

**19BCE2647**

**Aadarshbhushan Singh**

**VIT Vellore**

## **Smartinternz Assignment 03**

### **Data Analytics**

**23<sup>rd</sup> April 2022**

**Dataset Name: Diabetes.csv**

**Link to the dashboard:**

[https://eu-gb.dataplatfrom.cloud.ibm.com/canvas/flows/40078de2-3f00-4466-a866-be2e527d2ec5?context=cpdaas&project\\_id=6b7b24a3-cebb-4b29-b99c-77050327294e](https://eu-gb.dataplatfrom.cloud.ibm.com/canvas/flows/40078de2-3f00-4466-a866-be2e527d2ec5?context=cpdaas&project_id=6b7b24a3-cebb-4b29-b99c-77050327294e)

The screenshot shows the IBM Watson Studio interface. The top navigation bar includes 'IBM Watson Studio', a search bar, and user information. The main area is titled 'Projects / Predicting\_Diabetes\_21-04-2022'. The 'Assets' tab is selected, showing a list of assets. On the left, there's a sidebar for 'Asset types' with categories like Data, Flows, Data Refinery Flow, and SPSS Modeler flow. The 'SPSS Modeler flow' category is currently selected. The main panel displays a single asset named 'Classification Diabetes Dataset' under the 'SPSS Modeler flow' category. The asset details show it was last modified '1 day ago' by 'Aadarshbhushan Singh (You)'. A large blue button at the top right says 'New asset'. To the right, there's a section for 'Data in this project' with a placeholder for uploaded files.

**Fig: Create New Project**

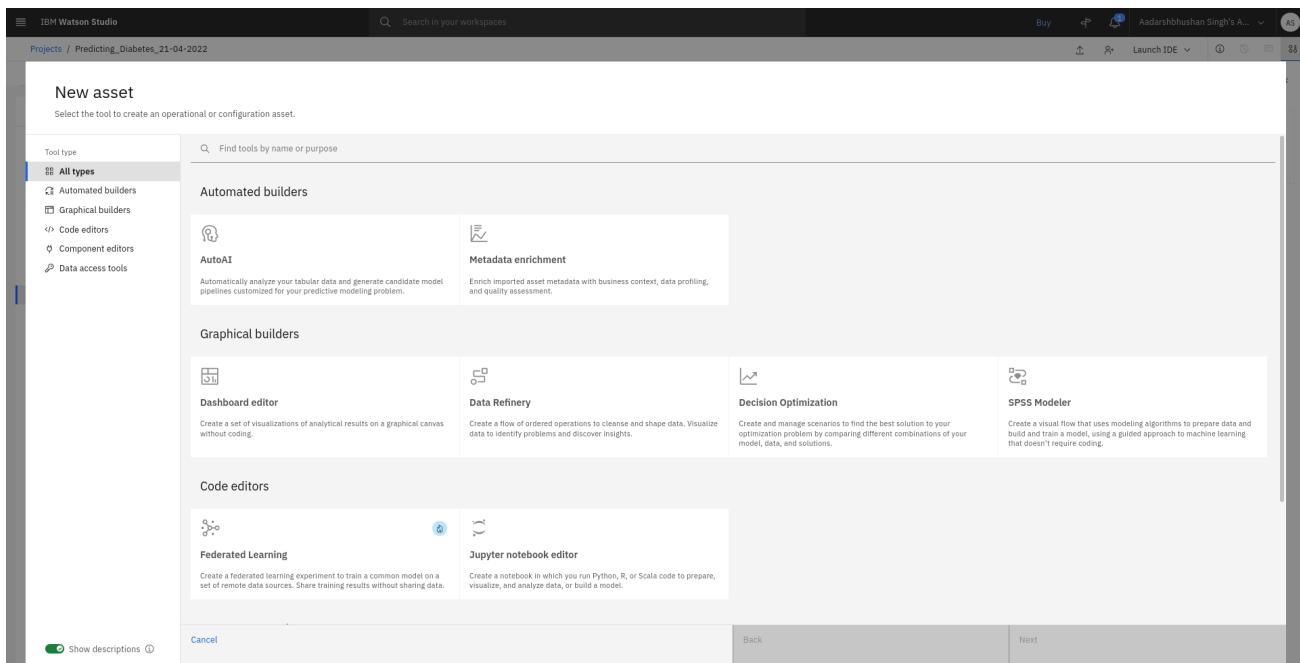


Fig: Go to Data Refinairy

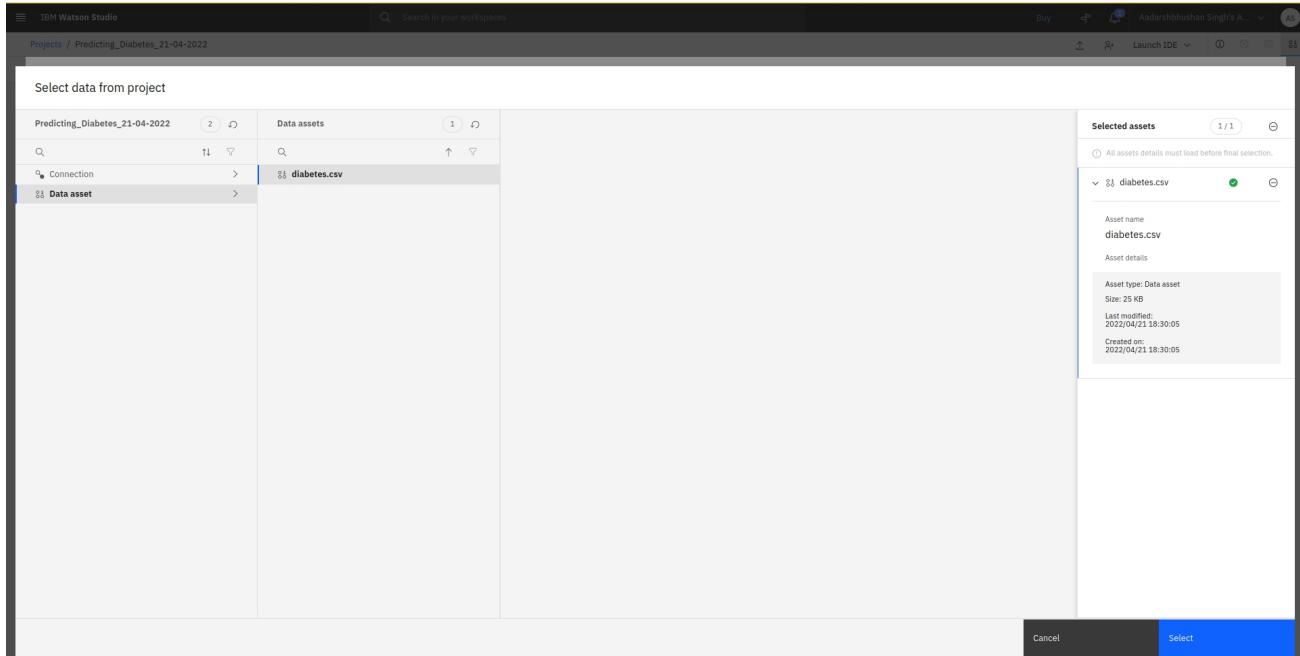


Fig: Select diabetes.csv data

IBM Watson Studio

Projects / Predicting\_Diabetes\_21-04-2022 / diabetes.csv / Refine data

Search operations

CLEANSE

Data Profile Visualizations

Pregnancies	Glucose	BloodPress...	SkinThickn...	Insulin	BMI	DiabetesPe...	Age
6	148	72	35	0	33.6	0.627	50
1	85	66	29	0	26.6	0.351	31
8	183	64	0	0	23.3	0.672	32
1	89	66	23	94	28.1	0.167	21
0	137	40	35	168	43.1	2.288	33
5	116	74	0	0	25.6	0.201	30
3	78	50	32	86	31	0.248	26
10	115	0	0	0	35.3	0.134	29
2	197	70	45	543	30.5	0.158	53
8	125	96	0	0	0	0.232	54
4	110	92	0	0	37.6	0.191	30
10	168	74	0	0	38	0.537	34
10	139	80	0	0	27.1	1.441	57
1	189	60	23	846	30.1	0.398	59
5	166	72	19	175	25.8	0.587	51
7	100	0	0	0	30	0.484	32
0	118	84	47	230	45.8	0.551	31
7	107	74	0	0	29.6	0.254	31
1	103	30	38	83	43.3	0.183	33
1	115	70	30	96	34.6	0.529	32
3	126	88	41	235	39.3	0.704	27
8	99	84	0	0	35.4	0.388	50
7	196	90	0	0	39.8	0.451	41

SOURCE FILE: diabetes.csv FULL DATA SET: 768 rows

Cancel Apply

Fig: Date Refining

IBM Watson Studio

Projects / Predicting\_Diabetes\_21-04-2022 / diabetes.csv / Refine data

Search operations

CLEANSE

Data Profile Visualizations

FREQUENCY

STATISTICS

Column	Interquartile Range	Minimum	Maximum	Median	Standard Deviation
Pregnancies	5	0	17	3	3.36957806269886
Glucose	41.25	0	199	117	31.9726181951362
BloodPressure	18	0	122	72	19.3558071706448
SkinThickness	18	0	72	72	19.3558071706448

LOCATION Predicting\_Diabetes\_21-04-2022

DATA REFINERY FLOW NAME diabetes.csv\_flow

Enter a description of the Data Refinery flow

STEPS 1

DATA REFINERY FLOW OUTPUT

Location Predicting\_Diabetes\_21-04-2022/...

