

Data Analytics Externship Program

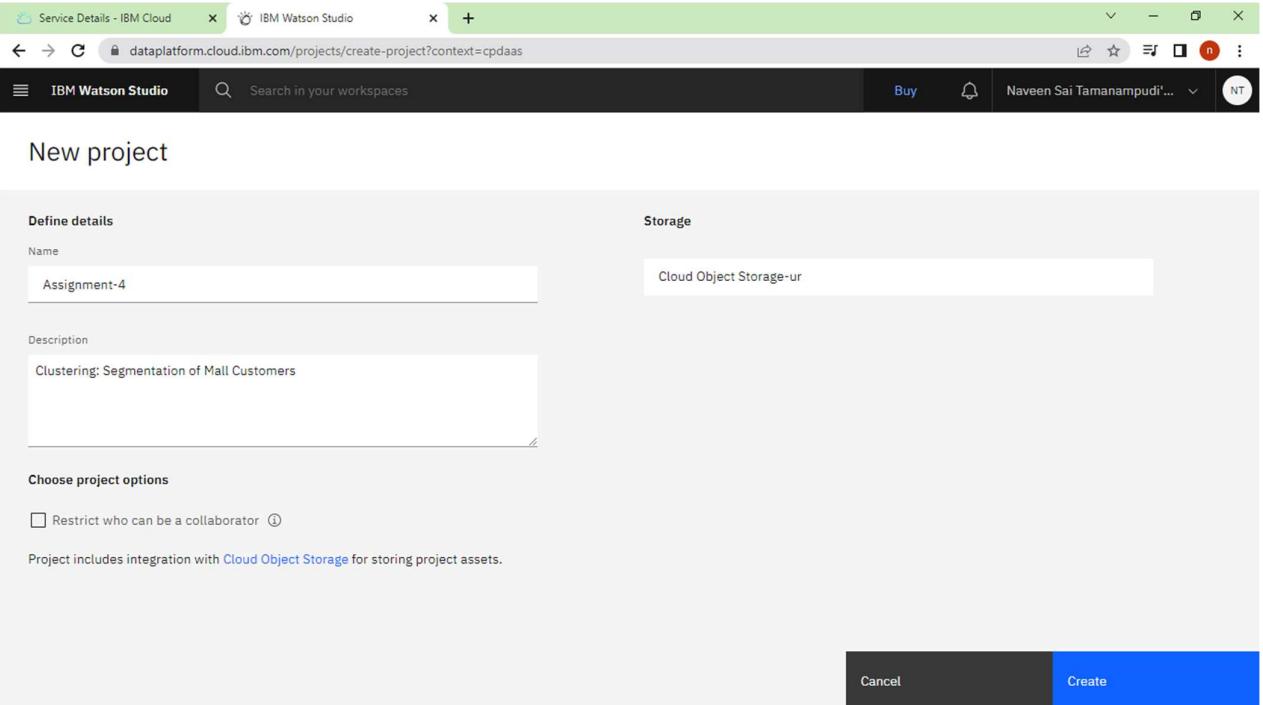
Assignment-4

Name: Tamanampudi Naveen Sai
Email ID: naveen.19bcd7003@vitap.ac.in

Clustering: Segmentation of Mall Customers

Dataset Used: Mall_Customers.csv

Creating the Project



The screenshot shows the 'Create Project' dialog box in IBM Watson Studio. The URL in the browser is dataplatform.cloud.ibm.com/projects/create-project?context=cpdaas. The form fields are as follows:

- Define details**:
 - Name**: Assignment-4
 - Description**: Clustering: Segmentation of Mall Customers
- Storage**: Cloud Object Storage-ur
- Choose project options**:
 - Restrict who can be a collaborator (i)
- Project includes integration with [Cloud Object Storage](#) for storing project assets.

At the bottom right, there are 'Cancel' and 'Create' buttons.

Uploading Data:

The screenshot shows the IBM Watson Studio interface with the 'Assets' tab selected. On the left, there's a sidebar for 'Asset types' with 'Data' selected, showing one data asset. The main area displays a table titled 'Data' with columns 'Name' and 'Last modified'. A single row is listed: 'Mail_Customers.csv' (CSV) last modified 'Now' by 'Naveen Sai Tamanampudi (You)'. To the right, a large panel titled 'Data in this project' contains a dashed box with the placeholder text 'Drop data files here or browse for files to upload'.

