Lab Exercise #1 - Linear Regression

Instruction:
Read the instructions carefully and follow each step.

Provide your best answers for all questions.

* Indicates required question

Email *

Record my email address with my response

Name *

Example: Dela Cruz, Juan C.

Your answer

Student Number *

Example: 24B1234

Your answer

Section *

- ○ 4A
- ○ 4B
- ○ 4C
- ○ 4D
- ○ 4E
- ○ 4F

Building a Linear Regression Model

Imagine a health insurance company has asked you to develop a machine learning model to **predict medical costs that will be billed to them.**

They explain that key factors affecting these costs include **age, sex, BMI, number of children, and smoking status.**

As someone familiar with machine learning, you know that **Linear Regression** is a suitable method for predicting continuous values like costs.

Follow these steps to build the model and answer the questions along the way.

Step 1: Download Orange Data Mining

Orange is a user-friendly machine learning and data mining suite.

Its intuitive interface allows users to drag and drop widgets, connect them, load datasets, and gain insights easily.

This makes building machine learning models simple.

To download Orange Data Mining, visit this website and click the button below "Suggested Download". This will initiate a download compatible with your operating system. *

A desktop computer is required for this process.

- [ ] Done

Install Orange and just keep clicking "Next". *

Once the installation is complete, leave "Start Orange" selected and click "Finish" to open Orange.

- [ ] Done

When Orange opens, a startup dialog will appear. *

Click "New" to start a new workflow.

- [ ] Done

To save your workflow, go to File > Save or press Ctrl + S. *

Choose an accessible directory and name the file using this format: **[Last_Name]_[Section]_Lab_Exercise_1** (e.g., Rey_4A_Lab_Exercise_1).

- [ ] Done

Take a screenshot of the window displaying your file name and upload it here. *

Example:

Add file

Step 2: Collect the Data

Suppose the insurance company provides you with past records of medical costs they've paid to hospitals.

These records also include information about each customer, such as their age, sex, BMI, number of children, and smoking status.

You'll use this data to build a linear regression model.

This model will help you find patterns in the data and predict future medical costs based on the customer's details.

Download the dataset here and save it in an accessible directory. *

- ◯ Done

Drag and drop the File widget into your workflow. *

Check the File widget.

In the Source section, select File and click the folder icon to locate the downloaded dataset on your computer.

In the Columns section, set the role of "charges" column to target (this will be used as your label—the value you want the model to predict).

Keep the other settings unchanged and click Apply.

- ◯ Done

To view the datasheet in a spreadsheet, connect the File widget to a Data Table widget. *

Rename the Data Table widget to "Insurance Data Table".

- ◯ Done

Check the Data Table widget. *

Check the Info section in the Insurance Data Table widget.

How many rows (or instances) are in the datasheet?

Your answer

Check the Info section in the Insurance Data Table widget. *

Is there any missing data?

- ◯ Yes
- ◯ No

Click the "Restore Original Order" button. *

Check the first 10 rows of the dataset.

What is the oldest age?

Your answer

Click the "Restore Original Order" button. *

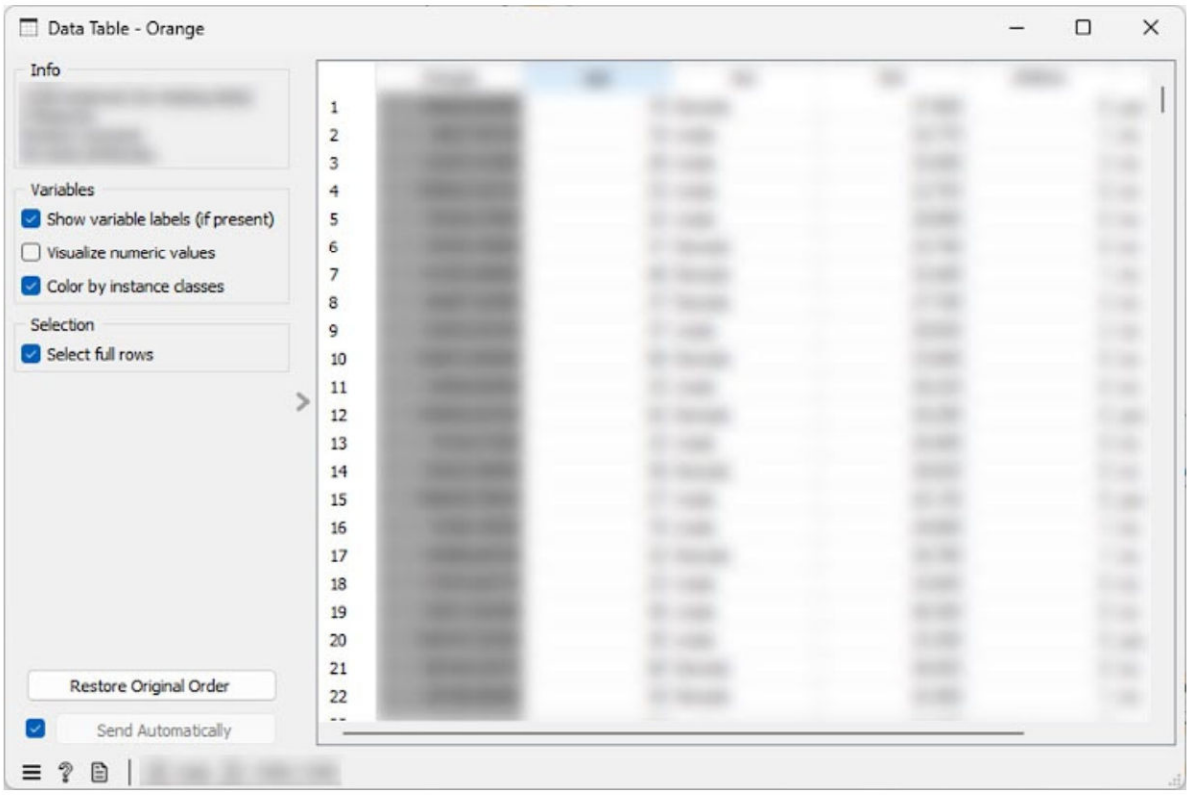
Check the first 10 rows of the dataset.