Data set name diabetes\_csv\_shaped

Fig: Reading profile of data through data refining

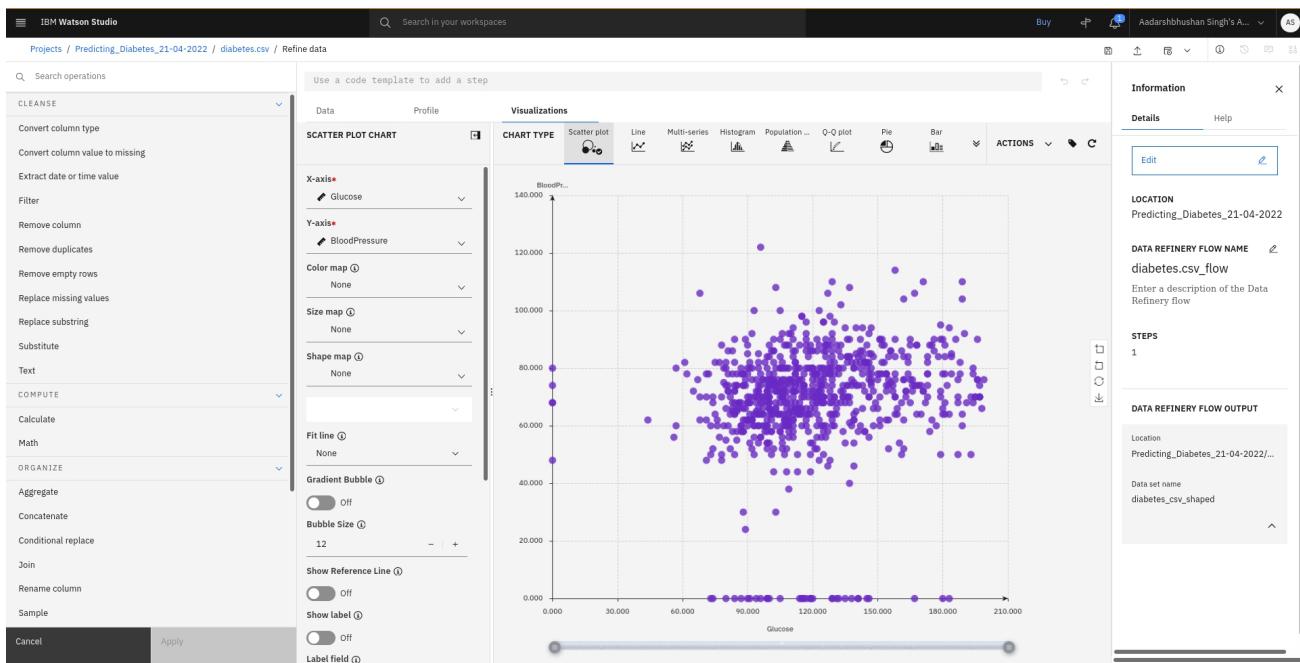


Fig: Visualising Glucose and Blood Pressure Data

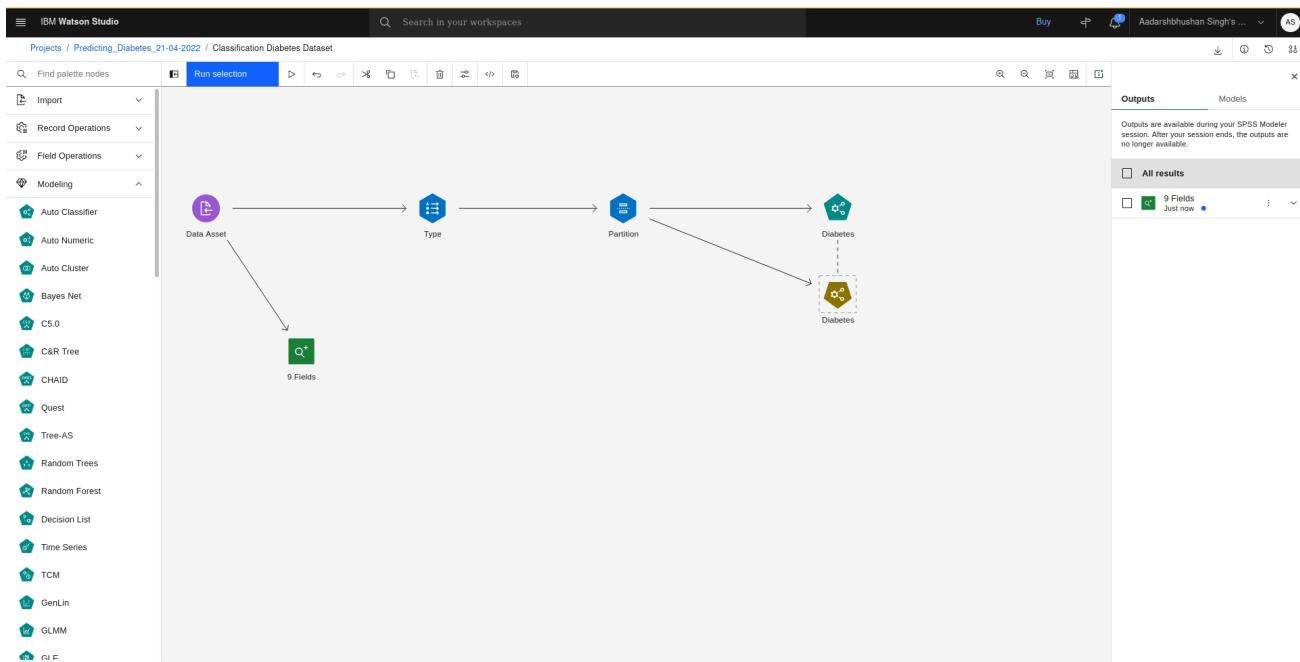


Fig: SPSS Modeler of diabetes

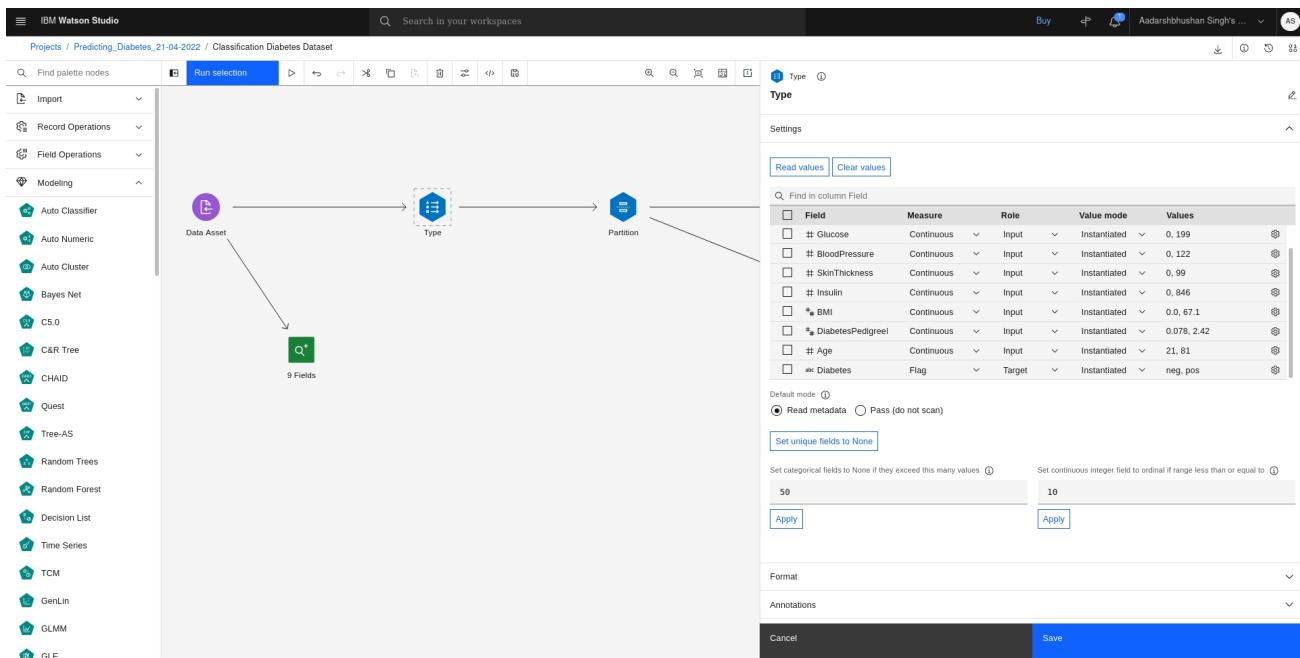


Fig: Setting the types of attribute

View Output: 9 Fields

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
1 Pregnancies		Continuous	0	17	3.845	3.370	0.902	--	768
2 Glucose		Continuous	0	199	120.895	31.973	0.174	--	768
3 BloodPressure		Continuous	0	122	69.105	19.356	-1.844	--	768
4 SkinThickness		Continuous	0	99	20.536	15.952	0.109	--	768
5 Insulin		Continuous	0	846	79.799	115.244	2.272	--	768
6 BMI		Continuous	0.000	67.100	31.993	7.884	-0.429	--	768
7 DiabetesPedigreeFunction		Continuous	0.078	2.420	0.472	0.331	1.920	--	768
8 Age		Continuous	21	81	33.241	11.760	1.130	--	768
9 Diabetes		Categorical	--	--	--	--	--	2	768

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
1 Pregnancies	Continuous	0	768	None	Never	Fixed	100.000	768	0	0	0	0
2 Glucose	Continuous	0	768	None	Never	Fixed	100.000	768	0	0	0	0

