

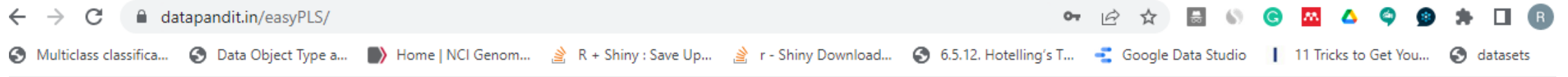
# DataPandit User Manual

Let's Excel Analytics Solutions LLP

# Import Data

- Important data requisites
  - Data should be in .CSV file format
  - Should have unique row names (it would be ideal to use unique serial numbers)
  - Avoid using space in column names (You can use . To separate two words)
  - Avoid characters such as %, \$, @, !, &, \*, (,) etc. in column names
  - For PCA and LDA you must have only one column that contains the categorical variable

# How to Import Data



## PLS

A screenshot of the PLS web application interface. The main content area is titled 'Data Inputs' and features a 'Choose CSV File with Data' section. A 'Browse...' button is circled in orange. A file icon with a green 'X' and 'a,' is being dragged over the interface. Below this are three checked checkboxes: 'Is this spectroscopic data?', 'Do you want to scale the data?', and 'Do you want to center the data?'. A '+ Copy' button is next to the first checkbox. The 'Create Training and Testing Samples' section has a checked checkbox 'Do you want to replace samples for training and test?' and a 'Training Set Probability' slider set to 0.59. On the right, a navigation menu includes 'Table', 'Correlation Matrix', 'Training Data', 'Testing Data', 'Spectra', 'Visualize', 'Model Summary', 'Scores', 'Loadings', 'Validation Plot', 'Predicted Vs. Actual', 'Prediction Summary', and 'Unknown Samples'. A copyright notice at the bottom reads '© 2022, Let's Excel Analytics Solutions LLP. All rights reserved.' An orange callout box with two bullet points is overlaid on the right side of the interface. The URL 'http://letsexcel.in' is visible at the bottom center, and a circular '+' icon is in the bottom right corner.

- Use the browse feature at the top left of screen
- Or simply drag and drop files



# Data Pretreatments

- Every application provides basic option of mean centering and scaling data in the sidebar layout
- As a thumb rule it is a good idea to mean center the data
- You can scale/ standardize data if the responses are measured in different units. (For example, some of the column values are measured in Kg while other column values are measured in liters)
- Pretreatments are easily applied by checking the respective boxes
- PCA application has more advanced data treatment options which can be used depending on the case
- Training and Testing data can be divided setting up probability

PLS

## Data Inputs

Choose CSV File with Data

Browse... No file

Is this spectroscopic data? [+ Copy](#)

Do you want to scale the data?

Do you want to center the data?

## Create Training and Testing Samples

Do you want to replace samples for training and test?

Training Set Probability

0 0.59 1

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

<http://letsexcel.in>

Always keep it checked

# Dependents and Independents-1

### Data Inputs

Choose CSV File with Data

Browse... Gasoline.CSV

Upload complete

Is this spectroscopic data?

Do you want to scale the data?

Do you want to center the data?

### Create Training and Testing Samples

Do you want to replace samples for training and test?

Training Set Probability

0 0.59 1

### Model Inputs

Number of Components


Loadings Validation Plot Predicted Vs. Actual Prediction Summary Unknown Samples

Prediction for Unknowns

Show 10 entries Search:

	octane	900	902	904	906	908	910	912	914	916
1	85.3	-0.050193	-0.045903	-0.042187	-0.037177	-0.033348	-0.031207	-0.030036	-0.031298	-0.034217
2	85.25	-0.044227	-0.039602	-0.035673	-0.030911	-0.026675	-0.023871	-0.022571	-0.02541	-0.02896
3	88.45	-0.046867	-0.04126	-0.036979	-0.031458	-0.02652	-0.023346	-0.021392	-0.024993	-0.029309
4	83.4	-0.046705	-0.04224	-0.038561	-0.034513	-0.030206	-0.02768	-0.026042	-0.02828	-0.03092
5	87.9	-0.050859	-0.045145	-0.041025	-0.036357	-0.032747	-0.031498	-0.031415	-0.034611	-0.037781
6	85.5	-0.048094	-0.042739	-0.038812	-0.034017	-0.030143	-0.02769	-0.026387	-0.028811	-0.031481
7	88.9	-0.049906	-0.044558	-0.040543	-0.035716	-0.031844	-0.029581	-0.027915	-0.030292	-0.03359
8	88.3	-0.049293	-0.043788	-0.039429	-0.034193	-0.029588	-0.026455	-0.025104	-0.028102	-0.031801
9	88.7	-0.049885	-0.044279	-0.040158	-0.034954	-0.031114	-0.02839	-0.027017	-0.029609	-0.032937
10	88.45	-0.051054	-0.045678	-0.041673	-0.036761	-0.033078	-0.030466	-0.029295	-0.031736	-0.034843

octane 900 902 904 906 908 910 912 914 +



- The dependents (Ys) can be selected by clicking on bottom header of the columns in PLS and PCR applications.