How many of them smoke?

Your answer

Take a screenshot of the Insurance Data Table window and upload it here. *

Example:



Add file

To show basic statistics for data features, connect the File widget to a Feature Statistics widget. *

- ◯ Done

Check the Feature Statistics widget. *

What is the mean age?

Your answer

What is the maximum number of children? *

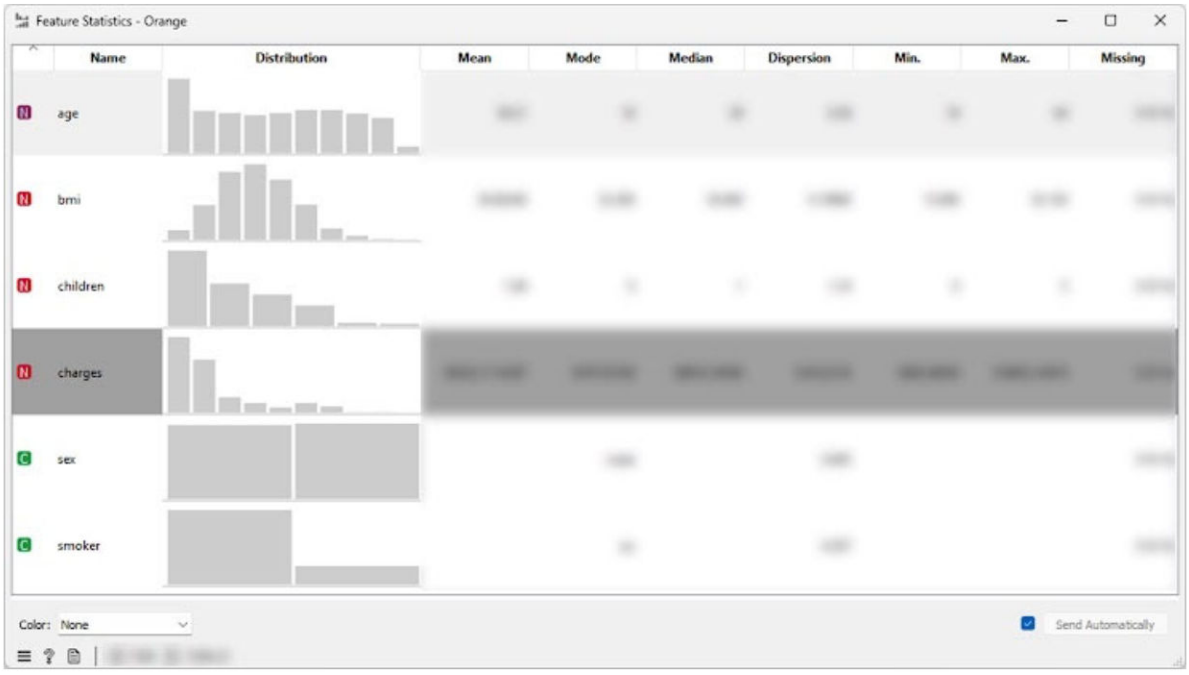
Your answer

Which sex appeared most frequently? *

- ◯ Male
- ◯ Female

Take a screenshot of the Feature Statistics window and upload it here. *

Example:



Add file

To provide a 2-dimensional scatter plot visualization, connect File widget to a Scatter Plot widget. *

- ◯ Done

In the Axes section, set "BMI" as the x-axis and "charges" as the y-axis, since "charges" is the variable the model aims to predict. *

- ◯ Done

Select "Show regression line" to display a line on the plot. *

- ◯ Done

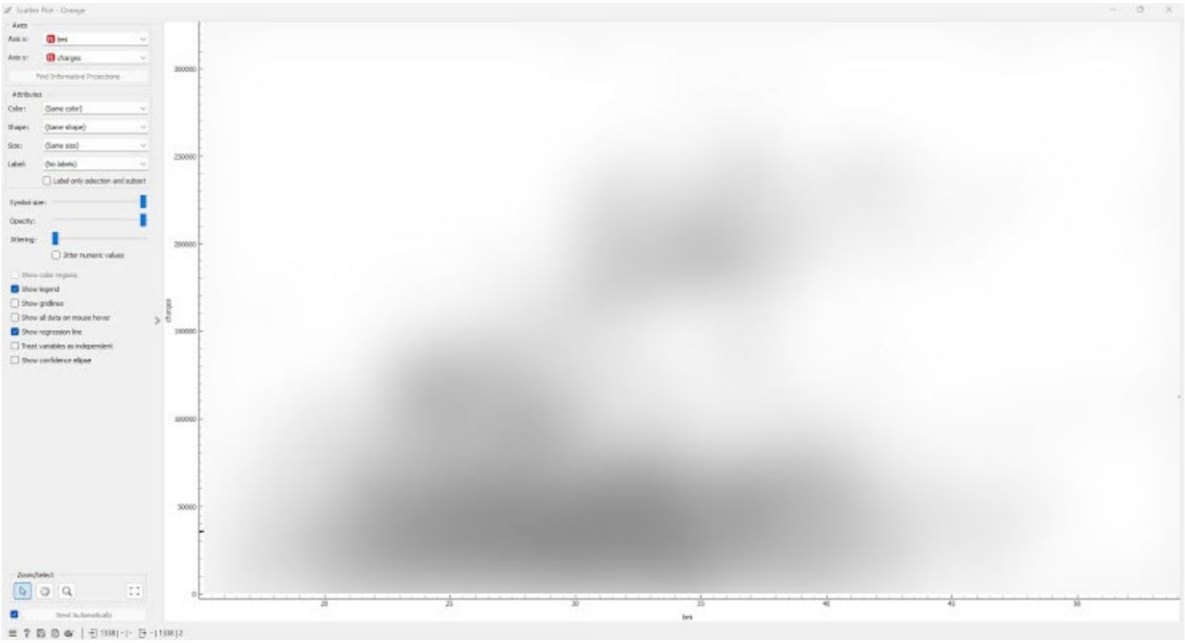
A regression line is increasing if it rises from left to right and decreasing if it falls. *

In the context of "bmi" versus "charges", how does the medical cost billed to the insurance company change as the body mass index of a customer increases?

