

Assignment 03

Data set: Diabetes.csv

Link: <https://dataplatform.cloud.ibm.com/projects/047cf7a-4874-4934-a9d1-a9f5d8071774/assets?context=cpdaas>

The screenshot shows the IBM Watson Studio interface. At the top, there are three tabs: 'SmartBridge', 'Service Details - IBM Cloud', and 'IBM Watson Studio'. The URL in the address bar is 'dataplatform.cloud.ibm.com/projects/047cf7a-4874-4934-a9d1-a9f5d8071774?context=cpdaas'. Below the tabs, there's a navigation bar with links like 'VTOP', 'CodeTantra Teach...', 'Moddle', 'CodeVIT', 'Forests And Their...', 'Front-End Web De...', 'HackWithInfy | Inf...', 'Select Exam Center', 'My home | eLitmu...', 'IBM Academic Init...', and a user account dropdown for 'Pranjal Gupta's Account'. The main area is titled 'Projects / Predictive_Analysis_diabetes'. It has four tabs: 'Overview' (which is selected), 'Assets', 'Jobs', and 'Manage'. The 'Overview' tab contains sections for 'Assets', 'Resource usage', and 'Project history'. The 'Assets' section says 'Assets that you create with tools show here. See data assets on the Assets page.' It features a small icon of a bar chart and a 'View all' link. The 'Resource usage' section shows '0 CUH' for the current month. The 'Project history' section says 'No notifications' and includes a note: 'You will see your most recent notifications here.'

Figure: Created the project in Watson Studio

This screenshot is identical to the one above, showing the 'Overview' tab of the 'Predictive_Analysis_diabetes' project in Watson Studio. The 'Assets' tab is now selected. The 'Assets' section displays the message 'Assets that you create with tools show here. See data assets on the Assets page.' and includes a 'View all' link. The 'Resource usage' and 'Project history' sections remain the same.

Figure: Adding the dataset

Pranjal Gupta

The screenshot shows the IBM Watson Studio interface. In the center, there is a data preview window titled "Previewing the first 50 rows" showing the first 50 rows of a CSV file named "diabetes.csv". The columns listed are Pregnancies, Glucose, BloodPress..., SkinThickn..., Insulin, BMI, and DiabetesPe. The data consists of various numerical values. To the right of the preview, there is a "Details" panel with tabs for "Edit" and "Help". The "Edit" tab is selected. It contains fields for "LOCATION" (set to "Predictive_Analysis_diabetes"), "DATA REFINERY FLOW NAME" (set to "diabetes.csv_flow"), and a description field. Below these are sections for "STEPS" (0) and "DATA REFINERY FLOW OUTPUT" (with a "Location" field set to "Predictive_Analysis_diabetes/Data as..."). At the bottom of the preview window, it says "SOURCE FILE: diabetes.csv SAMPLE SIZE: First 50 rows".

Figure: Created a data refinery asset

The screenshot shows the "Create an SPSS Modeler flow" dialog box. On the left, there is a sidebar with options "+ New", "Gallery sample", and "Local file". The main area is divided into two sections: "Define details" and "Define configuration". In "Define details", the "Name" field is filled with "Diabetes data modeler". In "Define configuration", the "Environment definition" dropdown is set to "Default SPSS Modeler S (2 vCPU 8 GB RAM)". A note below states: "To create additional runtime environments, view options in the Environments tab." At the bottom right, there are "Cancel" and "Create" buttons, with "Create" being highlighted in blue.