Creating a Data Refinery:

The screenshot shows a modal dialog titled 'Select data from project' over the IBM Watson Studio interface. In the modal, under the 'Assignment-4' section, a 'Data asset' named 'Mail_Customers.csv' is selected. On the right, a detailed view of the selected asset is shown in the 'Selected assets' panel. The asset details include: Asset name 'Mail_Customers.csv', Asset type 'Data asset', Size '4 KB', Last modified '2022/04/26 16:24:38', and Created on '2022/04/26 16:24:38'. At the bottom of the modal are 'Cancel' and 'Select' buttons.

Viewing the output of refinery:

Service Details - IBM Cloud

IBM Watson Studio

dataplatform.cloud.ibm.com/shaper?context=data&dataset_id=ec19d601-19da-46b2-a535-1fbfaba6e7b4&project_id=aee2a8b2-cebe-4e46-ab19-70b005d18190

IBM Watson Studio

Search in your workspaces

Buy

Naveen Sai Tamanampudi...

Projects / Assignment-4 / Mall_Customers.csv / Refine data

Steps (1)

Data Source: Mall_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

Use a code template to add a step

CustomerID
Integer

Gender
String

Age
Integer

Annual Inc...
Integer

1 1 Male 19 15

2 2 Male 21 15

3 3 Female 20 16

4 4 Female 23 16

5 5 Female 31 17

6 6 Female 22 17

7 7 Female 35 18

8 8 Female 23 18

9 9 Male 64 19

10 10 Female 30 19

11 11 Male 67 19

12 12 Female 25 10

SOURCE FILE: Mall_Customers.csv FULL DATA SET: 200 rows

Information

Details Help

Edit

LOCATION Assignment-4

DATA REFINERY FLOW NAME: Mall_Customers.csv_flow

Enter a description of the Data Refinery flow

STEPS 1

DATA REFINERY FLOW OUTPUT

Viewing Profile:

Service Details - IBM Cloud

IBM Watson Studio

dataplatform.cloud.ibm.com/shaper?context=data&dataset_id=ec19d601-19da-46b2-a535-1fbfaba6e7b4&project_id=aee2a8b2-cebe-4e46-ab19-70b005d18190

IBM Watson Studio

Search in your workspaces

Buy

Naveen Sai Tamanampudi...

Projects / Assignment-4 / Mall_Customers.csv / Refine data

Steps Use a code template to add a step

Data Profile Visualizations

CustomerID Integer

Gender String

Age Integer

Annual Income (k\$) Integer

FREQUENCY

[1,21]

[21,42]

[41,61]

[61,81]

[81,101]

[101,121]

[121,141]

[141,161]

[161,181]

[181,200]

0 5 10 15 20

FREQUENCY

Female

Male

0 20 40 60 80 100

FREQUENCY

[30,36]

[18,24]

[36,42]

[24,30]

[48,54]

[42,48]

[54,60]

[66,70]

[60,66]

0 10 20 30 40

FREQUENCY

[54,67]

[67,80]

[28,41]

[15,28]

[41,54]

[80,93]

[93,106]

[111,132]

[106,119]

[132,137]