- ◯ Increases
- ◯ Decreases

Take a screenshot of the Scatter Plot window and upload it here. *

Example:



Add file

Step 2: Clean the Data

Since the dataset has no missing data, there's no need to address imputation (e.g., fill in any gaps with average values).

Next, we should check for duplicate rows and convert any categorical data into numerical values so the machine learning algorithm can process it effectively.

To filter instances unique by specified key attribute(s), connect the File widget to a Unique widget. *

- ◯ Done

Check the Unique widget. *

In the Group by section, either click and drag to select all the attributes or press Ctrl + A.

- ◯ Done

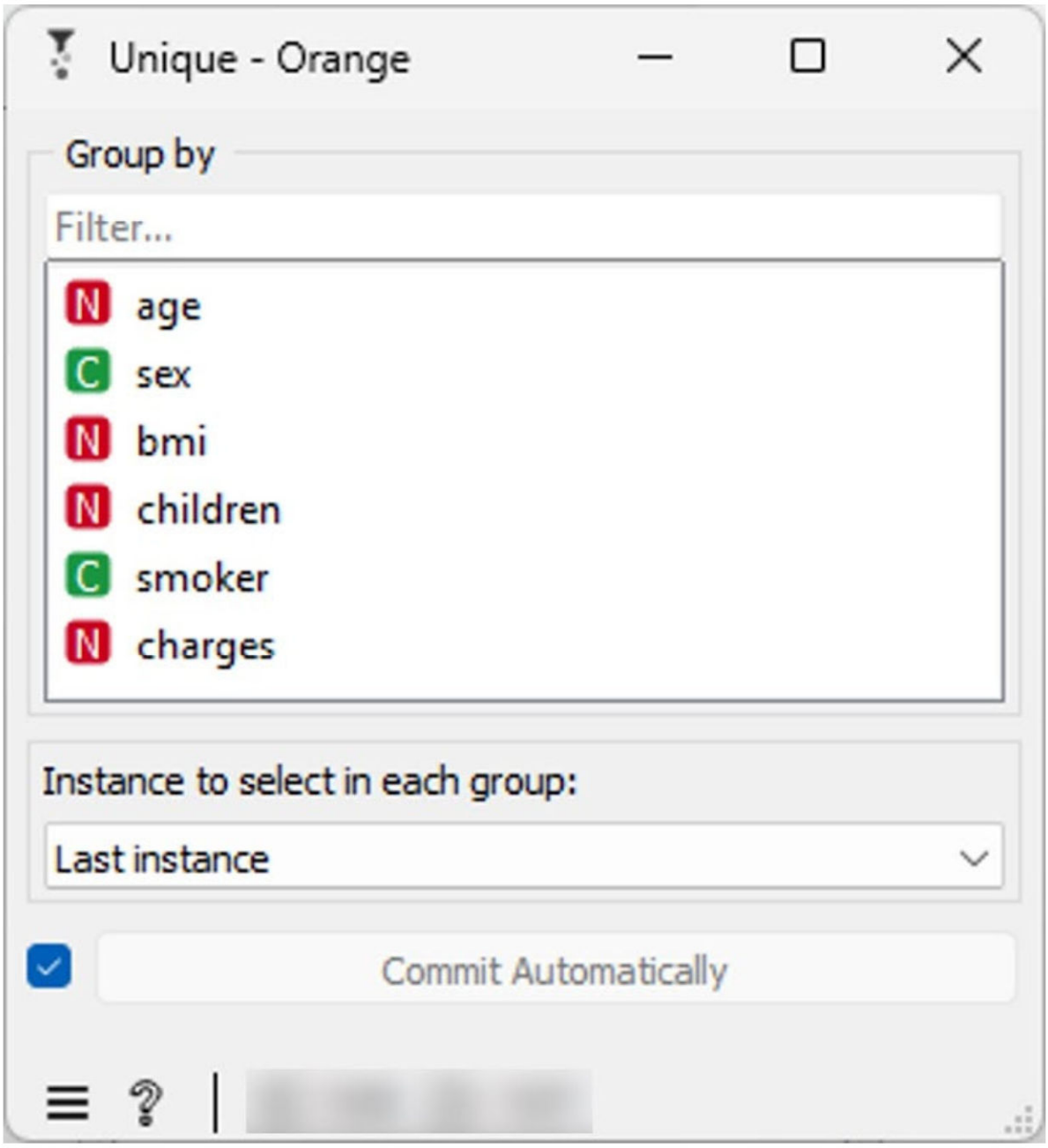
At the bottom of the Unique widget, you'll see two numbers: the original number of rows and the updated number after removing duplicates. *

Are there duplicate rows in the dataset?

- ◯ Yes
- ◯ No

Take a screenshot of the Unique window and upload it here. *

Example:



Add file

The dataset includes categorical columns "sex" and "smoker". *

To convert these categorical values into numeric, use one-hot encoding.

To construct a data preprocessing pipeline, connect the Unique widget to a Preprocess widget.

