

Database Management Systems

Lecture 6

Azure Machine Learning*

Azure Stream Analytics*

* not among the exam topics

Azure Machine Learning

Car Price Prediction

* preparing the data

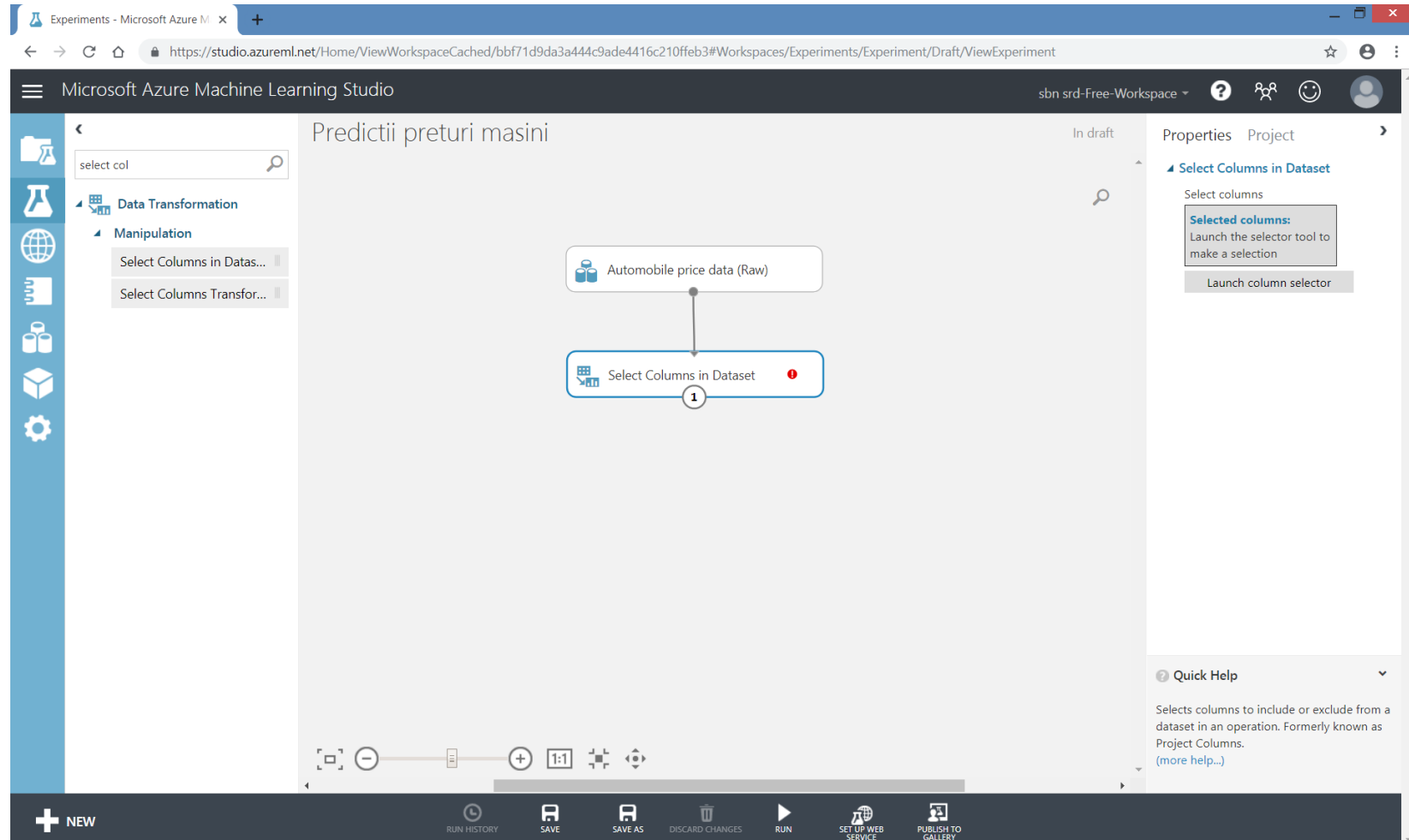
- eliminate column with missing values - *Select Columns in Dataset* module