Figure: Creating SPSS modeler

Pranjal Gupta

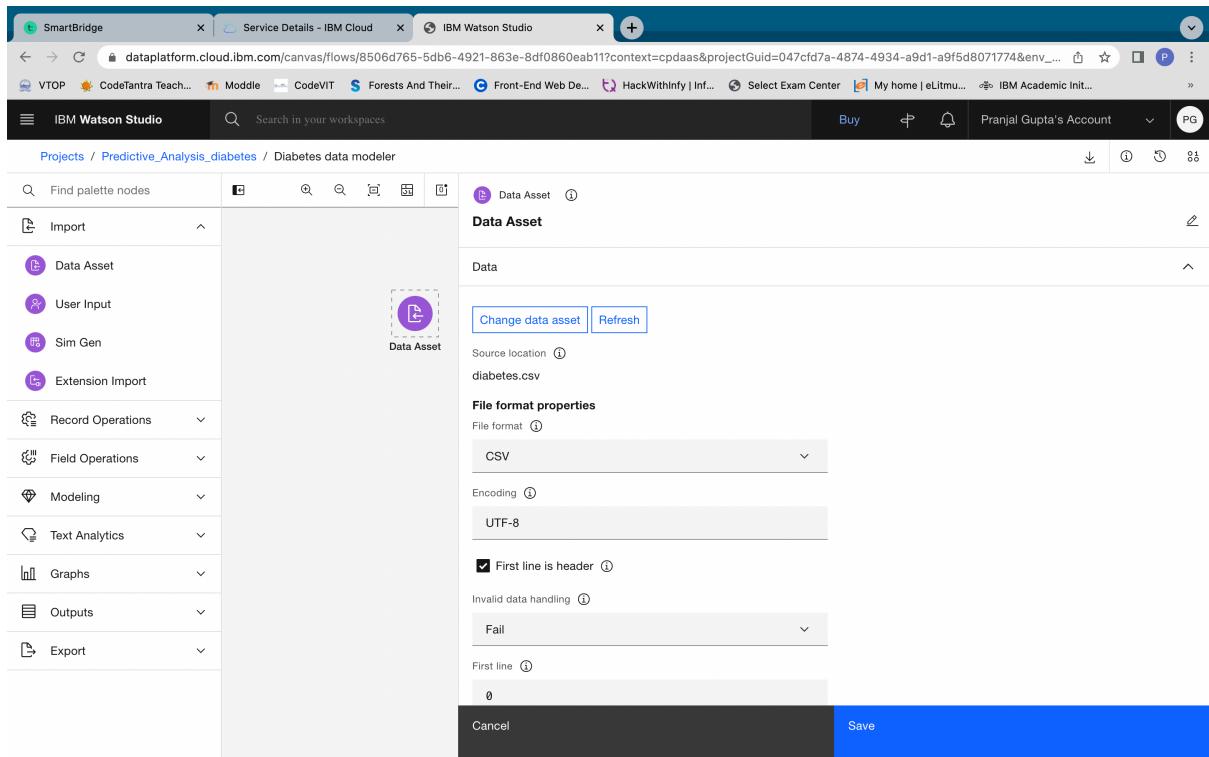


Figure: Imported Data asset

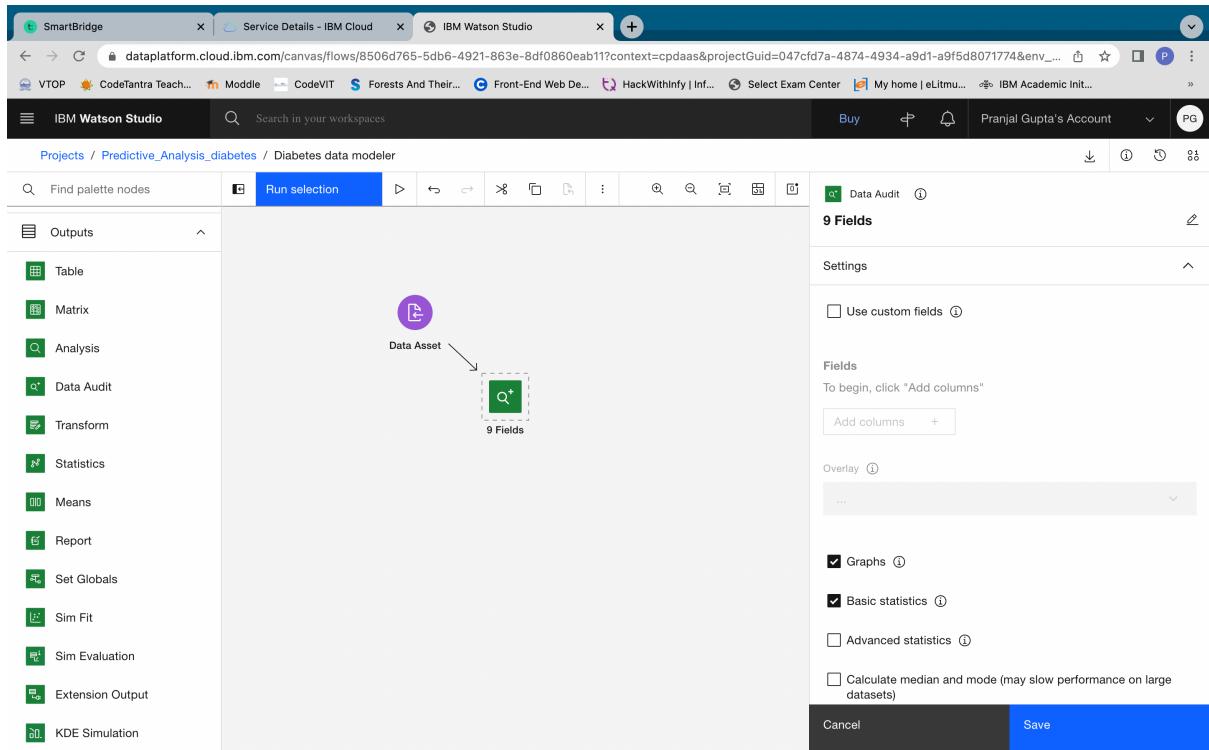


Figure: Performed Data audit

Pranjal Gupta

The screenshot shows the IBM Watson Studio interface for a project titled "Predictive_Analysis_diabetes". The current workspace is "Diabetes data modeler". On the left, there is a sidebar with various nodes categorized under "Field Operations" and "Graphs". The main panel displays a flow diagram where a "Data Asset" node (represented by a purple folder icon) is connected to a "Type" node (represented by a hexagonal icon). The "Type" node is then connected to a "Partition" node (represented by a blue hexagonal icon). Below the flow diagram, it is indicated that there are "9 Fields". To the right of the flow diagram, the "Type" node's configuration pane is open. It includes sections for "Type", "Settings", and "Values". Under "Values", there is a table with columns: Field, Measure, Role, Value mode, and Values. The table lists nine fields: # Pregnancies, # Glucose, # BloodPressure, # SkinThickness, # Insulin, *# BMI, *# DiabetesPedigreeF, and # Age. All fields are set to Continuous, Input, Read mode. At the bottom of the configuration pane, there are "Cancel" and "Save" buttons.

Figure: Type of attributes specified

The screenshot shows the IBM Watson Studio interface for the same project and workspace. The sidebar now lists a "Partition" node under "Field Operations". The main panel shows the same flow diagram: Data Asset → Type → Partition. The "Partition" node's configuration pane is open on the right. It includes sections for "Partition" and "Settings". Under "Settings", the "Derived Field Name" is set to "Partition". The "Training Partition(%)" is set to 70, and the "Testing Partition(%)" is set to 30. There are checkboxes for "Create validation partition", "Repeatable partition assignment" (which is checked), and "Use unique field to assign partitions". A "Seed" field contains the value 1234567. At the bottom of the configuration pane, there are "Cancel" and "Save" buttons.