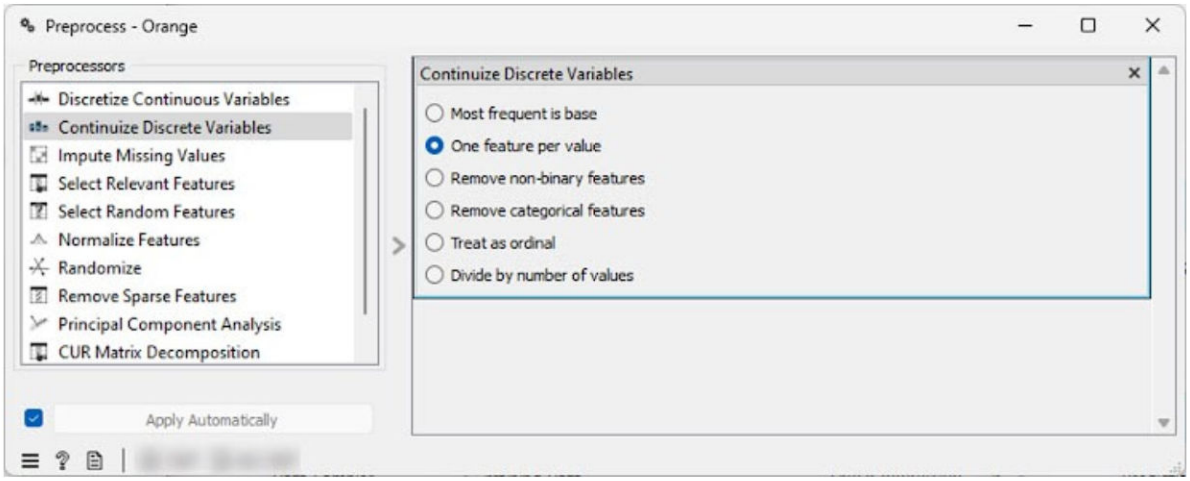
◯ Done

Check the Preprocess widget. *

In the Preprocessors section, double-click "Continuize Discrete Variables" to move it to the right side.

Then, select "One feature per value".

You should end up with something like this:



◯ Done

Let's check the dataset again by connecting the Preprocess widget to a new Data Table widget. *

Rename the Data Table widget to "Preprocess Data Table".

◯ Done

Verify that all columns are now in numeric format. *

Take a screenshot of the Preprocess Data Table window and upload it here.

Add file

Step 3: Split the Data

If the cleaned dataset is split 70% for training and 30% for testing, how many rows will be in the training data? *

Round it to a whole number.

Your answer

How many rows will be in the testing data? *

Round it to a whole number.

Your answer

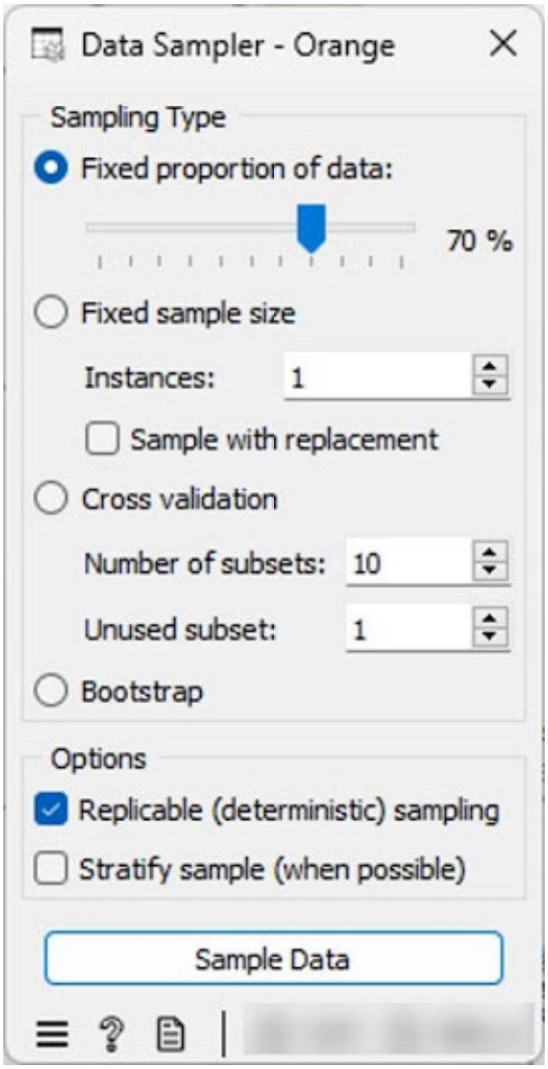
Verify your answers by connecting the Preprocess widget to a Data Sampler widget. *

- ◯ Done

Check the Data Sampler Widget. *

Set it up as shown in the picture, where the data is split 70/30.

Note: Ensure that "Replicable (deterministic) sampling" is selected; otherwise, you might not get correct answers to the next questions.



- [ ] Done

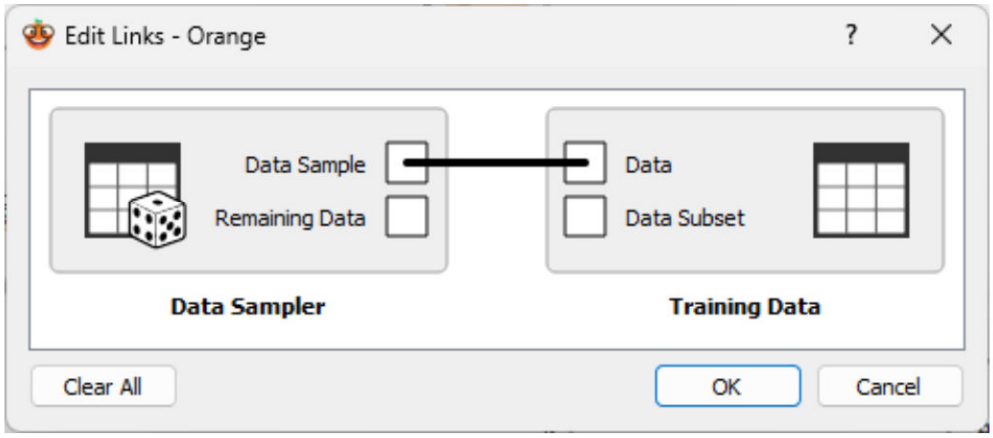
Let's check the training and testing datasets by connecting the Data Sampler widget to two Data Table widgets. *

Rename the Data Table widgets to "Training Data" and "Testing Data".

