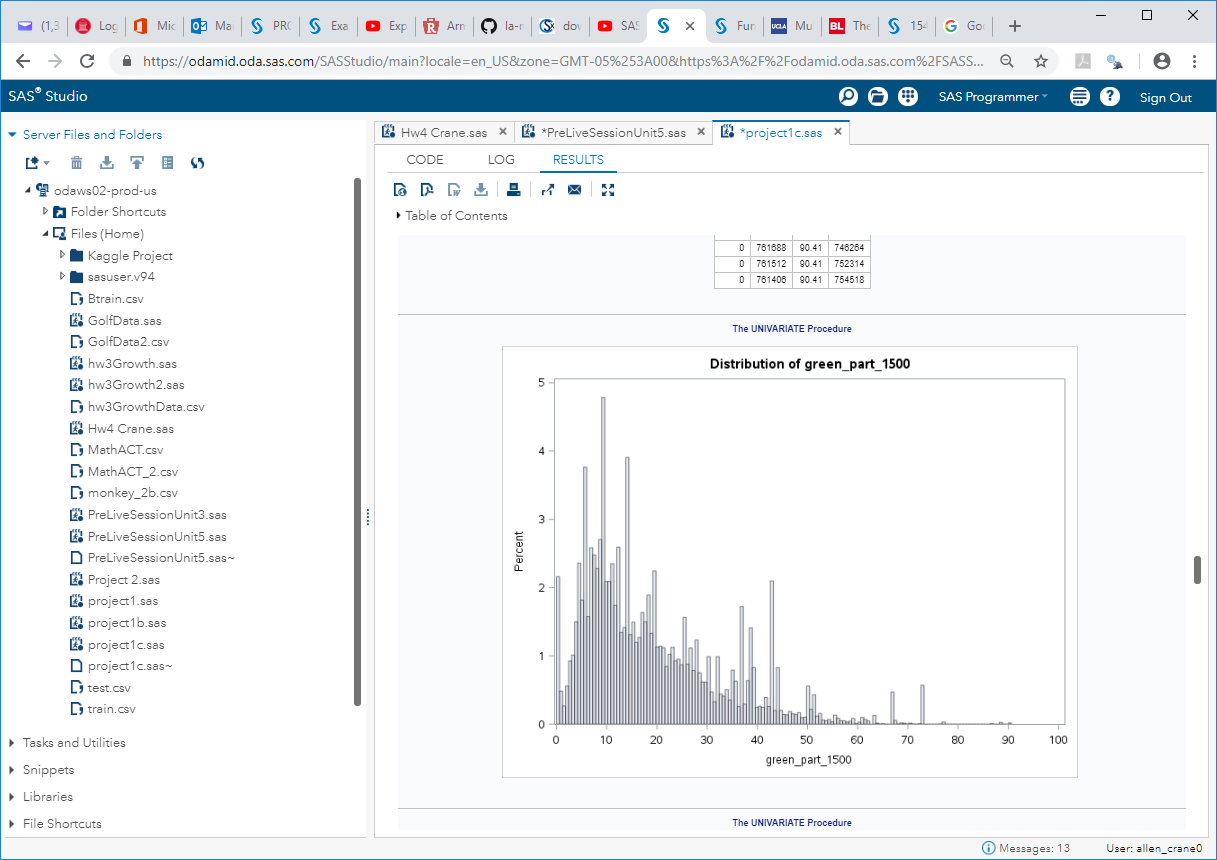
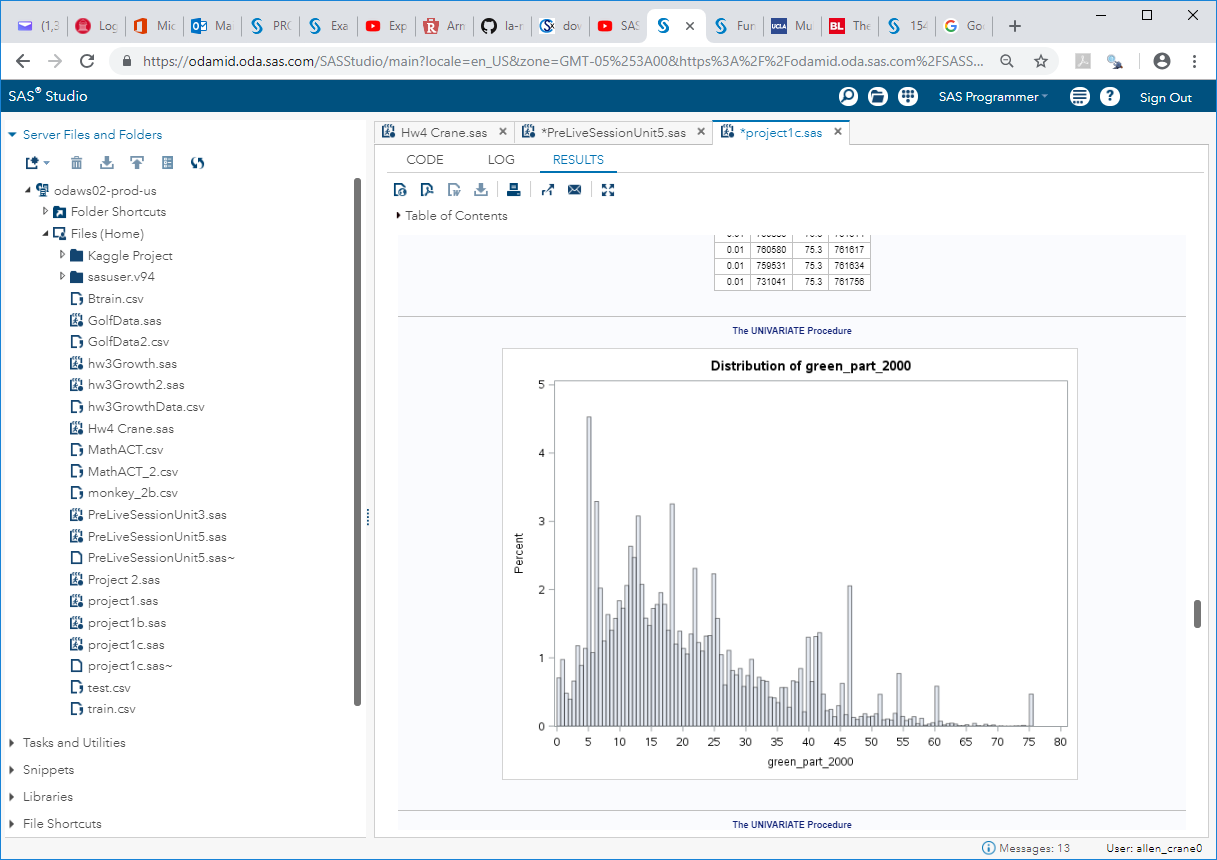
n\_prom\_part\_5000;