# Dependents and Independents-2

- The categorical variable/ dependent variable needs to be selected by using drop-down function in PCA, MLR and LDA applications

## Import Data

Choose CSV File with Data

Browse... Iris.csv

Upload complete

Select Categorical Data Columns

sepal.length  
sepal.width  
petal.length  
petal.width  
variety

For PCA select single categorical variable

## Model Input

Select Categorical Variable

variety

Select Numerical Variables

sepal.width petal.length petal.width

For LDA select single categorical variable and multiple numerical variables of choice

## Model Outputs

Select a Response Variable

b\_pres

Select Independent Variables

b\_temp b\_pres b\_time meth\_pct meth\_ec eth\_pct  
eth\_ec prop1\_pct

prop1\_ec

For MLR select single response variable and multiple independent variables of choice

Do you want to mean centre the data?

Create Training and Testing Samples

<http://retsexcel.in>

# Data Visualization Spectroscopic

← → ↻ [datapandit.in/easyPLS/](http://datapandit.in/easyPLS/)

Multiclass classifica... Data Object Type a... Home

## PLS

### Data Inputs

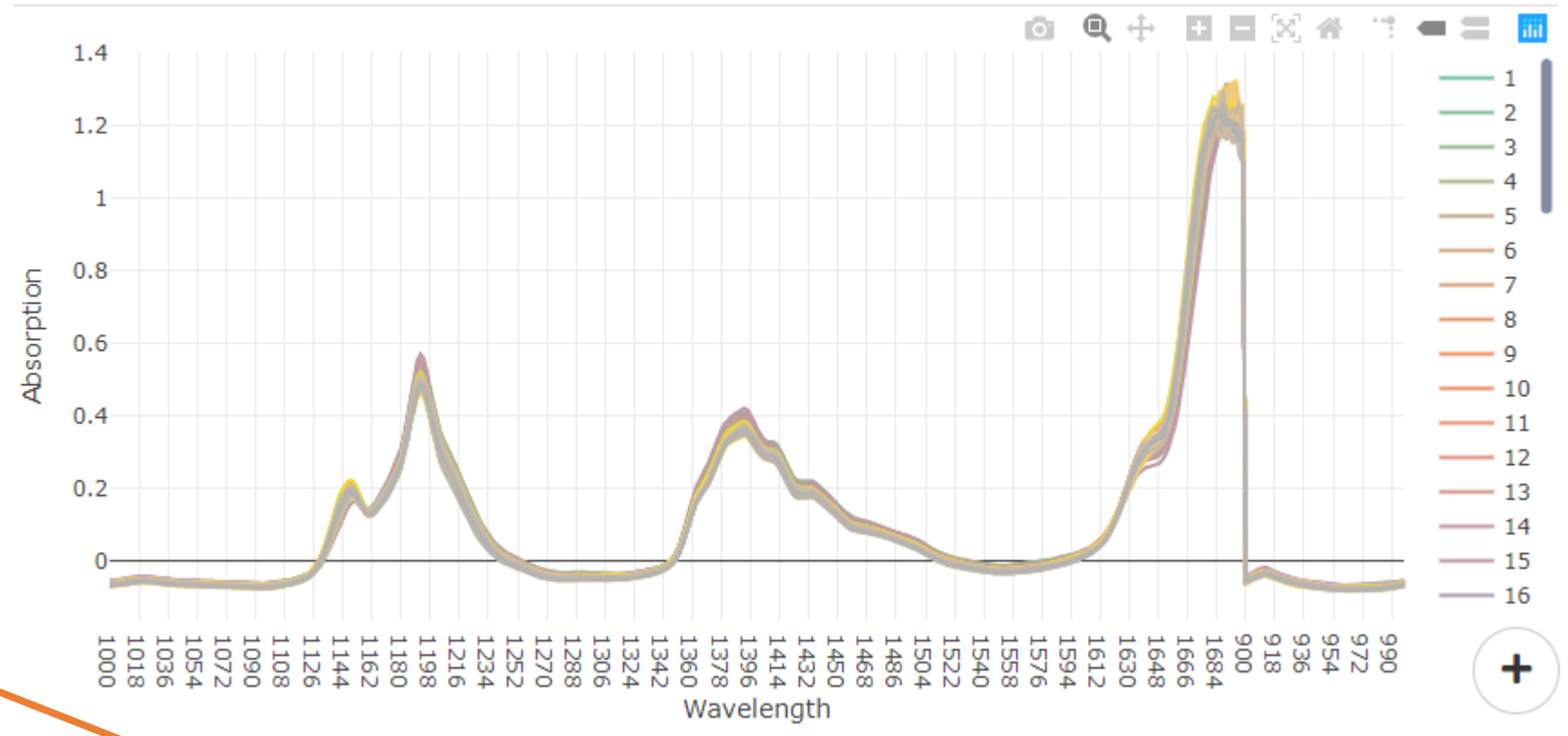
Choose CSV File with Data

Browse... No file

Is this spectroscopic data?  + Copy

Do you want to scale the data?

Do you want to center the data?



If data is spectroscopic, then check 'Is this spectroscopic data?' option in sidebar layout

# Data Visualization-non spectroscopic

Is this spectroscopic data?

Do you want to scale the data?

Do you want to center the data?

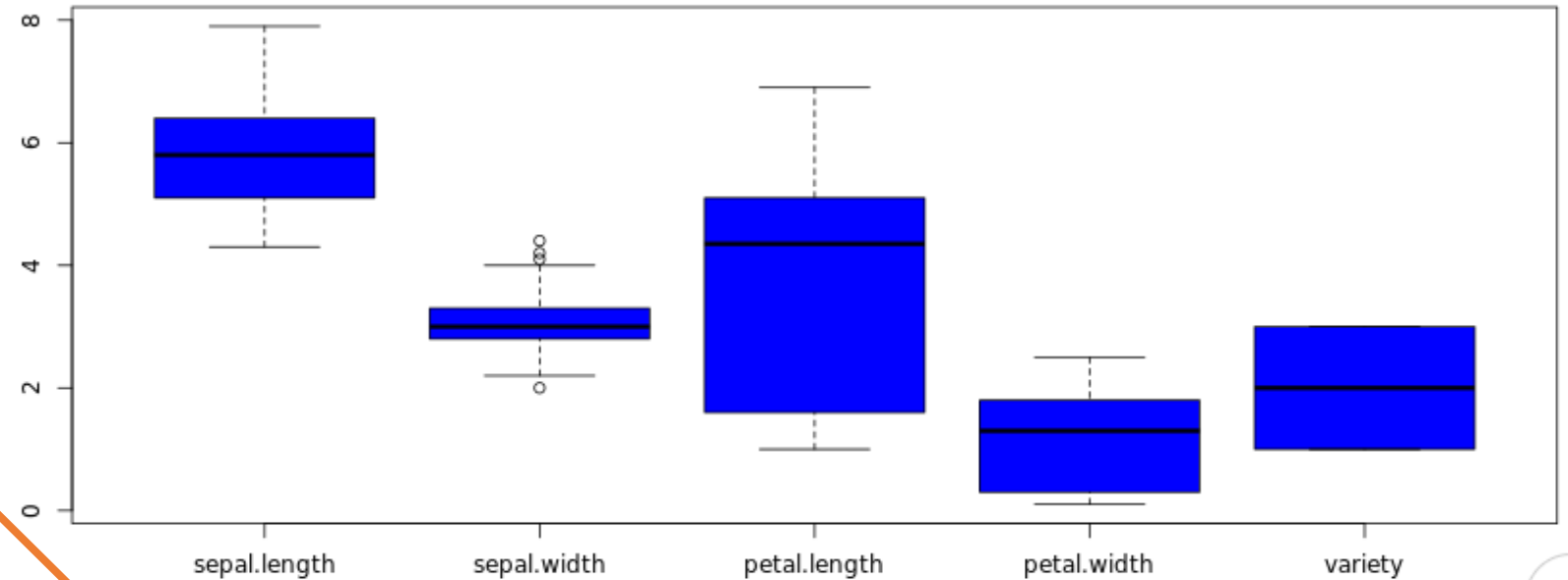
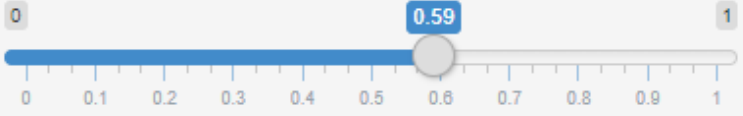
---

## Create Training and Testing Samples

Do you want to replace samples for training and test?

