

Housing Prices  
(Regression)

The goal of this study is to train a model in order to predict housing prices. The dataset used in this case study is found in <https://www.kaggle.com/datasets/yasserh/housing-prices-dataset/data> and has 13 features and 545 samples. This dataset contains information on certain factors like house area, bedrooms, furnished, nearness to main road, etc, aiming to predict housing prices in the Northeast states of USA.

The dataset contains no missing values and includes several categorical features. Categorical features contain multiple levels, and the data was transformed to corresponding numeric codes, as detailed below:

mainroad:

* No (0)
* Yes (1)

guestroom:

* No (0)
* Yes (1)

basement:

* No (0)
* Yes (1)

hotwaterheating:

* No (0)
* Yes (1)

airconditioning:

* No (0)
* Yes (1)

prefarea:

* No (0)
* Yes (1)

furnishingstatus:

* unfurnished (0)
* semi - furnished (1)
* furnished (2)
* Southwest (3)

# Step 1: Import data from file

A screenshot of a computer

Description automatically generatedA screenshot of a spreadsheet

Description automatically generatedRight click on the input spreadsheet and choose the option “Import from file”. Then navigate through your files to load the one with the housing price data.

# Step 2: Manipulate data

In order to use the data for training we have to exclude any columns that do not contain features. In our dataset there are no such columns. Therefore, we will include all columns in the training. We follow these steps to execute this:

* On the menu click on “Data Transformation” → “Data Manipulation” → “Select Column(s)”
* Select all columns.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The data will appear in the output spreadsheet.

# Step 3: Split data

Create a new tab by pressing the “+” button on the bottom of the page with the name “TRAIN\_TEST\_SPLIT” which we will use for splitting to create the train and test set.

A screenshot of a computer

Description automatically generatedImport data into the input spreadsheet of the “TRAIN\_TEST\_SPLIT” tab from the output of the “IMPORT” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

Split the dataset by choosing

“Data Transformation” → “Split” → “Random Partitioning”

A screenshot of a phone number

Description automatically generatedA screenshot of a computer

Description automatically generatedThen choose the “Training set percentage” and the column for the sampling as shown below:

A screenshot of a computer

Description automatically generatedThe results will appear on the output spreadsheet.

# Step 4: Normalize the training set

Create a new tab by pressing the “+” button on the bottom of the page with the name “NORMALIZE\_TRAIN\_SET”.

Import data into the input spreadsheet of the “NORMALIZE\_TRAIN\_SET” tab the train set from the output of the “TRAIN\_TEST\_SPLIT” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”. From the available Select input tab options choose “TRAIN\_TEST\_SPLIT: Training Set”

A screenshot of a computer

Description automatically generated

Normalize the data using Z-score:

“Data Transformation” → “Normalizers” → “Z-Score”

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedThen select all columns except “price” and click “Execute”.

The results will appear on the output spreadsheet.

# A screenshot of a computer Description automatically generatedStep 5: Normalize the test set

Create a new tab by pressing the “+” button on the bottom of the page with the name “NORMALIZE\_TEST\_SET”.

A screenshot of a computer

Description automatically generatedImport data into the input spreadsheet of the “NORMALIZE\_TEST\_SET” tab the test set from the output of the “TRAIN\_TEST\_SPLIT” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”. From the available Select input tab options choose “TRAIN\_TEST\_SPLIT: Test Set”.

Normalize the test set using the existing normalizer of the training set:

“Analytics” → “Existing Model Utilization” → “Model (from Tab: ) NORMALIZE\_TRAIN\_SET”.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThe results will appear on the output spreadsheet.

# Step 6: Feature selection

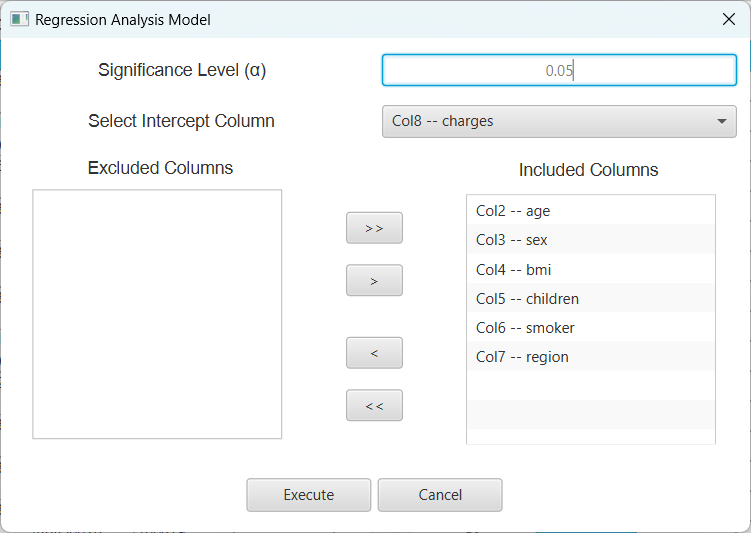
Create a new tab by pressing the “+” button on the bottom of the page with the name “FEATURE\_SELECTION\_REGRESSION”.

