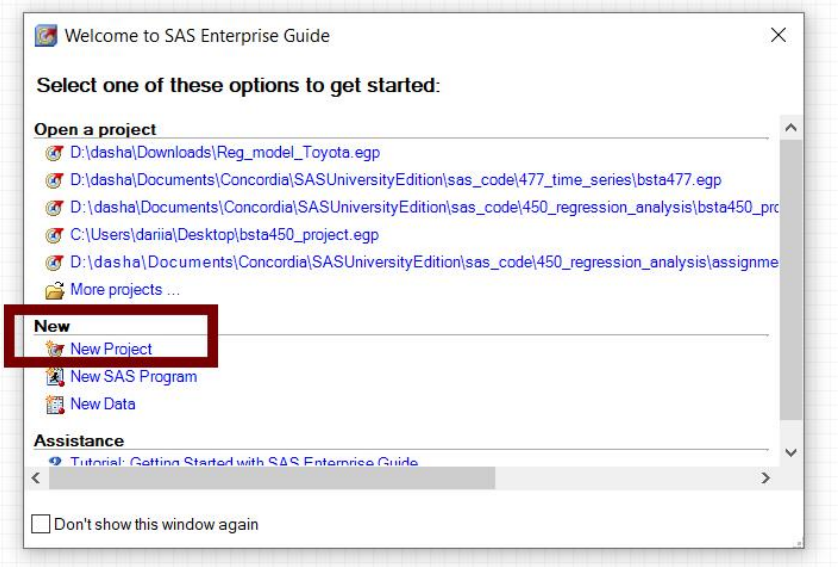
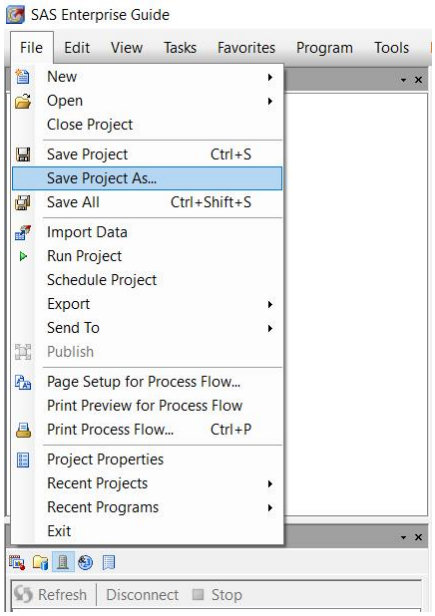
## ***OPEN A PROJECT***

**The dataset used for this project has been taken from:**
https://www.kaggle.com/harlfoxem/housesalesprediction

**1.** Open SAS EG: **NEW PROJECT**



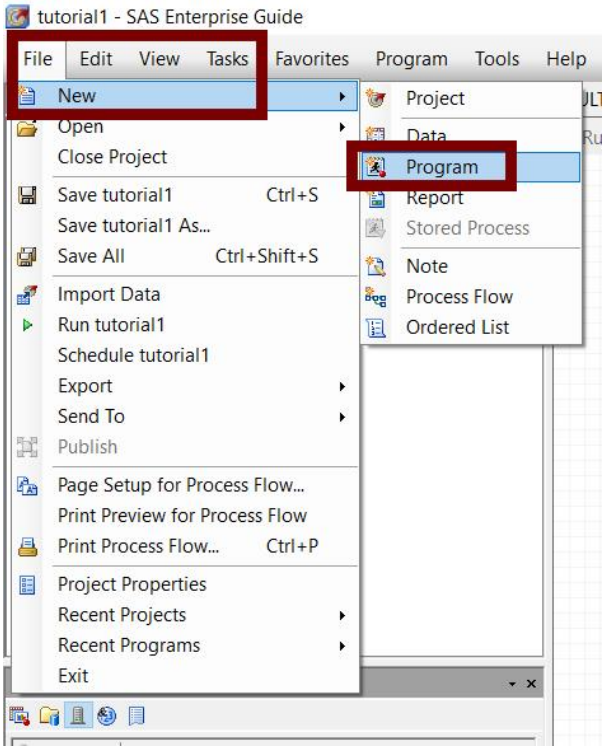

**2.** if you want to save it and have access to it, then save it in a certain place on your computer.

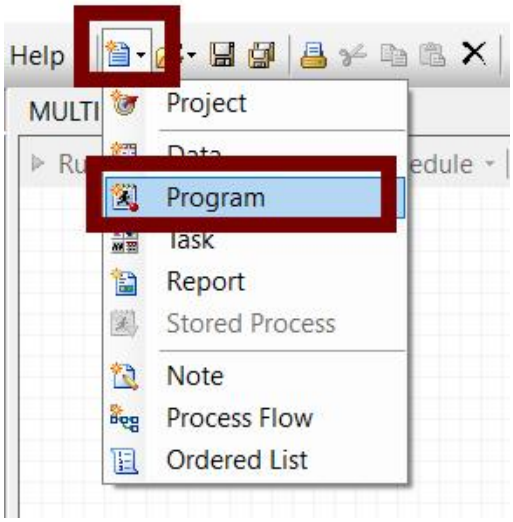


## ***ASSIGN A LIBRARY***

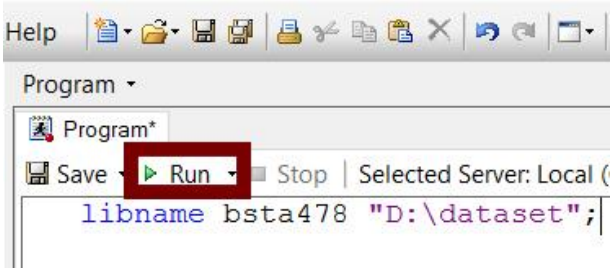
**3.** Assign a library (a place where to retrieve your datafiles from or where to save them): two options.

**A) FILE → NEW → PROGRAM** **OR** **B) ICON → PROGRAM**




**WRITE THE FOLLOWING CODE:**

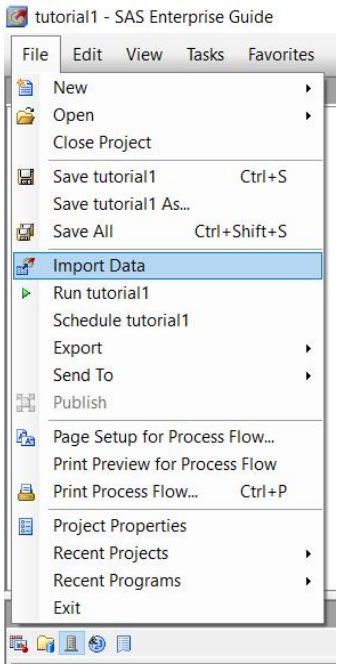


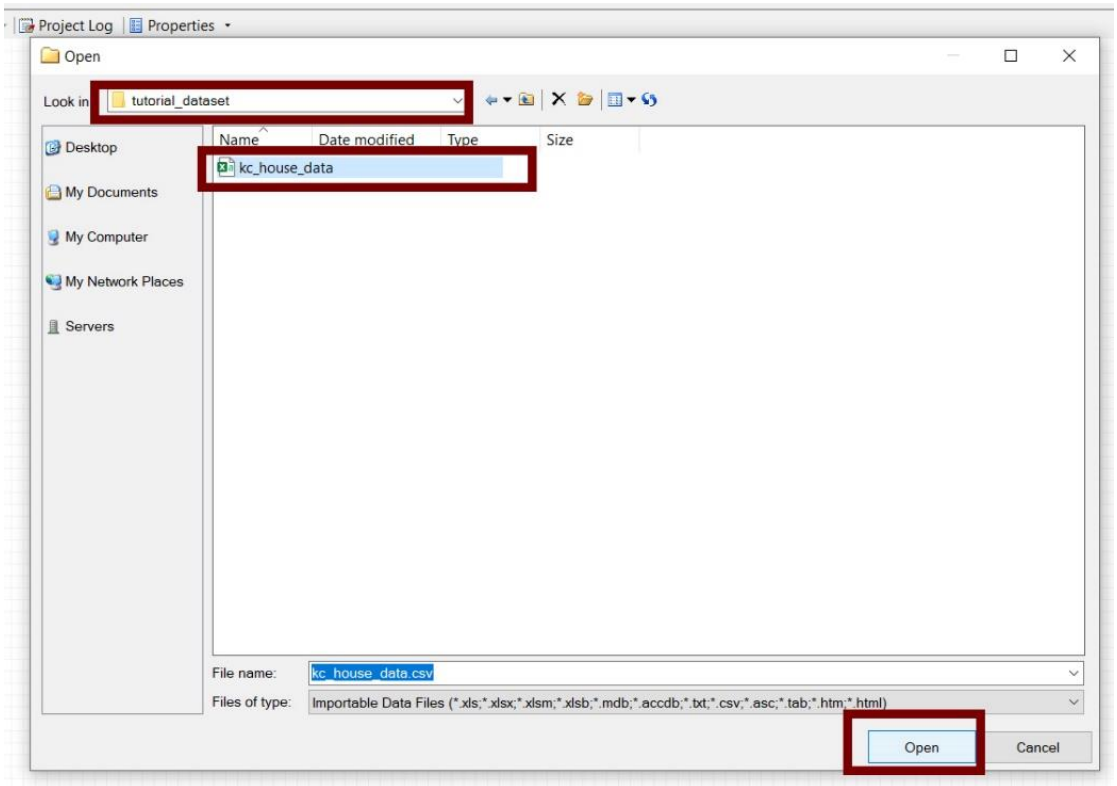

- **libname** is used to assign a library
- **bsta478** is a library name, you can assign any name you wish; it has to be max 8 characters long, should include digits and character symbols only
- **set a path in quotation marks** – where the data files are located, i.e. “C:\datafiles\”
- **put a semicolon at the end of the statement;**
- **run the program**

## ***IMPORT A FILE TO START WORKING***

**4.** Import a file:

**FILE → IMPORT DATA → LOCATE THE FILE ON YOUR COMPUTER → OPEN**



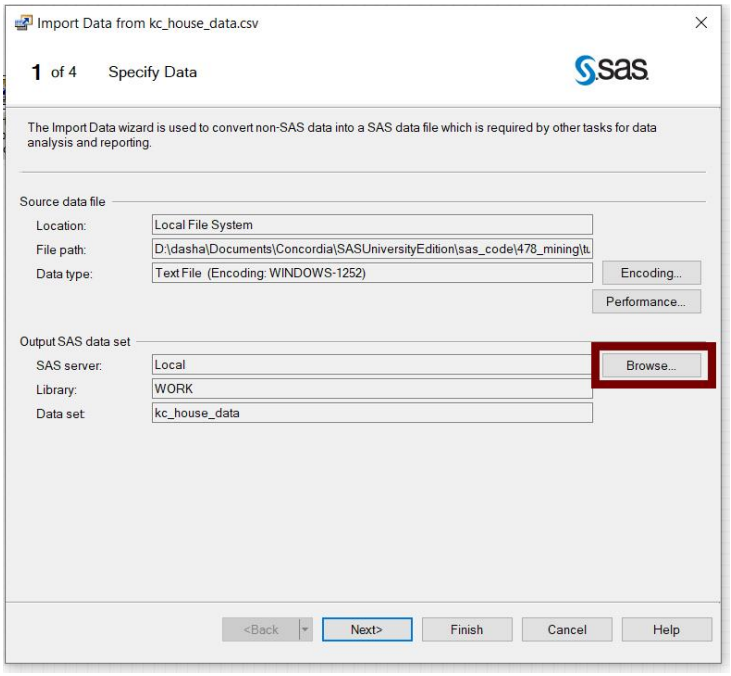



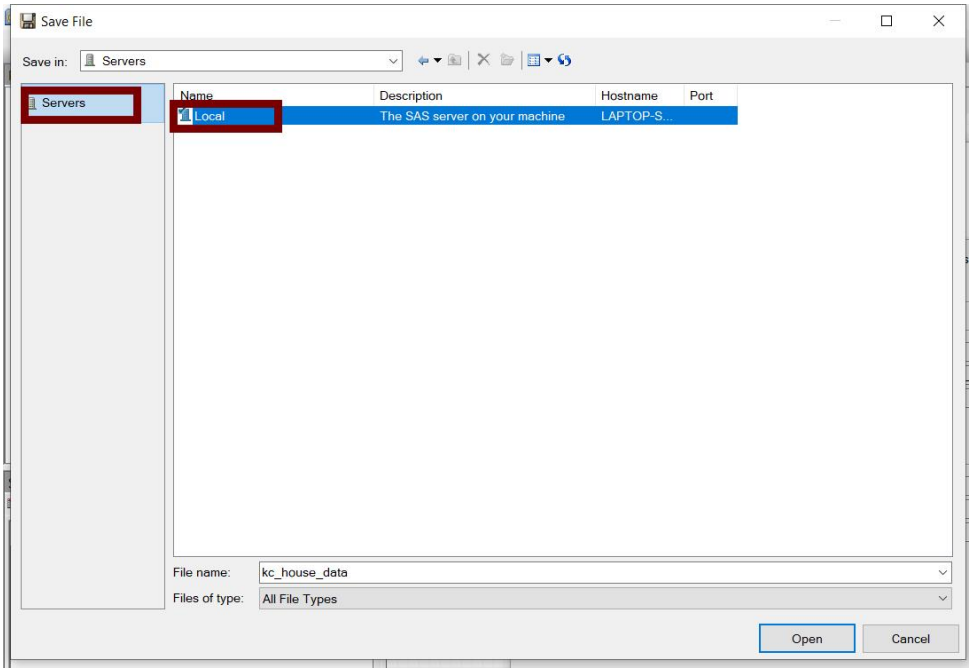
If you want to save a file in your library so that you could access it later, then you have to save it in the assigned library.