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The top navigation bar shows the workspace name 'sbn srd-Free-Workspace'. The left sidebar contains a search bar and a list of modules under 'Data Transformation' and 'Manipulation'. The main workspace area is titled 'Predictii preturi masini' and shows a workflow with two modules: 'Automobile price data (Raw)' and 'Select Columns in Dataset'. The 'Select Columns in Dataset' module is highlighted with a blue border and a red error icon, indicating it is the current focus. The right sidebar shows the 'Properties' tab for the 'Select Columns in Dataset' module, with a 'Launch column selector' button. The bottom status bar includes icons for 'NEW', 'RUN HISTORY', 'SAVE', 'SAVE AS', 'DISCARD CHANGES', 'RUN', 'SET UP WEB SERVICE', and 'PUBLISH TO GALLERY'.

Car Price Prediction

* preparing the data

- eliminate column with missing values



Car Price Prediction

* preparing the data

- eliminate column with missing values

- *Select Columns in Dataset*

- *Launch column selector*

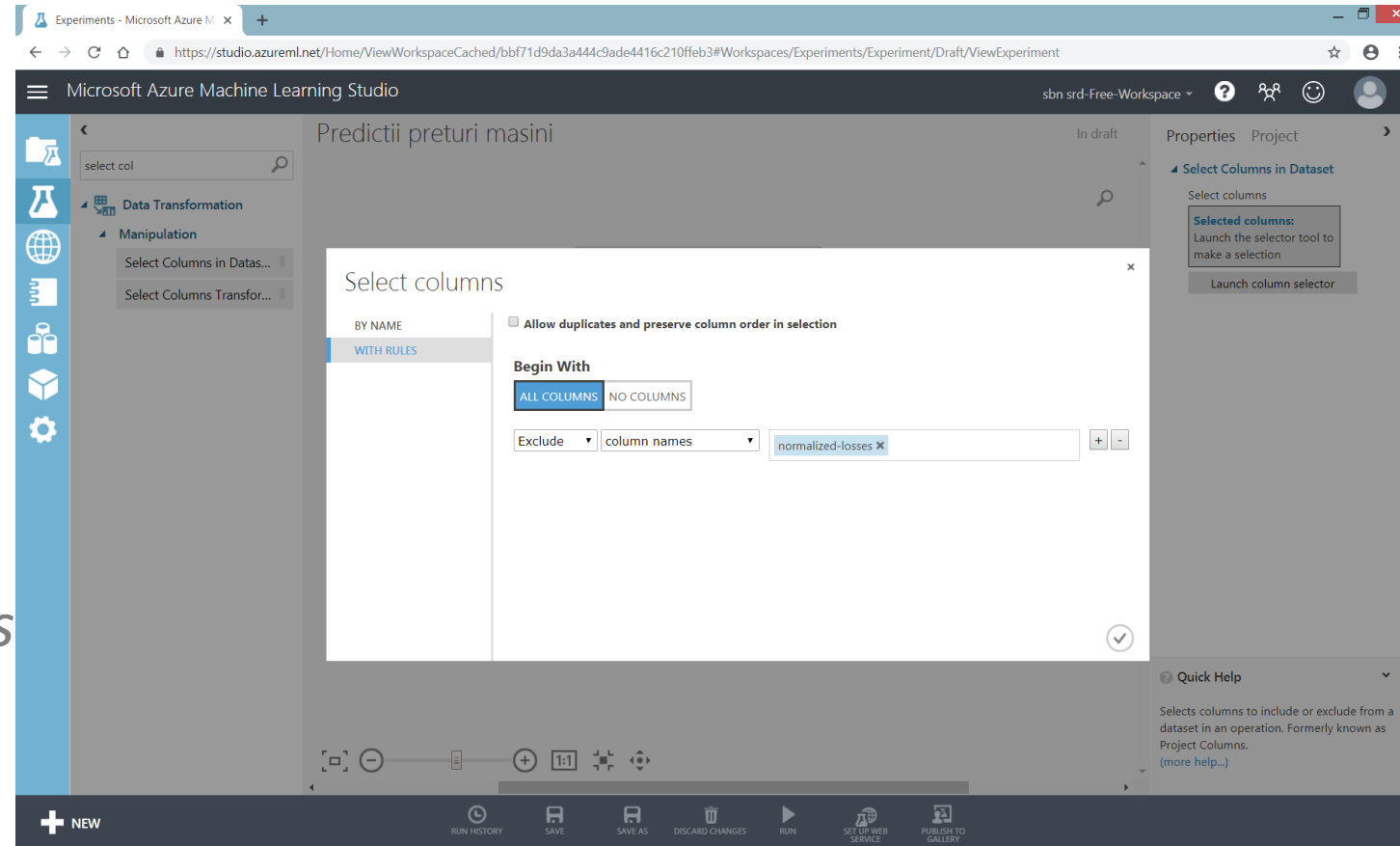
- *With Rules*

- *Begin With*

- *All Columns*

- *Exclude*

- *normalized-losses*



Car Price Prediction

* preparing the data

- eliminate rows with missing values – *Clean Missing Data* module

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The browser address bar shows the URL: <https://studio.azureml.net/Home/ViewWorkspaceCached/bbf71d9da3a444c9ade4416c210ffeb3#Workspaces/Experiments/Experiment/Draft/ViewExperiment>. The workspace is titled "Predictii preturi masini" and is in "In draft" status. The left sidebar shows the "Data Transformation" section expanded, with the "Clean Missing Data" module selected under the "Manipulation" category. The main canvas shows a data pipeline with two modules: "Automobile price data (Raw)" and "Select Columns in Dataset" (labeled with a circled '1'). The "Select Columns in Dataset" module is currently selected, and its properties are shown on the right. The "Properties" pane for "Select Columns in Dataset" includes a "Select columns" section with "Selected columns: All columns" and "Exclude column names: normalized-losses". A "Launch column selector" button is also visible. The bottom of the interface features a toolbar with icons for "NEW", "RUN HISTORY", "SAVE", "SAVE AS", "DISCARD CHANGES", "RUN", "SET UP WEB SERVICE", and "PUBLISH TO GALLERY".

Car Price Prediction

* preparing the data

- eliminate rows with missing values

- *Clean Missing Data*

- *Cleaning mode*

- *Remove entire row*