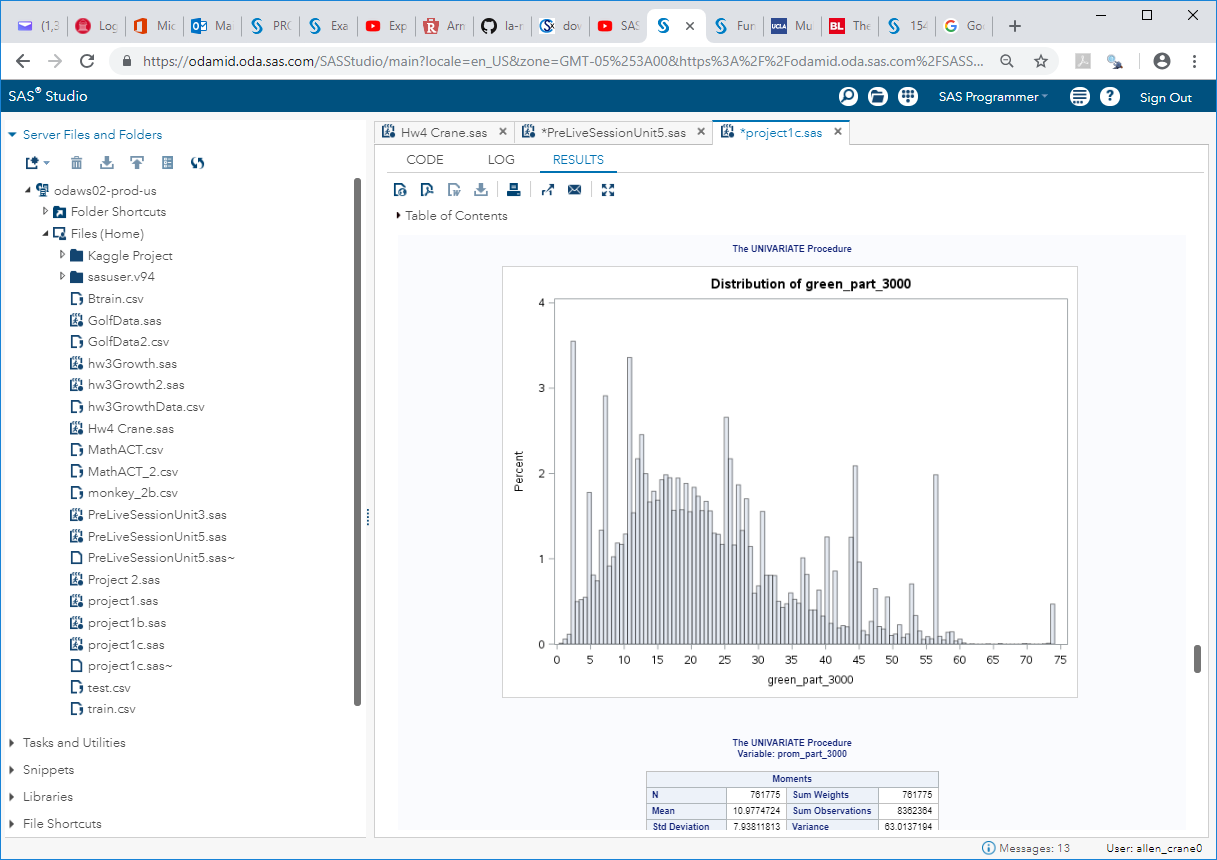
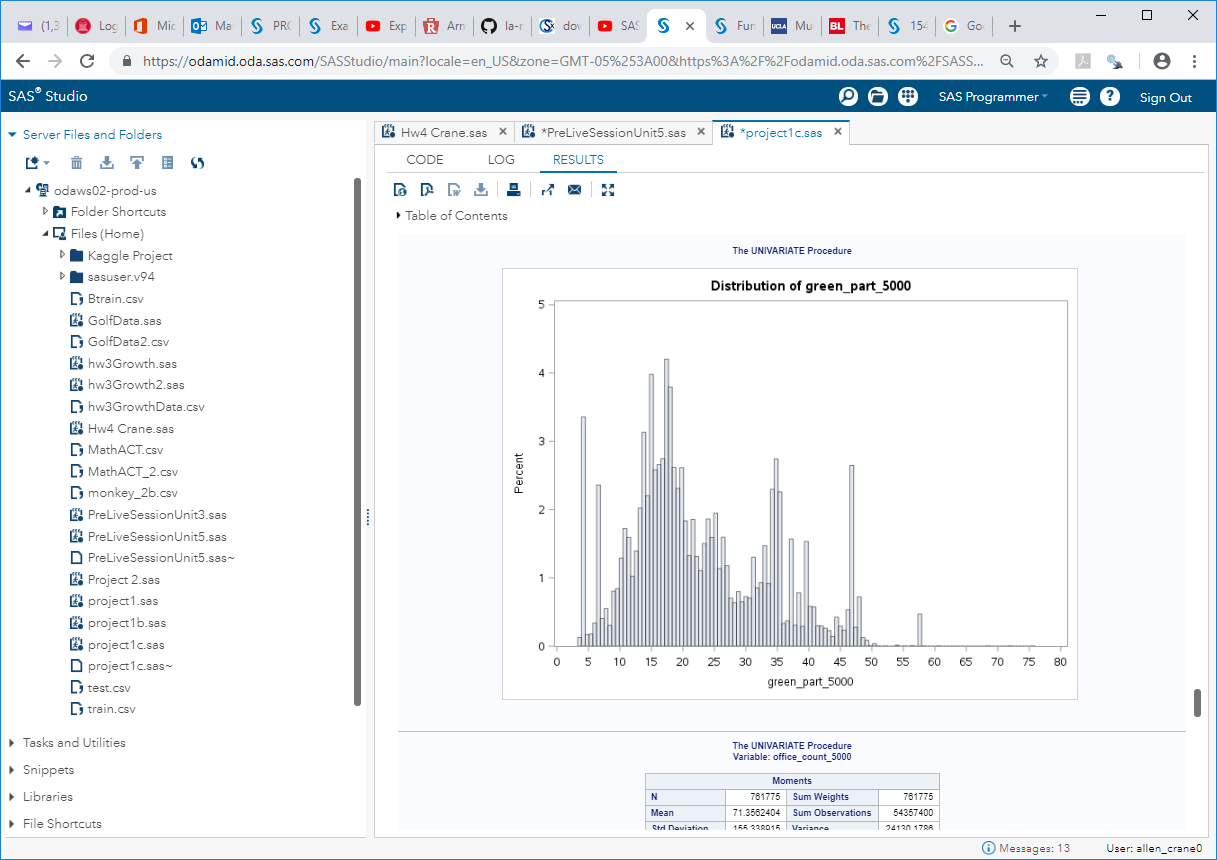
run;