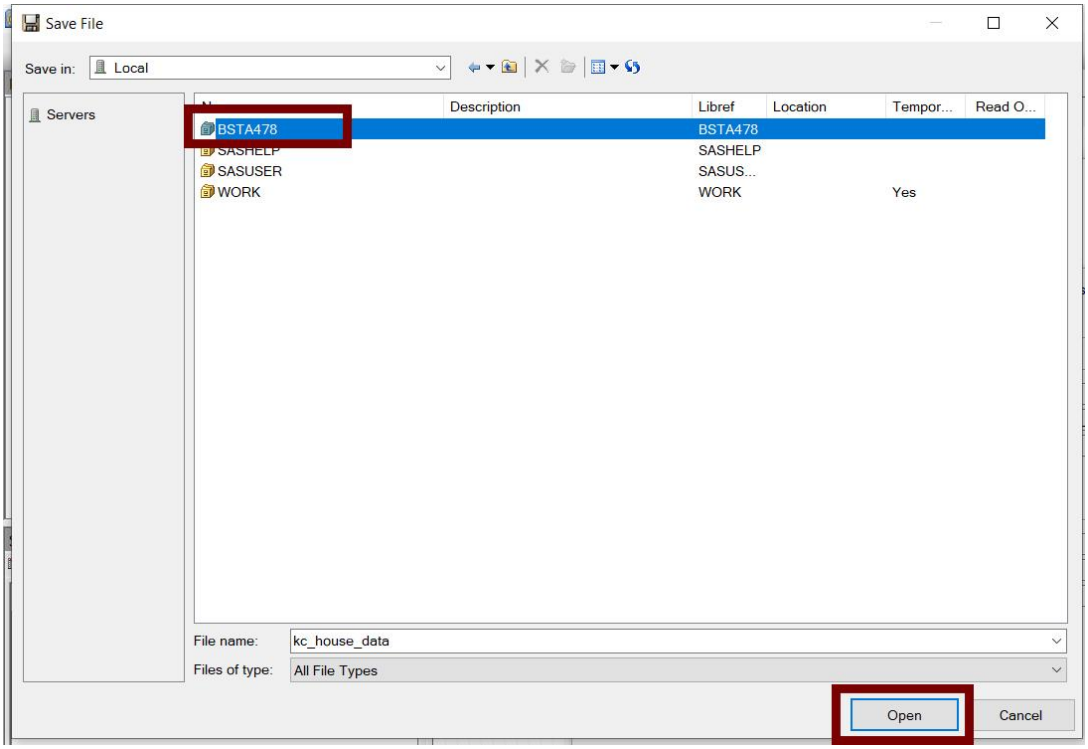
**Note:** If you don't assign a library into which to save the file, it will get deleted when you close your session and you will have to start from scratch.

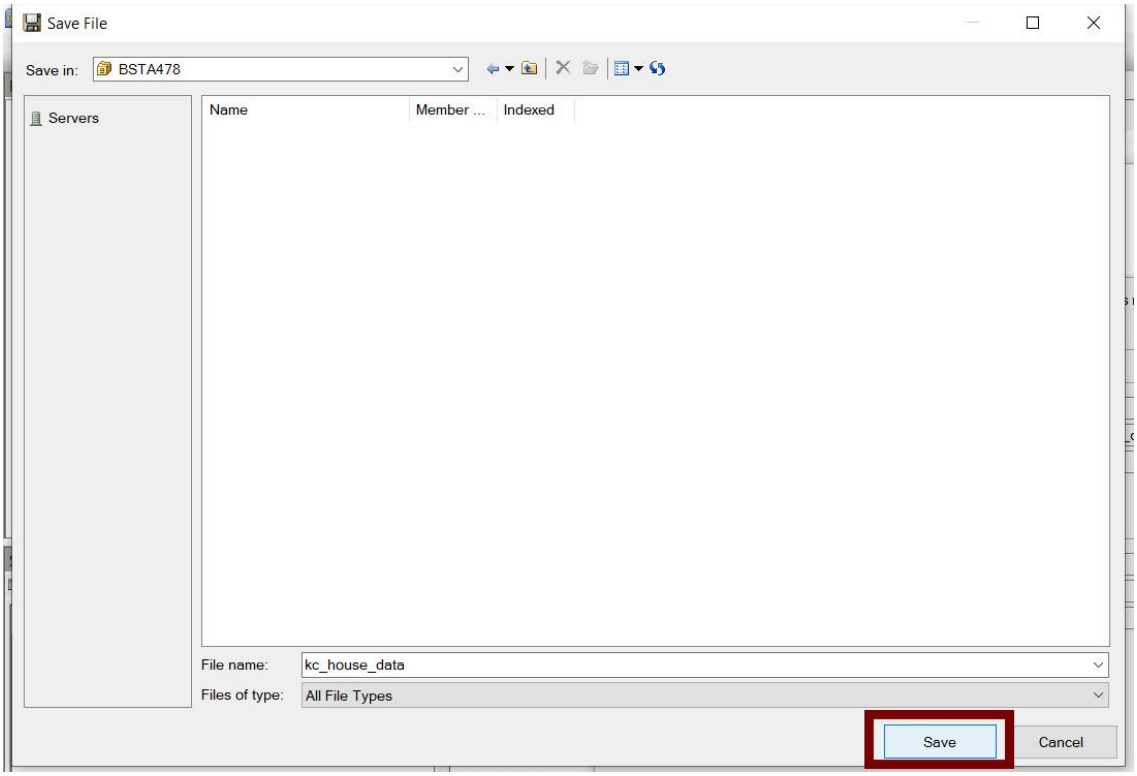
**BROWSE → SERVERS → LOCAL → BSTA478 (YOUR LIBRARY) → OPEN → SAVE → NEXT →**

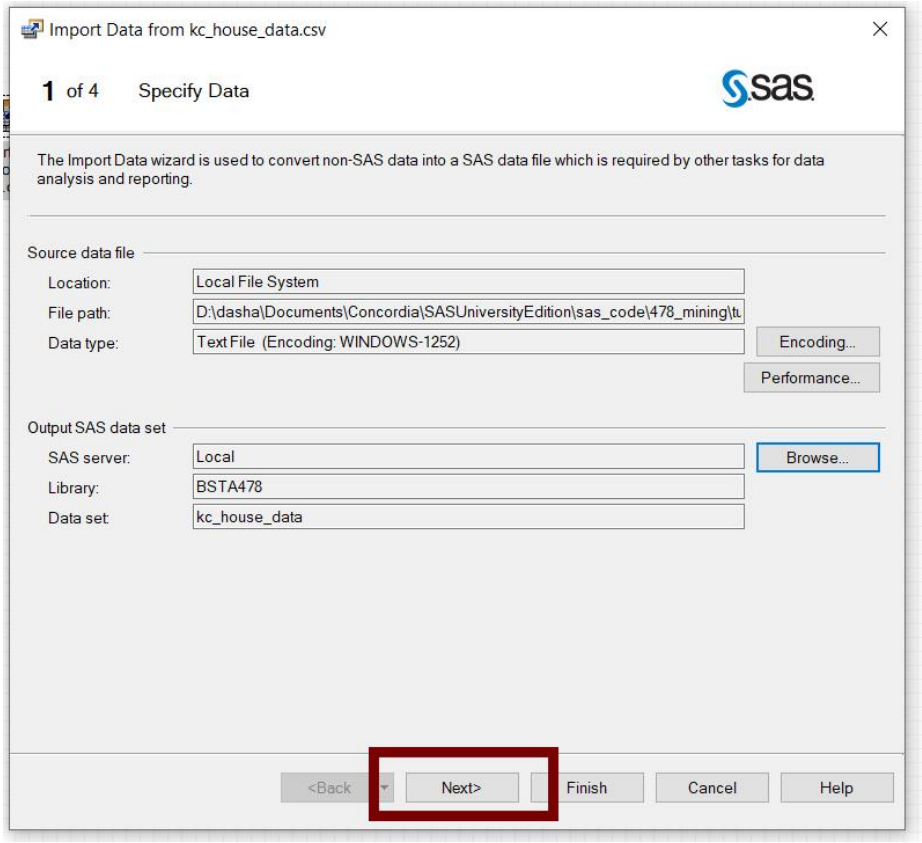
**TICK RENAME COLUMNS TO COMPLY WITH SAS NAMING CONVENTIONS → NEXT → NEXT → FINISH**

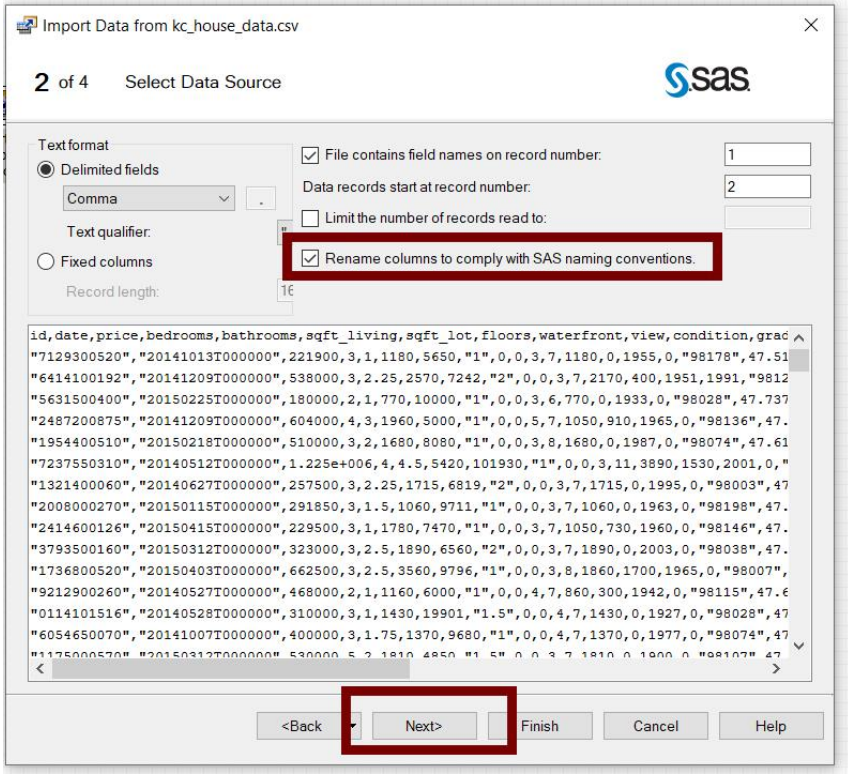


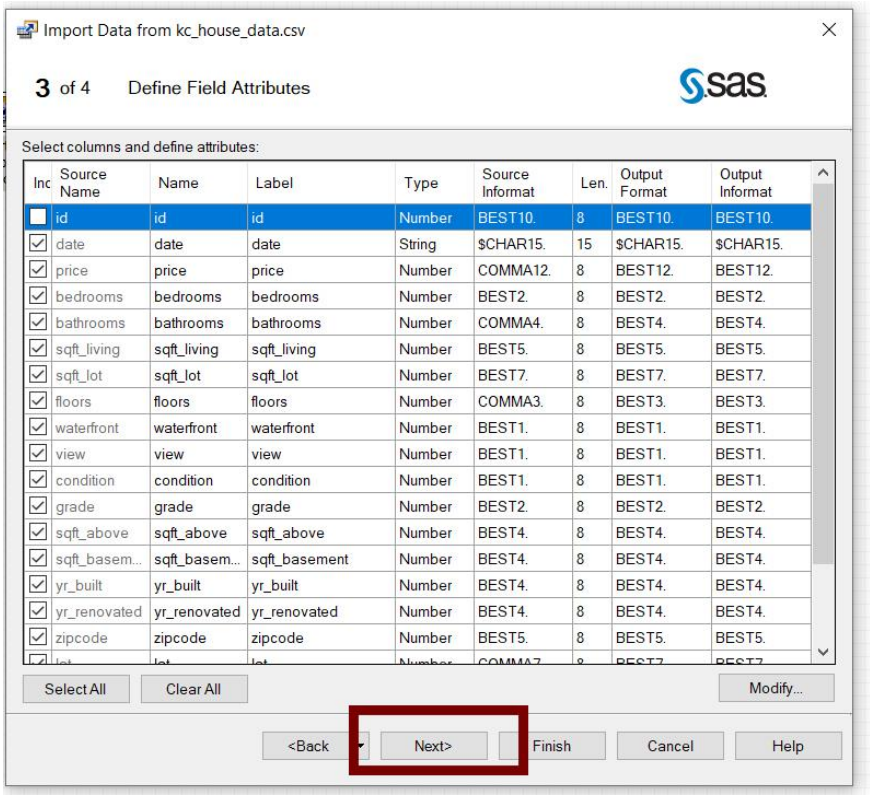



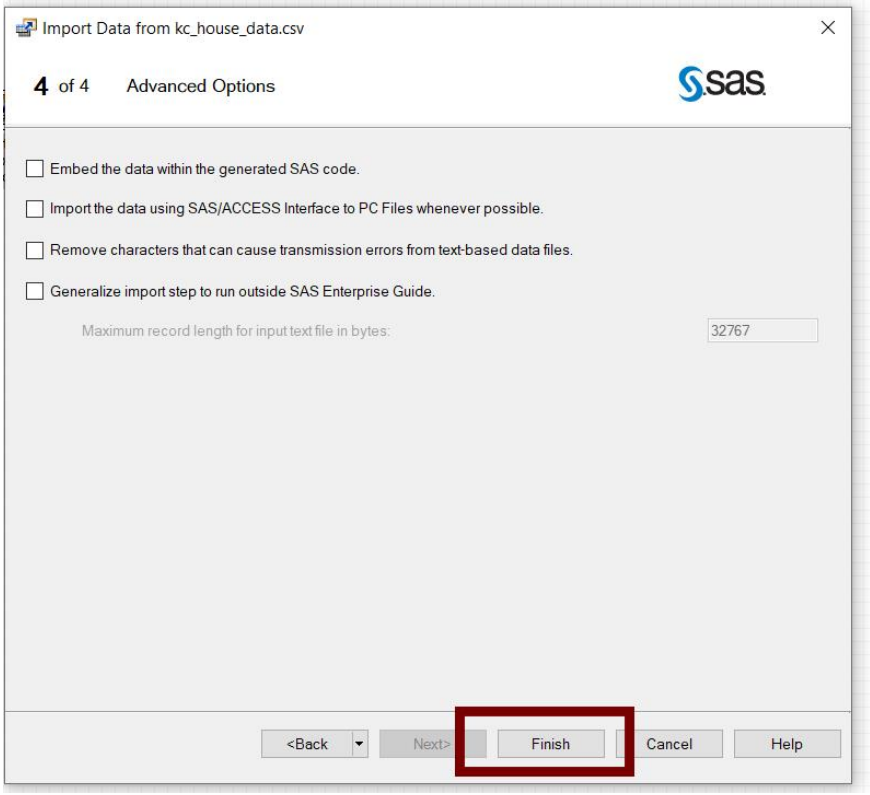

Save File
Save in: Local
Servers
Description
Libref
Location
Tempor...
Read O...
BSTA478
BSTA478
SASHELP
SASHELP
SASUSER
SASUS...
WORK
WORK
Yes
File name: kc_house_data
Files of type: All File Types
Open
Cancel