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The main workspace shows a workflow titled "Predictii preturi masini" in draft mode. The workflow consists of three steps: "Automobile price data (Raw)", "Select Columns in Dataset", and "Clean Missing Data". The "Clean Missing Data" step is highlighted with a blue border and numbered 1 and 2. On the left, a sidebar shows the "Data Transformation" section with "Clean Missing Data" selected under "Manipulation". On the right, the "Properties" pane for "Clean Missing Data" is visible, showing "Selected columns: All columns", "Minimum missing value range: 0", "Maximum missing value range: 1", and "Cleaning mode: Remove entire row". A "Quick Help" section at the bottom right explains that this mode specifies how to handle values missing from a dataset.

Car Price Prediction

* running the experiment

- *Run*

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The main workspace shows a workflow titled "Predictii preturi masini" (Car Price Prediction) with three steps: "Automobile price data (Raw)", "Select Columns in Dataset", and "Clean Missing Data". The workflow is marked as "Finished running" with a green checkmark. The right sidebar contains the "Properties" and "Project" tabs, with "Experiment Properties" showing "START TIME", "END TIME", "STATUS CODE" (Finished), and "STATUS DETAILS" (None). The "Summary" and "Description" sections are also visible. The bottom toolbar includes icons for "NEW", "RUN HISTORY", "SAVE", "SAVE AS", "DISCARD CHANGES", "RUN", "SET UP WEB SERVICE", and "PUBLISH TO GALLERY".

Microsoft Azure Machine Learning Studio

Predictii preturi masini

Finished running ✓

Properties Project

Experiment Properties

START TIME 4/25/20...

END TIME 4/25/20...

STATUS CODE Finished

STATUS DETAILS None

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

NEW RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

Car Price Prediction

* displaying the data

- *Clean Missing Data* module -> left output port -> *Visualize*

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The main workspace shows a workflow titled "Predictii preturi masini" with three modules: "Automobile price data (Raw)", "Select Columns in Dataset", and "Clean Missing Data". The "Clean Missing Data" module is selected, and a context menu is open, highlighting the "Visualize" option. The right sidebar shows the "Properties" tab with experiment details: START TIME 4/25/20..., END TIME 4/25/20..., STATUS CODE Finished, and STATUS DETAILS None. The bottom toolbar includes icons for NEW, RUN HISTORY, SAVE, SAVE AS, DISCARD CHANGES, RUN, SET UP WEB SERVICE, and PUBLISH TO GALLERY.