- [ ] Done

Edit the link between the Data Sampler and Training Data by double-clicking the line connecting them. *

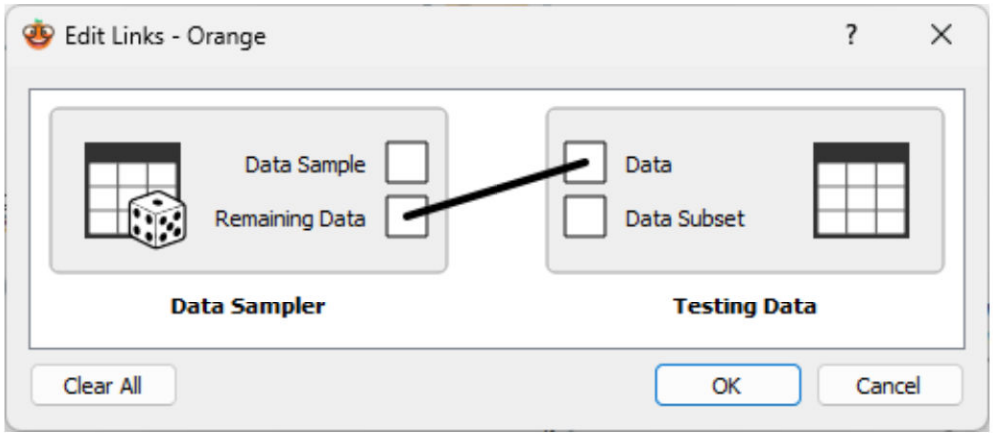
Set it up as shown in the picture.



◯ Done

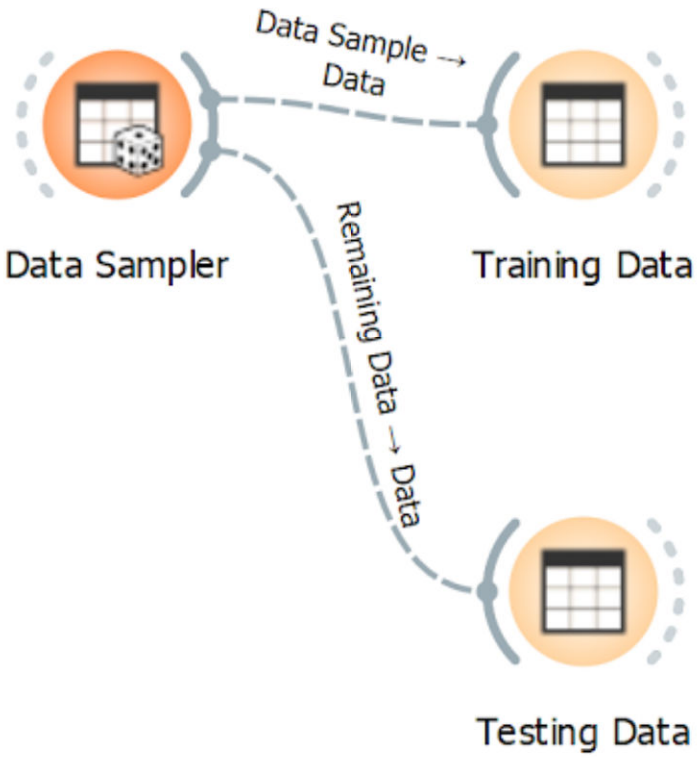
Edit the link between the Data Sampler and Testing Data by double-clicking the line connecting them. *

Set it up as shown in the picture.



◯ Done

You should end up with something like this: *



- [ ] Done

Verify the number of rows for your training data in the Info section. *

Take a screenshot of the Training Data window and upload it here.

Add file

Verify the number of rows for your testing data in the Info section. *

Take a screenshot of the Testing Data window and upload it here.

Add file

Step 4: Training the Model

Connect your training data to a Linear Regression widget. *

Keep the settings unchanged.

- ◯ Done

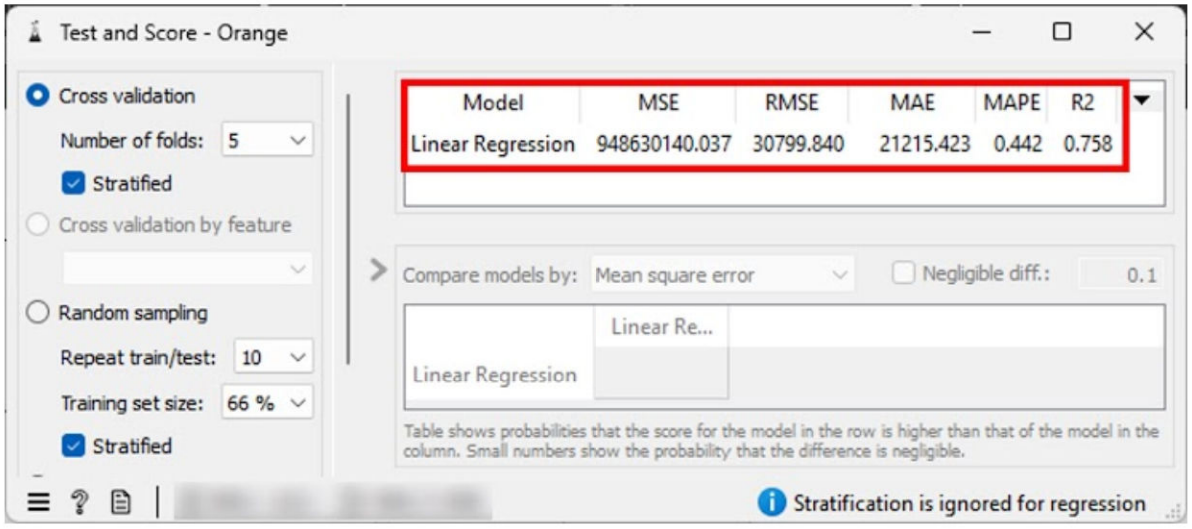
To perform cross-validation for accuracy estimation, connect your training data and the Linear Regression widget to a Test and Score widget. *

Keep the settings unchanged.

- ◯ Done

These are the performance metrics you should obtain for the model through cross-validation. *

If you do not achieve these metrics, you should review your process.