Figure: Performing data partitioning

Pranjal Gupta

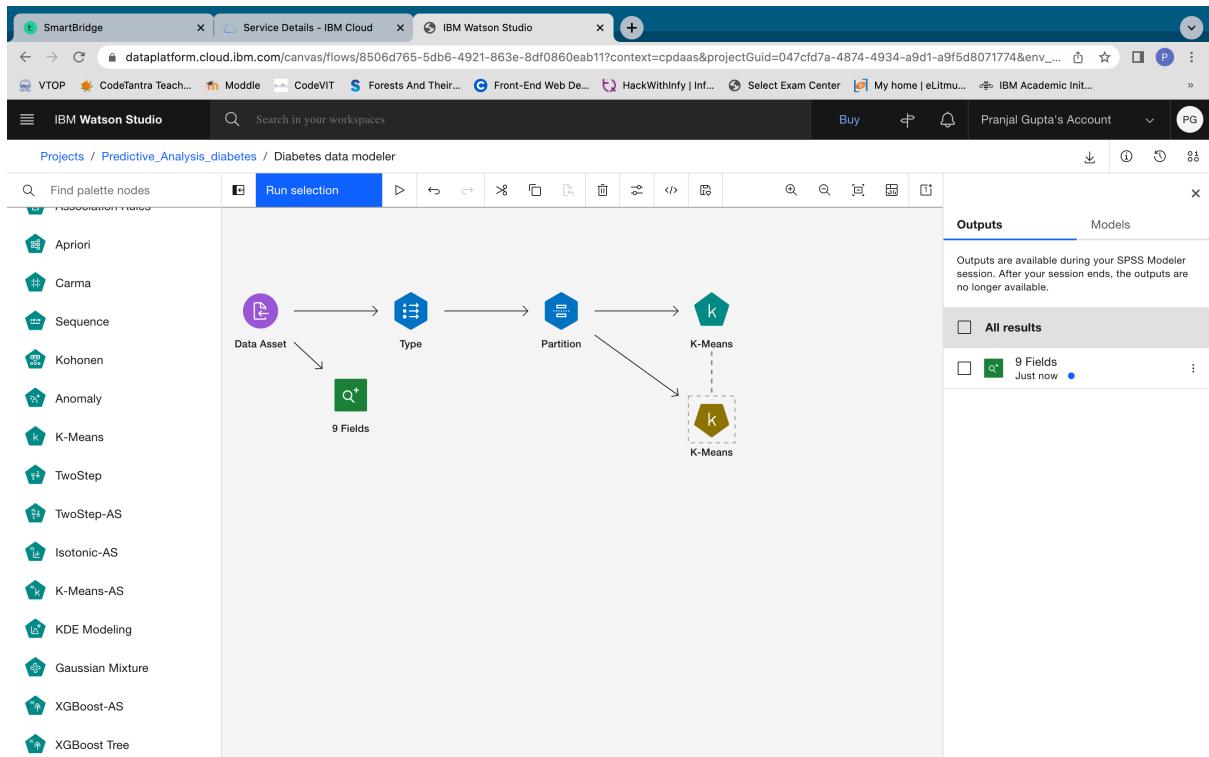


Figure: Performed K-Means Clustering

The screenshot shows the "View Model: K-Means" panel in IBM Watson Studio. The left sidebar lists model evaluation metrics: Cluster Quality, Feature Importance, Cluster Sizes, Cluster Comparison, Clusters, Cell Distributions (Absolute), Cell Distributions (Relative), Build Settings, and Training Summary. The "Model Information" section is currently selected. It displays the following details:

Distance Measure	Euclidean
Number of Clusters	5
Number of instances in each cluster	Cluster 1: 114 (21.15%) Cluster 2: 164 (30.43%) Cluster 3: 118 (21.89%) Cluster 4: 85 (15.77%) Cluster 5: 58 (10.76%)
Ratio of sizes (Largest to smallest)	2.828

Figure: Model Information

Pranjal Gupta

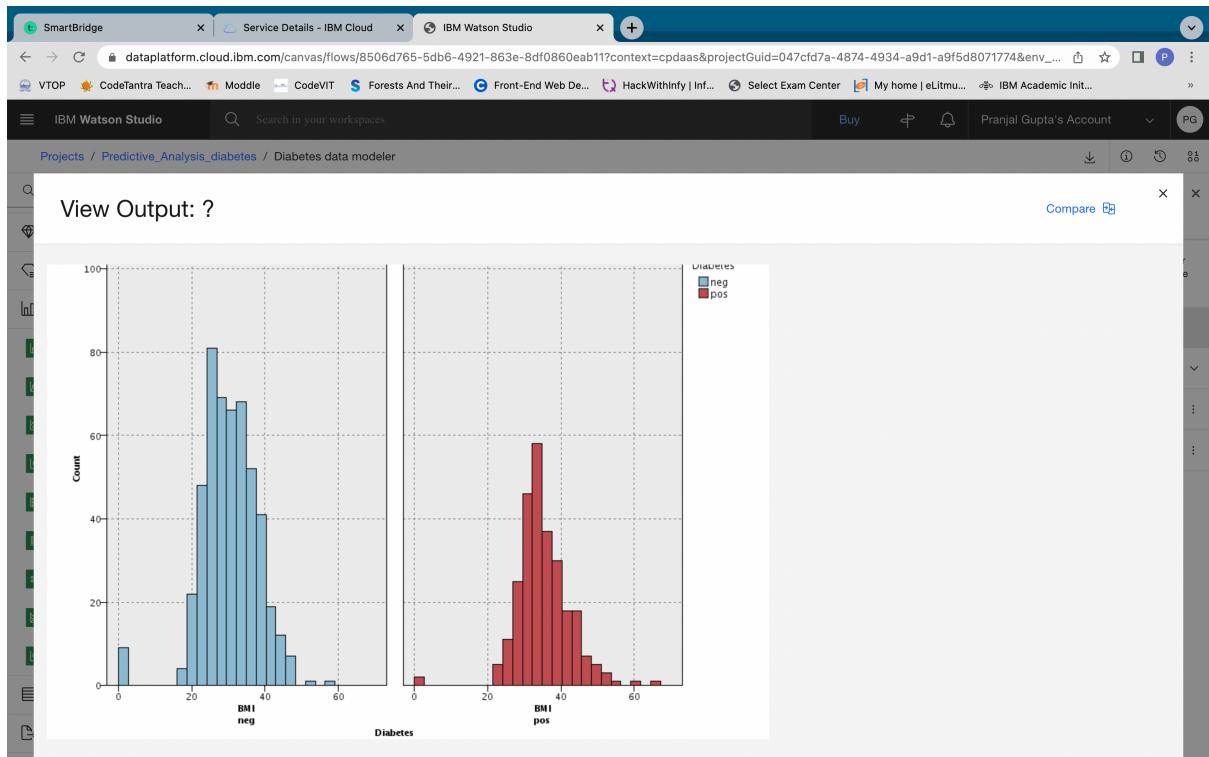


Figure: Histogram analysis

Data Set: Insurance.csv

Link: <https://dataplatform.cloud.ibm.com/projects/dc4c292b-f4c9-47bd-a802-c854178b2449/assets?context=cpdaas>

The screenshot shows the IBM Watson Studio interface. At the top, there are three tabs: 'SmartBridge', 'Service Details - IBM Cloud', and 'IBM Watson Studio'. The URL in the address bar is <https://dataplatform.cloud.ibm.com/projects/dc4c292b-f4c9-47bd-a802-c854178b2449?context=cpdaas>. The main header says 'IBM Watson Studio' with a search bar and a user account dropdown for 'Pranjal Gupta's Account'. Below the header, the page title is 'Projects / Predictive_Analysis_Insurance'. There are four tabs at the top of the main content area: 'Overview' (which is selected), 'Assets', 'Jobs', and 'Manage'. The 'Overview' tab has three sections: 'Assets' (with a note about creating assets and a 'View all' button), 'Resource usage' (showing 0 CUH), and 'Project history' (with a note about notifications). A sidebar on the left shows recent projects: VTOP, CodeTantra Teach..., Moddle, CodeVIT, Forests And Their..., Front-End Web De..., HackWithInfy | Inf..., Select Exam Center, My home | eLitmu..., and IBM Academic Init... .

Figure: Created project for Insurance data set

This screenshot is identical to the one above, showing the 'Predictive_Analysis_Insurance' project in IBM Watson Studio. The interface, tabs, and sidebar are the same. The main content area shows the 'Overview' tab with its three sections: Assets, Resource usage, and Project history. The 'Assets' section includes a note about creating assets and a 'View all' button. The 'Resource usage' section shows 0 CUH. The 'Project history' section notes that there are no notifications. The sidebar on the left lists the same recent projects.