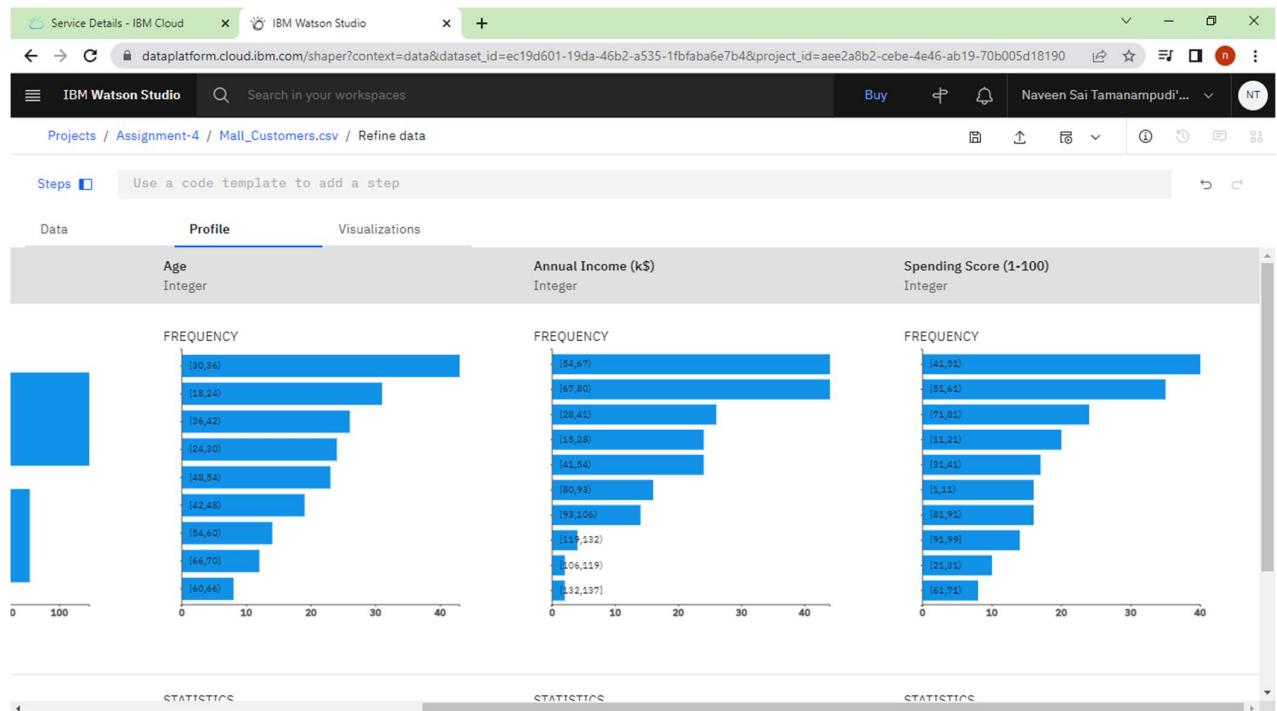
0 10

STATISTICS

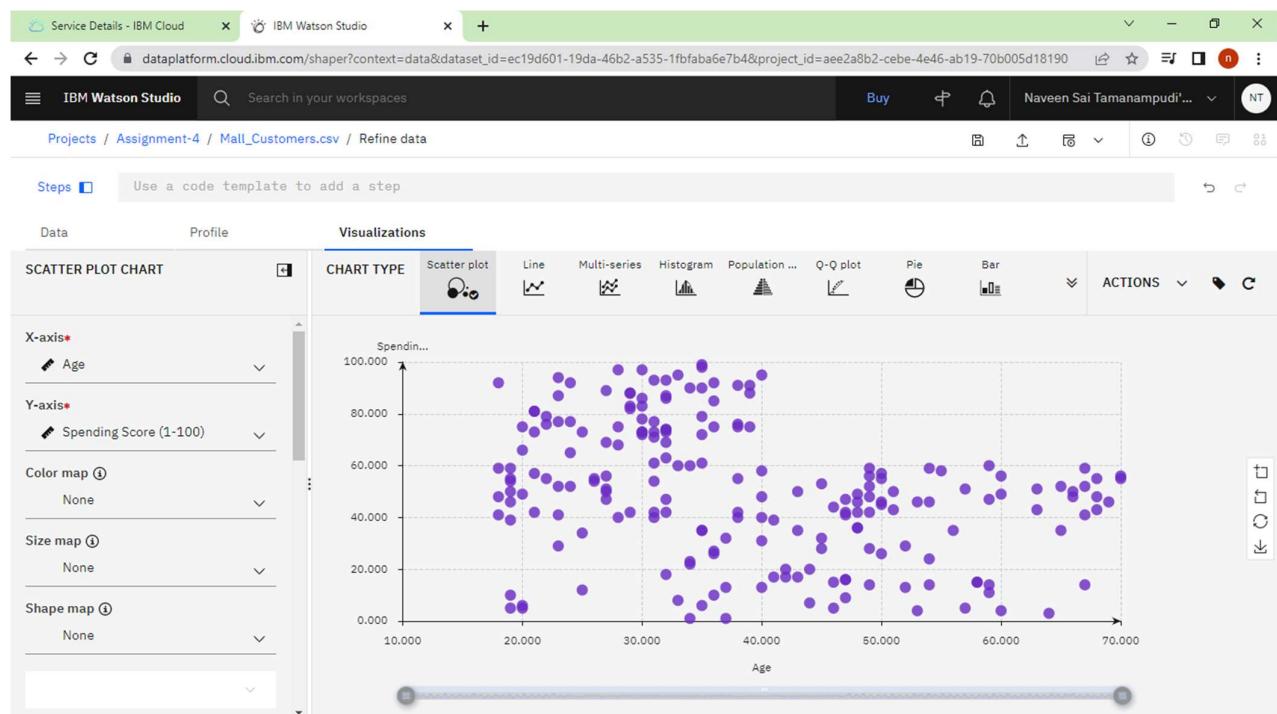
STATISTICS

STATISTICS