1. Multicollinearity (is there reason to believe it is present? You don’t have to address every potential pair of variables that may be collinear. Just provide a plot and or other evidence of a single occurrence of multicollinearity if at least one exists and then mention possible other occurrences.

In looking at the data, we can see several groups of variables that appear similar. Many of these are in the form of a distance or countable measure between the property (measured in increments of 500, 1000, 1500, 2000, 3000, and 5000 meters) and some item of significance (distance to a cafe with a certain min/max/average price, the number of large churches within an area, distance to leisure areas and markets, etc). An example of one series (“green\_part\_X”) is shown below:







To address these, we experimented with several methods of variable reduction, including summing the variables in each group, choosing only one variable to represent each group, etc. While this seemed to improve the explanation of the model, we found that it actually reduced the R2 value, compared with loading all variables and letting the model test for significance (stepwise, forward, backward, stepwise AIC, stepwise SBC, stepwise AICC). In some cases, some of the grouped variables are dropped, while in other cases they are kept. We can speculate whether the kept variables are due to multicollinearity or if some of these distance differences are legitimate (for example, if distance to market is small, that may be more attractive to a prospective property owner, and hence values like cafes less than 1.5km are material, but those that are more than 5km are immaterial). As a result, we ran the model many different ways on variable selection and produced consistent R2 values around 0.64.