Save File
Save in: BSTA478
Servers
Name
Member ...
Indexed
File name: kc_house_data
Files of type: All File Types
Save
Cancel


Import Data from kc_house_data.csv
1 of 4 Specify Data
The Import Data wizard is used to convert non-SAS data into a SAS data file which is required by other tasks for data analysis and reporting.
Source data file
Location: Local File System
File path: D:\dasha\Documents\Concordia\SASUniversityEdition\sas_code\478_mining\t
Data type: Text File (Encoding: WINDOWS-1252)
Encoding...
Performance...
Output SAS data set
SAS server: Local
Library: BSTA478
Data set: kc_house_data
Browse...
<Back
Next>
Finish
Cancel
Help


Import Data from kc_house_data.csv
2 of 4 Select Data Source
Text format
Delimited fields
Comma
Text qualifier:
Fixed columns
Record length:
File contains field names on record number: 1
Data records start at record number: 2
Limit the number of records read to:
Rename columns to comply with SAS naming conventions.
id,date,price,bedrooms,bathrooms,sqft_living,sqft_lot,floors,waterfront,view,condition,grad
"7129300520","20141013T000000",221900,3,1,1180,5650,"1",0,0,3,7,1180,0,1955,0,"98178",47.51
"6414100192","20141209T000000",538000,3,2.25,2570,7242,"2",0,0,3,7,2170,400,1951,1991,"9812
"5631500400","20150225T000000",180000,2,1,770,10000,"1",0,0,3,6,770,0,1933,0,"98028",47.737
"2487200875","20141209T000000",604000,4,3,1960,5000,"1",0,0,5,7,1050,910,1965,0,"98136",47.
"1954400510","20150218T000000",510000,3,2,1680,8080,"1",0,0,3,8,1680,0,1987,0,"98074",47.61
"7237550310","20140512T000000",1.225e+006,4,4.5,5420,101930,"1",0,0,3,11,3890,1530,2001,0,"
"1321400060","20140627T000000",257500,3,2.25,1715,6819,"2",0,0,3,7,1715,0,1995,0,"98003",47
"2008000270","20150115T000000",291850,3,1.5,1060,9711,"1",0,0,3,7,1060,0,1963,0,"98198",47.
"2414600126","20150415T000000",229500,3,1,1780,7470,"1",0,0,3,7,1050,730,1960,0,"98146",47.
"3793500160","20150312T000000",323000,3,2.5,1890,6560,"2",0,0,3,7,1890,0,2003,0,"98038",47.
"1736800520","20150403T000000",662500,3,2.5,3560,9796,"1",0,0,3,8,1860,1700,1965,0,"98007",
"9212900260","20140527T000000",468000,2,1,1160,6000,"1",0,0,4,7,860,300,1942,0,"98115",47.6
"0114101516","20140528T000000",310000,3,1,1430,19901,"1.5",0,0,4,7,1430,0,1927,0,"98028",47
"6054650070","20141007T000000",400000,3,1.75,1370,9680,"1",0,0,4,7,1370,0,1977,0,"98074",47
<Back
Next>
Finish
Cancel
Help

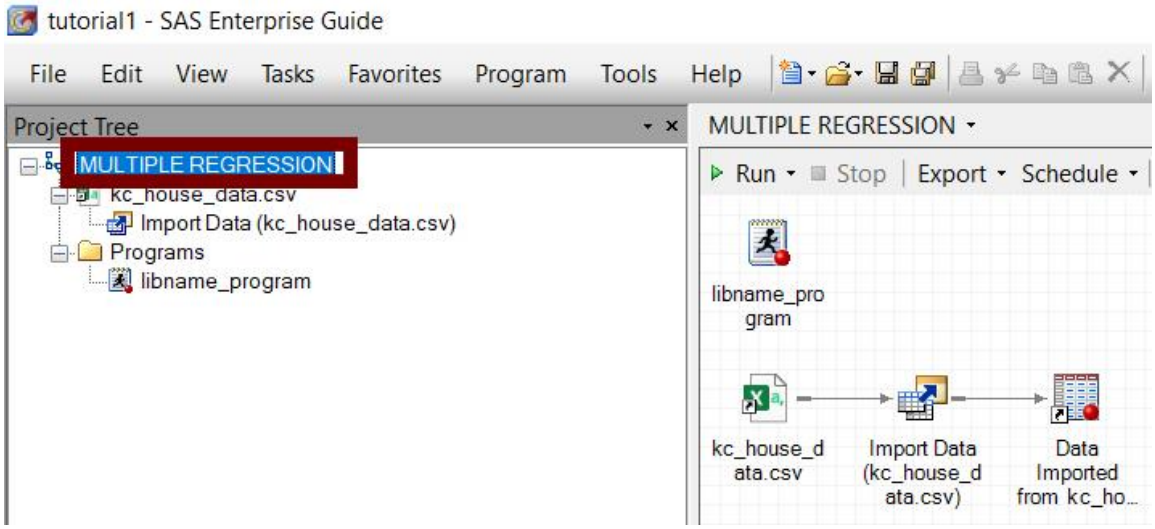


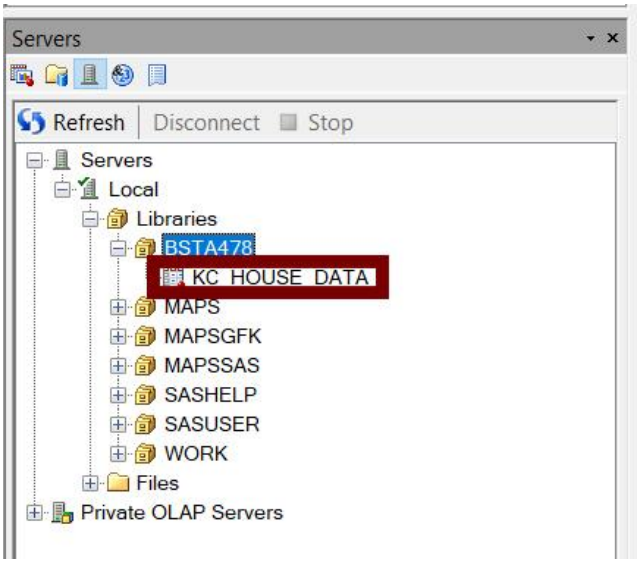

You have imported the file, and you can start analyzing your data now 😊

## ***HOW TO CHECK WHAT YOU HAVE DONE SO FAR***

Double-click on the **PROCESS FLOW**. You will see a diagram of all your work.




You can access your files here (the lower panel of the screen on the left).

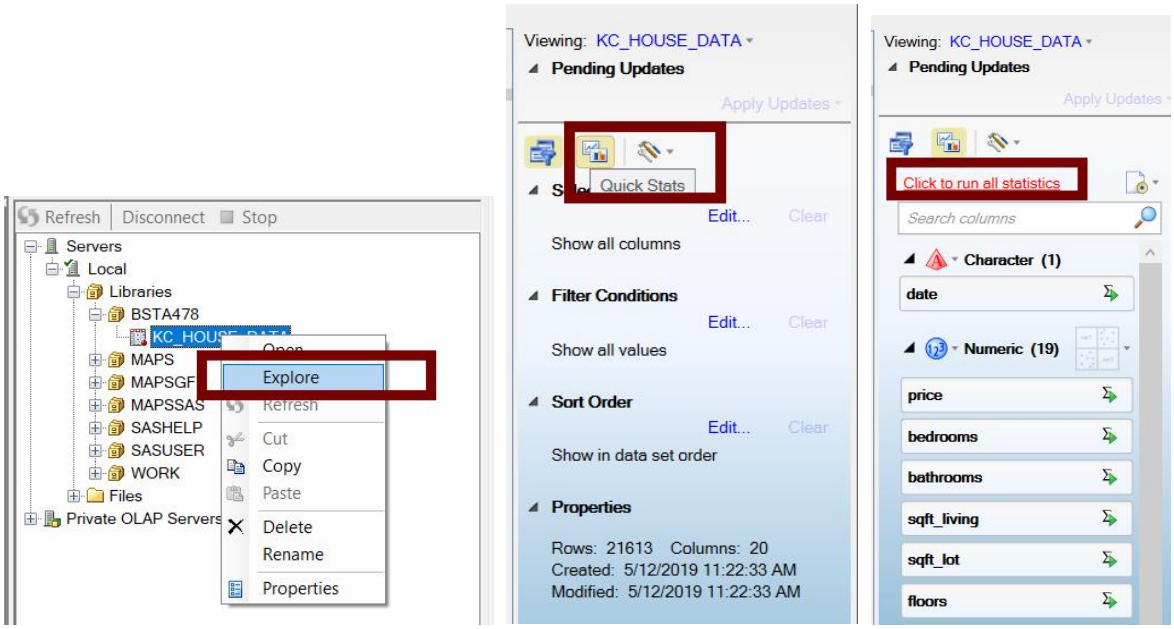


## ***DATA EXPLORATION***

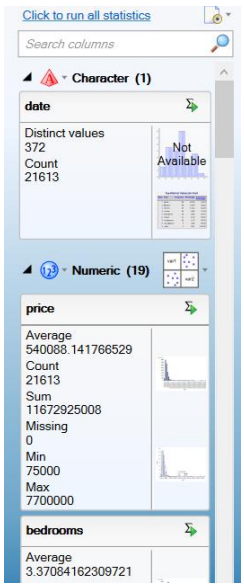
### ***1. QUICK STATS***

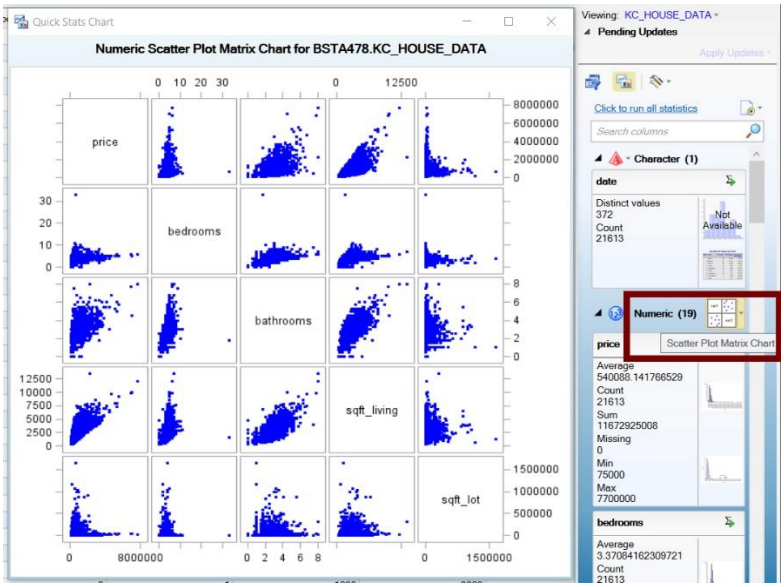
For a quick exploration of your data, go to the left bottom corner, double-click on servers, then local, libraries and your assigned library: BSTA478, for example. Right-click on the dataset you want to explore and left-click on Explore.

**SERVERS → LOCAL → LIBRARIES → BSTA478 → RIGHT-CLICK ON THE DATASET → EXPLORE → QUICK STATS (the right side of the screen) → Click to run all statistics**




You will see descriptive statistics for each of the variables you chose. You can also enlarge graphs by hovering over them, you'll see a magnifying glass icon, click on the graph you want to enlarge.

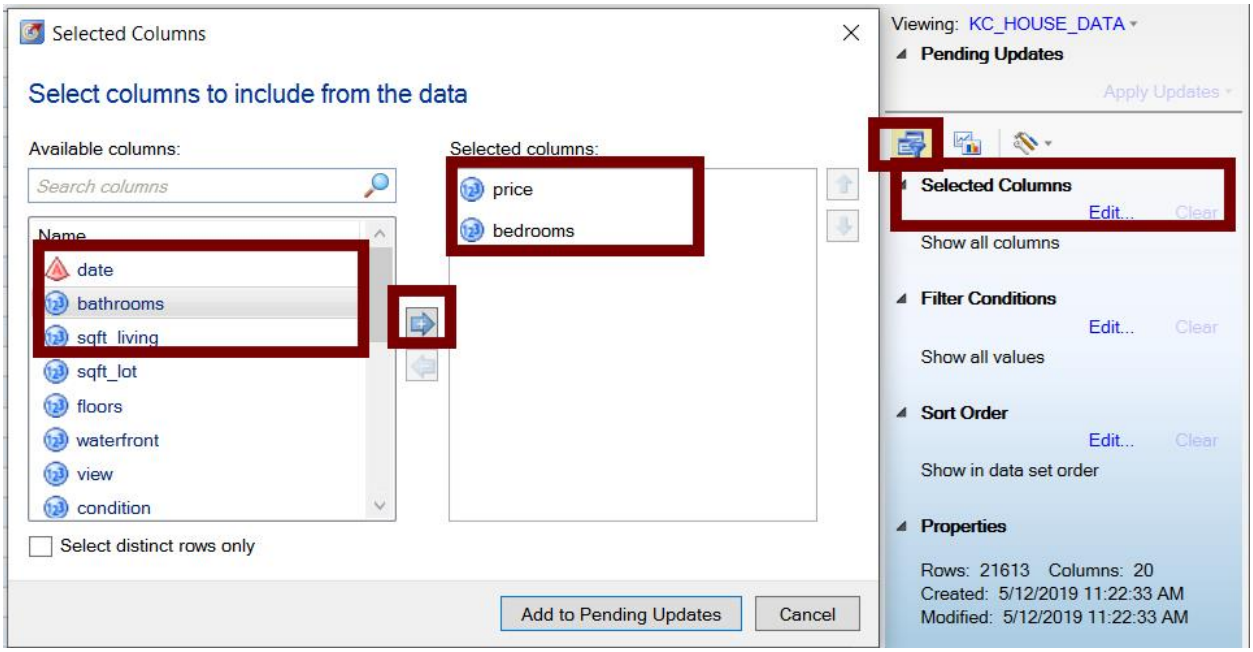


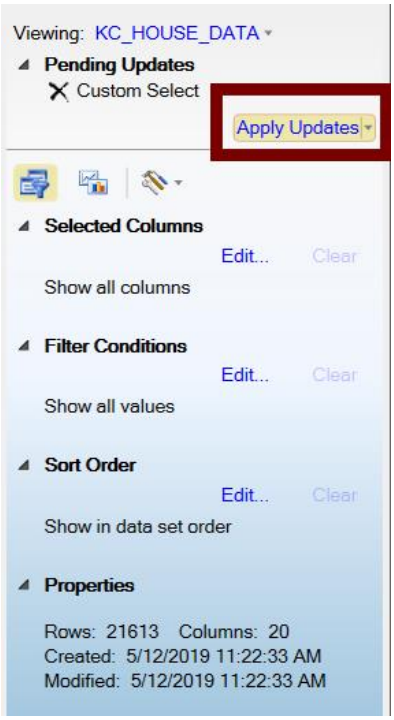



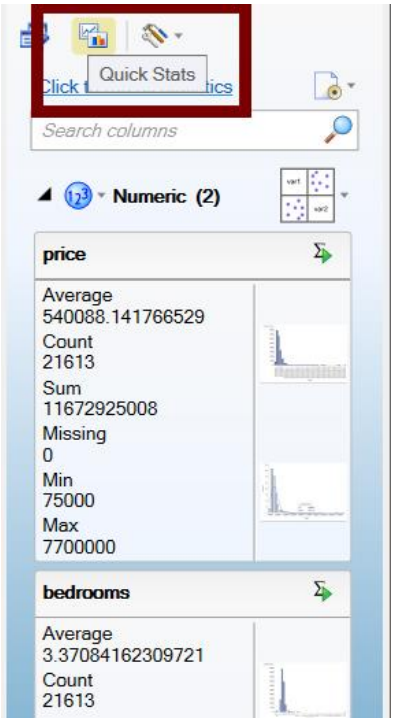
If you have too many variables, you can select only some columns:

**SELECTED COLUMNS → DOUBLE-CLICK ON THE VARIABLES YOU WANT TO EXPLORE →**

**ADD TO PENDING UPDATES → APPLY UPDATES → QUICK STATS → Click to run all statistics (if needed)**



You can also explore the dataset in 2 ways:

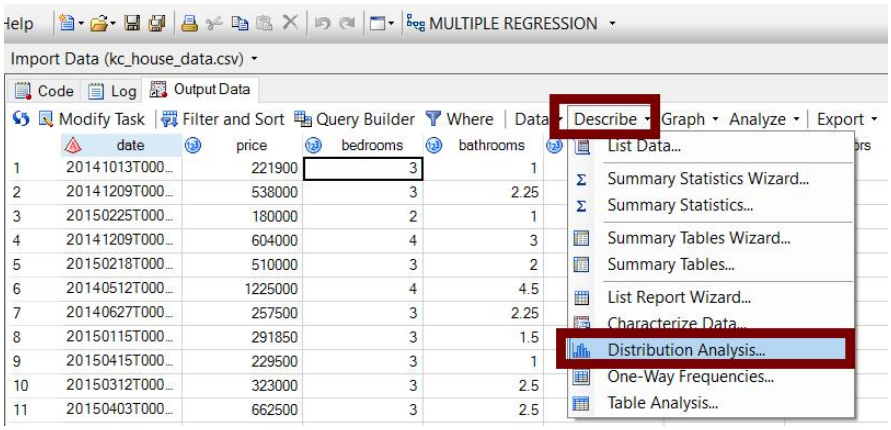
**1) Describe → The statistics option you want**

OR

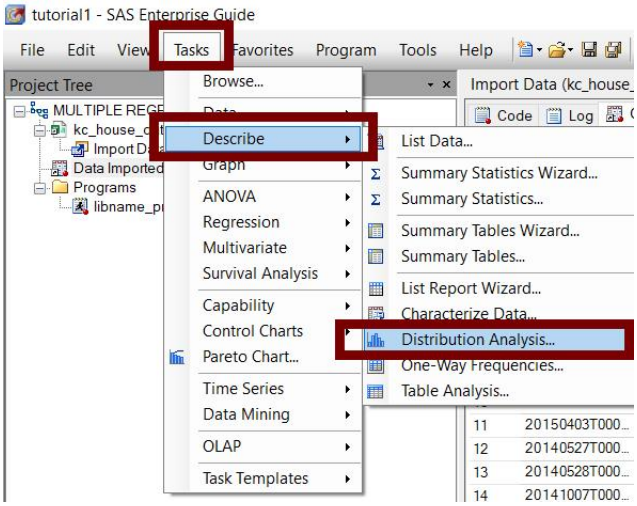
**2) Tasks → Describe → The statistics option you want**

## *2. UNIVARIATE ANALYSIS*

### 1. HISTOGRAM

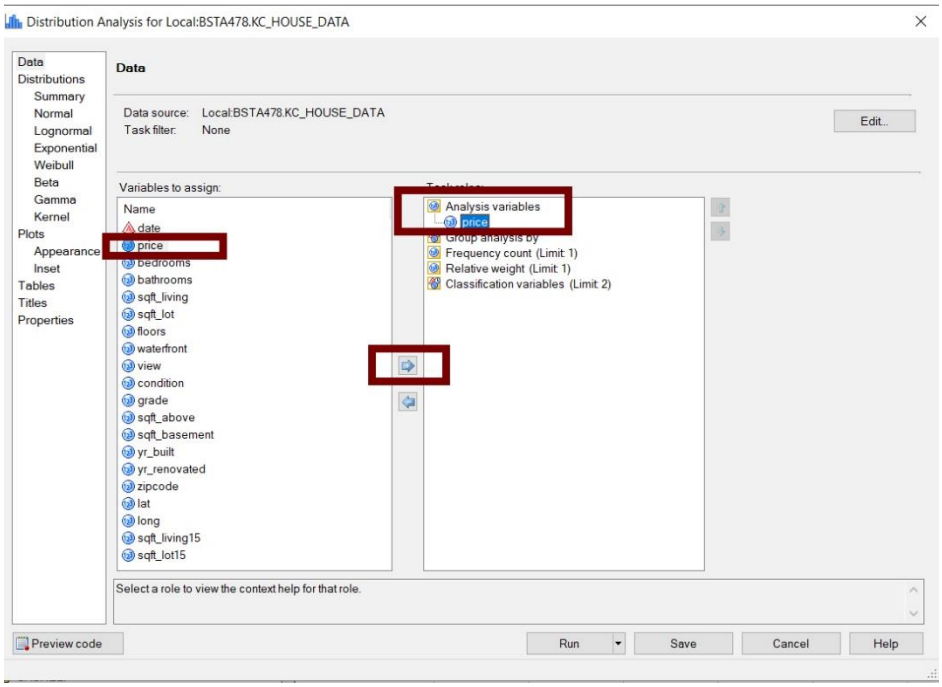



OR



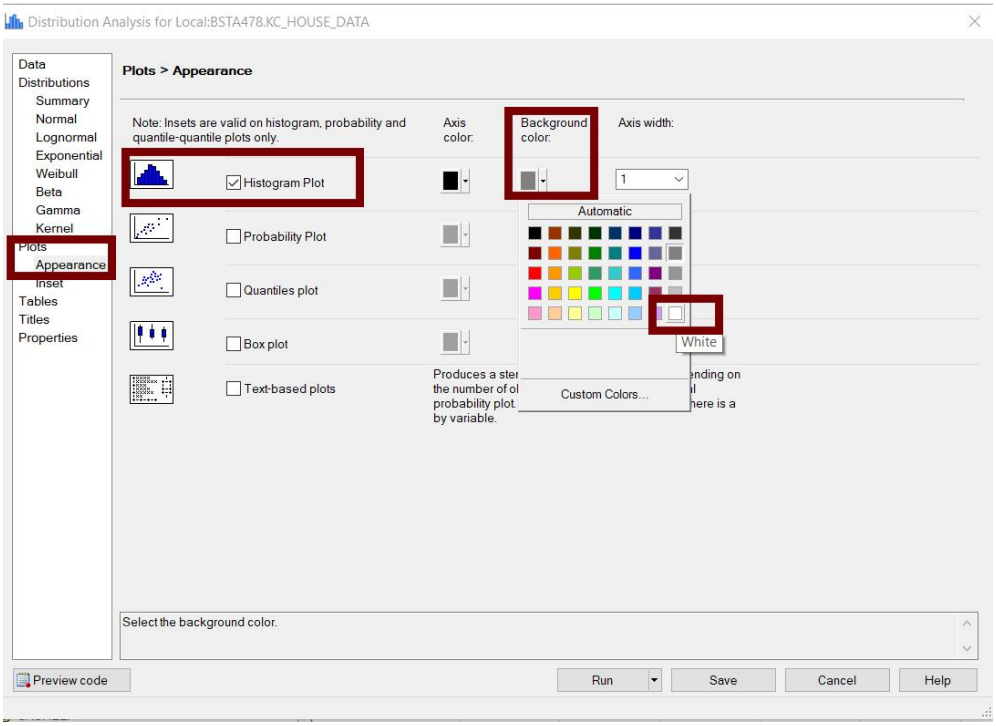
How to create the histogram:

**Data → Choose the variable and place it into the Analysis variables (you can simply drag and drop the variable of interest and move it from the left pane to the right one)**



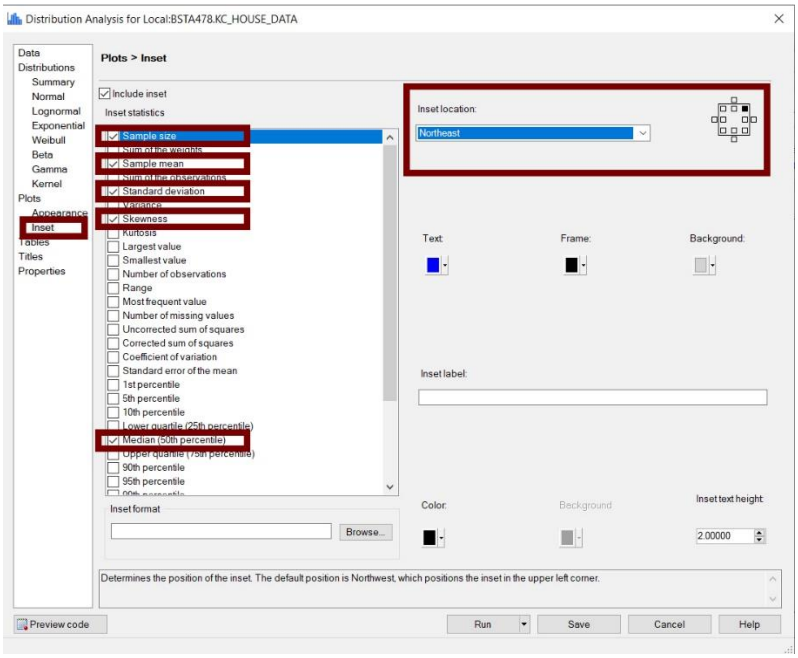

You can change the appearance of the plot:

**Plots → Appearance → Histogram Plot → Background color (choose it by yourself)**



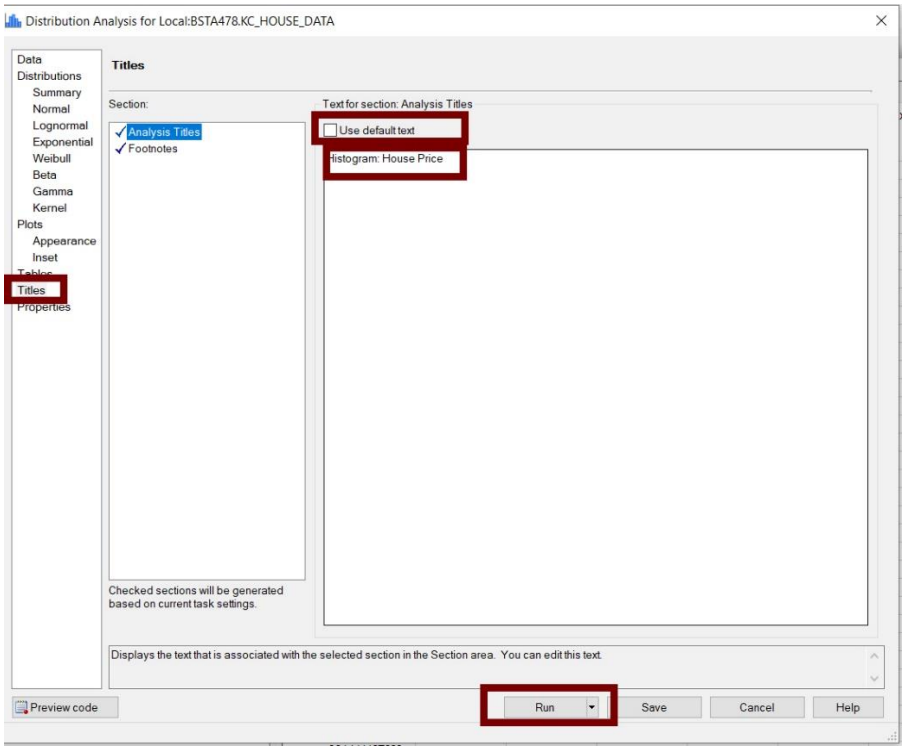
You can also place an inset with important stats next to the histogram.