**Training Set Probability**

0 0.59 1

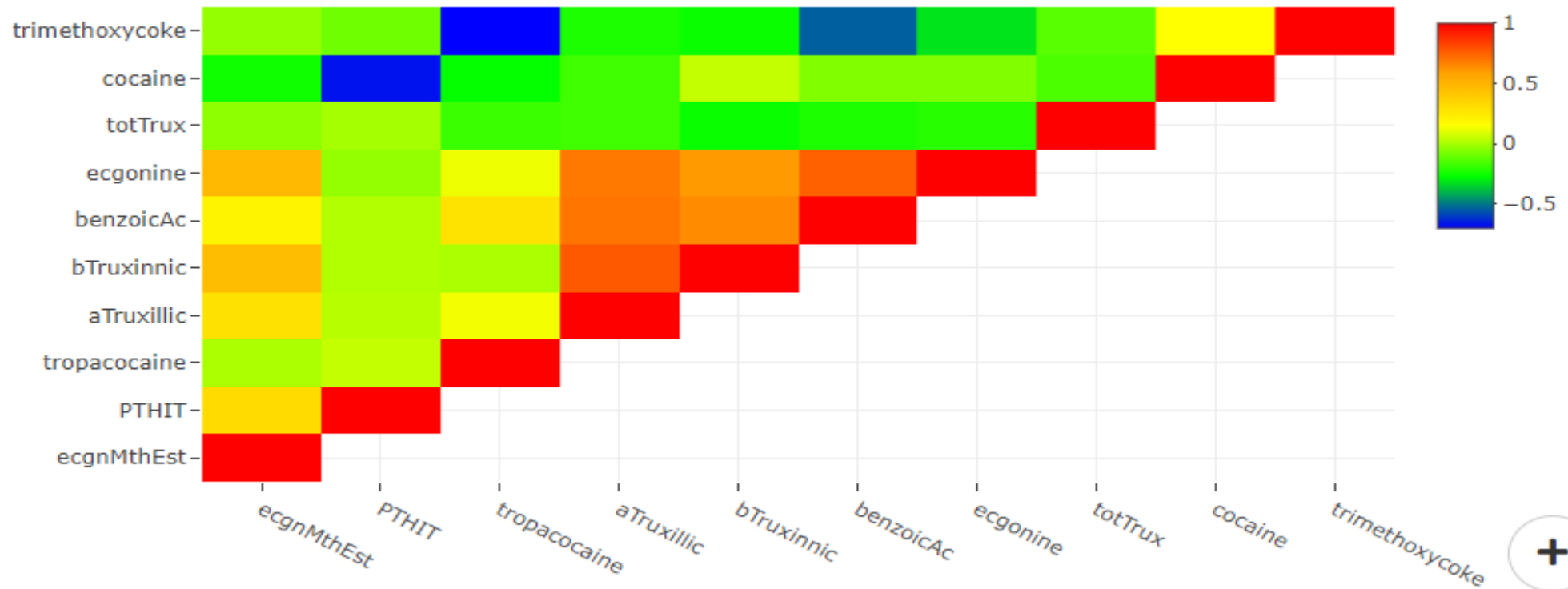


- If data is non-spectroscopic, then uncheck 'Is this spectroscopic data?'
- You can only visualize raw data in LDA, PLS, PCR and MLR applications
- In case of PCA application data pretreatments will be reflected in box plots or spectroscopic graphs



# Correlation Matrix

- Correlation matrix shown Pearson's correlation between each column in the dataset with every other column in the data set. This plot will not be calculated for PCA and LDA application unless categorical data is selected using dropdown menu



# Model Outputs & Graphs

- All the model outputs and graphs will be automatically populated as soon as you separate dependents and independents Or numerical data and categorical data

```
Data:  X dimension: 16 9
       Y dimension: 16 1
Fit method: kernelpls
Number of components considered: 5
TRAINING: % variance explained
      1 comps  2 comps  3 comps  4 comps  5 comps
X      19.79   49.96   66.90   77.63   88.32
y()    58.93   71.56   76.78   79.85   81.22
```

# SIMCA Model-Saving Files

- You need to prepare independent PCA model for each class in SIMCA model
- All the models need to be uploaded to get train and test SIMCA classification output

The screenshot displays the SIMCA software interface. On the left, the 'Model Inputs' section includes a dropdown menu for 'Setosa' and a 'Select Number of Components' knob set to 2. A blue callout bubble points to the knob with the text 'Knob to select component for each model'. The central 'SIMCA Summary' panel shows parameters for a 'Setosa' model: 2 components, ddmoments limits, Alpha: 0.05, Gamma: 0.05, and a performance table. A blue callout bubble points to the 'Save File' button in the 'Save Model' section with the text 'Save each model file one by one'. Another blue callout bubble points to the 'Browse...' button in the 'Upload Model' section with the text 'Browse to upload files'.