Figure: Adding insurance data set

Pranjal Gupta

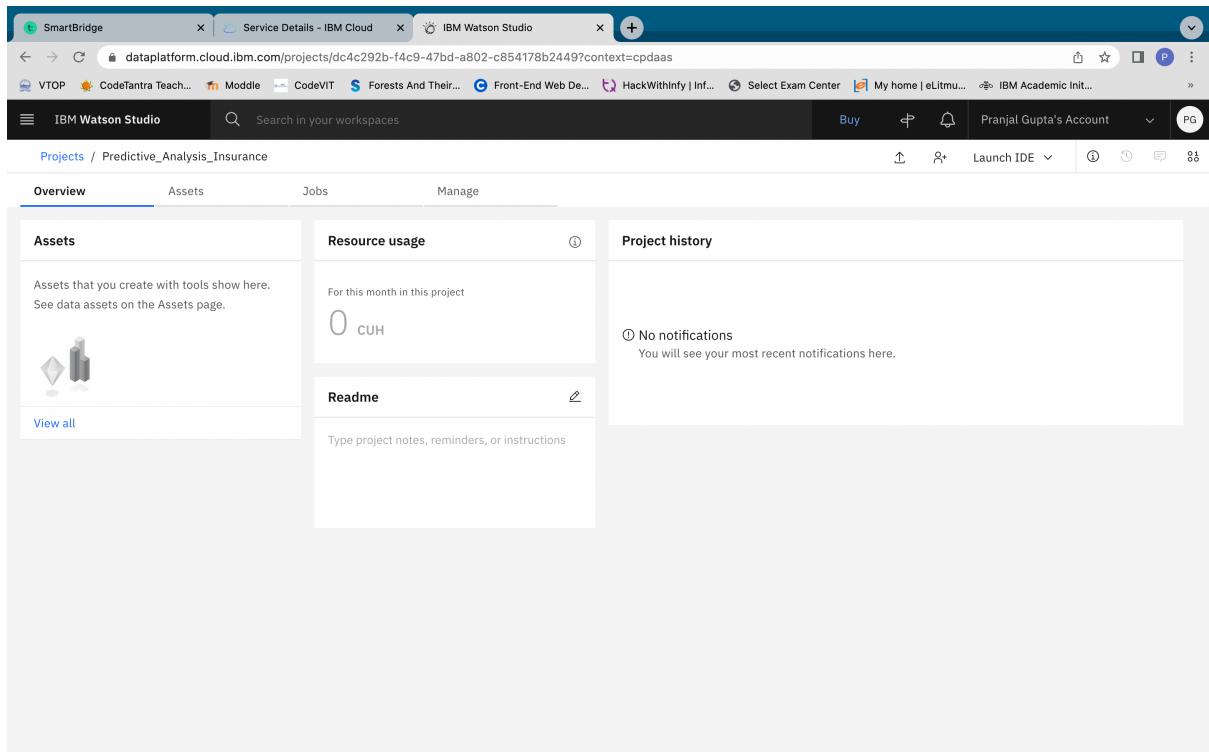


Figure: Created a data refinery flow

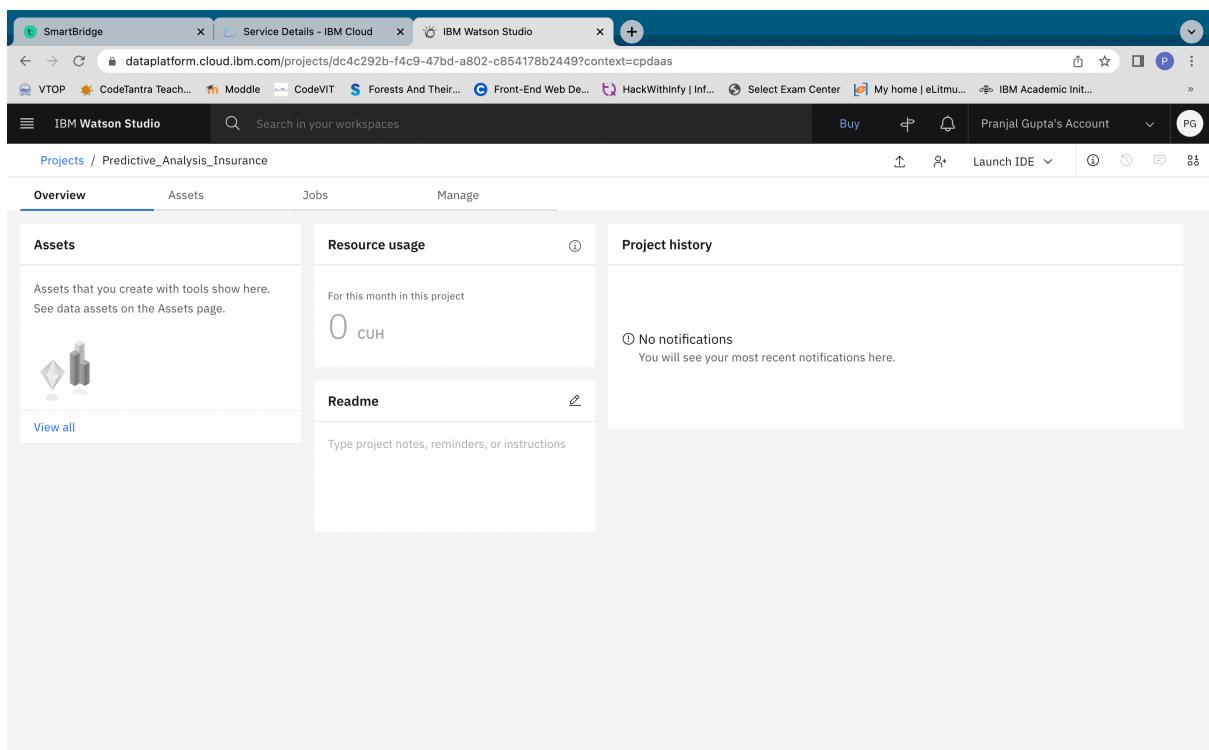


Figure: Creating a SPSS modeler for insurance data set

Pranjal Gupta

The screenshot shows the IBM Watson Studio interface. The top navigation bar includes tabs for SmartBridge, Service Details - IBM Cloud, and IBM Watson Studio. The main header displays the URL dataplatform.cloud.ibm.com/projects/dc4c292b-f4c9-47bd-a802-c854178b2449?context=cpdaas. The user's account, "Pranjal Gupta's Account", is visible in the top right. The main content area is titled "Overview" and shows sections for "Assets", "Resource usage", and "Project history". The "Assets" section indicates there are no assets yet. The "Resource usage" section shows 0 CUH (Cloud Unit Hours) for the current month. The "Project history" section shows 0 notifications.

Figure: Importing data to the data asset node

The screenshot shows the "Data Audit" interface within IBM Watson Studio. The left sidebar lists various operations: Modeling, Text Analytics, Graphs, Outputs, Table, Matrix, Analysis, Data Audit (which is selected), Transform, Statistics, Means, Report, Set Globals, and Sim Fit. The main workspace shows a "Data Asset" node with a green icon and a dashed box containing "7 Fields". To the right, the "7 Fields" panel is expanded, showing settings like "Use custom fields" (unchecked) and a "Fields" section with a note to "Add columns". Below this are sections for "Overlay" (with an empty dropdown), "Graphs" (checked), "Basic statistics" (checked), and "Advanced statistics" (unchecked). At the bottom are "Cancel" and "Save" buttons.