1. Checking assumptions: homoscedasticity, normal distributions of the response for fixed values of the explanatory variable(s), linear relationship between the mean of the response and each explanatory variable, etc. This is where you would apply transformation (log, square root, etc.)

For this analysis, we will use the White test for homoscedasticity, which is represented generically below. A value of > alpha = 0.05 indicates homoscedasticity, and therefore no transformation is needed. Alternatively, a value of < alpha = 0.05 indicates heteroscedasticity, which would indicate that some form of transformation is needed.

proc model data=data;

parms b0 b1;

y = b0 + b1\* x;

fit y/white;

Run;

The following variables were determined to be heteroscedastic (value < alpha = 0.05):

1. High level variable selection would be included in the EDA. (Example: There are many potential explanatory variables. Running stepwise variable selection with a high entry and exit threshold will not provide a plausible final model, but may leave you with a smaller set of potential explanatory to work with.)

High level variable selection is included in the model detail.

1. Anything else that might be appropriate in learning about the data before getting started. (Example: You might try interactions between explanatory variables in the EDA.)

No interactions were tested.

**Modeling (fit at least 3 candidate models)**

1. A model with OLS parameter estimates. You may choose the variables with or without the use of a variable selection technique (forward, backward, stepwise) (You may have done this already in the EDA.)

**See Appendix 3: Model 1.1 OLS detail**

1. A model with LASSO estimation and selection.

**See Appendix 3: Model 1.2 LASSO detail**

1. A model of your choice. This may be using another OLS or LASSO model or custom model, etc.

**See Appendix 3: Model 1.2 LASSO detail**

\*Notes:

In at least one of the models above, or in the EDA, you should test the significance of at least one term that involves an interaction.

You need to address the assumptions with respect to the residuals. (Normally distributed around 0 with constant standard deviation.)

For each model you need to conduct an internal and external cross validation.

You should compare the models using the AIC, SBC, Interval k-fold cross validation (you pick k), external cross validation. For external cross validation you will have to subset the train data set into modeling and test data sets.

**Prediction**

Identify which model you feel is the best and discuss why.

Use this model to predict the price of each of the 7662 properties in the prediction data set (*predictionData.csv*).

Submit these predictions in the zip file with your final paper. Please see the sample submission file. *SampleSubmission.csv*

\*Note: It is understood that all the material above should supported with plots, tables, charts, discussions, etc. where appropriate. All plots and tables should be clearly labeled and referenced in the discussion and all tables referenced in the discussion should exist in the paper. It is never appropriate to have a plot or table in the paper that it not described / referenced in the writing. *Your team could lose points or gain points depending on how well your professor feels you described research / findings.*

**Goal 2: 3 Pages (*30 pts*)**

1. Introduction (Brief introduction to the problem … 2 to 3 sentences).
2. Data Wrangling: Here you will have to manipulate/wrangle the data to produce the data you will use in this analysis. Your goal is to get the mean price for each month/year combination. Hints:

Data Wrangling Use SAS, Python or R (hint: dplyr and tidyr packages) to separate the timestamp column into 3 columns. These three columns should be called “Day”, “Month” and “Year” respectively (any order is fine). To be clear, your task is to turn:

 into 

USE R,Python or SAS to aggregate the data by year and month. This means that the new dataset should have the mean of all the properties for a given month in a given year. You may do this using any method you like. The “aggregate” function in R is one way of doing this. As a reference, the first four months should match the values below.



1. Plot the time series of the price\_doc versus the numbers 1 – 47 on the x-axis. You should simply name the x-axis “months”.
2. Model the residual series.
   1. Fit a simple linear regression model with price\_doc as the response variable and Month\_Number as the explanatory variable.
   2. Obtain the residuals from this model and plot them against the month number. Make sure this plot is a series rather than a scatter plot. This simply means make sure the points on the plot are connected by a line. Keep the x-axis label: “Month”.
   3. Fit the residual and time data using proc autoreg and investigate the autocorrelation structure based on the Durbin-Watson statistic, partial autocorrelation plots, AIC and SBC. You studied the AR(1) model in depth in Unit 4, it is possible you may find that an AR(p) with p > 1 is a better fit to this data. Explore what this means in terms of intuition, model construction and forecasting.
   4. Use your model from part c to forecast the residual for the next year (June 2015 – June 2016) with 95% confidence intervals.
3. Using your model from the last question and the series of residuals, forecast the next year: July 2015 – June 2016 and of course include 95% confidence intervals. You can do this by obtaining an estimate from the model in part b for the mean price and then add to it the forecast residuals from the model in part e. The confidence interval won’t be exact but a very conservative estimate will be to add the upper and lower bound from each model.
4. Submit your forecasts (with confidence intervals) as well as your final csv data set you used in the analysis (wrapped in a zip file.)