- ◯ Done

Connect the Linear Regression widget to a Data Table widget. *

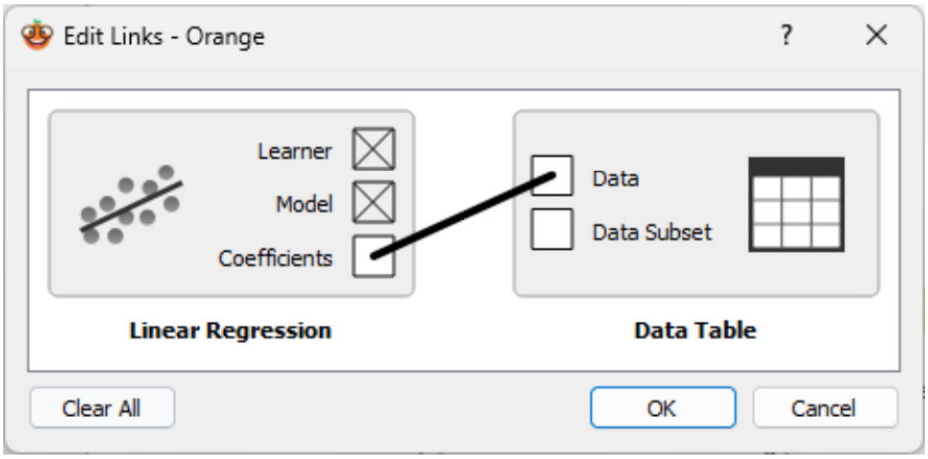
Rename the Data Table widget as "Coefficients Data Table".

- ◯ Done

Edit the link between the Linear Regression and Coefficients Data Table widgets by double-clicking the line connecting them. *

Set it up as shown in the picture.

This will allow you to view the slopes and y-intercept of the linear surface created by your model.



- ◯ Done

Based on the Coefficients Data Table, what is the value of the intercept? *

Include the negative sign if applicable.

Your answer

What is the coefficient (slope) for "age"? *

Include the negative sign if applicable.

Your answer

What is the coefficient (slope) for "sex=female"? *

Include the negative sign if applicable.

Your answer

What is the coefficient (slope) for "sex=male"? *

Include the negative sign if applicable.

Your answer

What is the coefficient (slope) for "bmi"? *

Include the negative sign if applicable.

Your answer

What is the coefficient (slope) for "children"? *

Include the negative sign if applicable.

Your answer

What is the coefficient (slope) for "smoker=no"? *

Include the negative sign if applicable.

Your answer

What is the coefficient (slope) for "smoker=yes"? *

Include the negative sign if applicable.

Your answer

Step 5: Testing the Model

Let's evaluate how the model you developed performs on new, unseen data.

Connect your test data and Linear Regression model to a Predictions widget. *

- ◯ Done

Check the Predictions widget. *

Select "Absolute difference" as the regression error metric.

Sort the "Error" column from smallest to largest.

- ◯ Done

What is the smallest absolute difference error? *

Your answer

What is the largest absolute difference error? *

Your answer

Check the bottom of the Predictions widget. *

What is the Root Mean Squared Error (RMSE) value of the model using the test data?

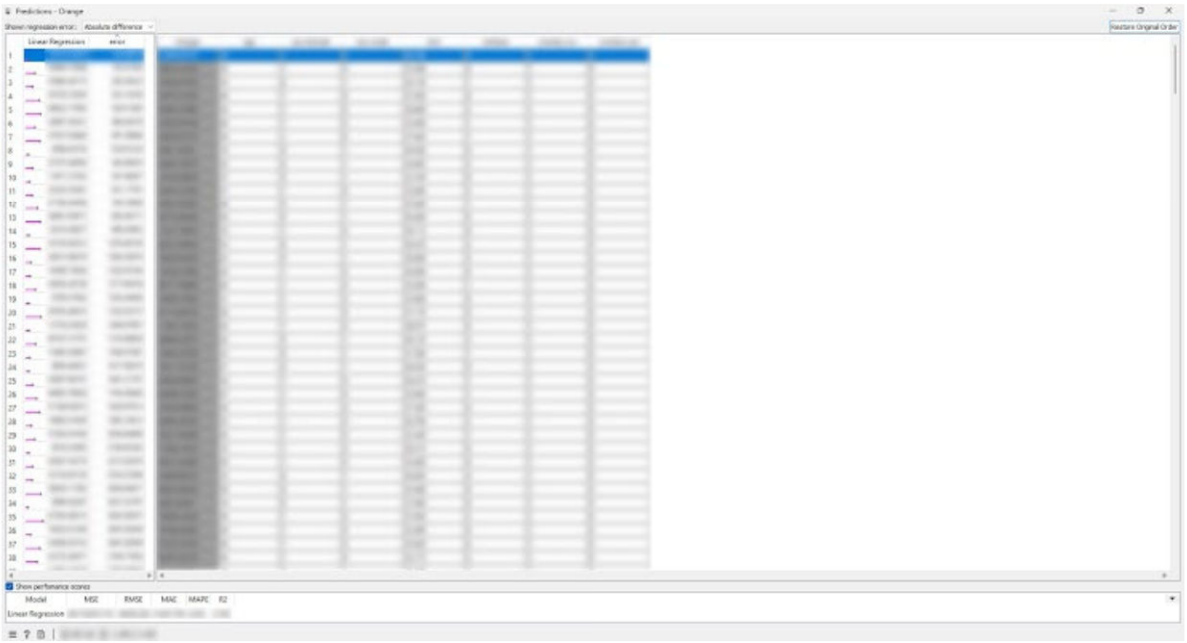
Your answer

What is the R^2 value of the model using the test data? *

Your answer

Take a screenshot of the Predictions window, where the errors are sorted from smallest to largest, and upload it here. *

Example:



Add file

Step 6: Test it Yourself

Utilize your model to determine the amount the insurance company would need to pay the hospital based on your factors.

Generate a CSV file with column names that exactly match those in your preprocessed dataset. *

The columns should be in this order: **age, sex=female, sex=male, bmi, children, smoker=no, smoker=yes**.

Populate the columns with values based on your own data.