Fig: View Output of fields

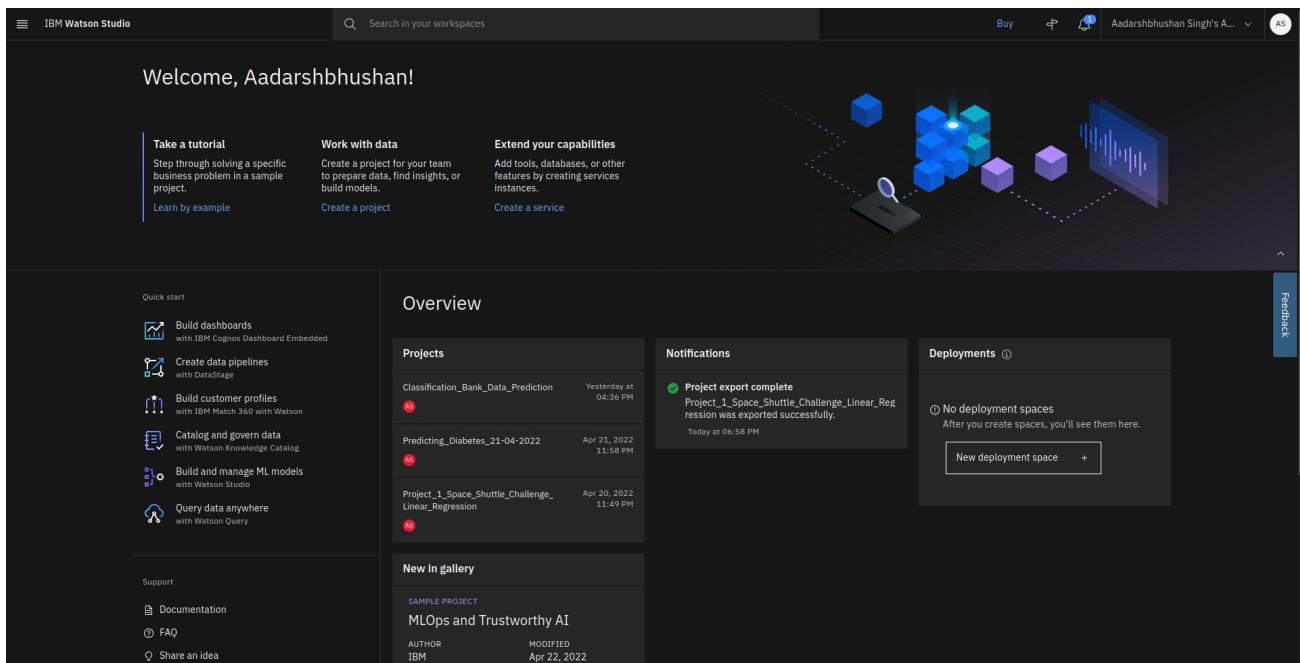
USE	MODEL NAME	ESTIMATOR	BUILD TIME (MINS)	NO. FIELDS USED	ACCURACY	ACCUMULATED ACCURACY	AREA UNDER CURVE	ACCUMULATED AUC	RECALL	PRECISION	ACTIONS
<input checked="" type="checkbox"/>	Logistic_Regression_1	Nominal Regression	< 1	8	68.293	68.293	0.779	0.779	0.412	0.700	
<input checked="" type="checkbox"/>	Discriminant_1	Discriminant	< 1	8	73.171	73.171	0.776	0.776	0.588	0.714	
<input checked="" type="checkbox"/>	Tree-AS_1	CHAID	< 1	4	73.171	73.171	0.766	0.766	0.471	0.800	
<input checked="" type="checkbox"/>	CHAID_1	CHAID	< 1	4	73.171	73.171	0.749	0.749	0.529	0.750	
<input checked="" type="checkbox"/>	C5_1	C5.0	< 1	7	69.512	69.512	0.712	0.712	0.529	0.667	

Fig: Result of Auto Classifier

## Dataset Name: insurance.csv

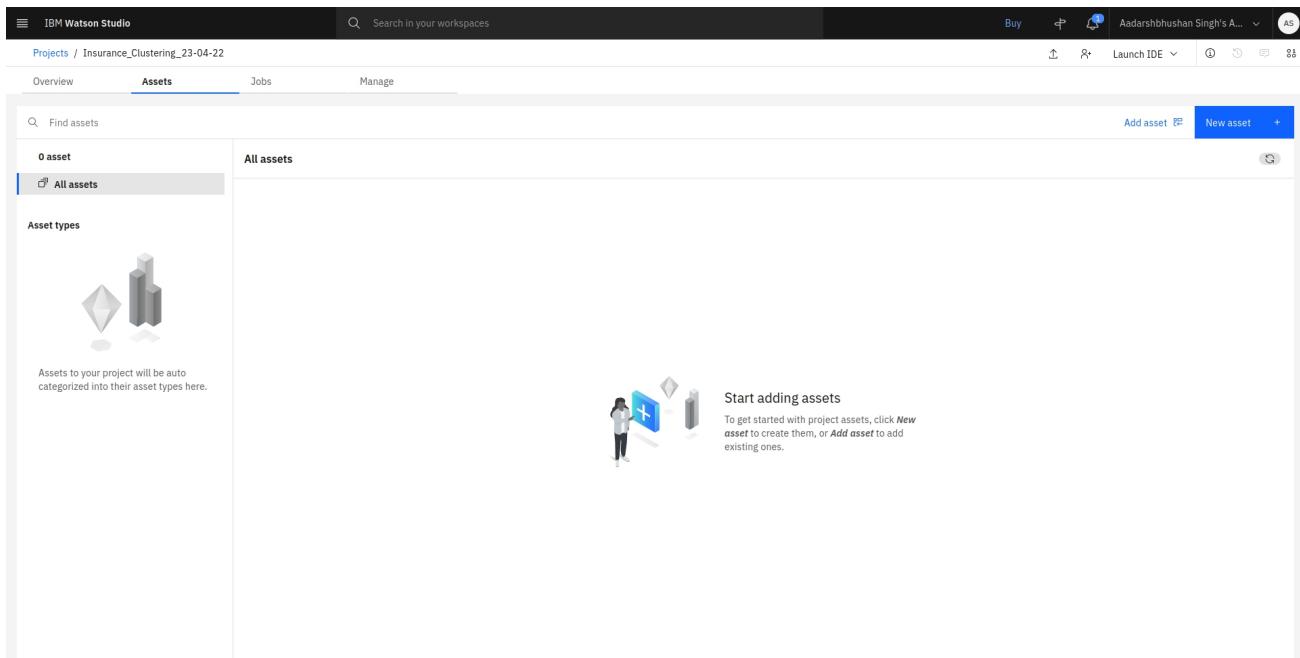
**Link to the dashboard:**

[https://eu-gb.dataplatform.cloud.ibm.com/canvas/flows/3162bb54-7f3c-4fce-a6a8-da074a9ebc93?context=cpdaas&projectGuid=6aa29757-cd4e-4b19-9b87-d6e390c25342&env\\_id=mods-6aa29757-cd4e-4b19-9b87-d6e390c25342](https://eu-gb.dataplatform.cloud.ibm.com/canvas/flows/3162bb54-7f3c-4fce-a6a8-da074a9ebc93?context=cpdaas&projectGuid=6aa29757-cd4e-4b19-9b87-d6e390c25342&env_id=mods-6aa29757-cd4e-4b19-9b87-d6e390c25342)



The screenshot shows the IBM Watson Studio interface. At the top, there's a navigation bar with 'IBM Watson Studio', a search bar, and user information 'Aadarshbhushan Singh's A...'. Below the header is a banner with the text 'Welcome, Aadarshbhushan!' and three sections: 'Take a tutorial', 'Work with data', and 'Extend your capabilities'. The 'Work with data' section includes links to 'Create a project' and 'Create a service'. To the right is a 3D visualization of data cubes and a graph. The main area is titled 'Overview' and contains sections for 'Projects', 'Notifications', and 'Deployments'. The 'Projects' section lists three projects: 'Classification\_Bank\_Data\_Prediction', 'Predicting\_Diabetes\_21-04-2022', and 'Project\_1\_Space\_Shuttle\_Challenge\_Linear\_Regression'. The 'Notifications' section shows a success message: 'Project export complete Project\_1\_Space\_Shuttle\_Challenge\_Linear\_Reg' was exported successfully. The 'Deployments' section indicates 'No deployment spaces' and has a button for 'New deployment space'. On the left sidebar, under 'Quick start', there are links for building dashboards, creating data pipelines, building customer profiles, cataloging data, managing ML models, and querying data. Under 'Support', there are links for documentation, FAQ, and sharing ideas.

Fig: Create New Project



The screenshot shows the 'Assets' tab in the IBM Watson Studio interface for the project 'Insurance\_Clustering\_23-04-22'. The top navigation bar includes 'Buy', 'Aadarshbhushan Singh's A...', and a 'Feedback' button. The 'Assets' tab is selected, showing tabs for 'Overview', 'Assets', 'Jobs', and 'Manage'. The 'Assets' tab has a search bar 'Find assets' and a button 'Add asset'. The main area is divided into two sections: 'Asset types' on the left and 'All assets' on the right. The 'Asset types' section shows icons for a diamond and a bar chart, with the text 'Assets to your project will be auto categorized into their asset types here.' The 'All assets' section features a 'Start adding assets' callout with an icon of a person interacting with a screen, and the text 'To get started with project assets, click **New asset** to create them, or **Add asset** to add existing ones.'