Figure: Performed data audit

Pranjal Gupta

The screenshot shows the IBM Watson Studio interface with the project 'Predictive_Analysis_Insurance / Insurance_SPSS_Modeler'. On the left, a sidebar lists nodes: 'Type', 'Filler', 'Multiplot', and 'Distribution'. The 'Type' node is selected and highlighted with a green border. In the main workspace, a 'Data Asset' node is connected to a 'Type' node, which is then connected to a 'Partition' node. A '7 Fields' node is also present. The 'Type' node configuration panel is open, showing a table of fields with their measures, roles, and value modes. The 'Premium' field is set as the target attribute ('Role: Target'). Buttons for 'Read values' and 'Clear values' are at the top of the panel.

Field	Measure	Role	Value mode	Values
# age	Continuous	Input	Read	
# sex	Categorical	Input	Read	
# bmi	Continuous	Input	Read	
# children	Continuous	Input	Read	
# smoker	Categorical	Input	Read	
# region	Categorical	Input	Read	
# premium	Continuous	Target	Read	

Figure: Specified the target attribute

The screenshot shows the IBM Watson Studio interface with the same project. The 'Partition' node is selected and highlighted with a green border. In the main workspace, a 'Data Asset' node is connected to a 'Type' node, which is then connected to a 'Partition' node. A '7 Fields' node is also present. The 'Partition' node configuration panel is open, showing settings for training and testing partitions. The 'Training Partition(%)' is set to 80 and the 'Testing Partition(%)' is set to 20. The 'Repeatable partition assignment' checkbox is checked. A seed value '1234567' is entered. Buttons for 'Cancel' and 'Save' are at the bottom of the panel.

Figure: Fixed the test and training data partition

Pranjal Gupta

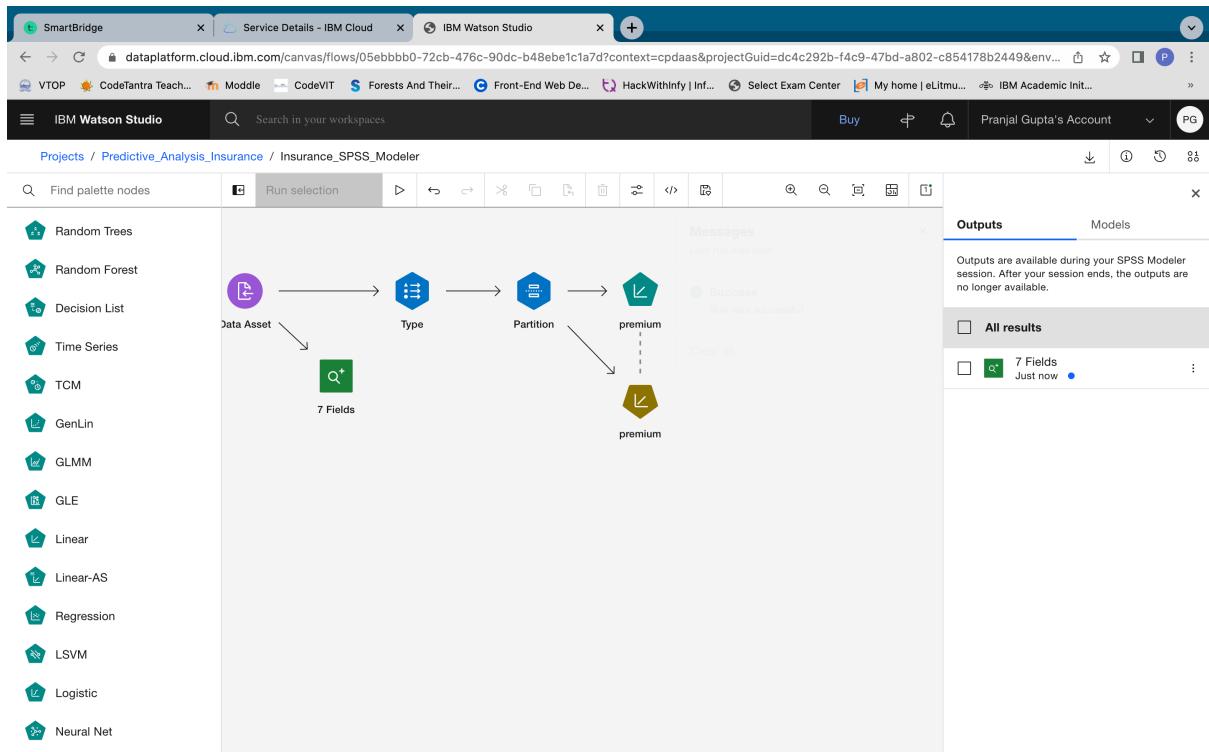


Figure: Performed linear regression

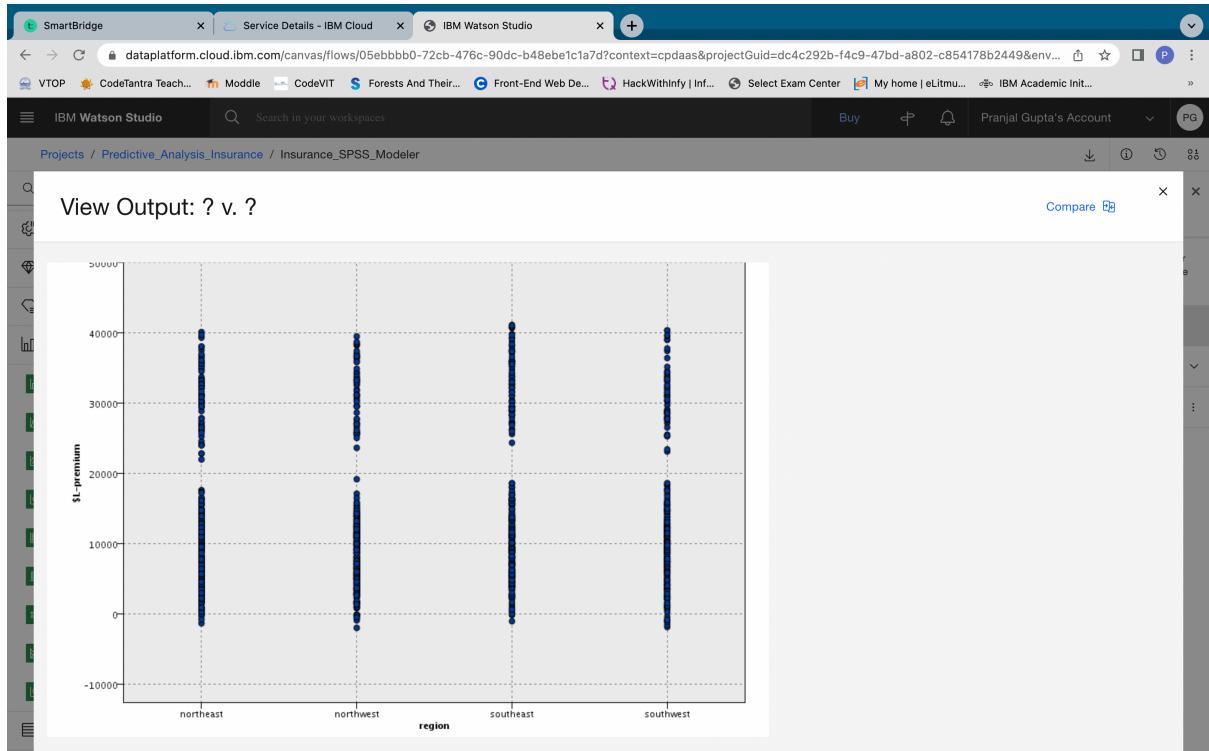


Figure: Linear plot for modelized premium over region

Data Set: MallCustomers.csv

Link: <https://dataplatform.cloud.ibm.com/projects/430ec4a6-234b-4189-b26c-9a1d4178eb09/assets?context=cpdaas>