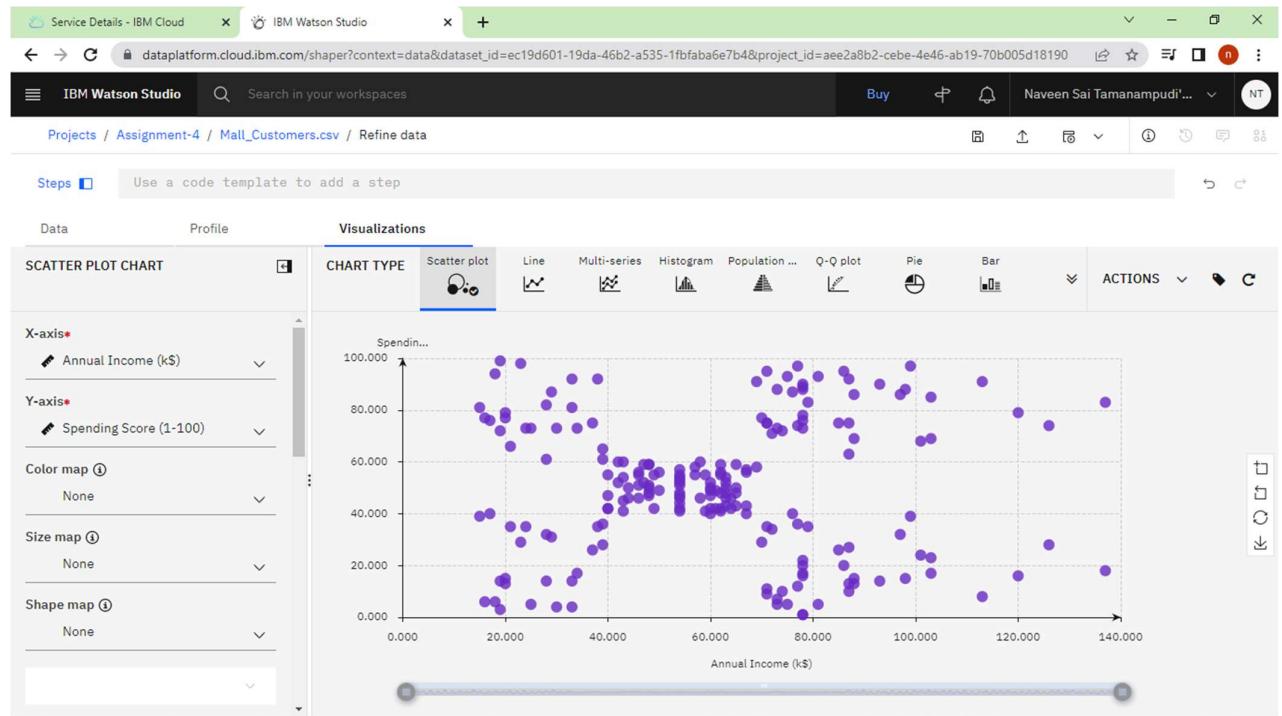
STATISTICS



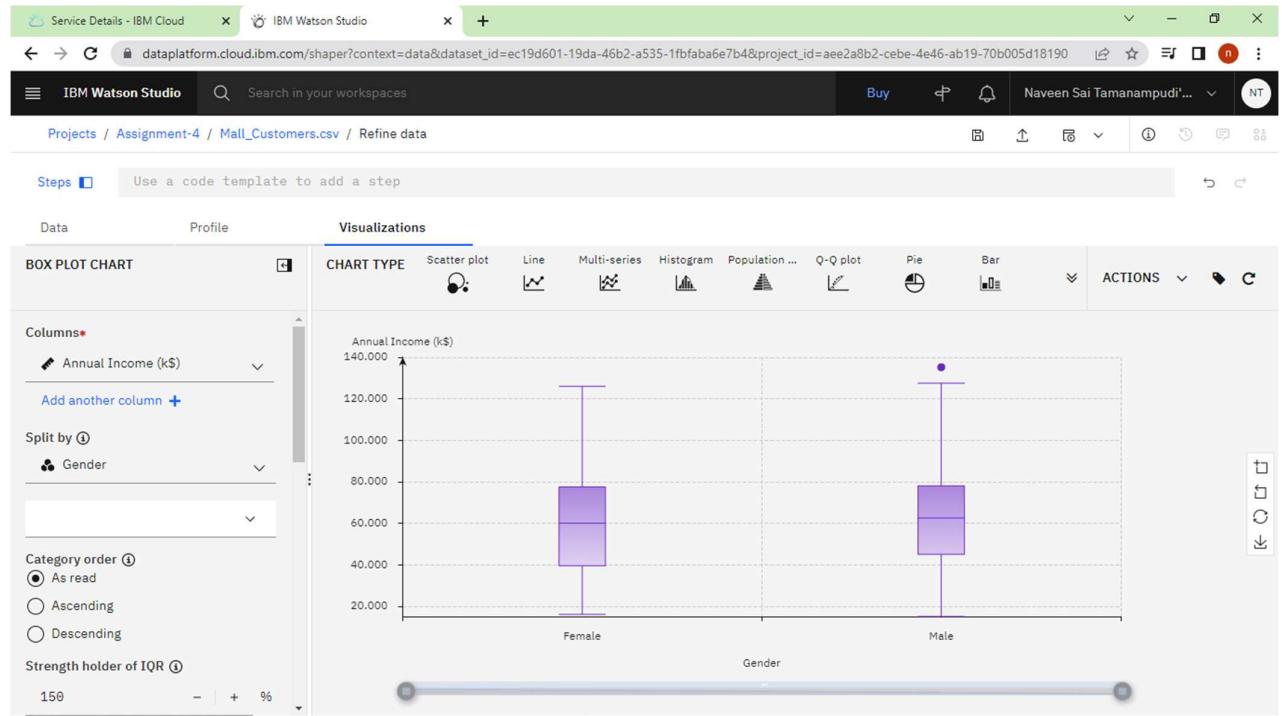
Viewing Scatterplot of Age vs Spending Score(1-100):