Fig: Go to Asset and Click on new asset

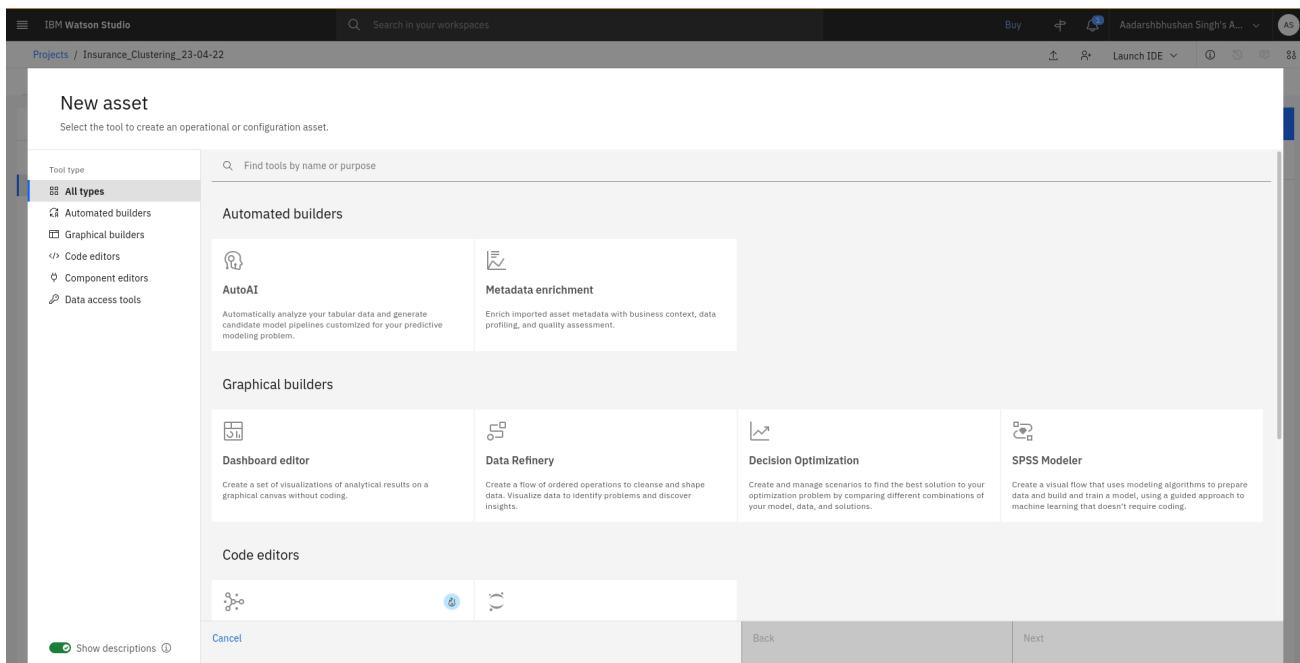


Fig: Data Refinery

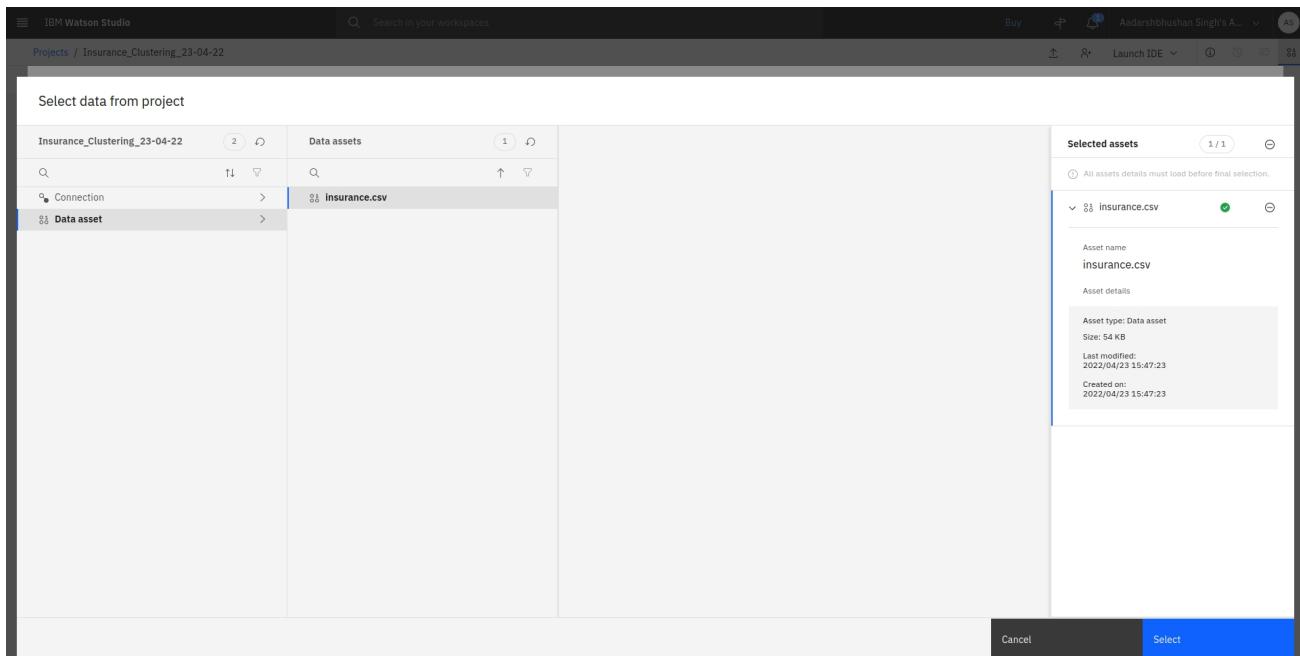


Fig: Select the dataset

IBM Watson Studio

Search in your workspaces

Buy Aadarshbhusan Singh's A... AS

Projects / Insurance\_Clustering\_23-04-22 / insurance.csv / Refine data

Steps (1)

Data Source: insurance.csv

1. Convert column type

Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.

Auto-generated

New step +

Use a code template to add a step

Data Profile Visualizations

	age	sex	bmi	children	smoker	region	premium
1	19	female	27.9	0	yes	southwest	16884.924
2	18	male	33.77	1	no	southeast	1725.5523
3	28	male	33	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.47061
5	32	male	28.88	0	no	northwest	3866.8552
6	31	female	25.74	0	no	southeast	3756.6216
7	46	female	33.44	1	no	southeast	8240.5896
8	37	female	27.74	3	no	northwest	7281.5056
9	37	male	29.83	2	no	northeast	6406.4107
10	60	female	25.84	0	no	northwest	28923.13692
11	25	male	26.22	0	no	northeast	2721.3208
12	62	female	26.29	0	yes	southeast	27808.7251
13	23	male	34.4	0	no	southwest	1826.843
14	56	female	39.82	0	no	southeast	11090.7178
15	27	male	42.13	0	yes	southeast	39611.7577
16	19	male	24.6	1	no	southwest	1837.237
17	52	female	30.78	1	no	northeast	10797.3362
18	23	male	23.845	0	no	northeast	2395.17155
19	56	male	40.3	0	no	southwest	10602.385
20	30	male	35.3	0	yes	southwest	36837.467
21	60	female	36.005	0	no	northeast	13228.84695

SOURCE FILE: insurance.csv FULL DATA SET: 1338 rows

Information

Details Help

Edit

LOCATION Insurance\_Clustering\_23-04-22

DATA REFINERY FLOW NAME insurance.csv\_flow

Enter a description of the Data Refinery flow

STEPS 1

DATA REFINERY FLOW OUTPUT

Location Insurance\_Clustering\_23-04-22/Da...

Data set name insurance\_csv\_shaped

Fig: Data Refining

IBM Watson Studio

Search in your workspaces

Buy Aadarshbhusan Singh's A... AS

Projects / Insurance\_Clustering\_23-04-22 / insurance.csv / Refine data

Steps (1)

Data Source: insurance.csv

1. Convert column type

Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.

Auto-generated

New step +

Use a code template to add a step

Data Profile Visualizations

age sex bmi children

FREQUENCY FREQUENCY FREQUENCY FREQUENCY

STATISTICS STATISTICS STATISTICS STATISTICS

	age	sex	bmi	children	
Interquartile Range	24	Maximum length	6	Interquartile Range	8.3975
Minimum	18	Minimum length	4	Minimum	15.96
Maximum	64	Mean length	4.98953662182362	Maximum	53.13
Median	39	Unique	2	Median	30.4
Standard Deviation	14.0499603792162			Standard Deviation	6.09818691167902

Information

Details Help

Edit

LOCATION Insurance\_Clustering\_23-04-22

DATA REFINERY FLOW NAME insurance.csv\_flow

Enter a description of the Data Refinery flow

STEPS 1

DATA REFINERY FLOW OUTPUT

Location Insurance\_Clustering\_23-04-22/Da...