**Appendix 1: Data Dictionary *3 pts***

The initial TRAIN data set contains property references references

**Feature**  **Description .**

id Property ID

timestamp Date of property entry

area\_m Area mun. 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

culture\_objects\_top\_25\_raion Number of objects of cultural heritage

shopping\_centers\_raion Number of malls and shopping centres 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 Female population younger than working age

work\_all Working population

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 highwys, 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 big market

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

o‑ce\_km Distance to business centers/ o‑ces

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

o‑ce\_count\_500 The number of o‑ce space in 500 meters zone

o‑ce\_sqm\_500 The square of o‑ce 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

cafe\_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

o‑ce\_count\_1000 The number of o‑ce space in 1000 meters zone

o‑ce\_sqm\_1000 The square of o‑ce 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 churchs 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

o‑ce\_count\_1500 The number of o‑ce space in 1500 meters zone

o‑ce\_sqm\_1500 The square of o‑ce 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

o‑ce\_count\_2000 The number of o‑ce space in 2000 meters zone

o‑ce\_sqm\_2000 The square of o‑ce 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 churchs in 2000 meters zone

church\_count\_2000 The number of churchs 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

o‑ce\_count\_3000 The number of o‑ce space in 3000 meters zone

o‑ce\_sqm\_3000 The square of o‑ce 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

o‑ce\_count\_5000 The number of o‑ce space in 5000 meters zone

o‑ce\_sqm\_5000 The square of o‑ce 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

price\_doc Sale price of the property (presumably in Rubles)

**Appendix 2: Code *7pts***

(Well commented … Remember: Reproducible Research!)

1. This Appendix does not count against your page count.
2. Simply cut and paste your well commented code here.

**Submissions:**

What to submit 2 DS in a single zip file:

1. Prediction from Goal 1. (csv file)
2. Wrangled data set from Goal 2. (48 rows including title row. / csv file)
3. Predictions from Goal 2. (csv file)
4. Line plot of predictions from Goal 2 with 95% confidence intervals. Image or cut and pasted into something like a word doc.
5. Final paper (No longer than 11 pages without appendix.) LaTex/Word/ etc.