**Model Inputs**

Select One Group for SIMCA Model

Setosa

Select Number of Components

Knob to select component for each model

**SIMCA Summary**

SIMCA model for class 'Setosa' summary

Number of components: 2  
Type of limits: ddmoments  
Alpha: 0.05  
Gamma: 0.05

	Expvar	Cumexpvar	TP	FP	TN	FN	Spec.	Sens.	Accuracy
Ca1	12.13	91.24	29	0	0	2	NA	0.935	0.935

© 2022, Let's Excel Analytics Solutions LLP. All rights reserved.

**Save Model:**

Model File

Save File

**Upload Model**

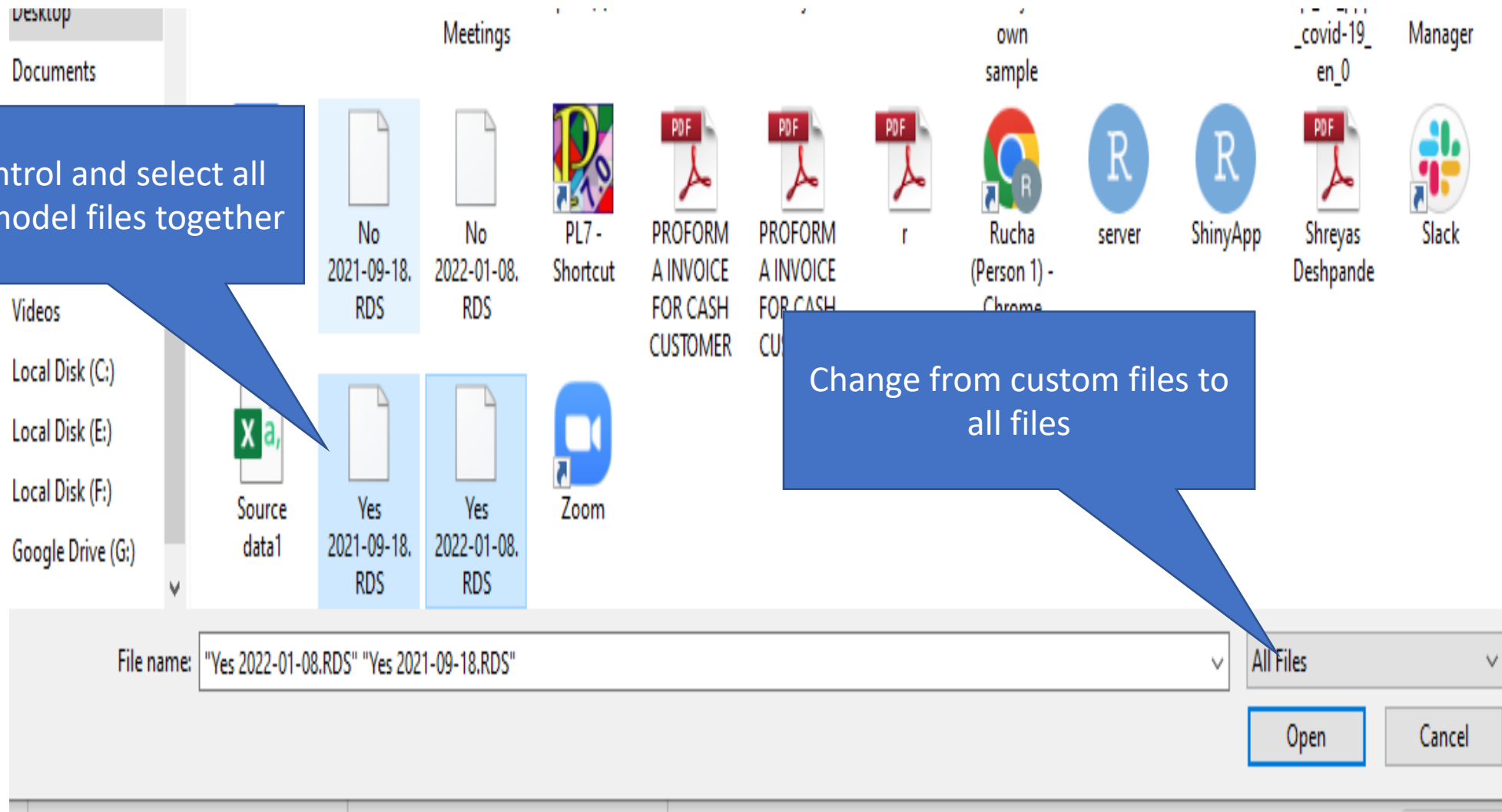
Choose Rdata File

Browse... No file selected

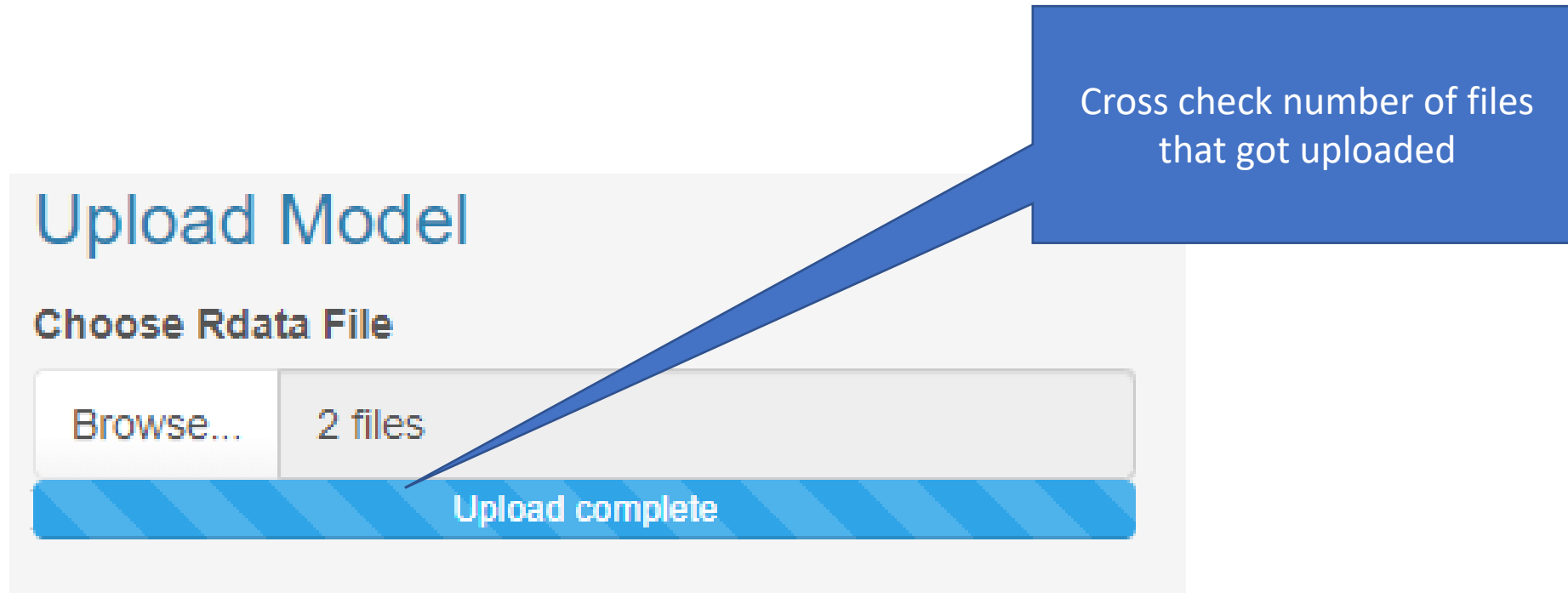
Save each model file one by one

Browse to upload files

# SIMCA Model-Uploading Files



# SIMCA Model-Uploading Files



As soon as model files are uploaded train and test SIMCA classification results will become visible

# Predicting Unknown

- Ensure to keep same column names and same column sequence
- Upload data without results
- Upload model for prediction
- Check the predicted outcome

# Thank you

- We prefer to keep it short and simple
- For reporting bugs or requesting more features write to [rucha@letsexcel.in](mailto:rucha@letsexcel.in) or [info@letsexcel.in](mailto:info@letsexcel.in)
- Follow on [Linkedin](#) and [Facebook](#) to stay connected