Example:

	A	B	C	D	E	F	G
1	age	sex=female	sex=male	bmi	children	smoker=no	smoker=yes
2	25	0	1	26	0	1	0

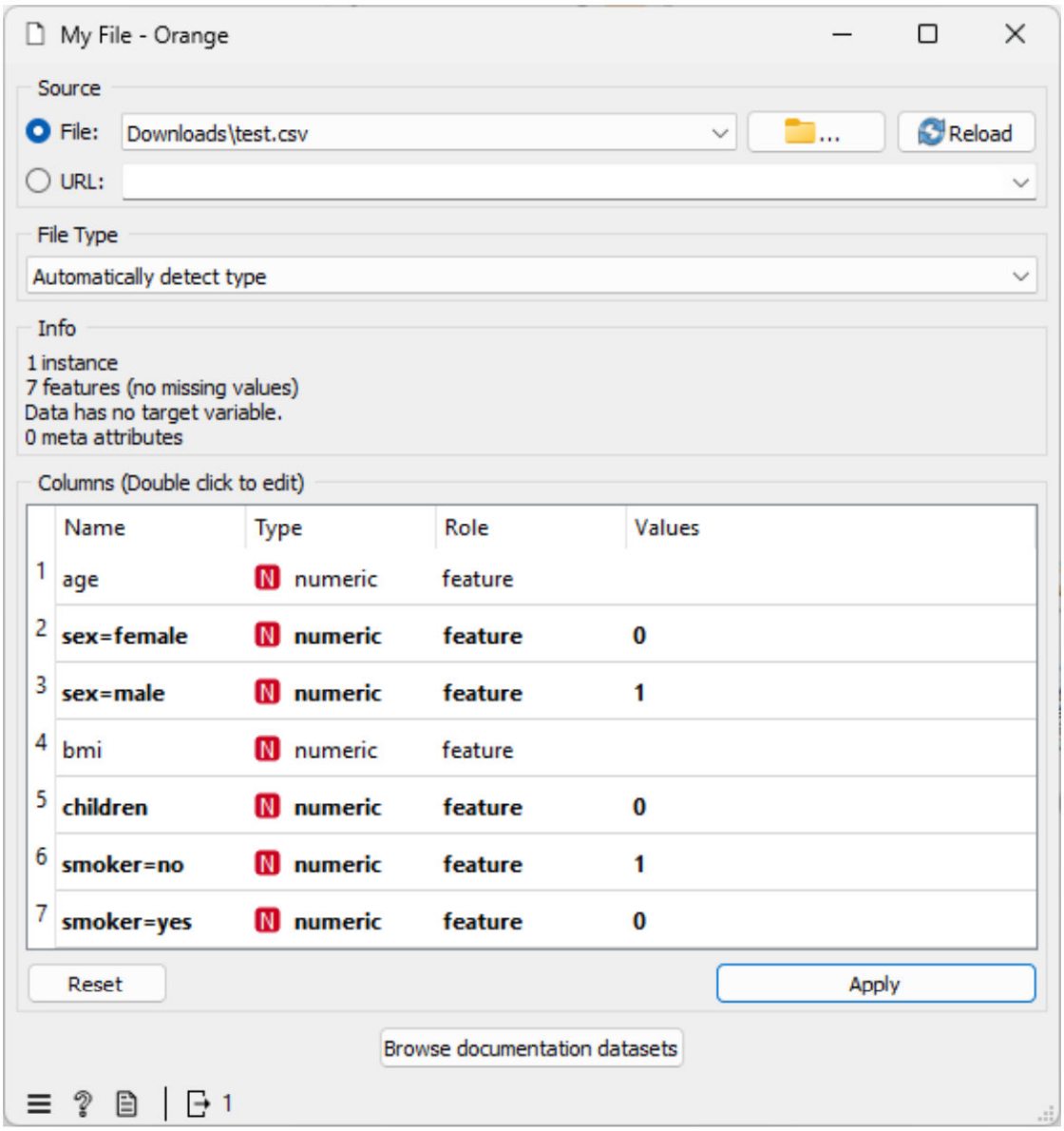
◯ Done

Drag and drop a File widget and rename it "My File". *

Use your CSV file as the source.

Set the type of all columns to "numeric" and their role to "feature".

Example:

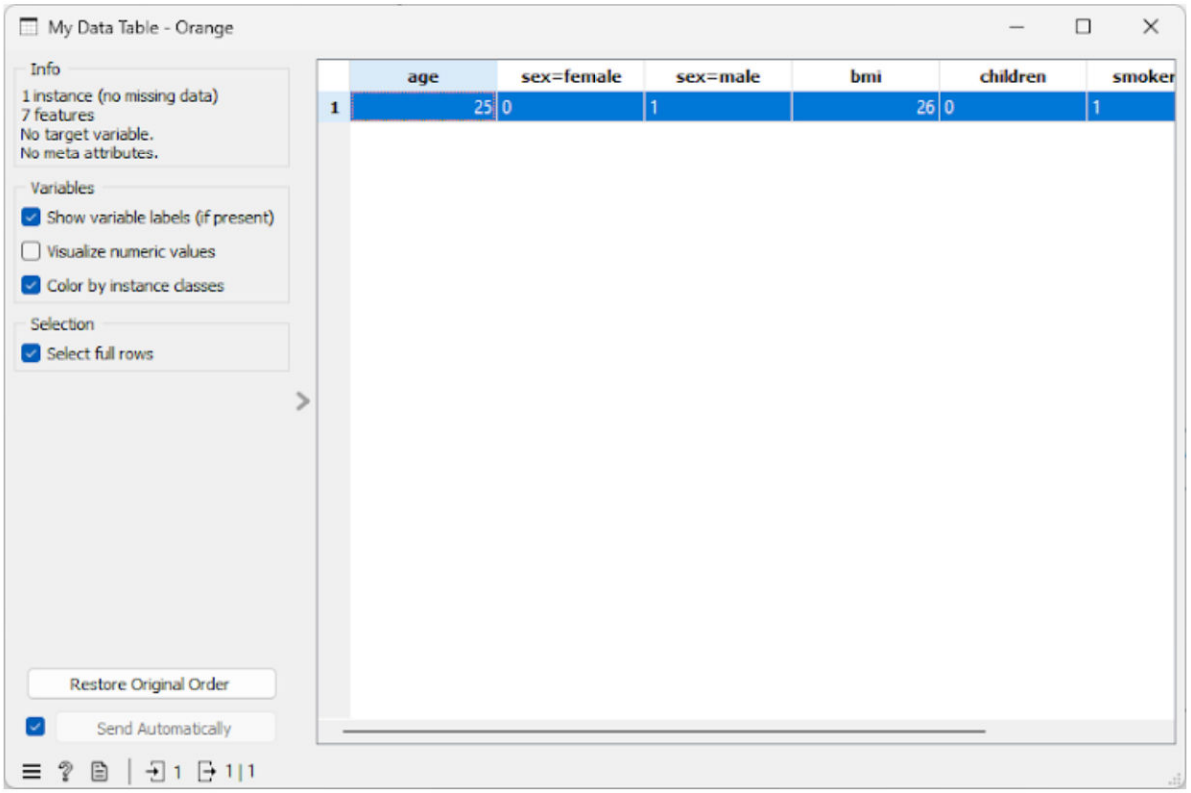


- [ ] Done

Connect the File widget to the Data Table widget. *

Rename the Data Table widget to "My Data Table".