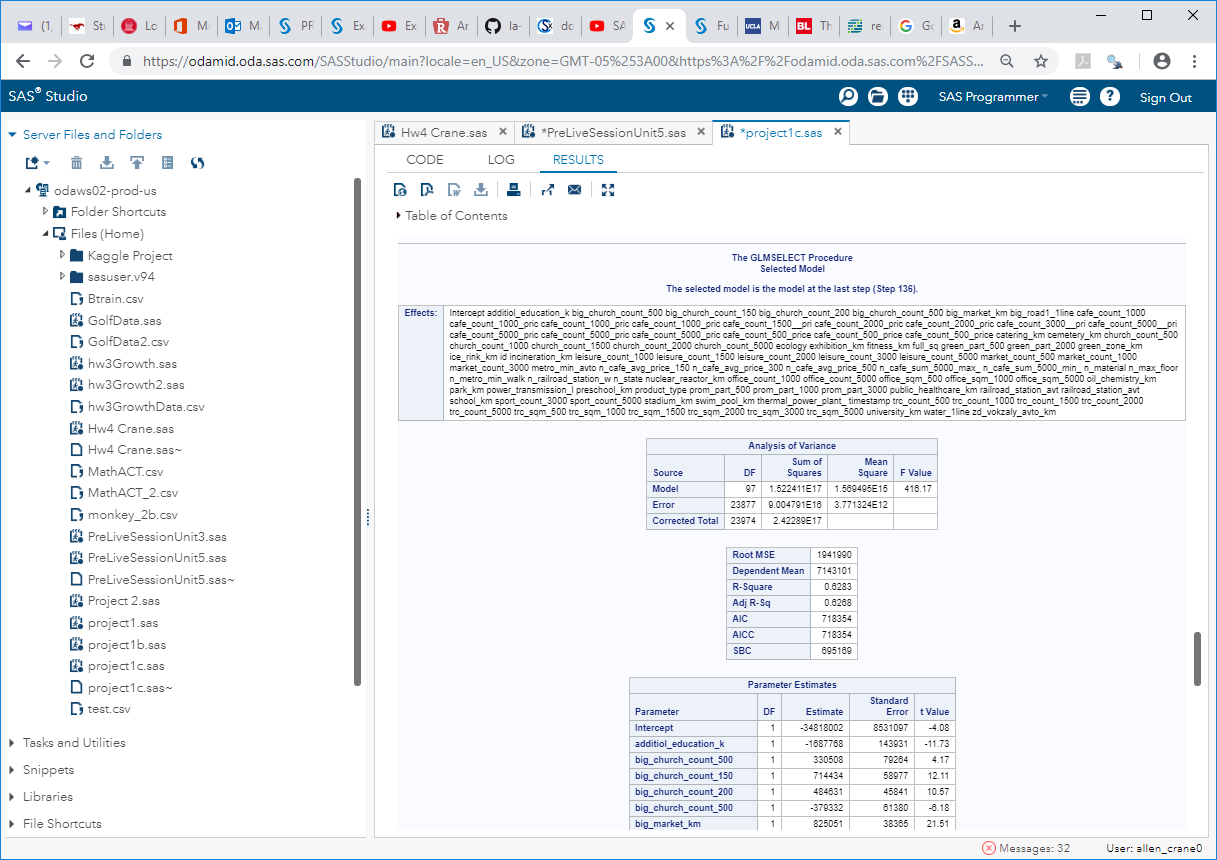
**Note: Data Wrangling:**

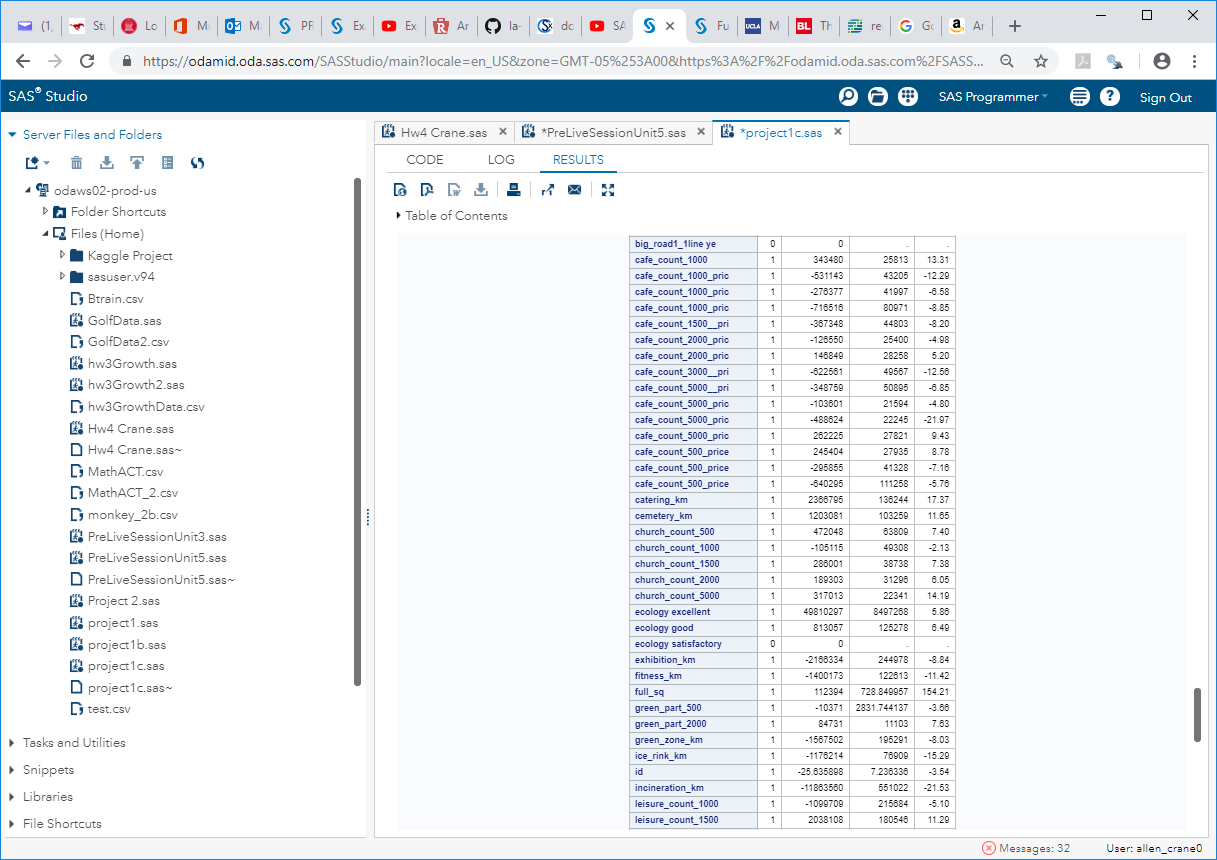
*Wrangling = having a long and complicated dispute.*