Data set name insurance\_csv\_shaped

Fig: Observing data through data profile

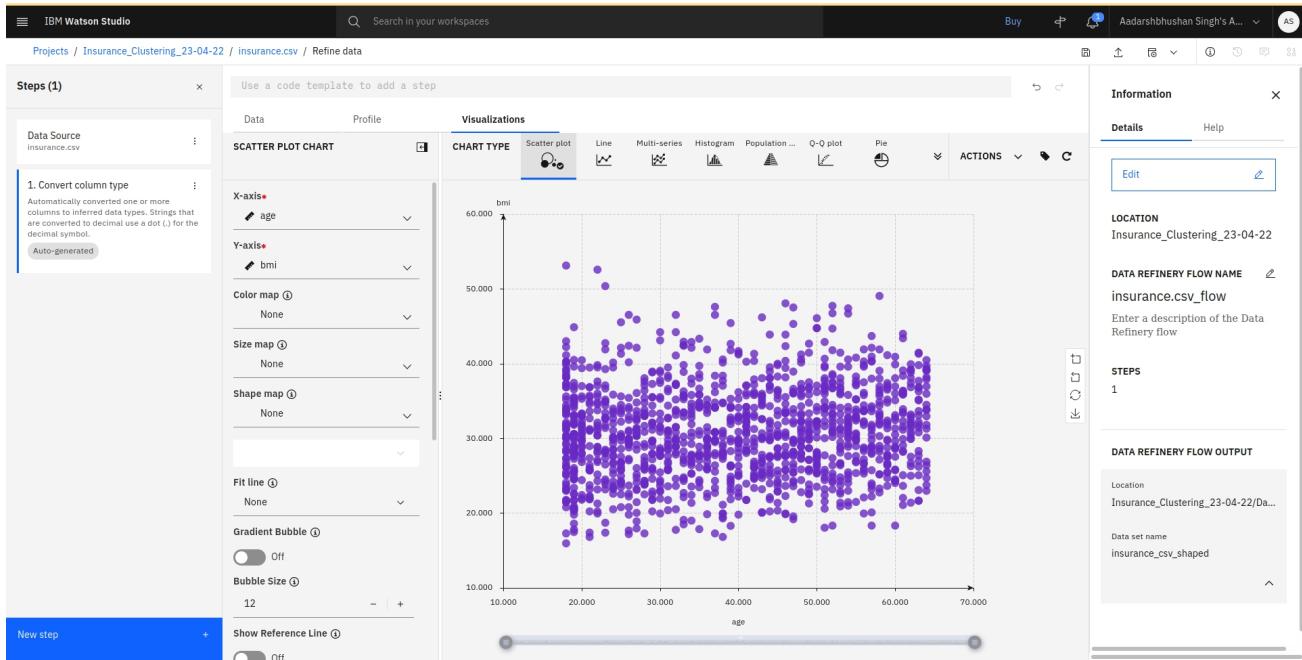


Fig: Observing bmi vs age graph

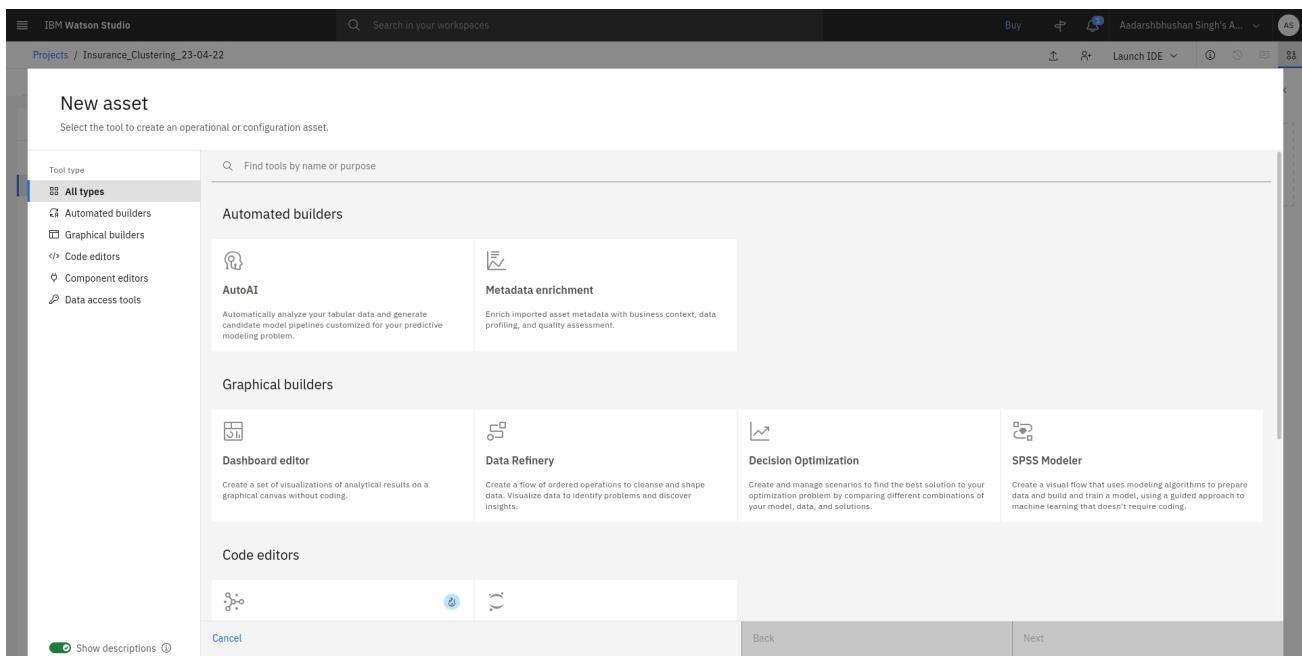


Fig: Creating new asset and selecting SPSS Modeler

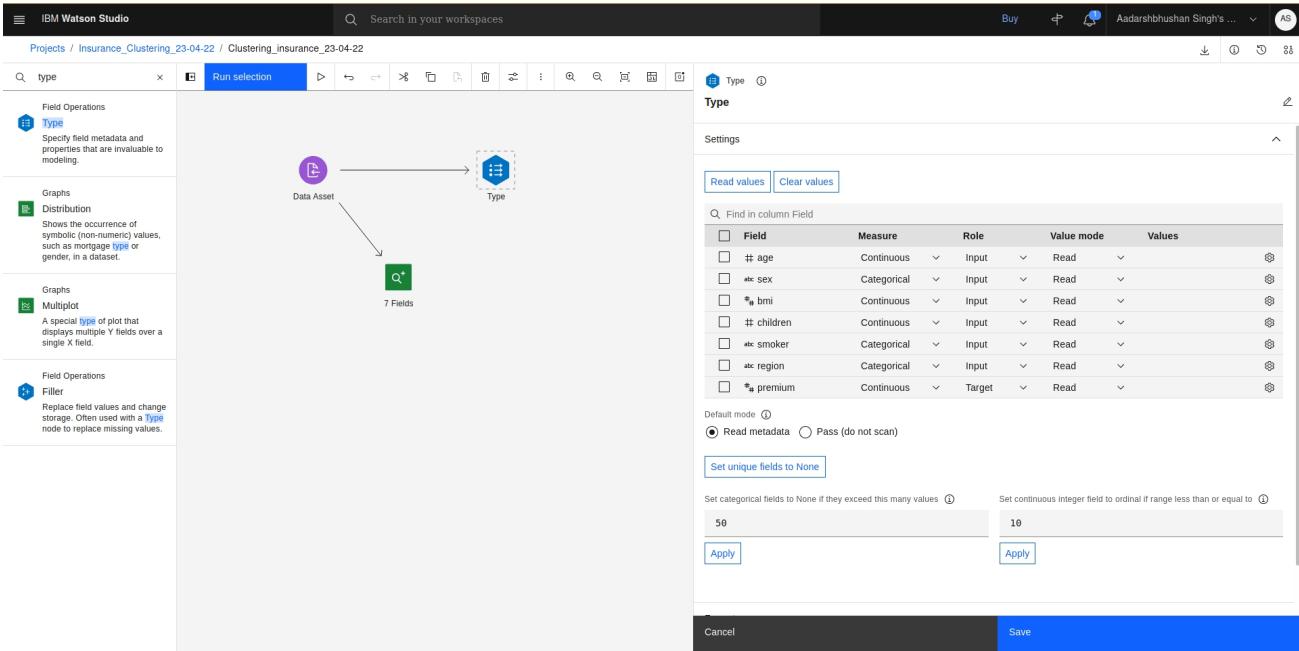


Fig: Setting up target and input attribute

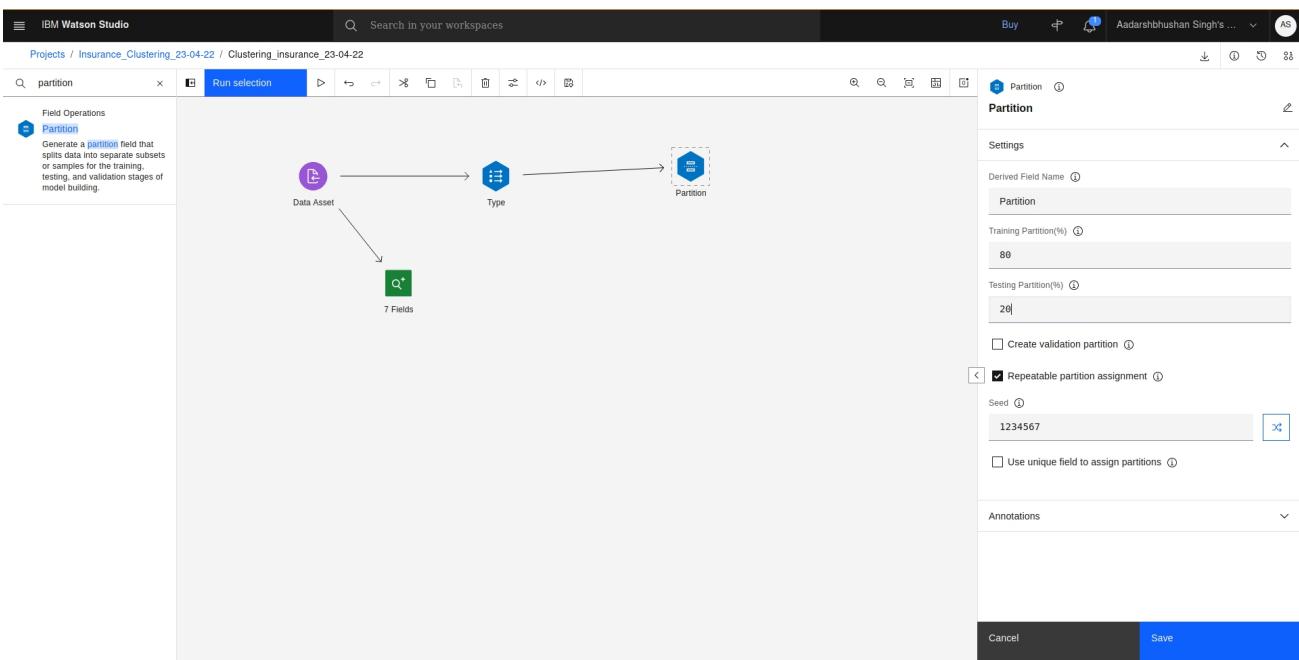


Fig: Creating partition

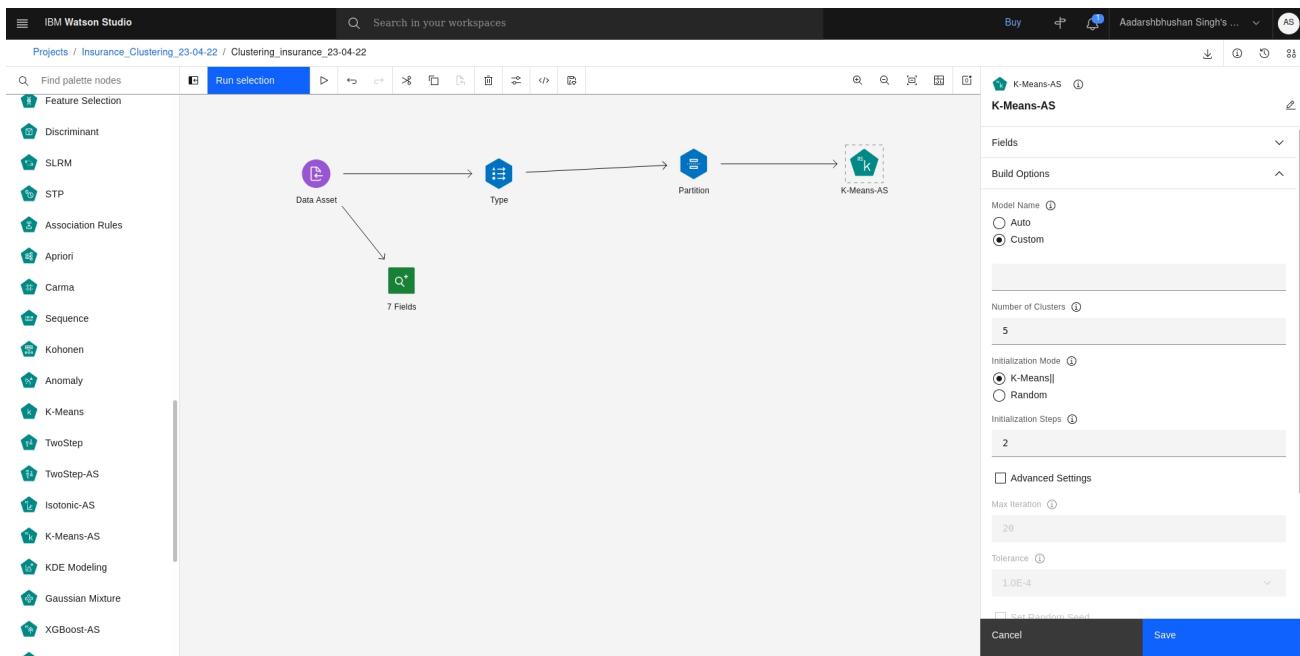


Fig: Selecting Number of clusters

**View Output: Data Audit of [7 fields]**

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
1 age		Continuous	18	64	39.207	14.050	0.056	--	1338
2 sex		Categorical	--	--	--	--	--	2	1338
3 bmi		Continuous	15.960	53.130	30.663	6.098	0.284	--	1338
4 children		Continuous	0	5	1.095	1.205	0.938	--	1338
5 smoker		Categorical	--	--	--	--	--	2	1338
6 region		Categorical	--	--	--	--	--	4	1338
7 premium		Continuous	1121.874	63770.428	13270.422	12110.011	1.516	--	1338

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
1 age	Continuous	0	0	None	Never	Fixed	100.000	1338	0	0	0	0
2 sex	Categorical	--	--	--	Never	Fixed	100.000	1338	0	0	0	0
3 bmi	Continuous	4	0	None	Never	Fixed	100.000	1338	0	0	0	0

Fig: Data Audit output

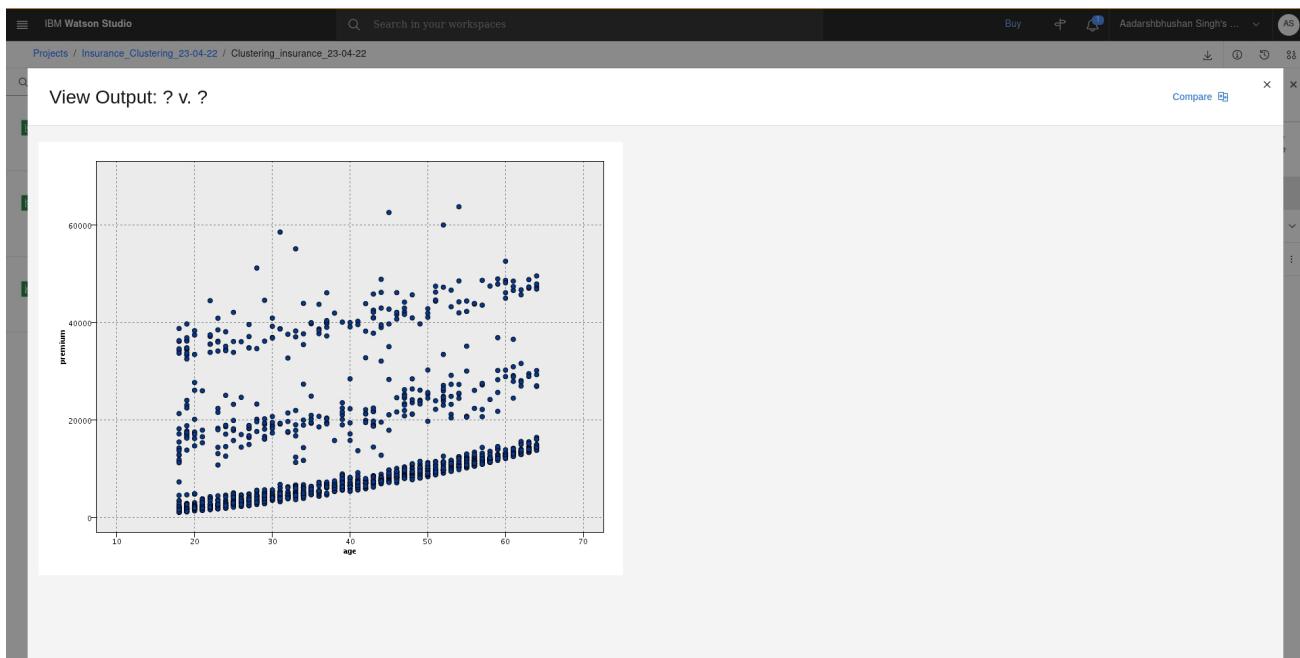


Fig: Ploting of Premium and age graph after clustering

## Dataset Name: Mall\_Customers.csv

**Link to the dashboard:**

[https://eu-gb.dataplatform.cloud.ibm.com/canvas/flows/8e0077ad-828d-4e1d-9495-09ad62a41229?context=cpdaas&projectGuid=ccdac624-e9ea-47b8-b33f-348063a32f87&env\\_id=mods-ccdac624-e9ea-47b8-b33f-348063a32f87](https://eu-gb.dataplatform.cloud.ibm.com/canvas/flows/8e0077ad-828d-4e1d-9495-09ad62a41229?context=cpdaas&projectGuid=ccdac624-e9ea-47b8-b33f-348063a32f87&env_id=mods-ccdac624-e9ea-47b8-b33f-348063a32f87)

Flg: Creating new project

The screenshot shows the IBM Watson Studio interface. The top navigation bar includes 'IBM Watson Studio', a search bar 'Search in your workspaces', and user information 'Aadarshbhushan Singh's A...'