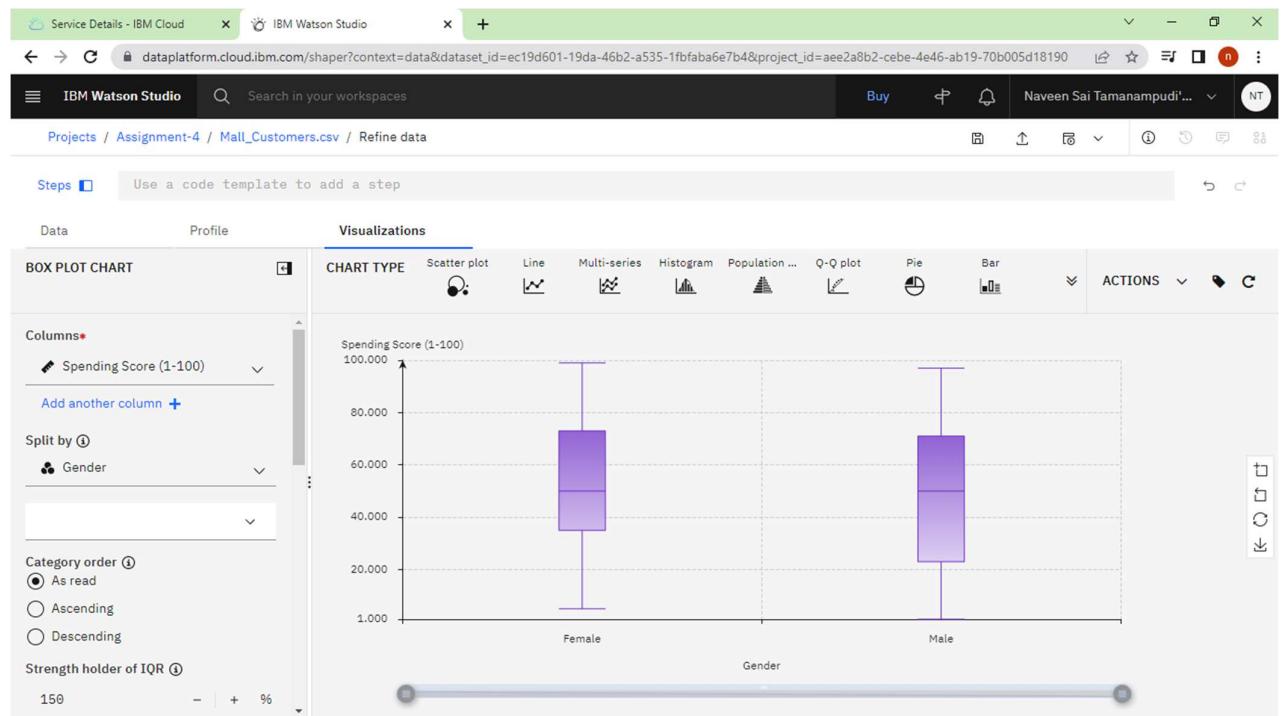
Viewing Scatterplot of Annual Income vs Spending Score(1-100):