The screenshot shows the IBM Watson Studio interface. At the top, there are tabs for 'SmartBridge', 'Service Details - IBM Cloud', and 'IBM Watson Studio'. The main title bar displays the URL: dataplatform.cloud.ibm.com/projects/430ec4a6-234b-4189-b26c-9a1d4178eb09?context=cpdaas. The header includes a search bar, user account information for 'Pranjal Gupta's Account', and a 'Launch IDE' button. Below the header, the page title is 'Projects / Mall_Customers_analysis'. The interface has a tab-based navigation with 'Overview' selected, followed by 'Assets', 'Jobs', and 'Manage'. The 'Overview' section contains three main panels: 'Assets' (with a note about creating assets), 'Resource usage' (showing 0 CUH), and 'Project history' (indicating no notifications). A 'Readme' panel is also present.

Figure: Created project for MallCustomers.csv

This screenshot shows the 'Data in this project' panel within the IBM Watson Studio interface. The panel lists various data assets, including 'insurance.csv', 'Mall_Customers.csv' (which is highlighted in blue), 'bank.csv', 'challengers.csv', 'diabetes.csv', 'sales1.xls', and 'Sample - Sup...x versions.xls'. The 'Mall_Customers.csv' entry is detailed below: it is a CSV document (4 KB) created yesterday at 5:05 PM. There are 'Cancel' and 'Open' buttons at the bottom right of the preview window.

Figure: Adding data set

Pranjal Gupta

The screenshot shows the IBM Watson Studio interface. On the left, there's a data preview table titled 'Data' with columns: CustomerID, Gender, Age, Annual Inc..., and Spending S... (partial view). The table contains 15 rows of sample data. On the right, the 'Details' panel is open for a 'Data Refinery Flow' named 'Mall_Customers.csv_flow'. It shows the location as 'Mall_Customers_analysis', a description field, and a step count of 0. Below that, the 'Data Refinery Flow Output' section shows the location as 'Mall_Customers_analysis/Data assets'.

Figure: Creating data refinery flow

The screenshot shows the creation of a new SPSS Modeler flow. The 'Define details' section has 'Name' set to 'Mail_Customers_Modeler'. The 'Define configuration' section has 'Environment definition' set to 'Default SPSS Modeler S (2 vCPU 8 GB RAM)'. The 'Create' button at the bottom is highlighted in blue.

Figure: Creating SPSS modeler

Pranjal Gupta

The screenshot shows the IBM Watson Studio interface. On the left, a sidebar lists various palette nodes: Import, Data Asset, User Input, Sim Gen, Extension Import, Record Operations, Field Operations, Modeling, Text Analytics, Graphs, Outputs, and Export. The 'Data Asset' node is selected and highlighted with a dashed box. The main panel displays the 'Data Asset' configuration for 'Mall_Customers.csv'. It includes sections for 'File format properties' (set to CSV), 'Encoding' (set to UTF-8), and 'First line is header' (checkbox checked). Below these are options for 'Invalid data handling' (Fail) and 'First line' (empty). At the bottom right of the panel is a blue 'Save' button.

Figure: Imported data asset

The screenshot shows the IBM Watson Studio interface. The sidebar now lists: Text Analytics, Graphs, Outputs, Table, Matrix, Analysis, Data Audit, Transform, Statistics, Means, Report, Set Globals, Sim Fit, and Sim Evaluation. The 'Data Audit' node is selected and highlighted with a dashed box. The main panel displays the results of the data audit, showing a green icon with a magnifying glass and the text '5 Fields'.