. Below the navigation is a toolbar with icons for 'Buy', 'Launch IDE', and other workspace management options. The main area is titled 'Projects / Mall\_Customers\_Clustering' and features a 'Assets' tab selected. On the left, a sidebar shows '1 assets' under 'Data' and 'Asset types'. The main content area displays a table titled 'Data' with one item: 'Name' (Mall\_Customers.csv), 'Last modified' (Now), and 'Created by' (Aadarshbhushan Singh (You)). At the bottom, there are pagination controls for 'Items per page: 20' and '1 of 1 pages'.

Fig: Adding dataset Mall Customers

This screenshot shows a modal dialog titled 'Select data from project' within the IBM Watson Studio interface. The dialog has three main sections: 'Project' (containing 'Mall\_Customers\_Clustering'), 'Data assets' (containing 'Mall\_Customers.csv'), and 'Selected assets' (containing 'Mall\_Customers.csv'). The 'Selected assets' section includes detailed asset information: Asset name 'Mall\_Customers.csv', Asset type 'Data asset', Size '4 KB', Last modified '2022/04/23 16:23:23', and Created on '2022/04/23 16:23:23'. At the bottom right of the dialog are 'Cancel' and 'Select' buttons.

Fig: Select dataset for data refinery

IBM Watson Studio

Projects / Mail\_Customers\_Clustering / Mail\_Customers.csv / Refine data

Steps (1)

Data Source: Mail\_Customers.csv

1. Convert column type

Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.

Auto-generated

New step +

SOURCE FILE: Mail\_Customers.csv FULL DATA SET: 200 rows

Information

Details Help

LOCATION: Mail\_Customers\_Clustering

DATA REFINERY FLOW NAME: Mail\_Customers.csv\_flow

Enter a description of the Data Refinery flow

STEPS: 1

DATA REFINERY FLOW OUTPUT

Location: Mail\_Customers\_Clustering/Data as...

Data set name: Mail\_Customers\_csv\_shaped

Fig: Data Refining

IBM Watson Studio

Projects / Mail\_Customers\_Clustering / Mail\_Customers.csv / Refine data

Steps (1)

Data Source: Mail\_Customers.csv

1. Convert column type

Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.

Auto-generated

New step +

SOURCE FILE: Mail\_Customers.csv FULL DATA SET: 200 rows

Information

Details Help

LOCATION: Mail\_Customers\_Clustering

DATA REFINERY FLOW NAME: Mail\_Customers.csv\_flow

Enter a description of the Data Refinery flow

STEPS: 1

DATA REFINERY FLOW OUTPUT

Location: Mail\_Customers\_Clustering/Data as...