Example:



- [ ] Done

Connect the My Data Table widget and your Linear Regression widget to a new Predictions widget. *

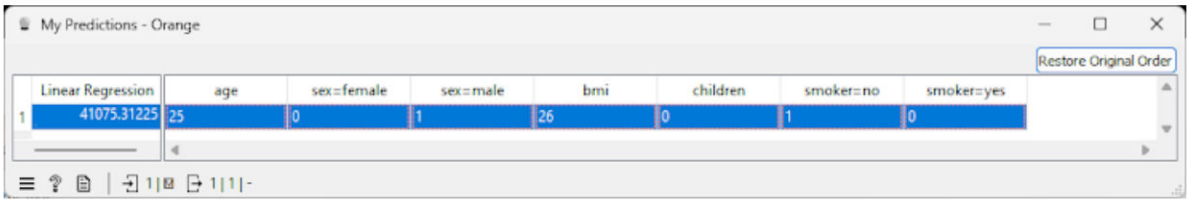
Rename the Predictions widget to "My Predictions".

- [ ] Done

Check the My Predictions widget. *

Take a screenshot of the My Predictions window and upload it here.

Example: If I was admitted to a hospital, the insurance company I'm with should cover up to P41,075 of my hospital expenses.



Add file

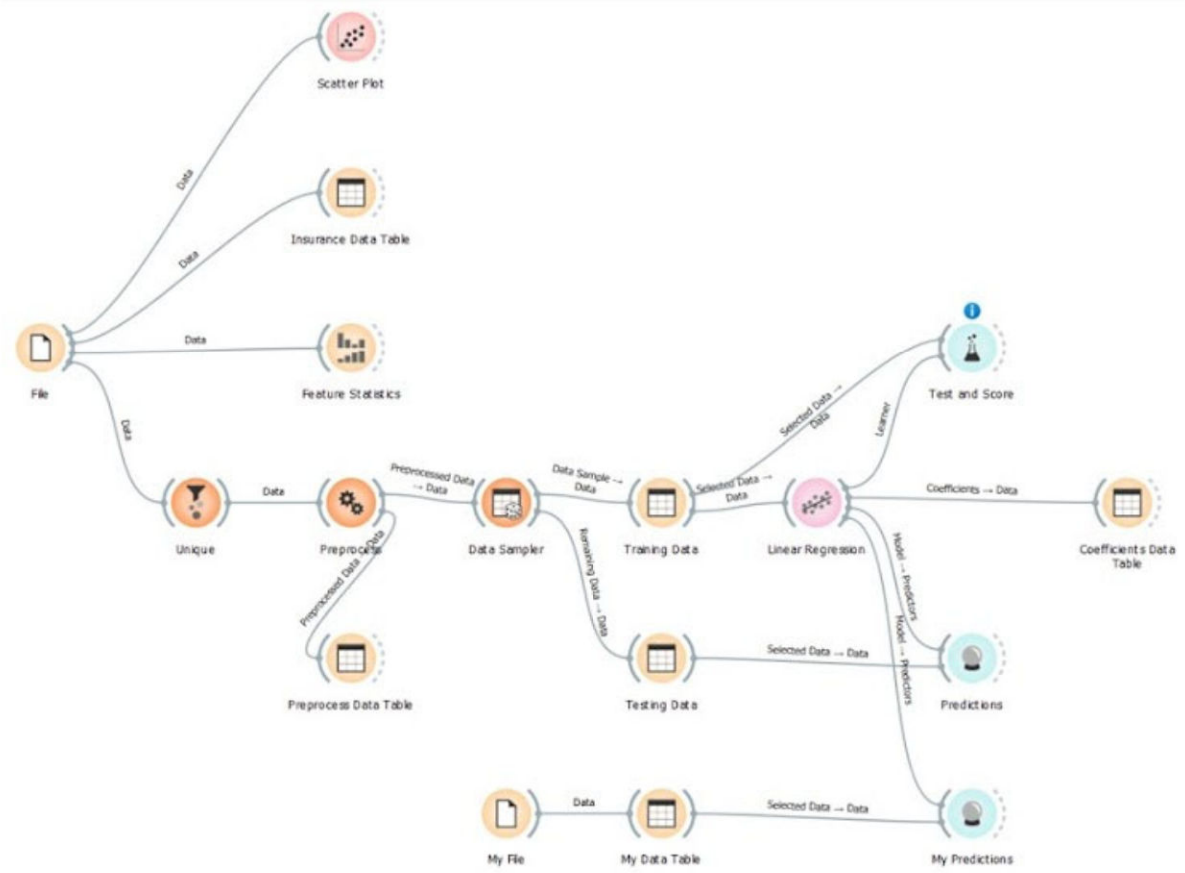
Step 7: Save and upload your workflow

Review your workflow. *

You should have a total of 17 widgets:

- 9 Data widgets (Yellow)
- 3 Transform widgets (Orange)
- 3 Evaluate widgets (Blue)
- 1 Visualize widget (Peach)
- 1 Model widget (Pink)

You should end up with something like this:



◯ Done

If your result is similar to the one above, save your workflow. *

Compress your workflow into a ZIP file, using the same name as your original workflow file (e.g., Rey_4A_Lab_Exercise_1).

Upload it here.

Add file

Get link

Never submit passwords through Google Forms.