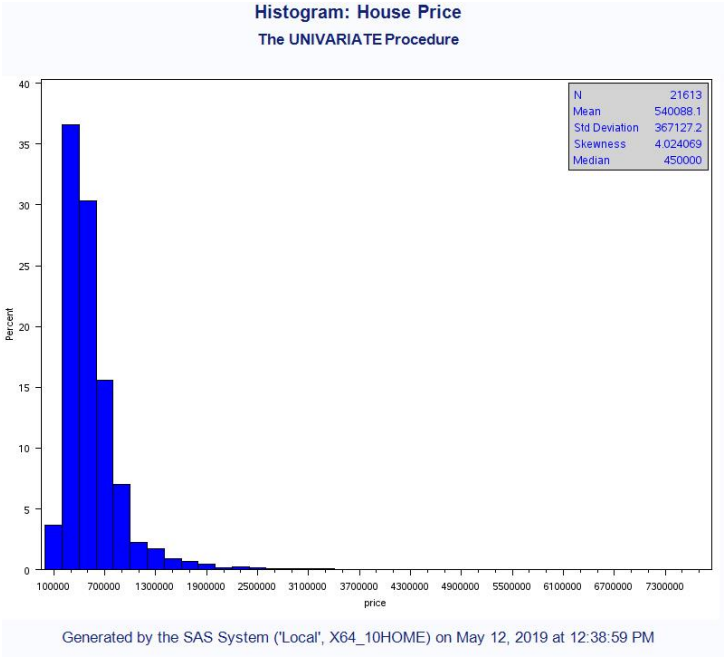
**Plots → Inset → Include inset → Select the stats data you want to be included into the inset**



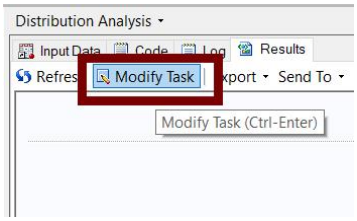

You can also change the title of the histogram by clicking:

**Titles → Analysis Titles → Remove a check from the "Use default text" → Place the text of your choice in the title window → RUN**



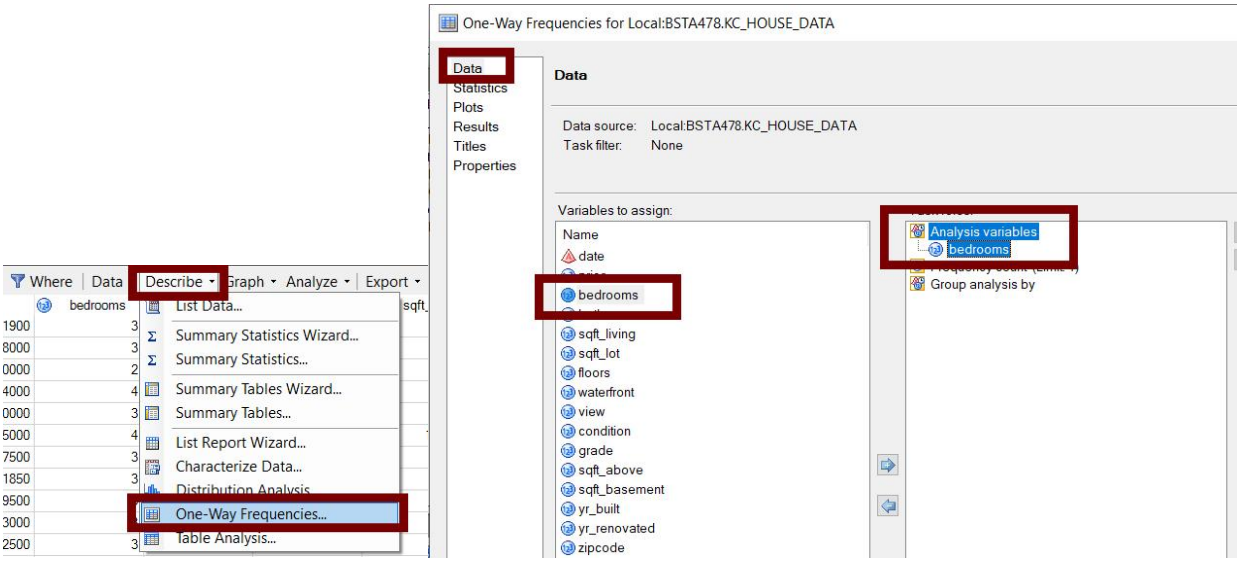



Are you not happy with the result? Click on **Modify Task** and adjust the settings.



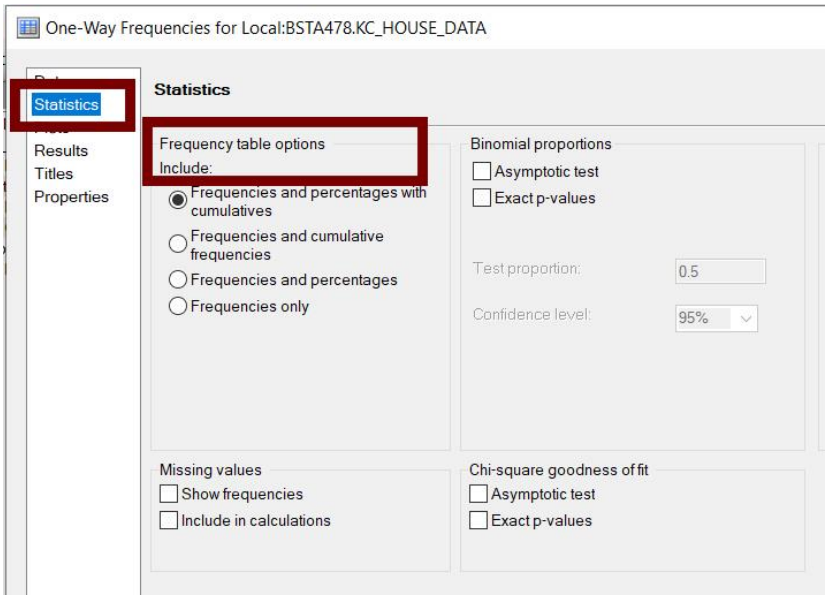

## 2. FREQUENCY TABLES (for categorical variables)

### Describe → One-Way Frequencies → Data → Drag and drop the variable of interest

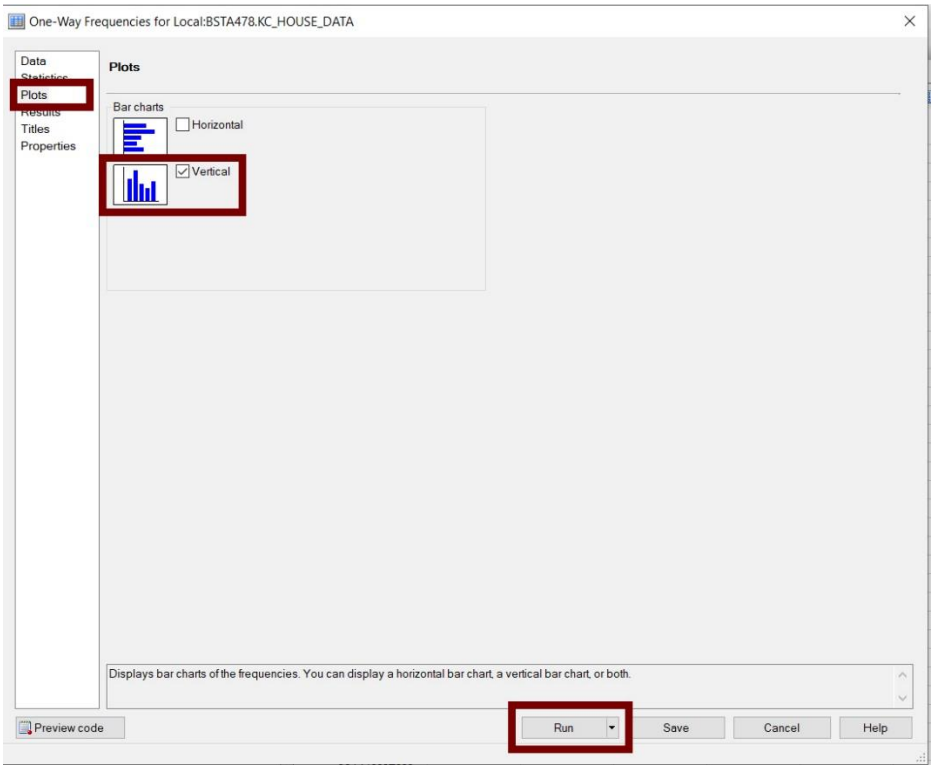


### Statistics → (modify if you like) Frequency table options Include

In this example we will keep the options at default, but you are free to change them including the stats you want to.




### Plots → Vertical




You can also change the title like done in the Histogram example.

Once you have selected all the options, click on **RUN**.

**One-Way Frequencies**

**Results**

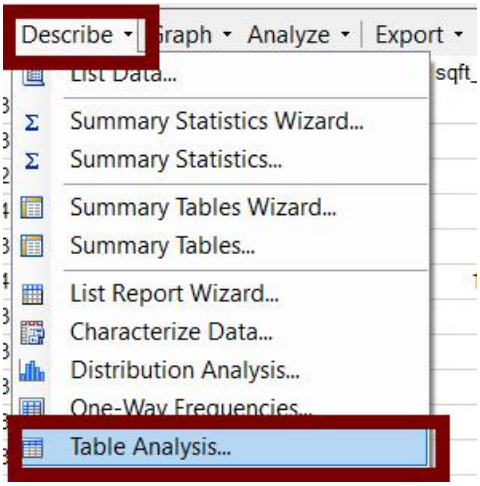
**The FREQ Procedure**

| bedrooms | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 13 | 0.06 | 13 | 0.06 |
| 1 | 199 | 0.92 | 212 | 0.98 |
| 2 | 2760 | 12.77 | 2972 | 13.75 |
| 3 | 9824 | 45.45 | 12796 | 59.21 |
| 4 | 6882 | 31.84 | 19678 | 91.05 |
| 5 | 1601 | 7.41 | 21279 | 98.45 |
| 6 | 272 | 1.26 | 21551 | 99.71 |
| 7 | 38 | 0.18 | 21589 | 99.89 |
| 8 | 13 | 0.06 | 21602 | 99.95 |
| 9 | 6 | 0.03 | 21608 | 99.98 |
| 10 | 3 | 0.01 | 21611 | 99.99 |
| 11 | 1 | 0.00 | 21612 | 100.00 |
| 33 | 1 | 0.00 | 21613 | 100.00 |

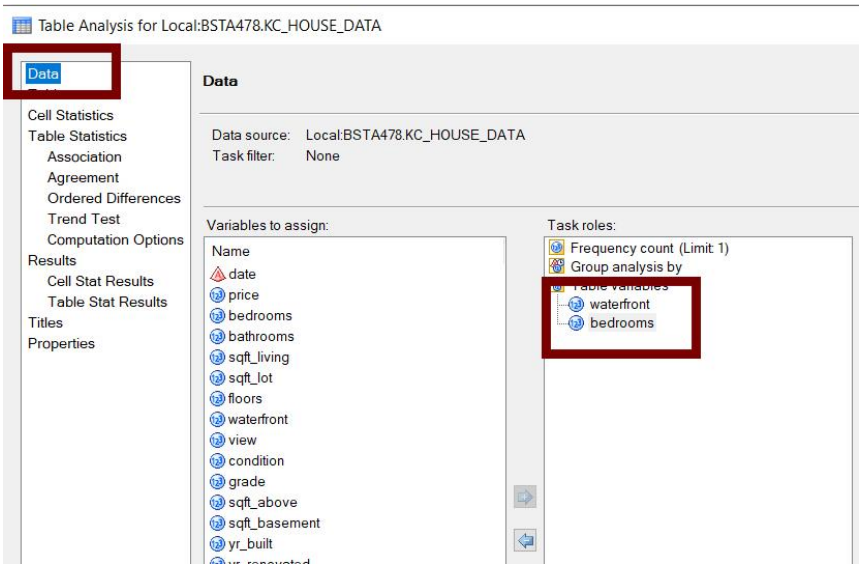
## ***3. BIVARIATE ANALYSIS***

For categorical variables, you can construct a two-way frequency table.

**Describe → Table Analysis**

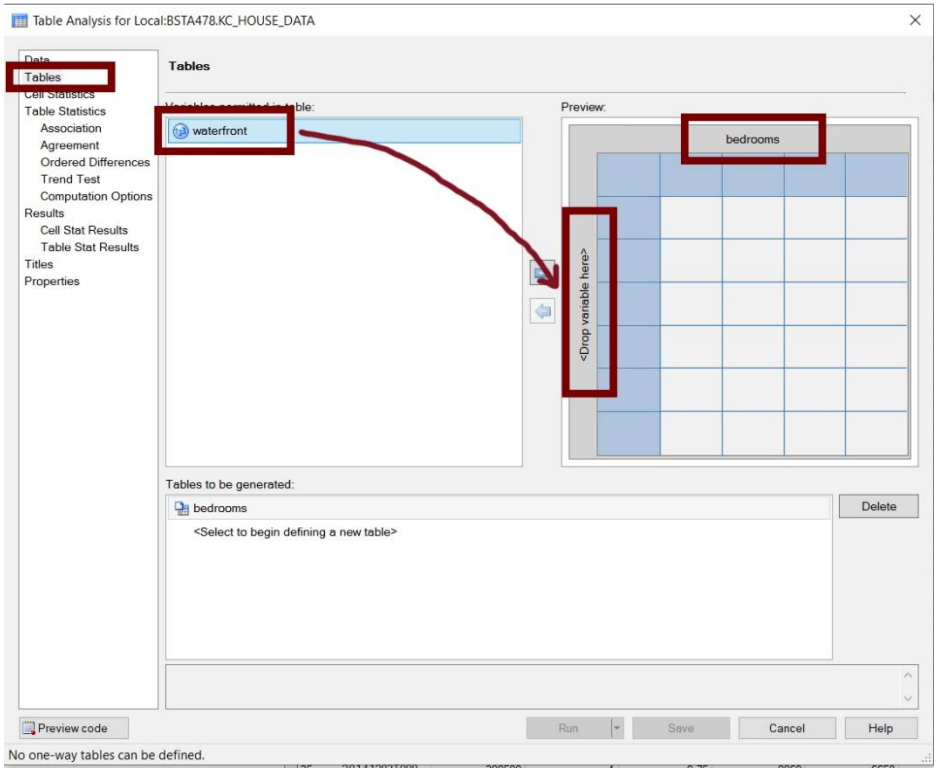


### Data → Drag and drop the variables into the Table variables



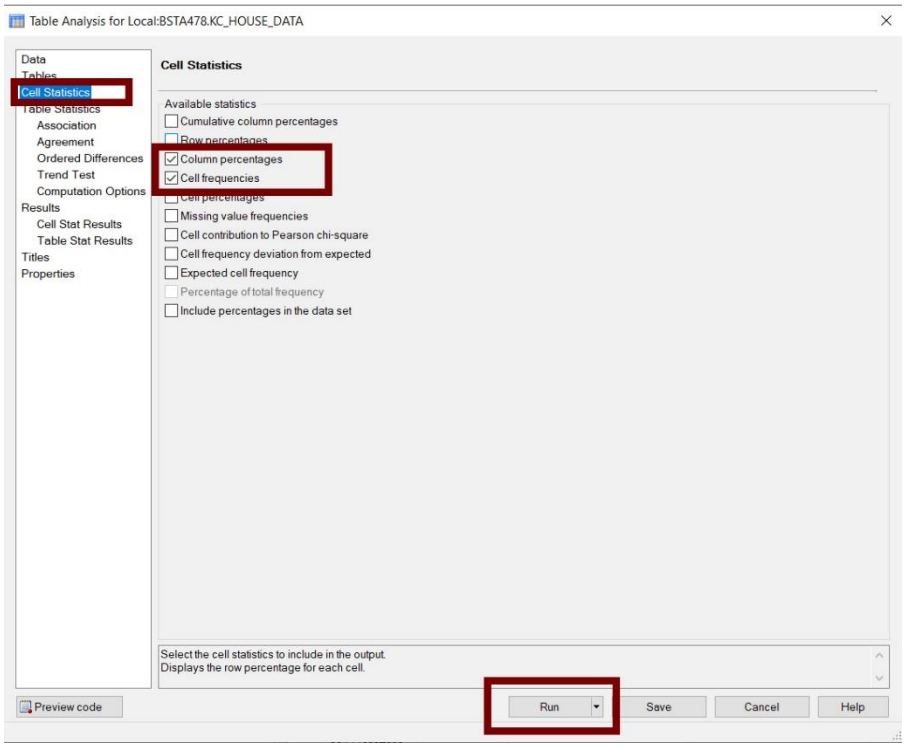

### Tables → Drag and drop variables into the table.

**Note!** The variable you drag first will end up in the column of the table.



### Cell statistics → Choose the stats you want to see in the table

You can also run a Chi-square test and others in the Table Statistics section. In addition, you can change the title of the table. Once you are done, click on **RUN**.