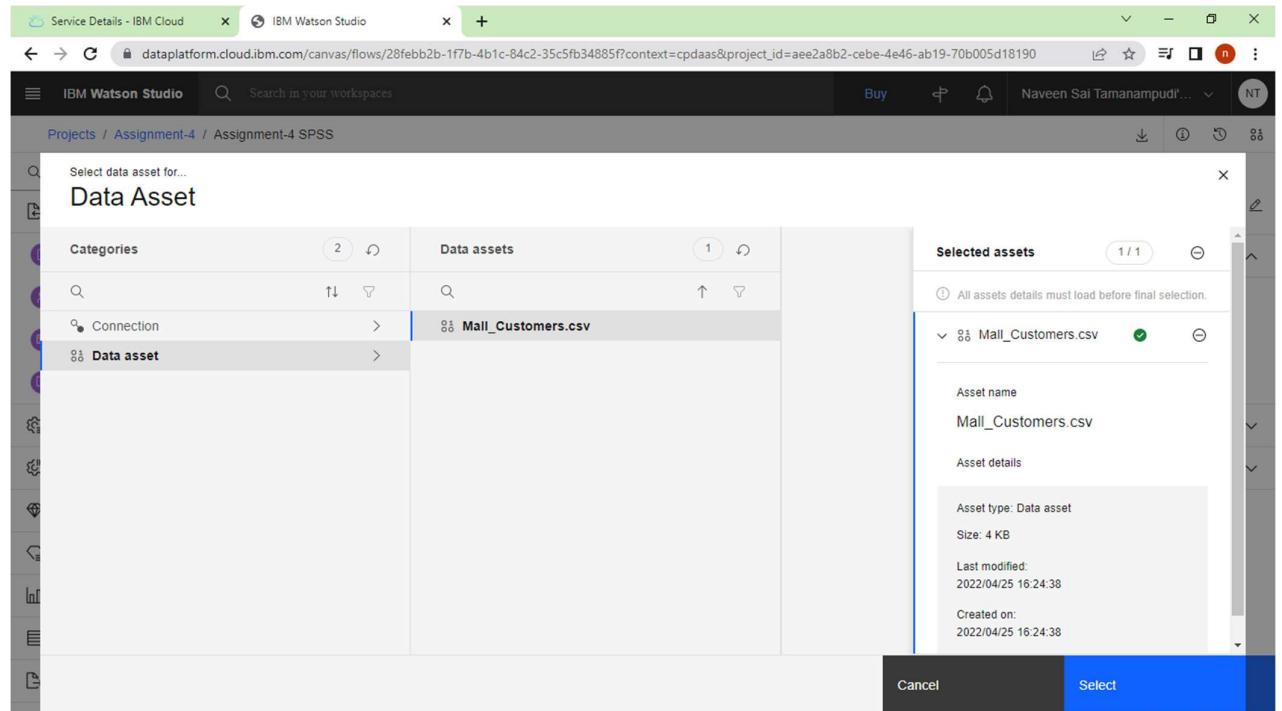
Viewing Boxplot of Annual Income vs Gender:



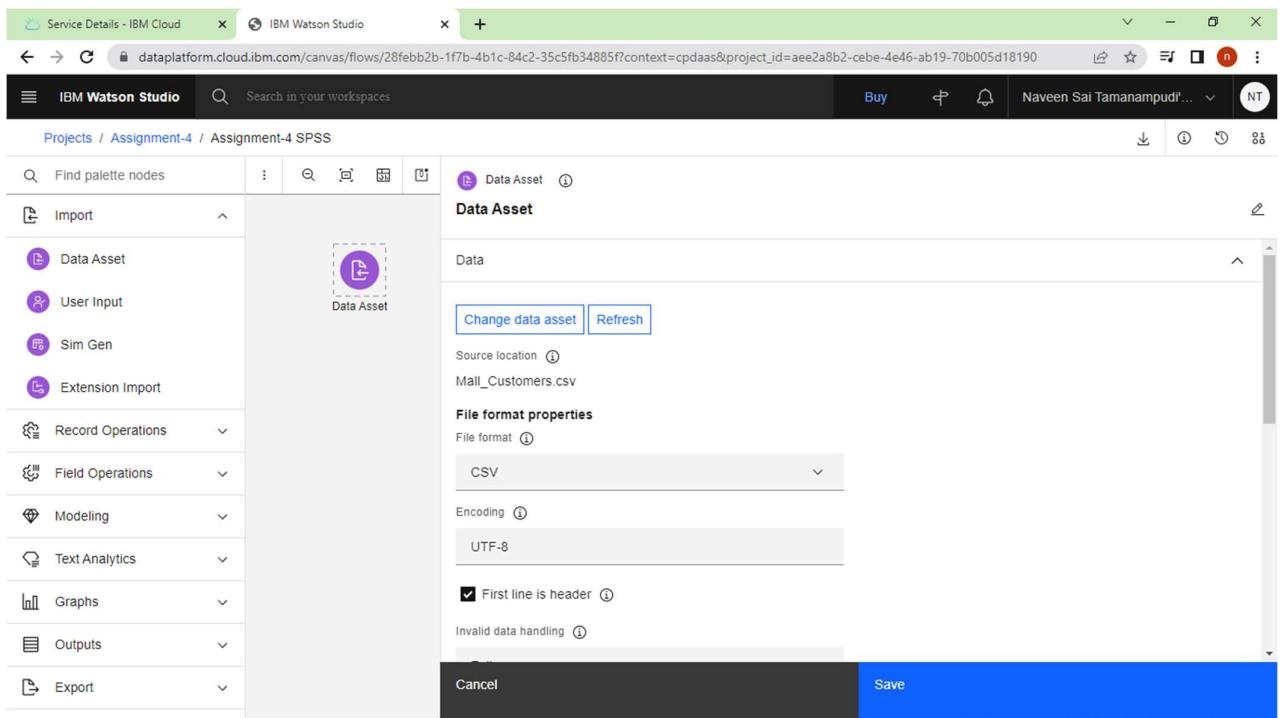
Viewing Boxplot of Spending Score(1-100) vs Gender:



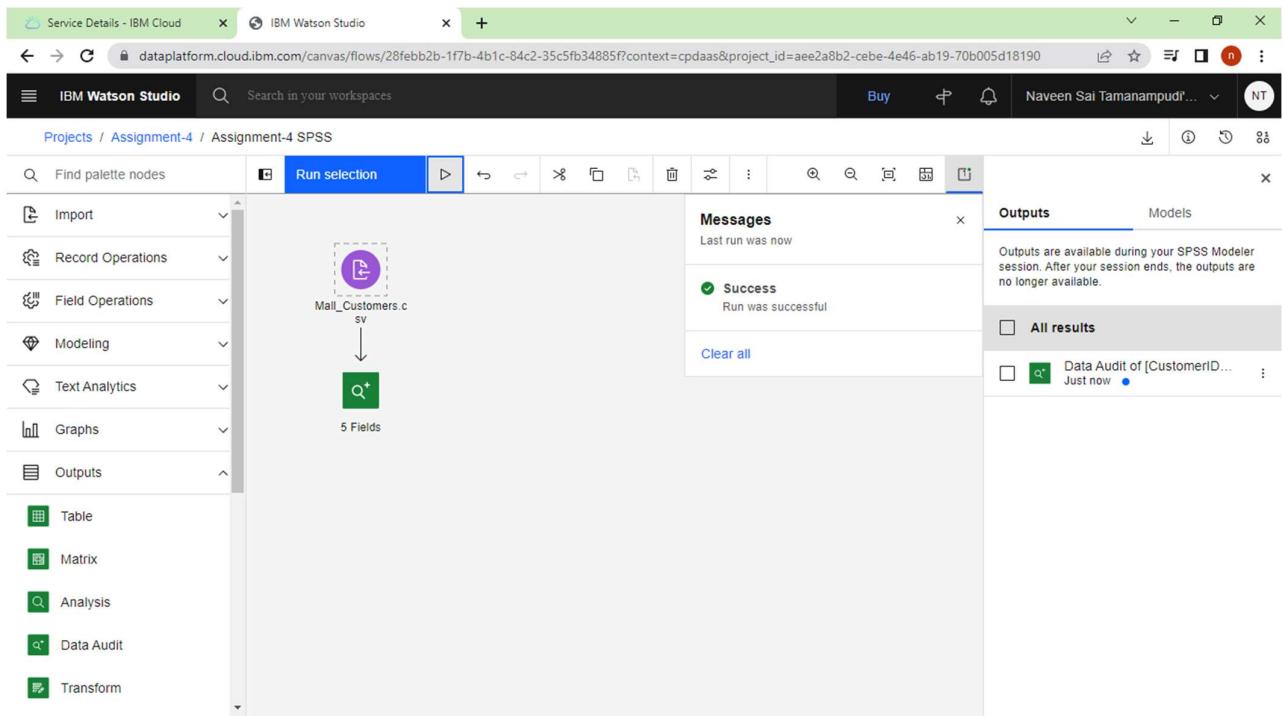
Creating an SPSS Modeler:



Uploading Dataset to Data Asset Node:



Creating Data Audit Node:



Output from Data Audit:

Service Details - IBM Cloud IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/28febb2b-1f7b-4b1c-84c2-35c5fb34885f?context=cpdaas&project_id=aee2a8b2-cebe-4e46-ab19-70b005d18190