Data set name: Mail\_Customers\_csv\_shaped

Fig: Understanding data through profile

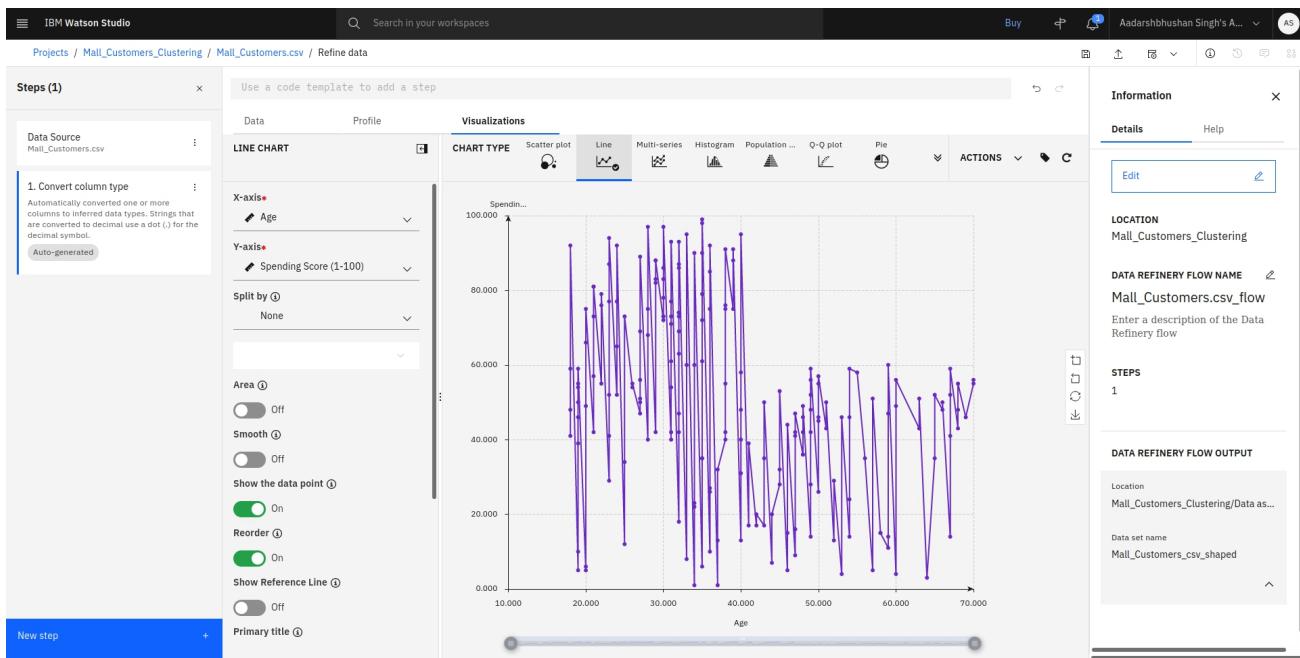


Fig: Visualising data through graph: Age vs Spending score

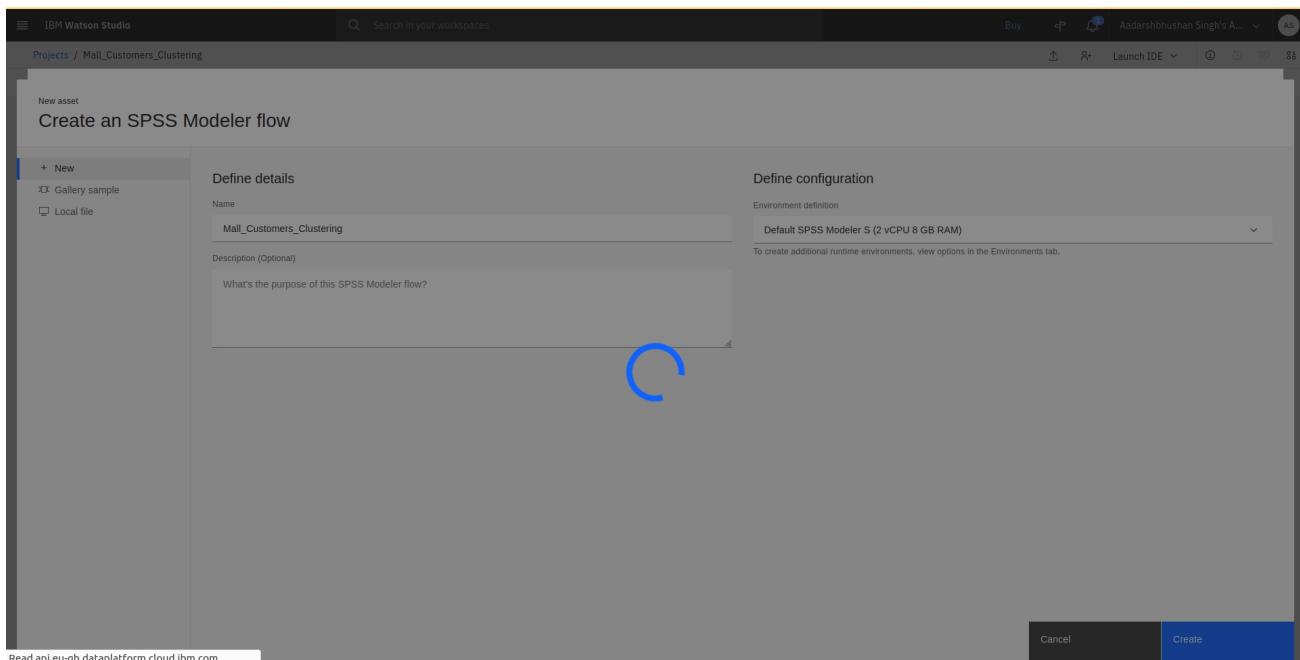


Fig: Creating SPSS Modeler

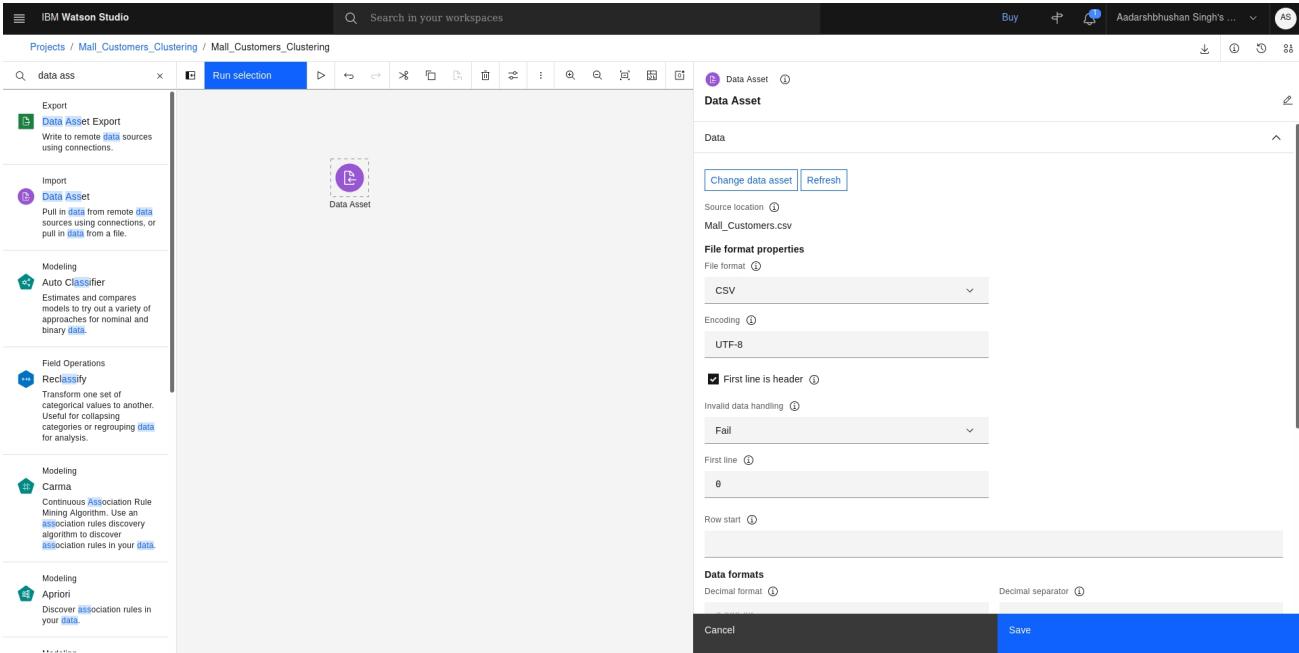


Fig: Selecting Data Asset

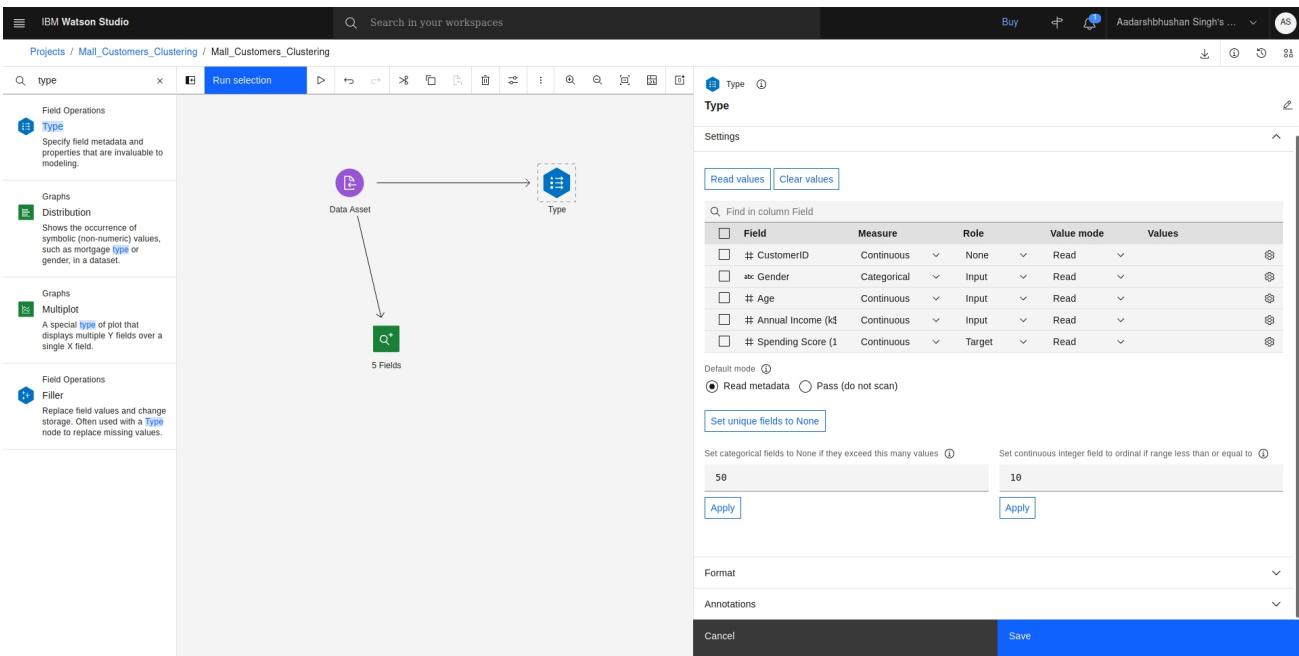


Fig: Fixing the target and input attribute

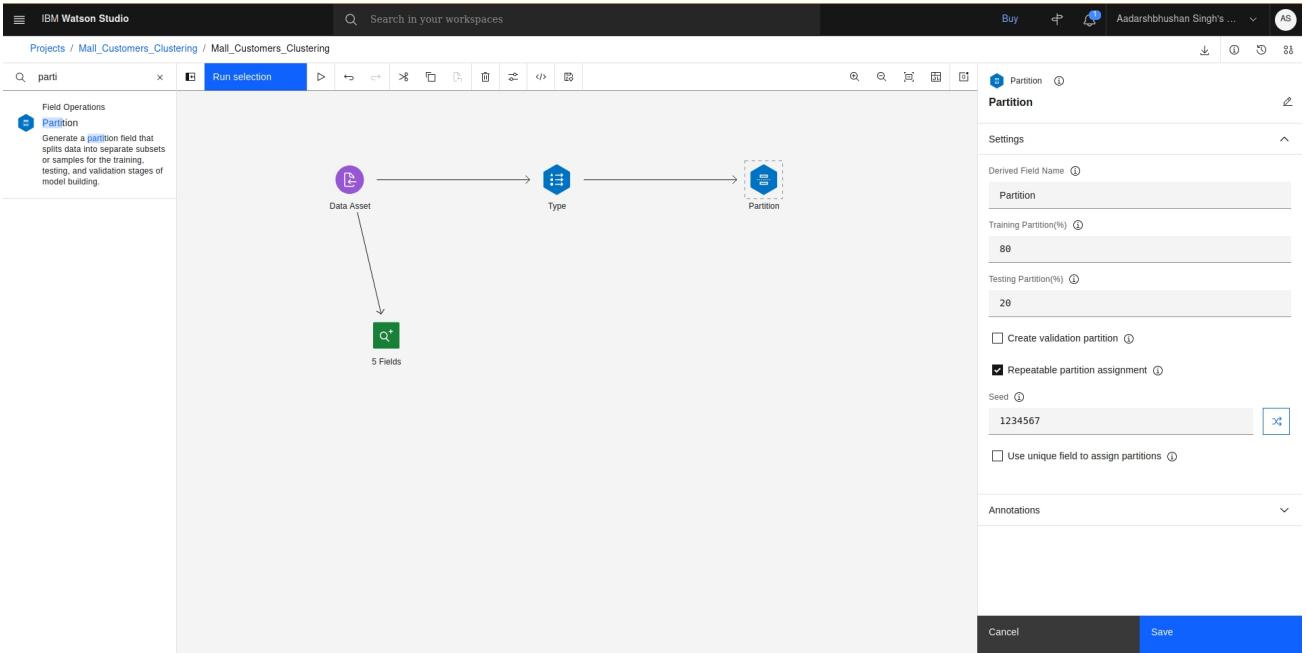
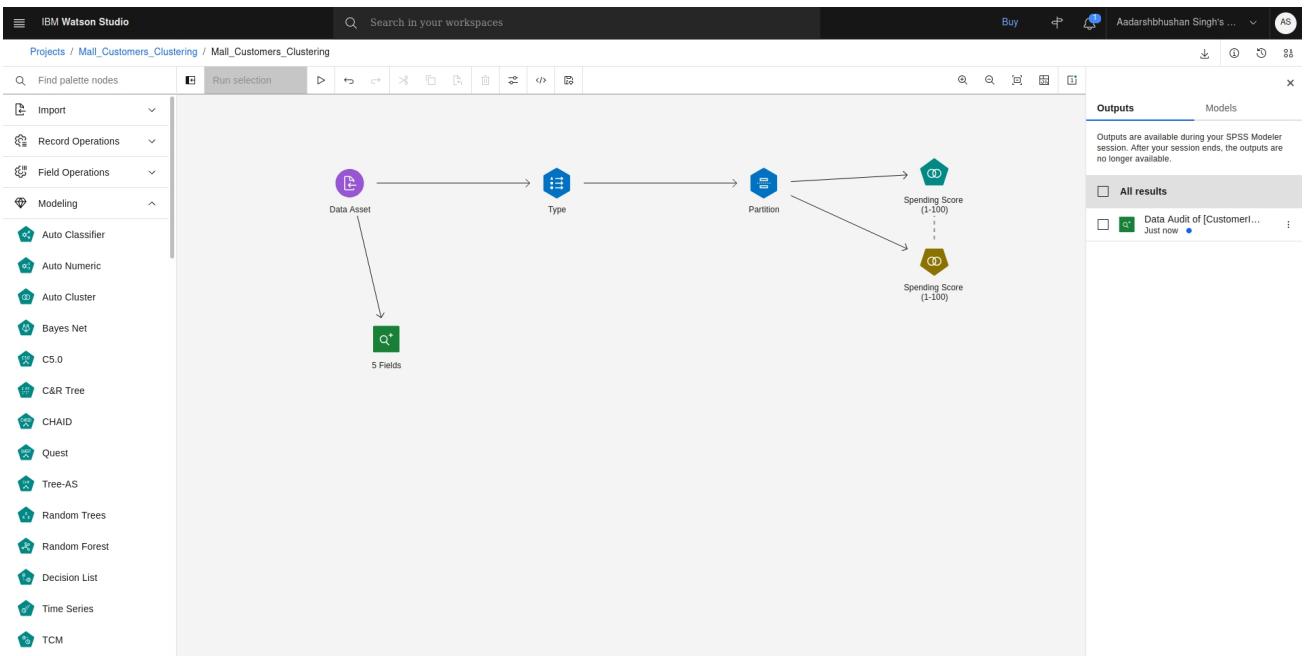


Fig: Setting the testing and training attribute



Final SPSS Modeler

IBM Watson Studio    Search in your workspaces    Buy    Aadarshbhusan Singh's ...    AS

Projects / Mall\_Customers\_Clustering / Mall\_Customers\_Clustering

View Output: Data Audit of [CustomerID Gender Age Annual Income (k\$) Spending Score (1-100)]    Compare

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
1 CustomerID		Continuous	1	200	100.500	57.879	0	--	200
2 Gender		Categorical	--	--	--	--	--	2	200
3 Age		Continuous	18	70	38.850	13.969	0.486	--	200
4 Annual Income (k\$)		Continuous	15	137	60.560	26.265	0.322	--	200
5 Spending Score (1-100)		Continuous	1	99	50.200	25.824	-0.047	--	200