**Waterfront and Bedroom: Two-way table**

**The FREQ Procedure**

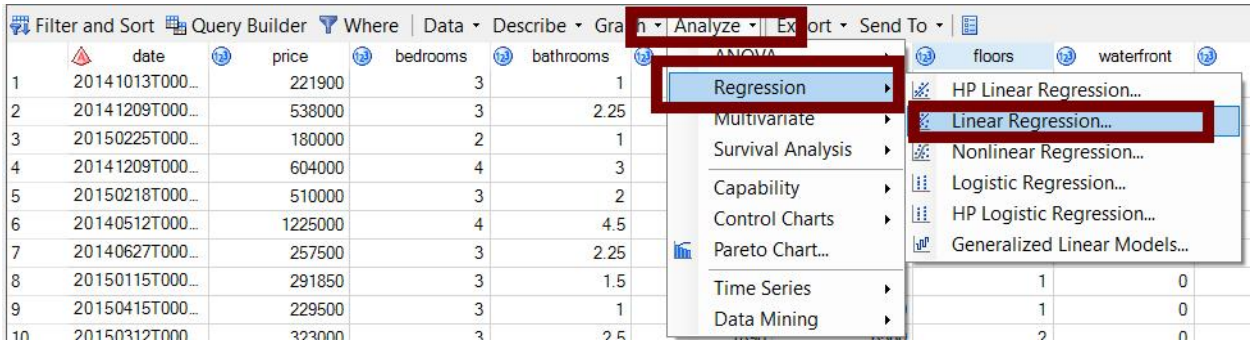
Table of waterfront by bedrooms

| waterfront | | bedrooms | | | | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 33 | |
| 0 | Frequency | 13 | 194 | 2729 | 9760 | 6842 | 1582 | 268 | 38 | 13 | 6 | 3 | 1 | 1 | 21450 |
| | Percent | 0.06 | 0.90 | 12.63 | 45.16 | 31.66 | 7.32 | 1.24 | 0.18 | 0.06 | 0.03 | 0.01 | 0.00 | 0.00 | 99.25 |
| | Row Pct | 0.06 | 0.90 | 12.72 | 45.50 | 31.90 | 7.38 | 1.25 | 0.18 | 0.06 | 0.03 | 0.01 | 0.00 | 0.00 | |
| | Col Pct | 100.00 | 97.49 | 98.88 | 99.35 | 99.42 | 98.81 | 98.53 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | |
| 1 | Frequency | 0 | 5 | 31 | 64 | 40 | 19 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 163 |
| | Percent | 0.00 | 0.02 | 0.14 | 0.30 | 0.19 | 0.09 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.75 |
| | Row Pct | 0.00 | 3.07 | 19.02 | 39.26 | 24.54 | 11.66 | 2.45 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| | Col Pct | 0.00 | 2.51 | 1.12 | 0.65 | 0.58 | 1.19 | 1.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| Total | Frequency | 13 | 199 | 2760 | 9824 | 6882 | 1601 | 272 | 38 | 13 | 6 | 3 | 1 | 1 | 21613 |
| | Percent | 0.06 | 0.92 | 12.77 | 45.45 | 31.84 | 7.41 | 1.26 | 0.18 | 0.06 | 0.03 | 0.01 | 0.00 | 0.00 | 100.00 |

Generated by the SAS System ('Local', X64_10HOME) on May 12, 2019 at 1:21:20 PM

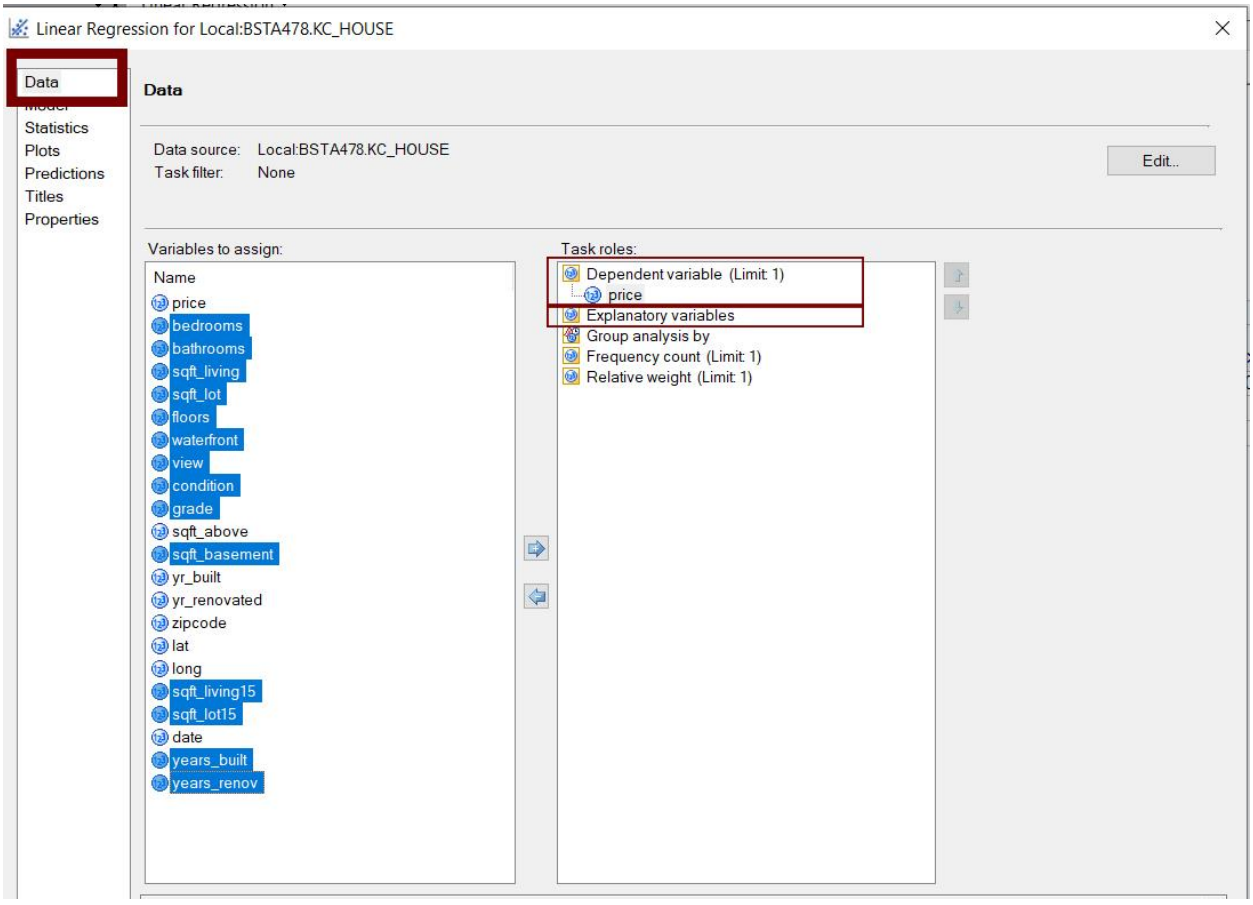
## ***MULTIPLE REGRESSION***

### ANALYZE → REGRESSION → LINEAR REGRESSION

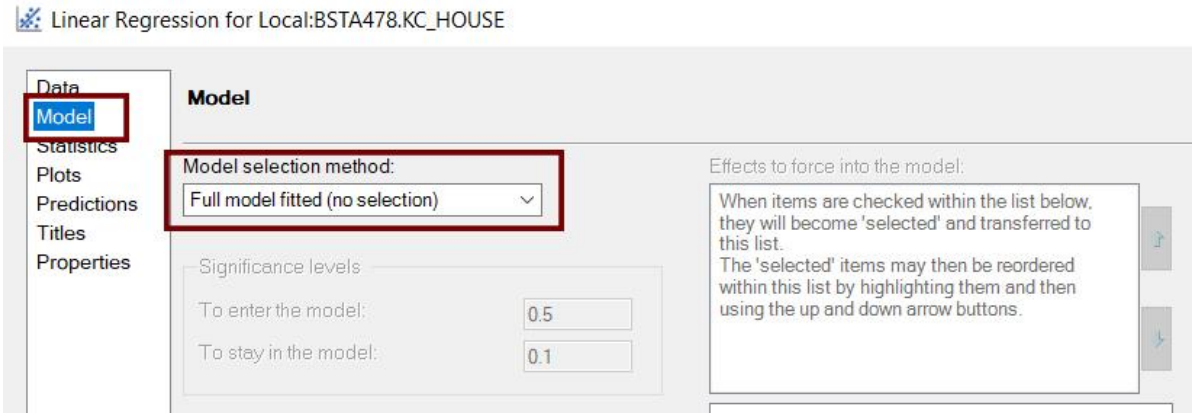


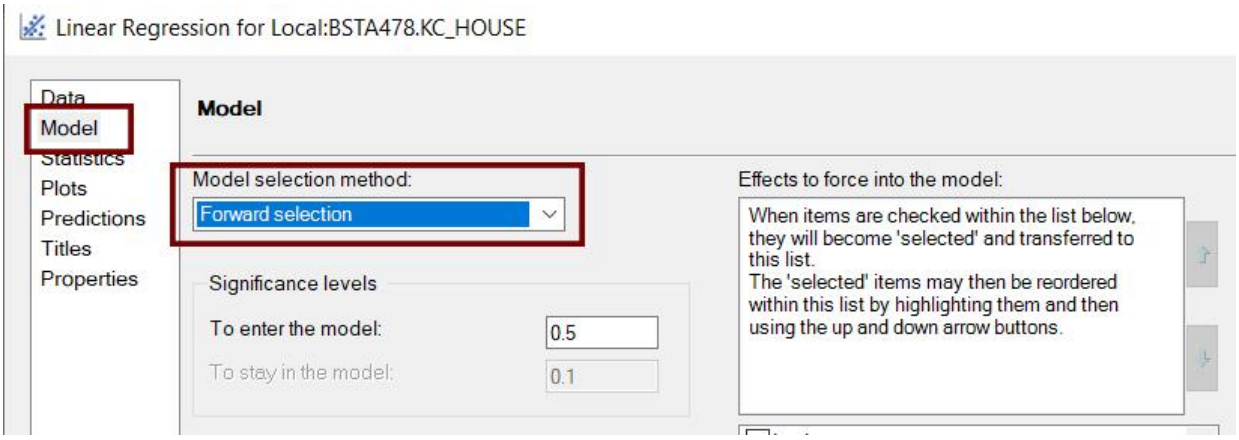

### Data → Dependent variable (can be only 1; must be numerical) → Explanatory variables (whichever ones you want to assign)

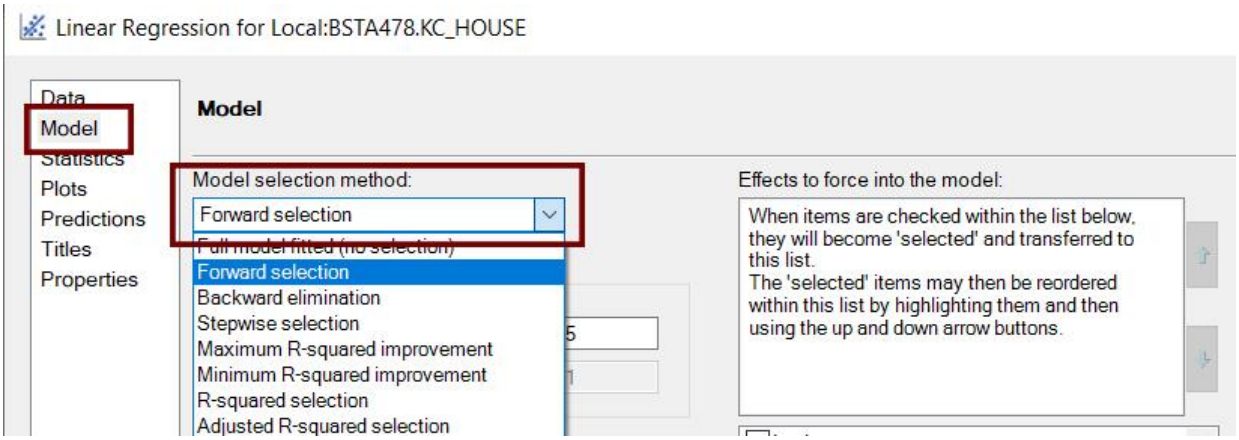
If you have categorical variables that have not been turned into dummy variables, then you have to use HP Linear Regression (Analyze → Regression → HP Linear Regression).