Microsoft Azure Machine Learning Studio

Predictii preturi masini

Finished running ✓

Properties Project

Experiment Properties

START TIME 4/25/20...

END TIME 4/25/20...

STATUS CODE Finished

STATUS DETAILS None

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

Automobile price data (Raw)

Select Columns in Dataset ✓

Clean Missing Data ✓

Download

Save as Dataset

Save as Trained Model

Save as Transform

Visualize

Generate Data Access Code...

Open in a new Notebook

NEW

RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

RUN

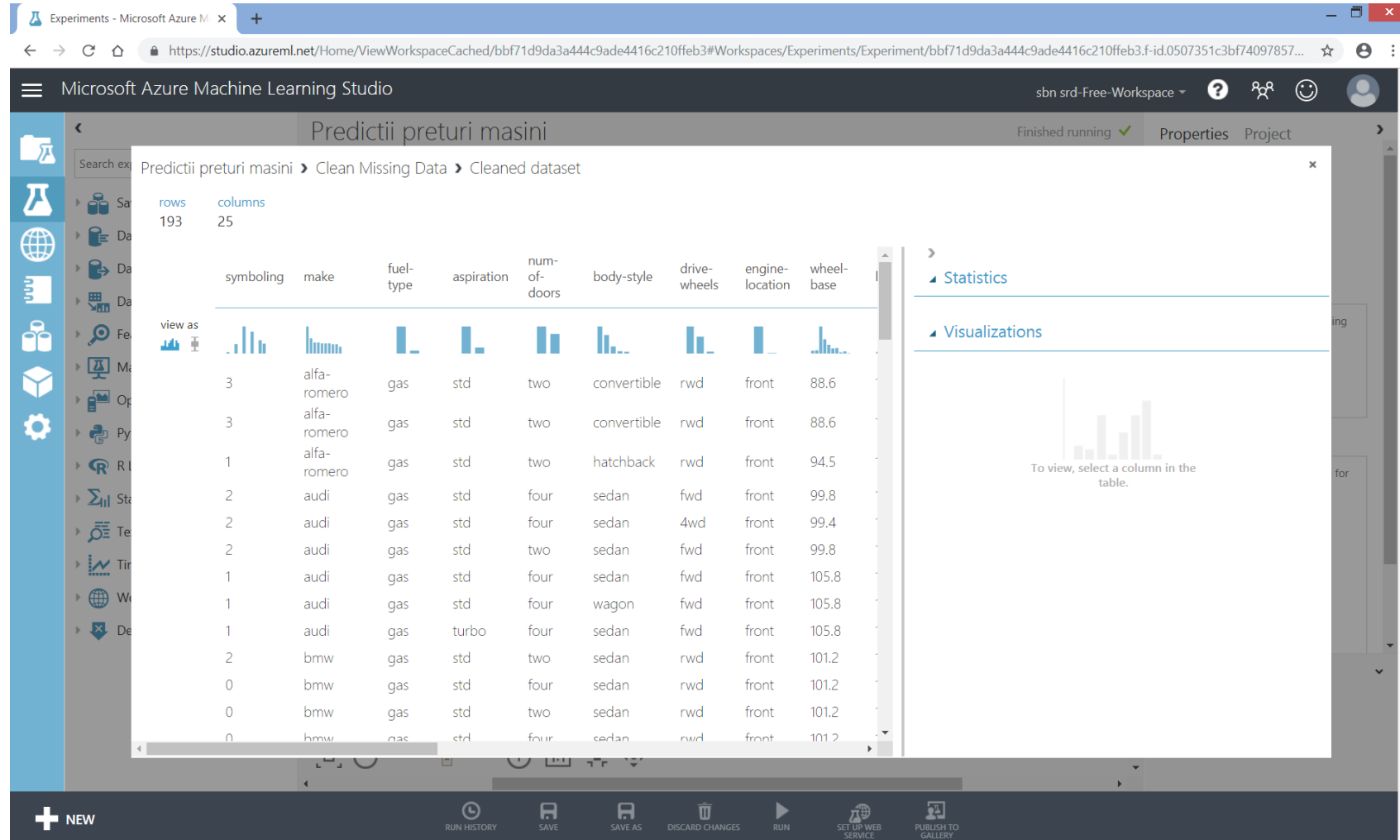
SET UP WEB SERVICE

PUBLISH TO GALLERY

Car Price Prediction

* displaying the data

- *Clean Missing Data* module -> left output port -> *Visualize*



Car Price Prediction

* defining the *features*

- used to create the predictive model
- *Select Columns in Dataset* module

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The main workspace shows a pipeline titled "Predictii preturi masini" (Car Price Predictions) in draft status. The pipeline consists of four modules connected sequentially:

- Automobile price data (Raw)**: The starting data source.
- Select Columns in Dataset**: A module with a green checkmark, indicating it has been successfully configured.
- Clean Missing Data**: A module with a green checkmark, indicating it has been successfully configured.
- Select Columns in Dataset**: A second instance of the module, currently in an error state (red exclamation mark) and labeled with a circled "1".

The left sidebar shows the "Data Transformation" section expanded, with the "Manipulation" sub-section selected. The "Select Columns in Dataset" module is highlighted in the list.

The right sidebar shows the "Properties" pane for the selected "Select Columns in Dataset" module. It includes a "Launch column selector" button and a "Quick Help" section explaining the module's function: "Selects columns to include or exclude from a dataset in an operation. Formerly known as Project Columns. (more help...)".