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
1 CustomerID	Continuous	0	0	None	Never	Fixed	100.000	200	0	0	0	0
2 Gender	Categorical	--	--	--	Never	Fixed	100.000	200	0	0	0	0
3 Age	Continuous	0	0	None	Never	Fixed	100.000	200	0	0	0	0
4 Annual Income (k\$)	Continuous	0	0	None	Never	Fixed	100.000	200	0	0	0	0
5 Spending Score (1-100)	Continuous	0	0	None	Never	Fixed	100.000	200	0	0	0	0

Fig: Data Audit Output

IBM Watson Studio    Search in your workspaces    Buy    Aadarshbhusan Singh's ...    AS

Projects / Mall\_Customers\_Clustering / Mall\_Customers\_Clustering

View Model: Spending Score (1-100)

Auto Cluster    Auto Cluster - Models

Models

USE	MODEL NAME	ESTIMATOR	GRAPH	SILHOUETTE	BUILD TIME (MINS)	NUMBER OF CLUSTERS	SMALLEST CLUSTER (N)	SMALLEST CLUSTER (%)	LARGEST CLUSTER (N)	LARGEST CLUSTER (%)	SMALLEST/LARGEST	IMPORTANCE	AC
○	Kohonen_1	Kohonen		0.414	< 1	8	9	0.061	35	0.238	0.257	0.044	
○	TwoStep_1	TwoStep		0.526	< 1	2	58	0.395	89	0.605	0.652	0.022	
○	K-means_1	KMeans		0.568	< 1	5	16	0.109	45	0.306	0.356	0.088	

Fig: Final Clustering

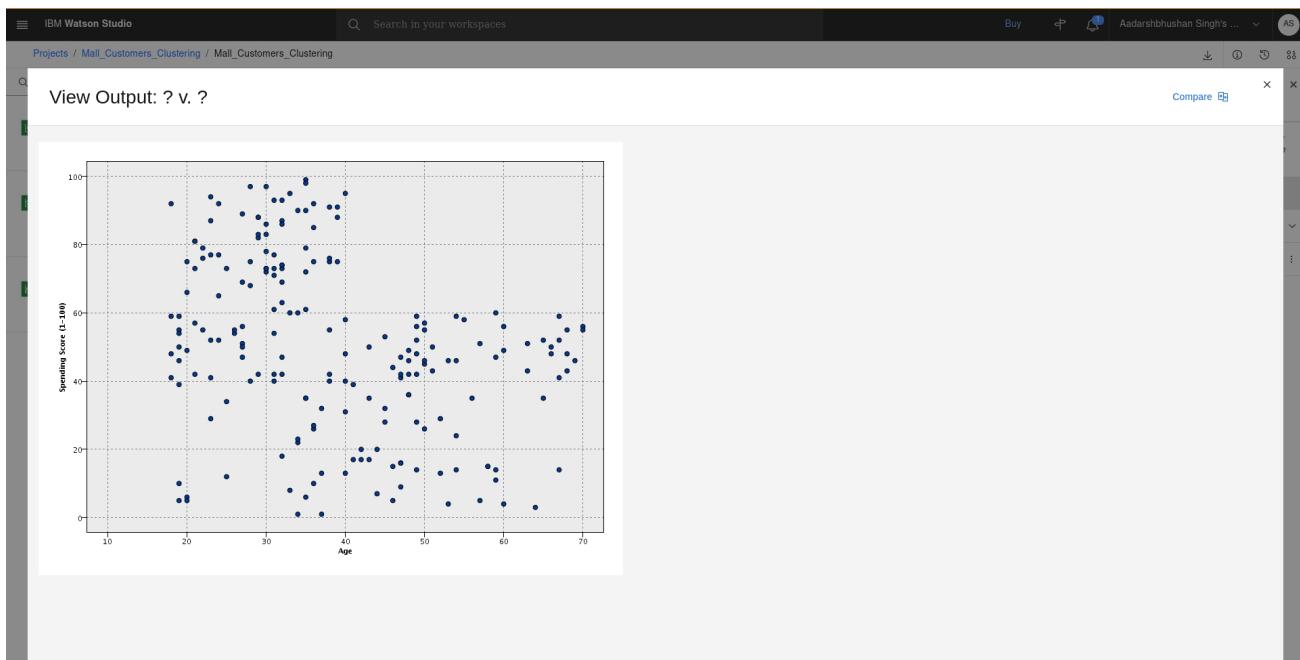


Fig: Data plotting of Spending score vs age