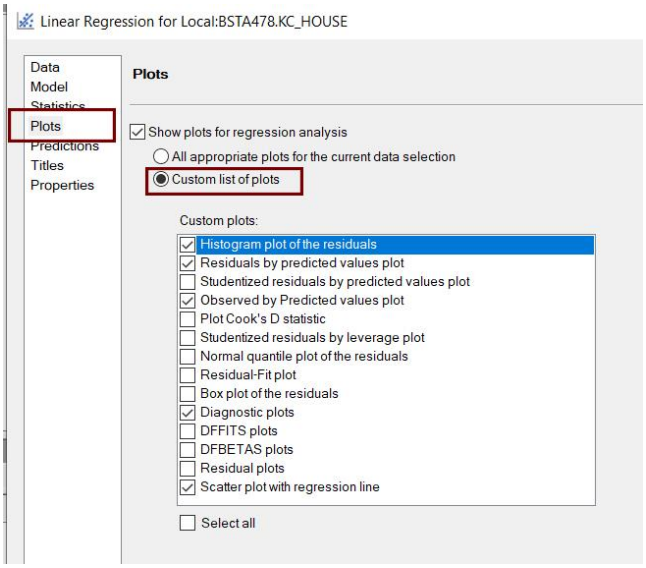
In the **MODEL** section you can choose a variable selection process: Full regression, forward, backward, stepwise etc.



In the **PLOTS** you can also choose diagnostic residual plots, i.e. residual distribution, their normality etc. Don't use this option if you have too many datapoints, i.e. over 5000.




Once you are done, click **RUN**.

**Full regression model results**

**Full Linear Regression Results**

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: price

| | |
|---|---|
| Number of Observations Read | 21613 |
| Number of Observations Used | 21613 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 14 | 1.9051E15 | 1.360786E14 | 2916.23 | <.0001 |
| Error | 21598 | 1.007817E15 | 46662492273 | | |
| Corrected Total | 21612 | 2.912917E15 | | | |

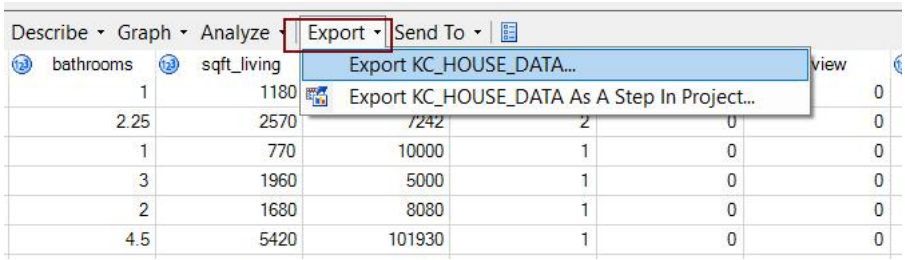
| | | | |
|---|---|---|---|
| Root MSE | 216015 | R-Square | 0.6540 |
| Dependent Mean | 540088 | Adj R-Sq | 0.6538 |
| Coeff Var | 39.99625 | | |

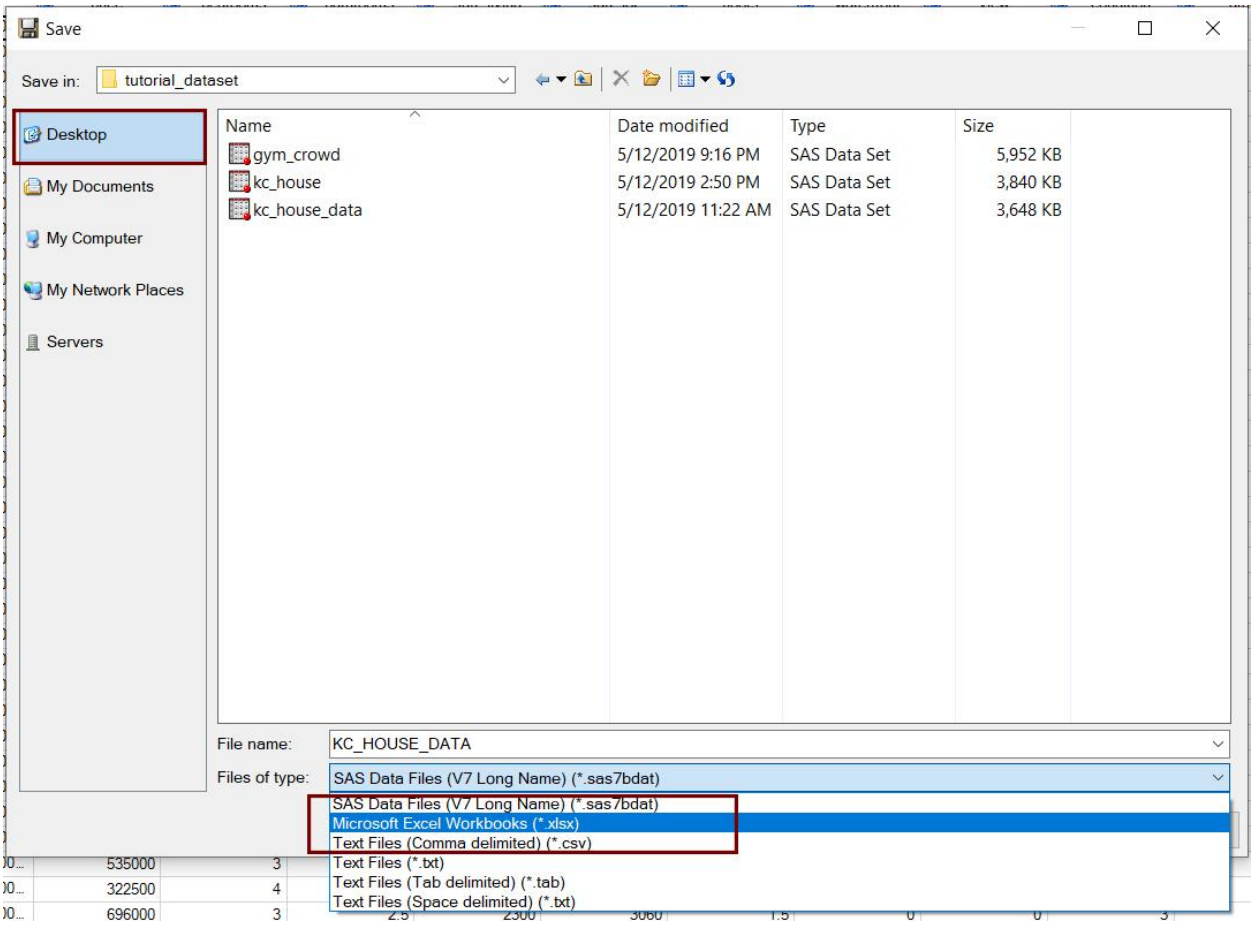
Parameter Estimates

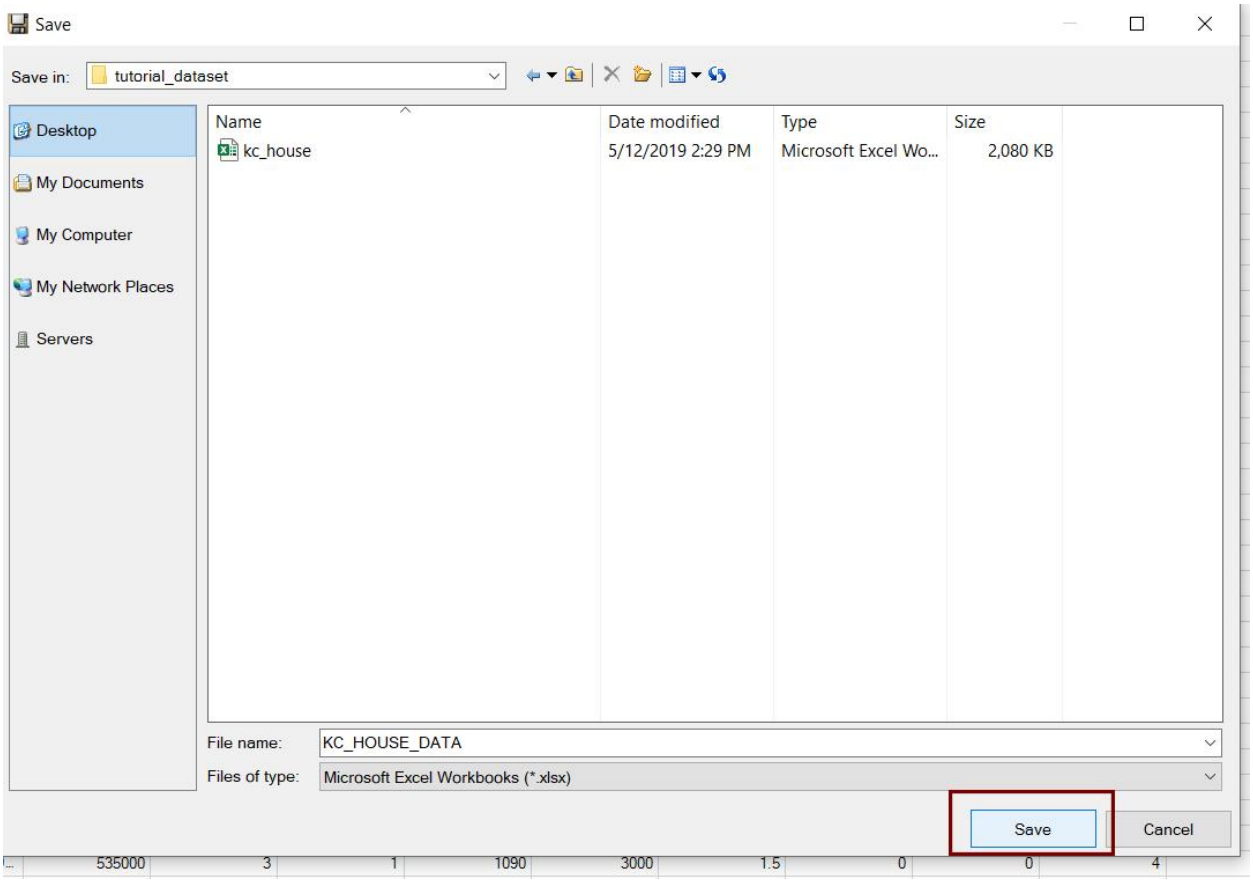
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -977754 | 19170 | -51.00 | <.0001 |
| bedrooms | 1 | -39352 | 2025.79104 | -19.43 | <.0001 |
| bathrooms | 1 | 45967 | 3490.89202 | 13.17 | <.0001 |
| sqft_living | 1 | 160.72433 | 3.89950 | 41.22 | <.0001 |
| sqft_lot | 1 | -0.00220 | 0.05126 | -0.04 | 0.9658 |
| floors | 1 | 27060 | 3782.48385 | 7.15 | <.0001 |
| waterfront | 1 | 579205 | 18629 | 31.09 | <.0001 |
| view | 1 | 43145 | 2272.32685 | 18.99 | <.0001 |
| condition | 1 | 19529 | 2494.77779 | 7.83 | <.0001 |
| grade | 1 | 119815 | 2248.31482 | 53.29 | <.0001 |
| sqft_basement | 1 | 6.20441 | 4.54242 | 1.37 | 0.1720 |
| sqft_living15 | 1 | 24.91674 | 3.59890 | 6.92 | <.0001 |
| sqft_lot15 | 1 | -0.54948 | 0.07833 | -7.02 | <.0001 |
| years_built | 1 | 3582.38887 | 70.94690 | 50.49 | <.0001 |
| years_renov | 1 | -10.02816 | 3.91127 | -2.56 | 0.0104 |

## ***EXPORT DATA***

**EXPORT → EXPORT (FILENAME) → CHOOSE A LOCATION → CHOOSE A FILE FORMAT (FILES OF TYPE) → SAVE**



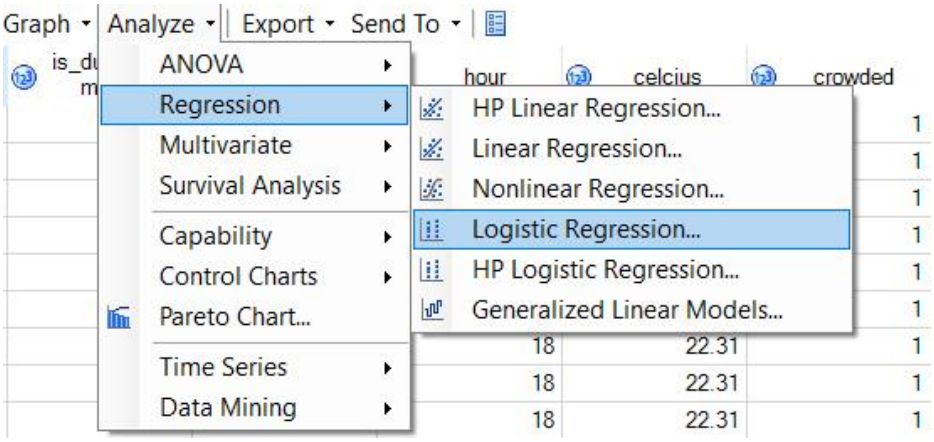




Save
Save in:
tutorial_dataset
Desktop
My Documents
My Computer
My Network Places
Servers
Name
Date modified
Type
Size
kc_house
5/12/2019 2:29 PM
Microsoft Excel Wo...
2,080 KB
File name:
KC_HOUSE_DATA
Files of type:
Microsoft Excel Workbooks (*.xlsx)
Save
Cancel
535000 3 1 1090 3000 1.5 0 0 4