Import data into the input spreadsheet of the “FEATURE\_SELECTION\_REGRESSION” tab from the output of the “NORMALIZE\_TRAIN\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

A screenshot of a computer

Description automatically generatedChoose the most important features using the Regression Analysis:

“Data Transformation” → “Variable Selection” → “Regression Analysis”.

A screenshot of a computer

Description automatically generatedThen choose the “price” column as the intercept column, the Significance level (α) as 0.05 and include all columns.

The results will appear on the output spreadsheet.

A screenshot of a computer

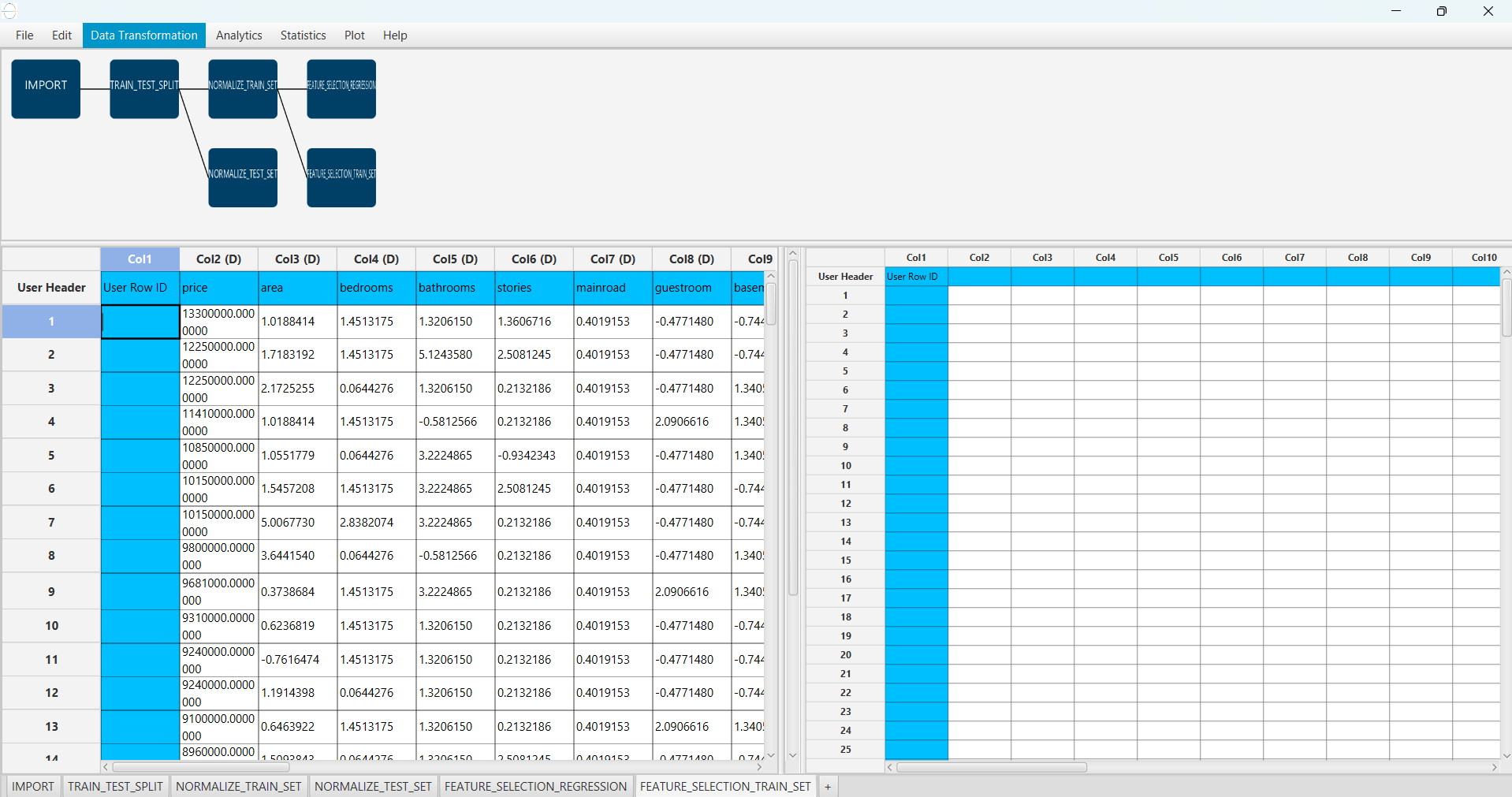
Description automatically generatedThe significant features according to the p-value are the following:

* price (p-value = 0.0)
* area (p-value = 0.0)
* bathrooms (p-value = 0.0)
* stories (p-value = 0.0)
* mainroad (p-value = 0.011078305445895668)
* guestroom (p-value = 0.010553358931198279)
* basement (p-value = 0. 01994612978933753)
* hotwaterheating (p-value = 4.1889240314438027E-4)
* airconditioning (p-value = 0.0)
* parking (p-value = 7.015384208333902E-5)
* prefarea (p-value = 1.0032449728224924E-6)
* furnishingstatus (p-value = 0.010503202918605318)

# Step 7: Feature selection: train set

Create a new tab by pressing the “+” button on the bottom of the page with the name “FEATURE\_SELECTION\_TRAIN\_SET”.