Figure: Performed data audit

Pranjal Gupta

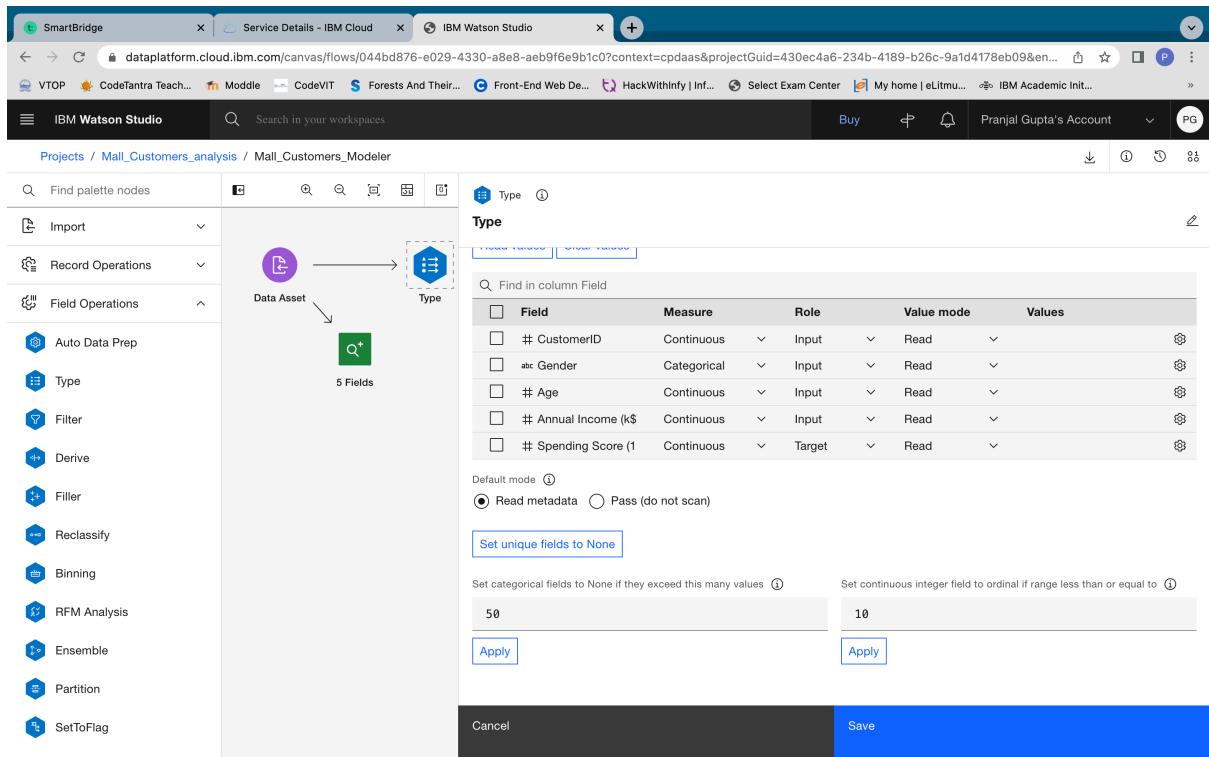


Figure: Specified the target attribute

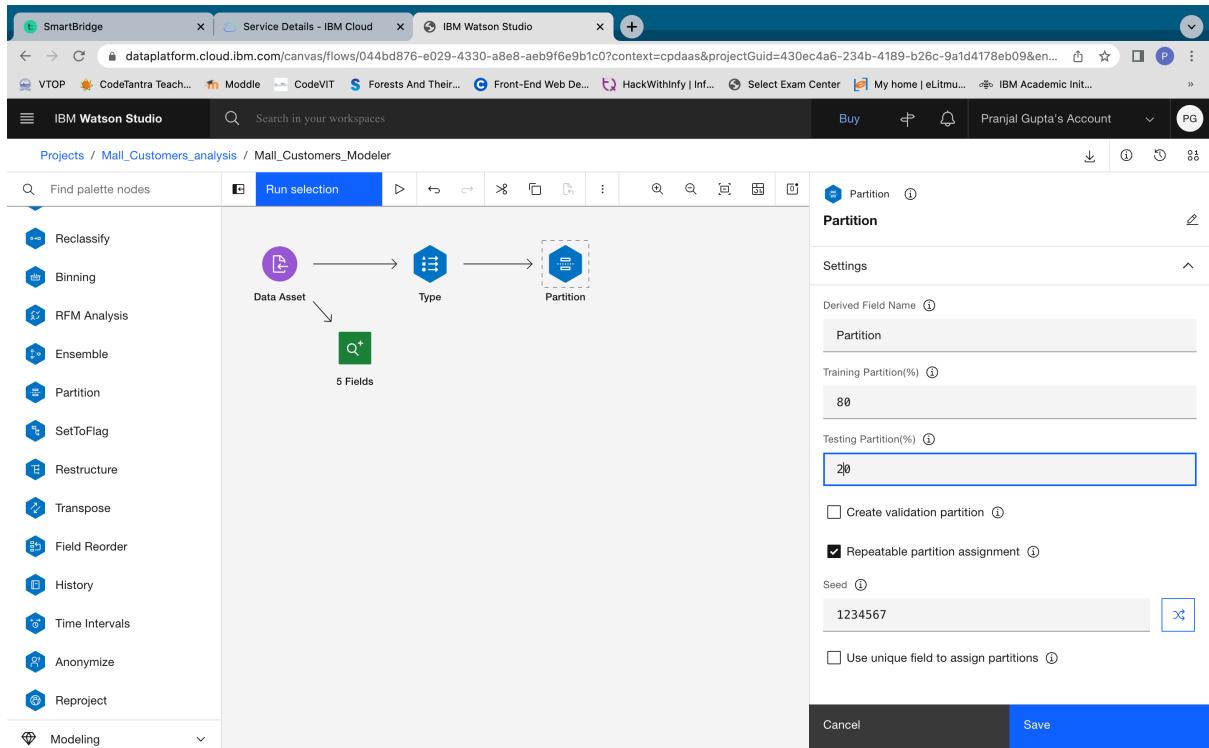


Figure: Fixed the training and test data

Pranjal Gupta

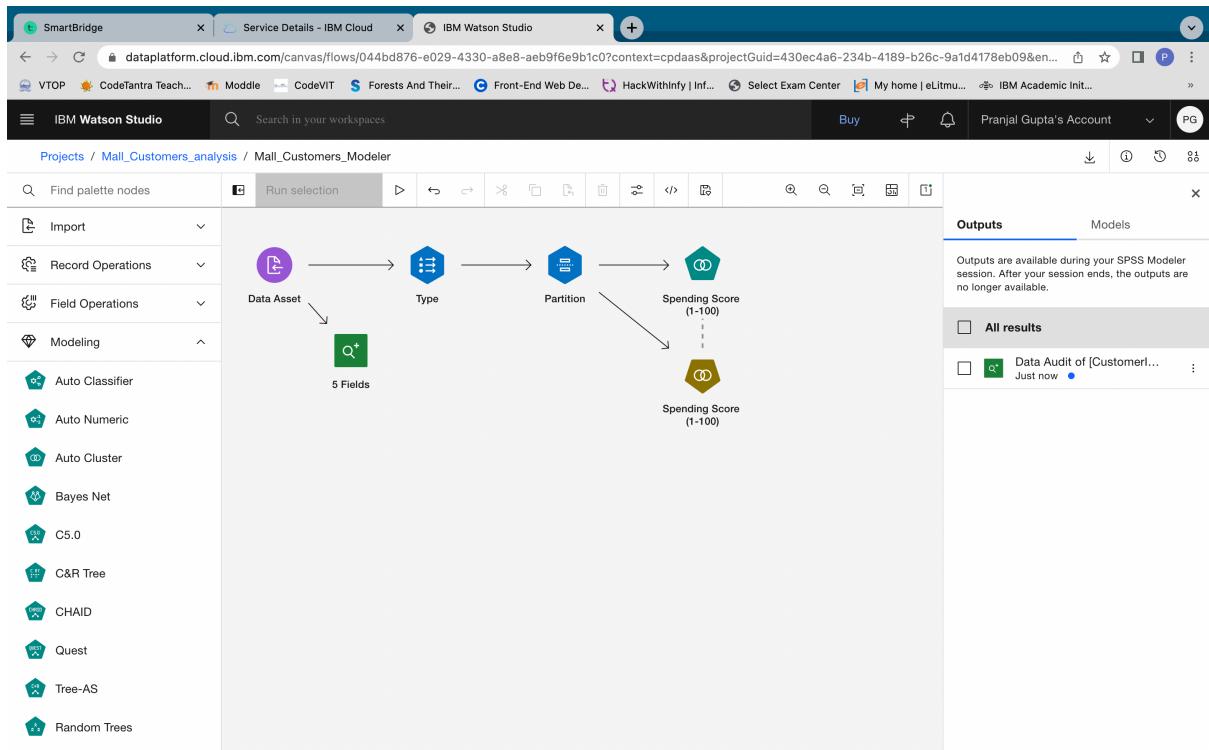


Figure: Performed auto clustering

The screenshot shows the 'View Model: Spending Score (1-100)' page in IBM Watson Studio. On the left, a sidebar shows 'Auto Cluster' and 'Models'. The main area is titled 'Auto Cluster - Models'. It displays a table with the following data:

USE	MODEL_NAME	ESTIMATOR	GRAPH	SILHOUETTE	BUILD TIME (MINS)	NUMBER OF CLUSTERS	SMALLEST CLUSTER (N)	SMALLEST CLUSTER (%)	LARGEST CLUSTER (N)	LARGEST CLUSTER (%)
○	TwoStep	TwoStep	[Bar chart]	0.428	< 1	2	58	0.395	89	100
●	K-means	KMeans	[Bar chart]	0.522	< 1	5	15	0.102	48	100
○	Kohonen	Kohonen	[Bar chart]	0.368	< 1	6	10	0.068	38	100

Figure: Model View

Pranjal Gupta

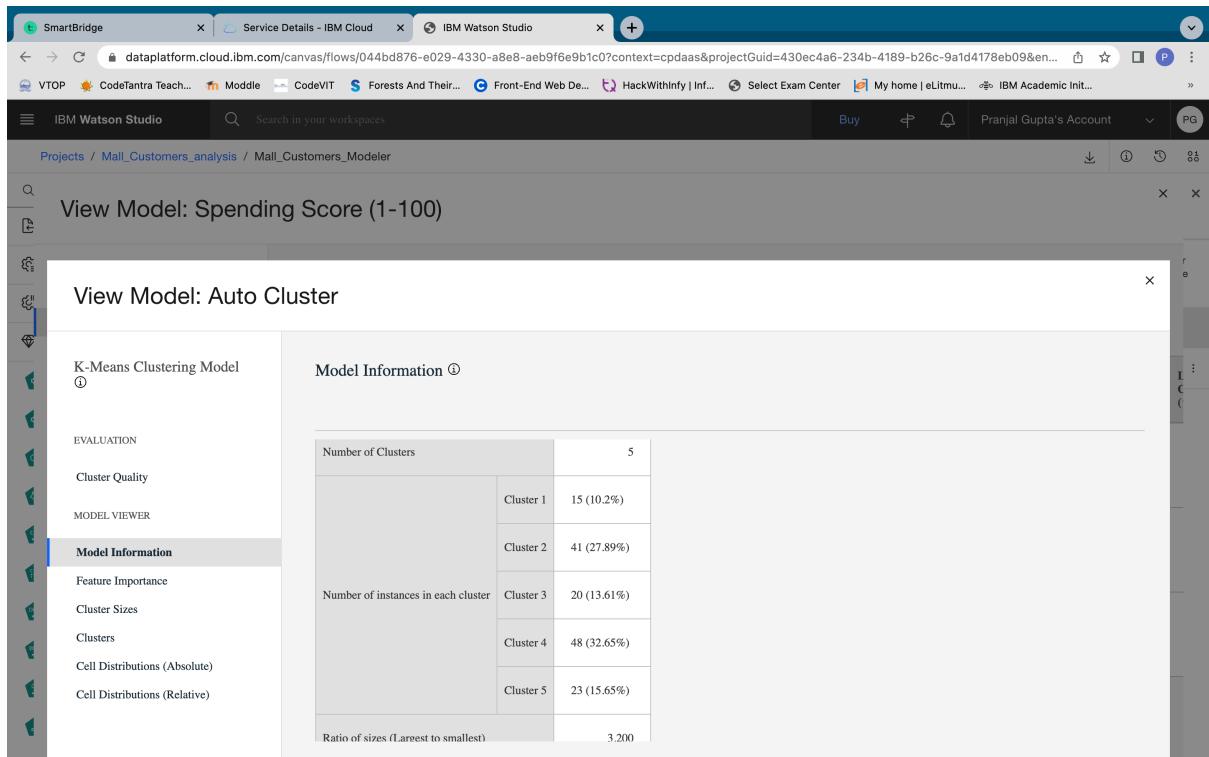


Figure: Model information for K-means clustering

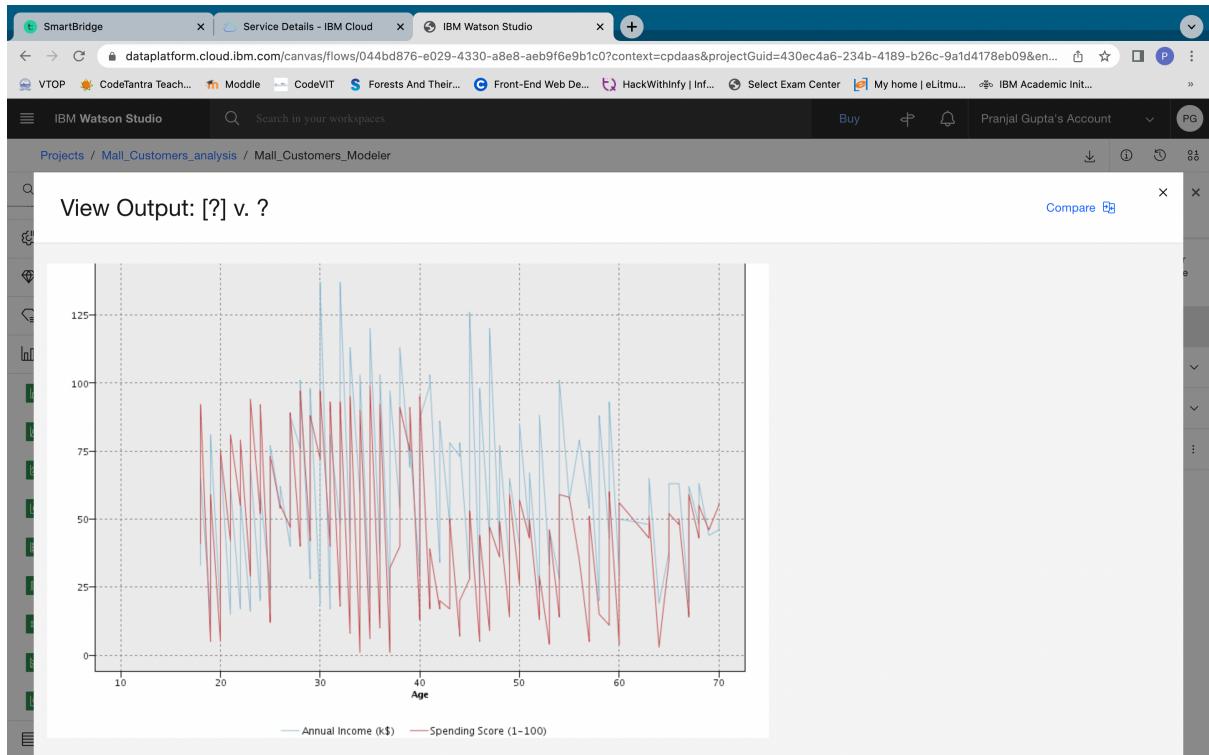


Figure: Linear plot for Age over annual income over spending score