Car Price Prediction

* defining the *features*

- *Select Columns in Dataset*

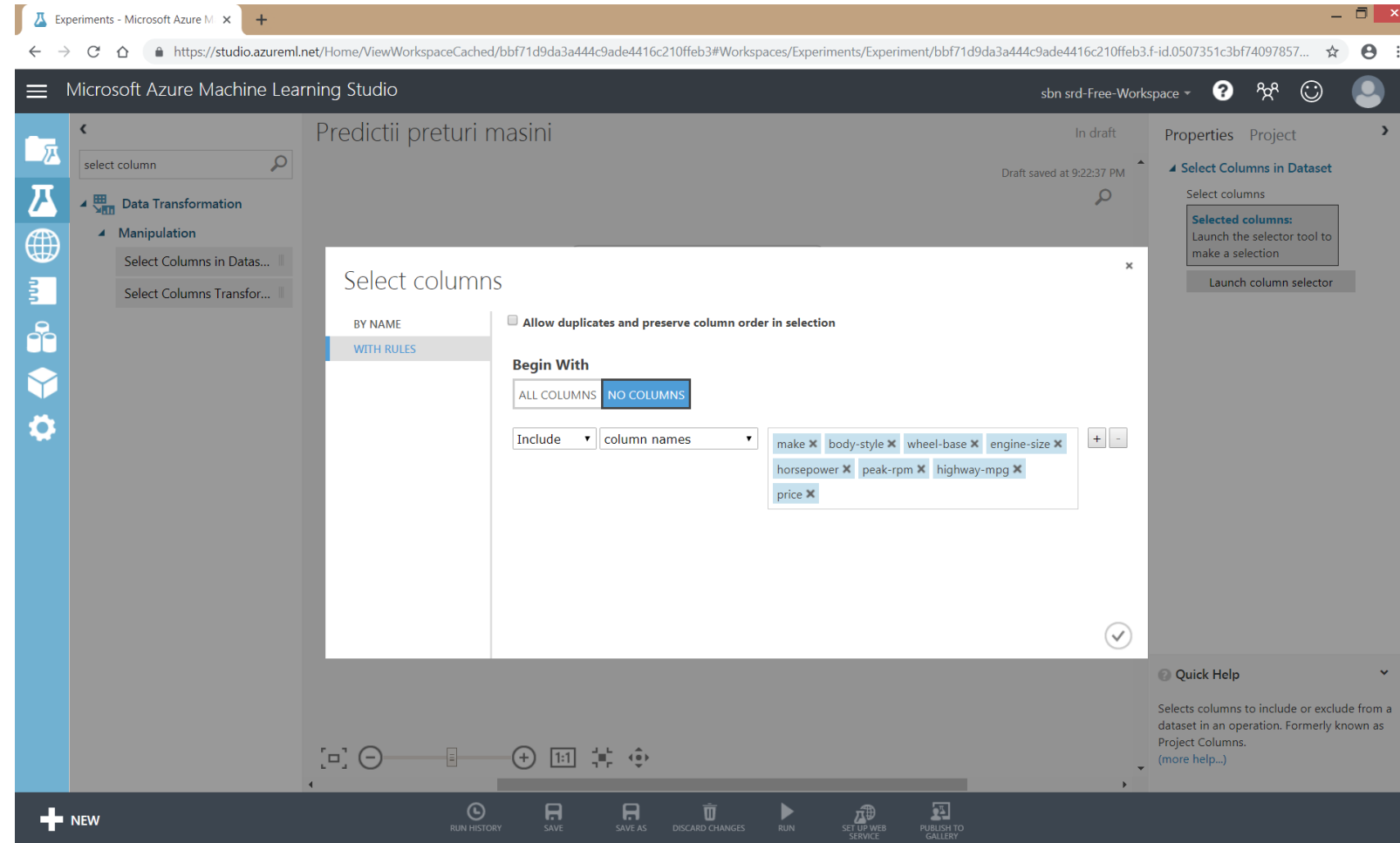
- *Launch column selector*

- *Begin With*

- *No columns*

- *Include*

- *make, body-style, wheel-base, engine-size, horsepower, peak-rpm, highway-mpg, price*



- goal: predict car price from selected features

Car Price Prediction

* choosing / applying the algorithm

- create the training dataset and the test dataset
- training dataset
 - dataset that includes the car price
 - the model is trained on this dataset
 - it searches for correlations between a car's features and its price
- test dataset
 - dataset that includes the car price
 - the model is tested on this dataset
 - the price estimated by the model for each car is compared with the real price

Car Price Prediction

* choosing / applying the algorithm

- create the training / test datasets - *Split Data* module

- *Split Data*

- *Fraction of rows in the first output dataset*
 - 0.75
 - i.e., training dataset - 75% of the data

- *Run experiment*

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The main workspace shows a workflow titled "Predictii preturi masini" (Car Price Predictions). The workflow consists of the following modules in sequence:

- Automobile price data (Raw)
- Select Columns in Dataset (with a green checkmark)
- Clean Missing Data (with a green checkmark)
- Select Columns in Dataset (with a green checkmark)
- Split Data (labeled with 1 and 2)

A "Mini Map" in the bottom left corner provides a visual overview of the entire workflow. The right-hand side of the interface shows the "Properties" pane for the "Split Data" module, which is currently in "Draft" mode. The properties include:

- Splitting mode: Split Rows
- Fraction of rows in...: 0.75
- Randomized split: ☒
- Random seed: 0
- Stratified split: False

At the bottom of the interface, there is a "Quick Help" section with the text: "Split the rows of a dataset into two distinct sets (more help...)". The bottom navigation bar includes icons for "NEW", "RUN HISTORY", "SAVE", "SAVE AS", "DISCARD CHANGES", "RUN", "SET UP WEB SERVICE", and "PUBLISH TO GALLERY".

Car Price Prediction

* choosing / applying the algorithm

- *Machine Learning -> Initialize Model -> Regression -> Linear Regression*

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The main workspace shows a workflow titled "Predictii preturi masini" (Car Price Predictions) in draft status. The workflow consists of the following steps:

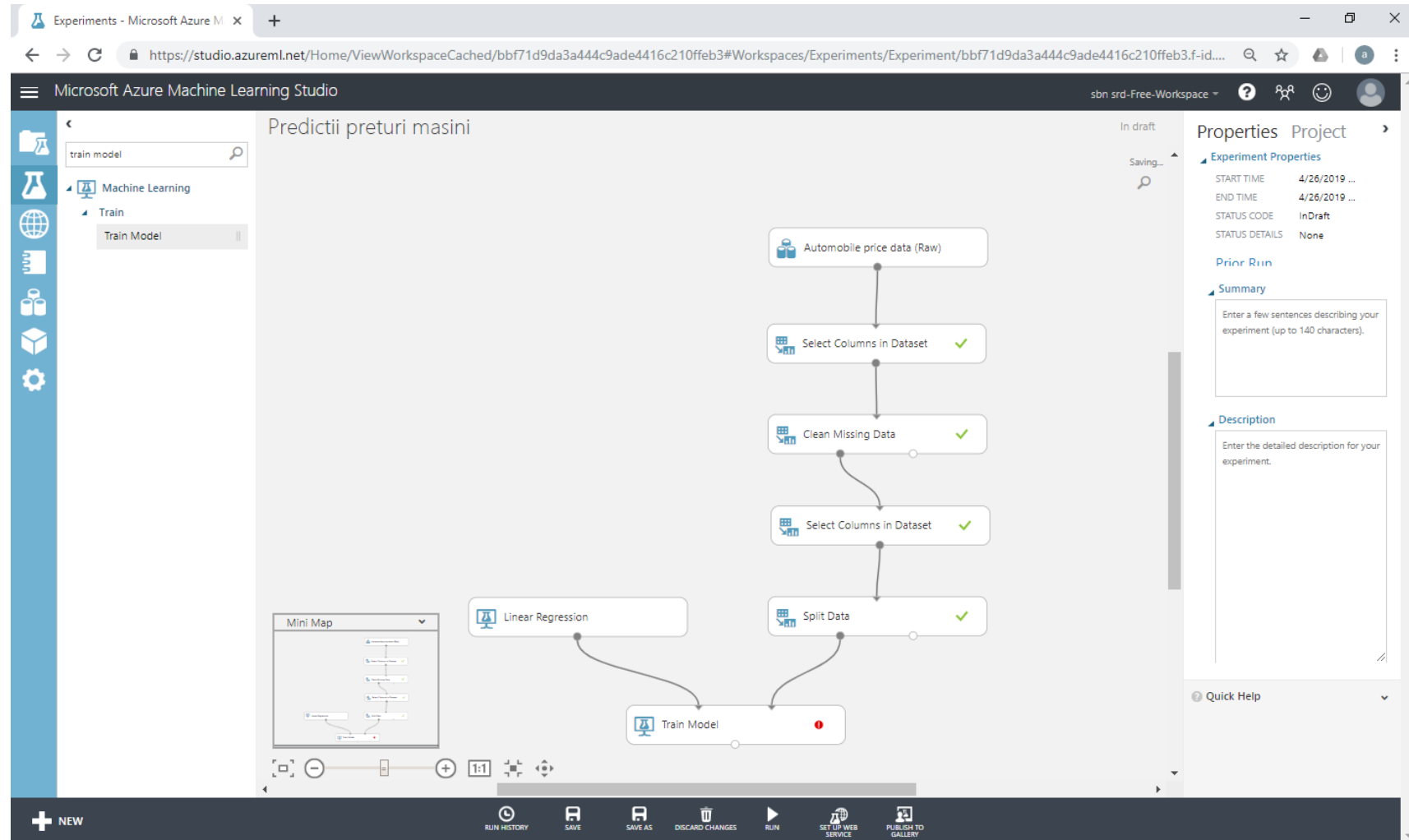
- Automobile price data (Raw)
- Select Columns in Dataset (checked)
- Clean Missing Data (checked)
- Select Columns in Dataset (checked)
- Split Data (checked)
- Linear Regression (highlighted with a blue box and a circled '1')

The left sidebar shows the "Machine Learning" category expanded, with "Initialize Model" > "Regression" > "Linear Regression" selected. The right sidebar shows the "Properties" panel for the "Linear Regression" model, with the "Solution method" set to "Ordinary Least Squares". Other properties include "L2 regularization w..." set to "0.001", "Include interce..." checked, "Random number s..." set to a default value, and "Allow unknown..." checked. The bottom status bar includes icons for "NEW", "RUN HISTORY", "SAVE", "SAVE AS", "DISCARD CHANGES", "RUN", "SET UP WEB SERVICE", and "PUBLISH TO GALLERY".

Car Price Prediction

* choosing / applying the algorithm

- *Train Model* module



Car Price Prediction

* choosing / applying the algorithm

- *Train Model*

- *Launch column selector*
 - move column *price* from *Available columns* to *Selected columns*

- *Run experiment*