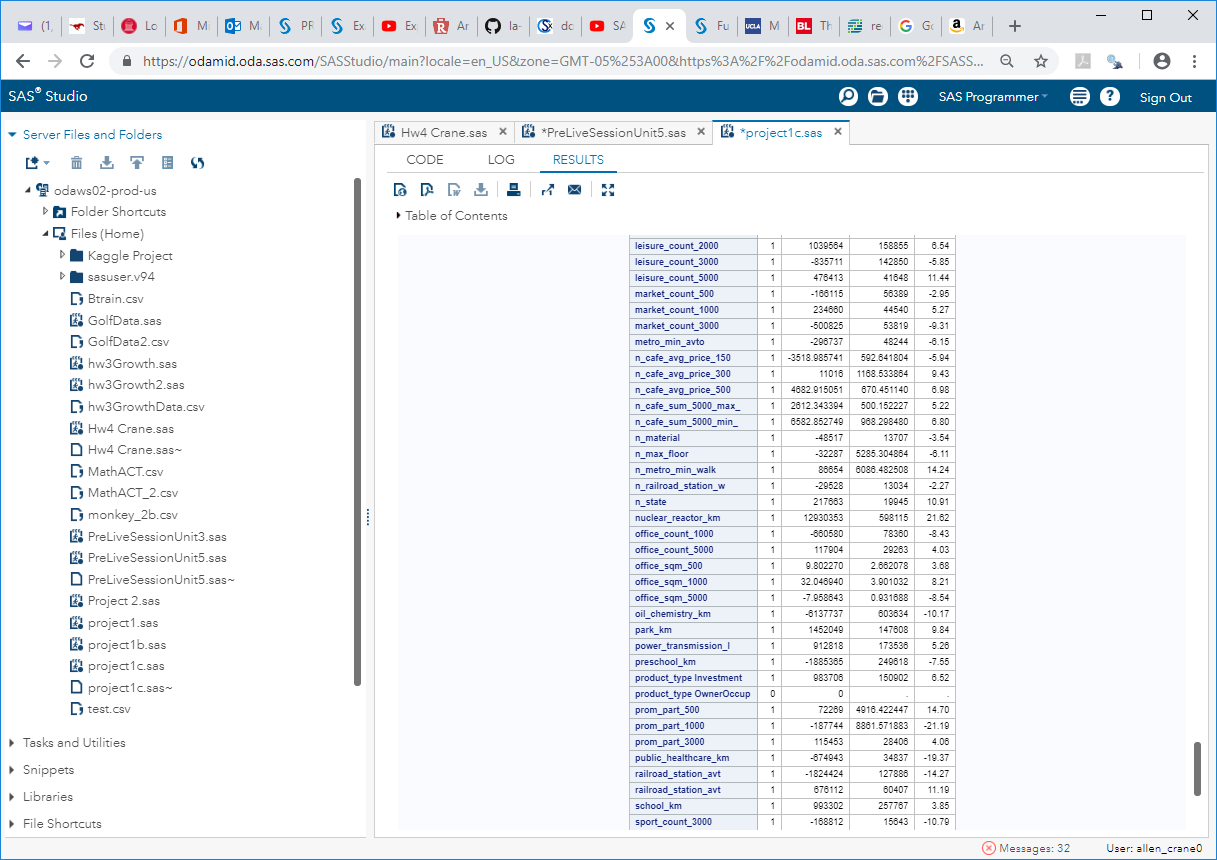
Part of this project is meant to have a significant data wrangling component. As an example, you will more than likely need to work with R or SAS or both to change data from character/string to integer/numeric so your models make the predictions that are required. This is only an example of the data wrangling you will need to conduct. It will help to start early and bring these issues up in live session and/or office hours.