Import data into the input spreadsheet of the “FEATURE\_SELECTION\_TRAIN\_SET” tab from the output of the “NORMALIZE\_TRAIN\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.



Manipulate the data by choosing the columns that correspond to the significant features (from the previous step):

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated “Data Transformation” → “Data Manipulation” → “Select Column(s)”.

The results will appear on the output spreadsheet.

A screenshot of a computer

Description automatically generated

# Step 8: Feature selection: test set

Create a new tab by pressing the “+” button on the bottom of the page with the name “FEATURE\_SELECTION\_TEST\_SET”.

A screenshot of a computer

Description automatically generatedImport data into the input spreadsheet of the “FEATURE\_SELECTION\_TEST\_SET” tab from the output of the “NORMALIZE\_TEST\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

Manipulate the data by choosing the columns that correspond to the significant features (from the step 6):

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated“Data Transformation” → “Data Manipulation” → “Select Column(s)”.

The results will appear on the output spreadsheet. A screenshot of a computer

Description automatically generated

# Step 9: Train the model

Create a new tab by pressing the “+” button on the bottom of the page with the name “TRAIN\_MODEL(.fit)”.  
Import data into the input spreadsheet of the “TRAIN\_MODEL(.fit)” tab from the output of the “FEATURE\_SELECTION\_TRAIN\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

A screenshot of a computer

Description automatically generated

Use the k Nearest Neighbors (kNN) method to train and fit the model:

“Analytics” → “Regression” → “k Nearest Neighbors (kNN)”

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedand set the “Target Column” as the column corresponding to “price” and the “Number of Neighbors” to 5.

The predictions will appear on the output spreadsheet.

A screenshot of a computer

Description automatically generated

# Step 10: Validate the model

Create a new tab by pressing the “+” button on the bottom of the page with the name “VALIDATE\_MODEL(.predict)”.

A screenshot of a computer

Description automatically generatedImport data into the input spreadsheet of the “VALIDATE\_MODEL(.predict)” tab from the output of the “FEATURE\_SELECTION\_TEST\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

To validate the model:

“Analytics” → “Existing Model Utilization”. Then choose Model “(from Tab:) TRAIN\_MODEL (.fit)”.

and transfer the “price” column in the output.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThe predictions will appear on the output spreadsheet.

# Step 11: Statistics calculation

Create a new tab by pressing the “+” button on the bottom of the page with the name “STATISTICS\_ACCURACIES”.

Import data into the input spreadsheet of the “STATISTICS\_ACCURACIES” tab from the output of the “VALIDATE\_MODEL(.predict)” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

A screenshot of a computer

Description automatically generated

Calculate the statistical metrics for the regression:

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated“Statistics” → “Model Metrics” → “Regression Metrics”.

The results will appear on the output spreadsheet.

A screenshot of a computer

Description automatically generated

# Step 12: Reliability check of each record of the test set

## Step 12.a: Create the domain

Create a new tab by pressing the “+” button on the bottom of the page with the name “EXCLUDE\_PRICE”.

A screenshot of a computer

Description automatically generatedImport data into the input spreadsheet of the “EXCLUDE\_PRICE” tab from the output of the “FEATURE\_SELECTION\_TRAIN\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

Manipulate the data to exclude the column that corresponds to the “price”

“Data Transformation” → “Data Manipulation” → “Select Columns”

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedThen select all the columns except the “price”.

The results will appear on the output spreadsheet.

Create a new tab by pressing the “+” button on the bottom of the page with the name “DOMAIN”.

A screenshot of a computer

Description automatically generatedImport data into the input spreadsheet of the “DOMAIN” tab from the output of the “EXCLUDE\_PRICE” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

Create the domain:

“Statistics” → “Domain APD”

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThe results will appear on the output spreadsheet.

## Step 12.b: Check the test set reliability

Create a new tab by pressing the “+” button on the bottom of the page with the name “EXCLUDE\_PRICE\_TEST\_SET”.

Import data into the input spreadsheet of the “EXCLUDE\_PRICE\_TEST\_SET” tab from the output of the “FEATURE\_SELECTION\_TEST\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

A screenshot of a computer

Description automatically generated

Filter the data to exclude the column that corresponds to the “price”

“Data Transformation” → “Data Manipulation” → “Select Columns”.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedThen select all the columns except “price”.

The results will appear on the output spreadsheet.

Create a new tab by pressing the “+” button on the bottom of the page with the name “RELIABILITY”.

Import data into the input spreadsheet of the “RELIABILITY” tab from the output of the “EXCLUDE\_PRICE\_TEST\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

Check the Reliability:

“Analytics” → “Existing Model Utilization”.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedThen select as Model “(from Tab:) DOMAIN”.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThe results will appear on the output spreadsheet.

There are no unreliable samples in the test set.

# Final Isalos Workflow

Following the above-described steps, the final workflow on Isalos will look like this:

A diagram of a train

Description automatically generated