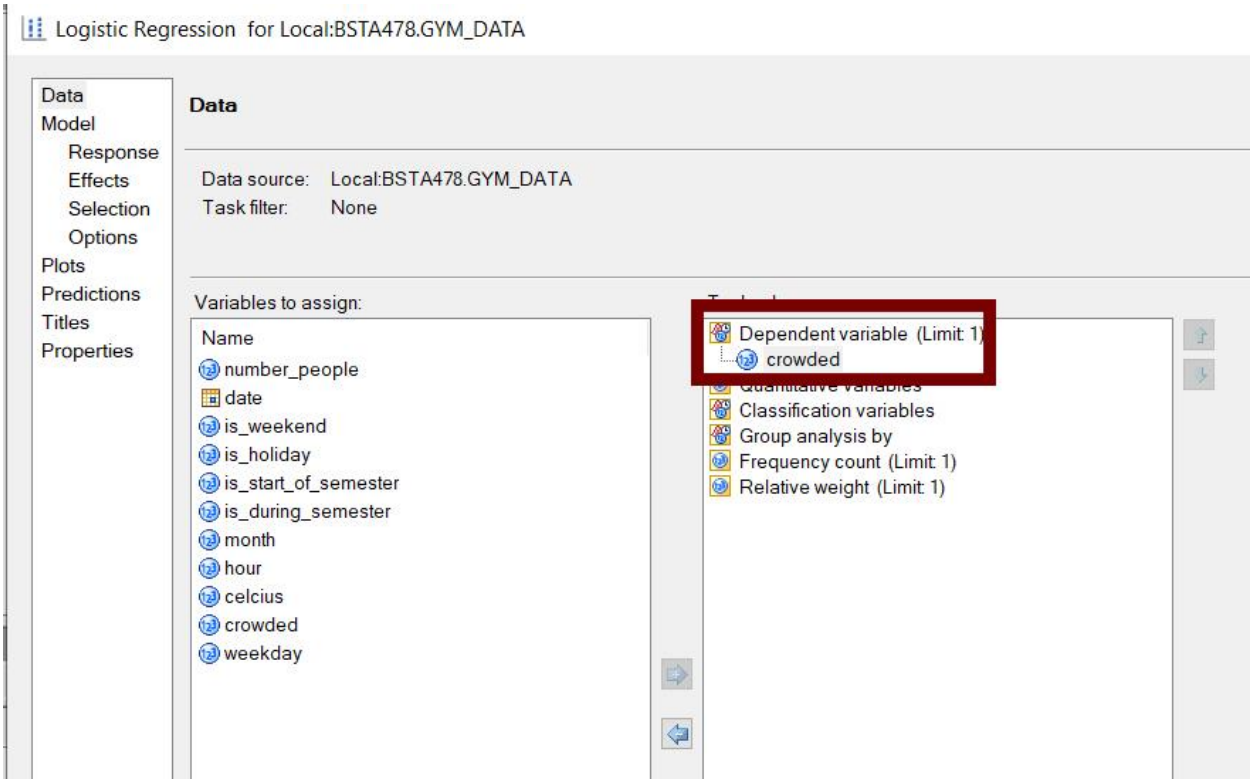
## ***LOGISTIC REGRESSION***

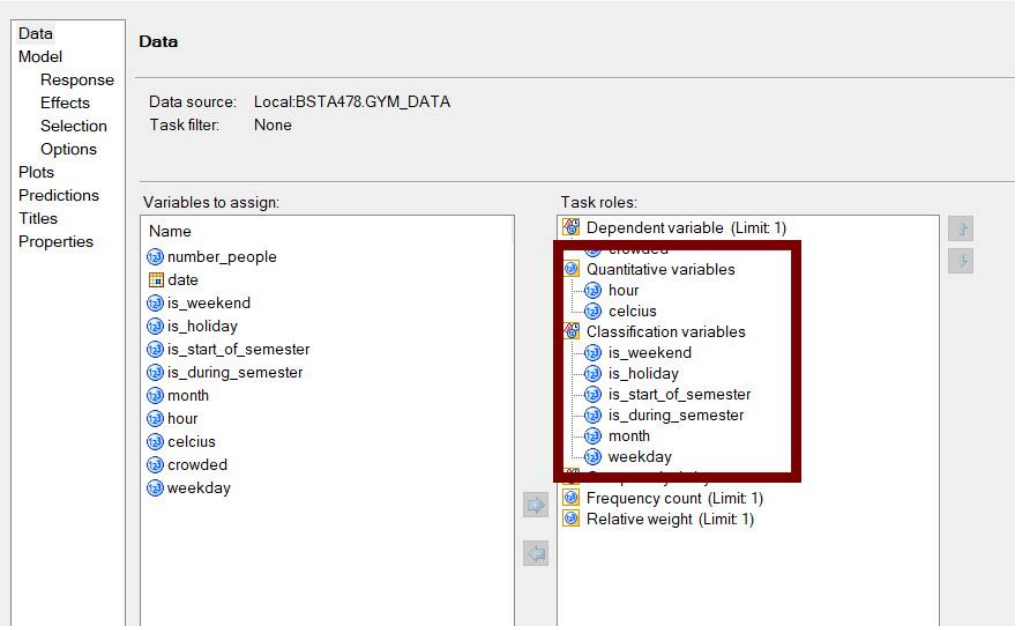
**The dataset used for this project has been taken from:**
https://www.kaggle.com/nsrose7224/crowdedness-at-the-campus-gym/downloads/crowdedness-at-the-campus-gym.zip/2

**ANALYZE → REGRESSION → LOGISTIC REGRESSION**

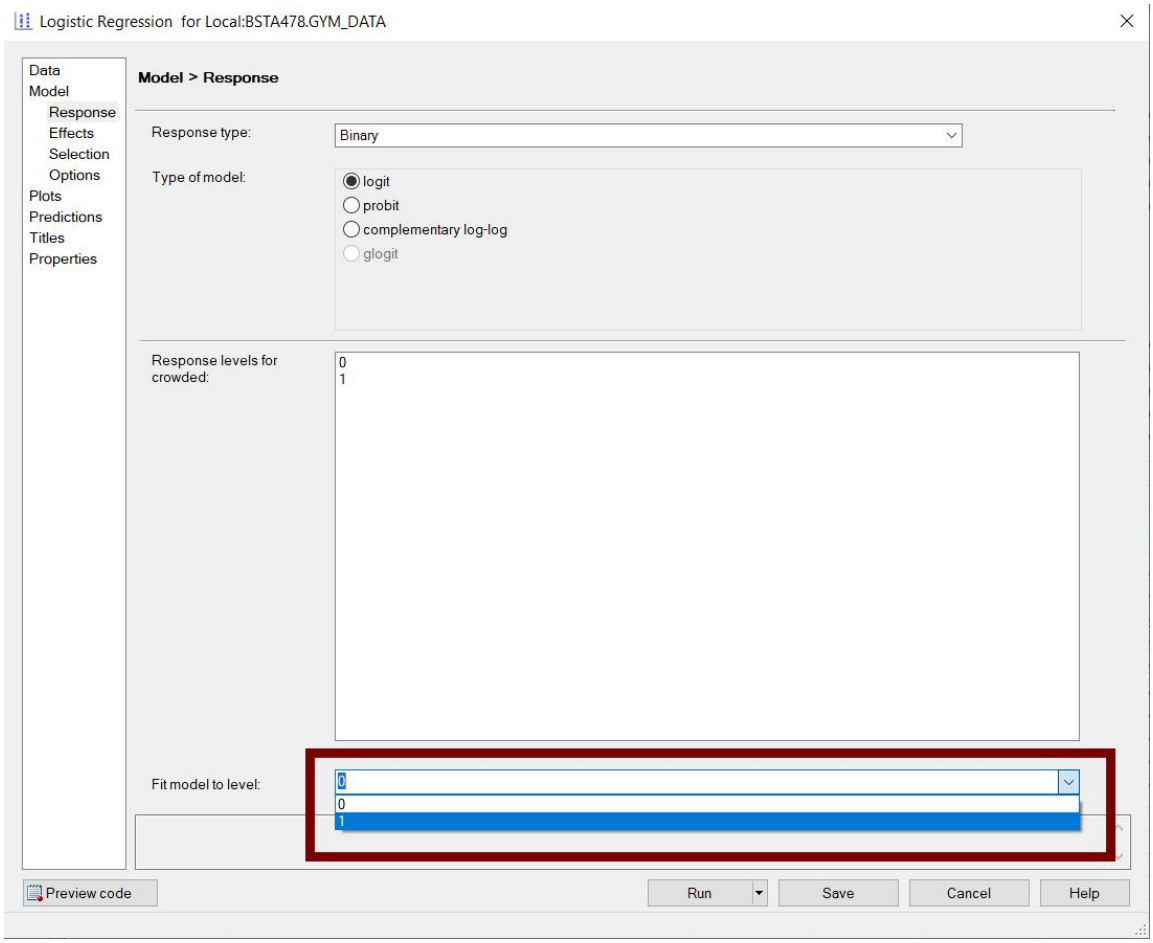



**DATA → SET A BINARY VARIABLE AS YOUR DEPENDENT VARIABLE (CROWDED IN THIS CASE) → ASSIGN QUANTITATIVE AND CLASSIFICATION VARIABLES**

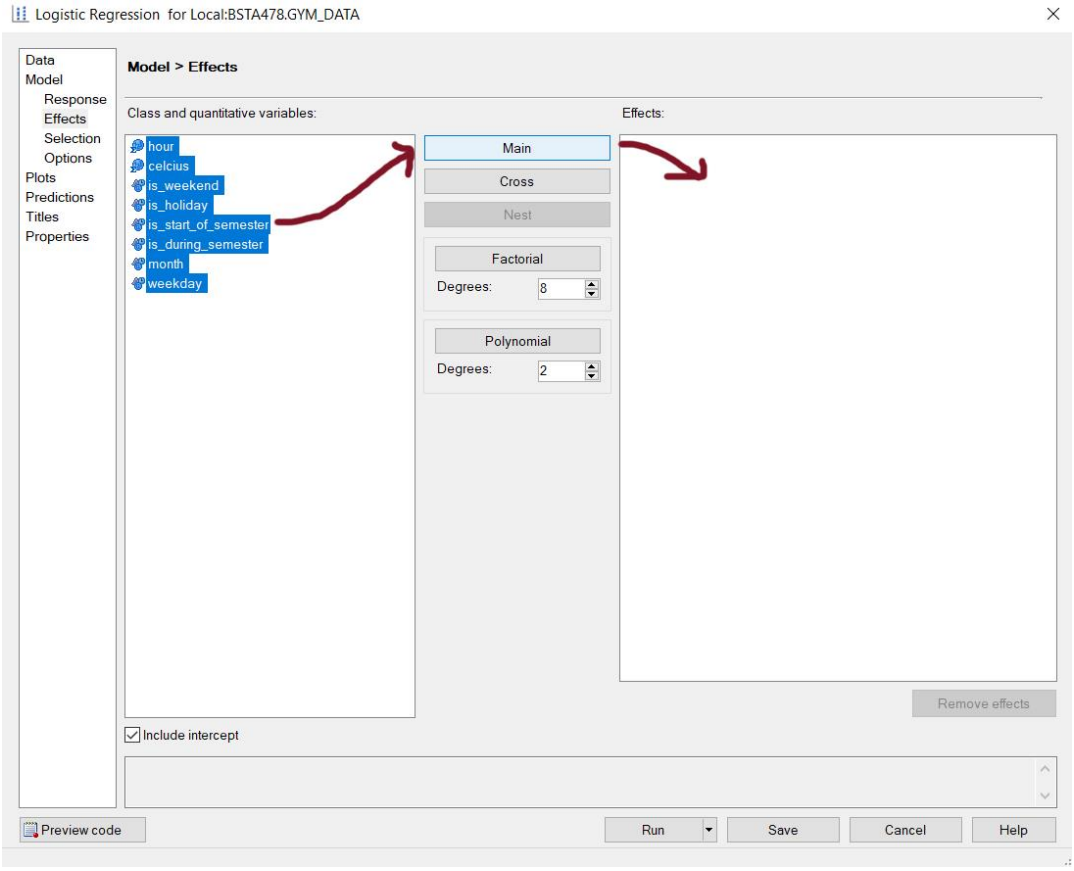




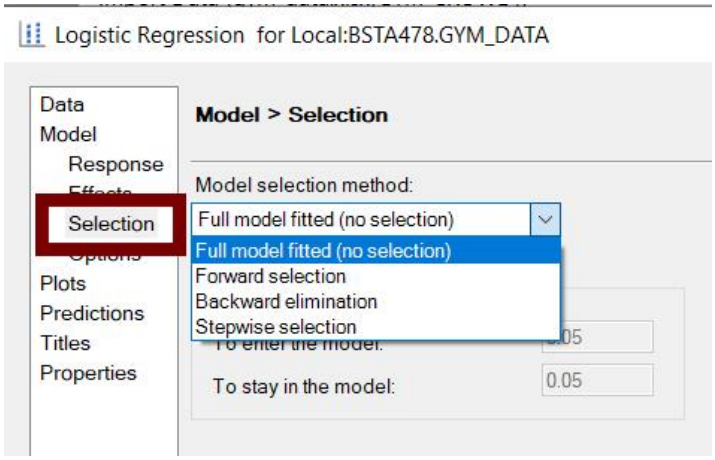

**RESPONSE → FIT MODEL TO LEVEL → 1**



## EFFECTS → SET THE VARIABLES TO MAIN EFFECTS




## SELECTION → MODEL SELECTION METHOD




You can also continue adjusting other settings. After you have finished click on **RUN**.

## Questions:

### Multiple linear regression

1. Import the data file called kc_house.xlsx
2. Explore the data and give answers to the following questions:
    a. What distribution do the house prices have? Does the price variable need any transformation?
    b. What is the maximum number of bedrooms?
    c. What is the minimum size of a lot?
    d. Build a frequency table for bedrooms, bathrooms, waterfront.
    e. Build a two-way table of waterfront by condition and waterfront by grade.
    f. How many variables and how many observations does the dataset have?
    g. Are there any variables that have missing observations? What should you do with the missing values?
    h. Which variables do you think should be excluded from the analysis since they will not contribute much to the explanation of the price of a house?
    i. Which variables do you need to predict a size of a lot?
    j. Should you include ZIP code into your model?
3. A) Build a multiple linear regression predicting the price of a house:
    a. Full model
    b. Forward selection
    c. Backward selection
    d. Stepwise

    B) Compare the models. Which one is the best out of the four? Why?

### Logistic regression

1. Import the datafile called gym_data.xlsx
2. Explore the data and answer the questions:
    a. What distribution does the number of visitors have?
    b. What is the minimum and maximum number of gym visitors?
    c. How many variables and how many observations does the dataset have?
    d. Are there any variables that have missing observations? What should you do with missing values?
    e. Which variables do you think should be excluded from the analysis since they will not contribute much to the explanation of the number of visitors to the gym?
    f. Are there any redundant variables?
3. A) Run a logistic regression for event = 1:
    a. Full
    b. Stepwise
    c. Forward
    d. Backward

    B) Analyze the odds likelihood ratio as well as the ROC curve. Which model is the best?