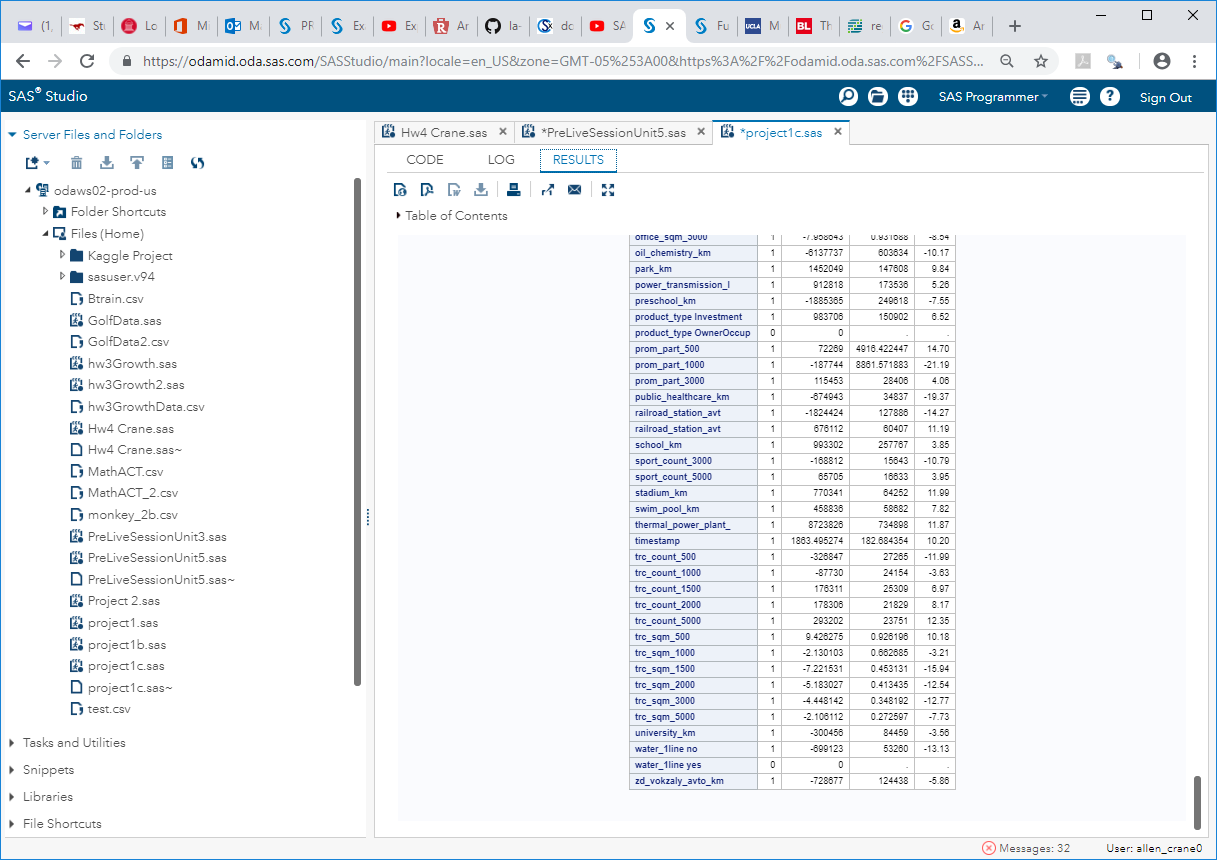
**Appendix 3. Models**

**Model 1.1 OLS (Adj R2 = 0.6268)**

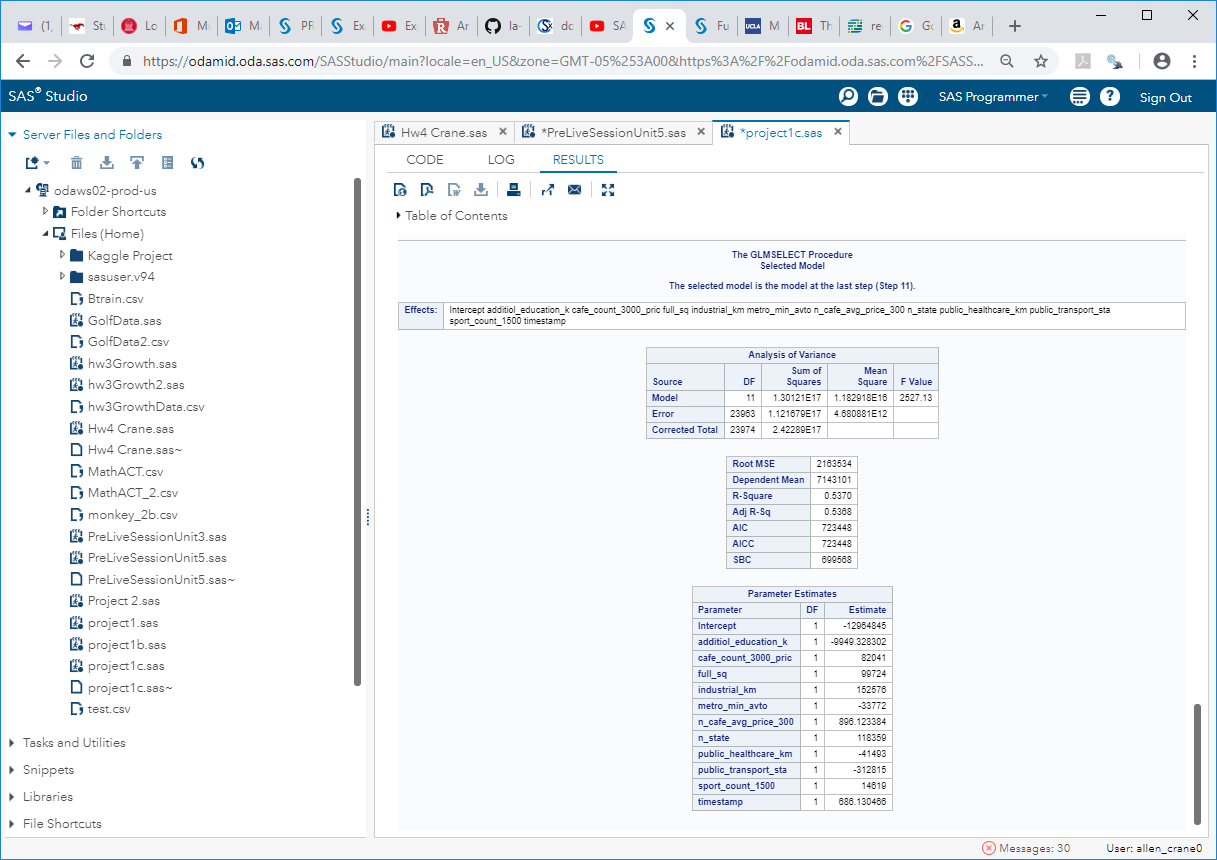








**Model 1.2 LASSO (Adj R2 = 0.5368)**



**Model 1.3 LARS (Adj R2 = 0.5368)**