IBM Watson Studio Search in your workspaces Buy Compare

Projects / Assignment-4 / Assignment-4 SPSS Naveen Sai Tamanampudi... NT

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

| 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 |

Service Details - IBM Cloud IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/28febb2b-1f7b-4b1c-84c2-35c5fb34885f?context=cpdaas&project_id=aee2a8b2-cebe-4e46-ab19-70b005d18190

IBM Watson Studio Search in your workspaces Buy Compare

Projects / Assignment-4 / Assignment-4 SPSS Naveen Sai Tamanampudi... NT

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

| 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 | |
| 1 CustomerID | Continuous | 0 | 0 | None | Never | Fixed | 100.000 | 200 | |
| 2 Gender | Categorical | -- | -- | -- | Never | Fixed | 100.000 | 200 | |
| 3 Age | Continuous | 0 | 0 | None | Never | Fixed | 100.000 | 200 | |
| 4 Annual Income (k\$) | Continuous | 0 | 0 | None | Never | Fixed | 100.000 | 200 | |
| 5 Spending Score (1-100) | Continuous | 0 | 0 | None | Never | Fixed | 100.000 | 200 | |

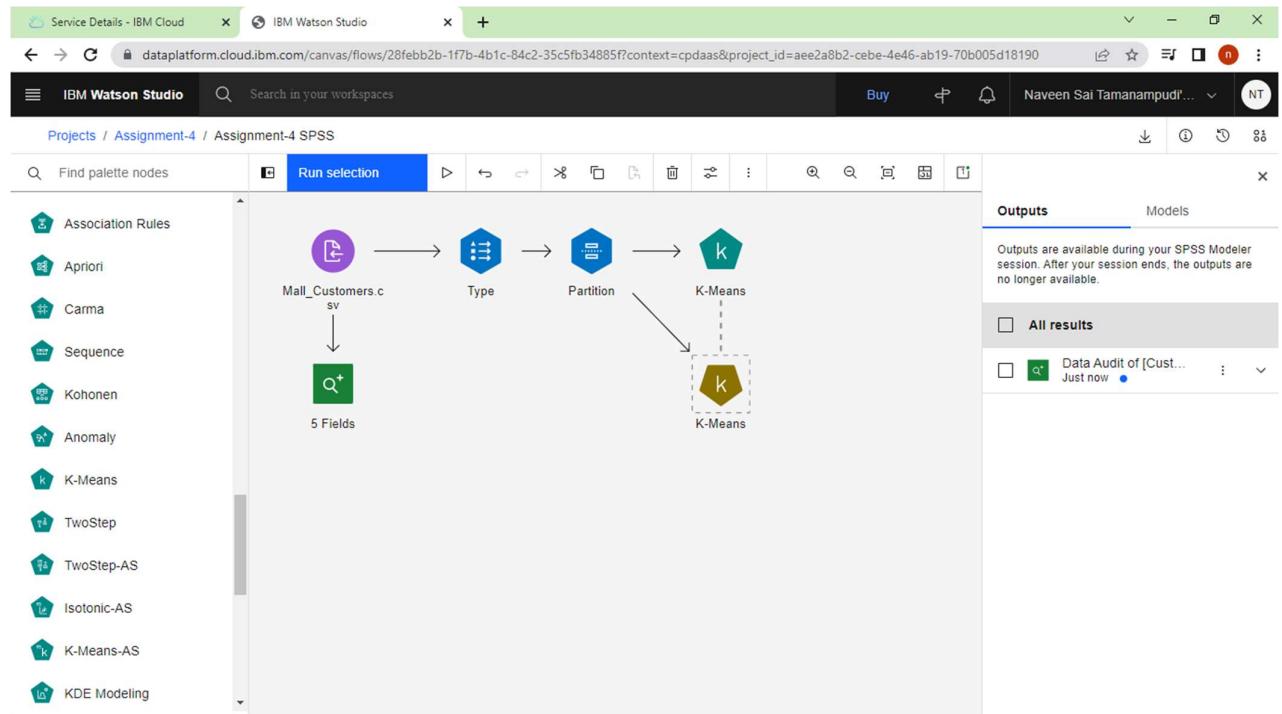
Creating a Type Node and setting Inputs, Targets:

The screenshot shows the IBM Watson Studio interface with a flow canvas. A 'Type' node is selected, indicated by a dashed border. On the right, the 'Type' configuration panel is open, showing settings for 'Read values' and a table of field roles. The table includes columns for Field, Measure, Role, Value mode, and Values. Fields listed include CustomerID, Gender, Age, Annual Income (K\$), and Spending Score (1). The 'Default mode' dropdown is set to 'Read metadata'. At the bottom, there are 'Cancel' and 'Save' buttons.

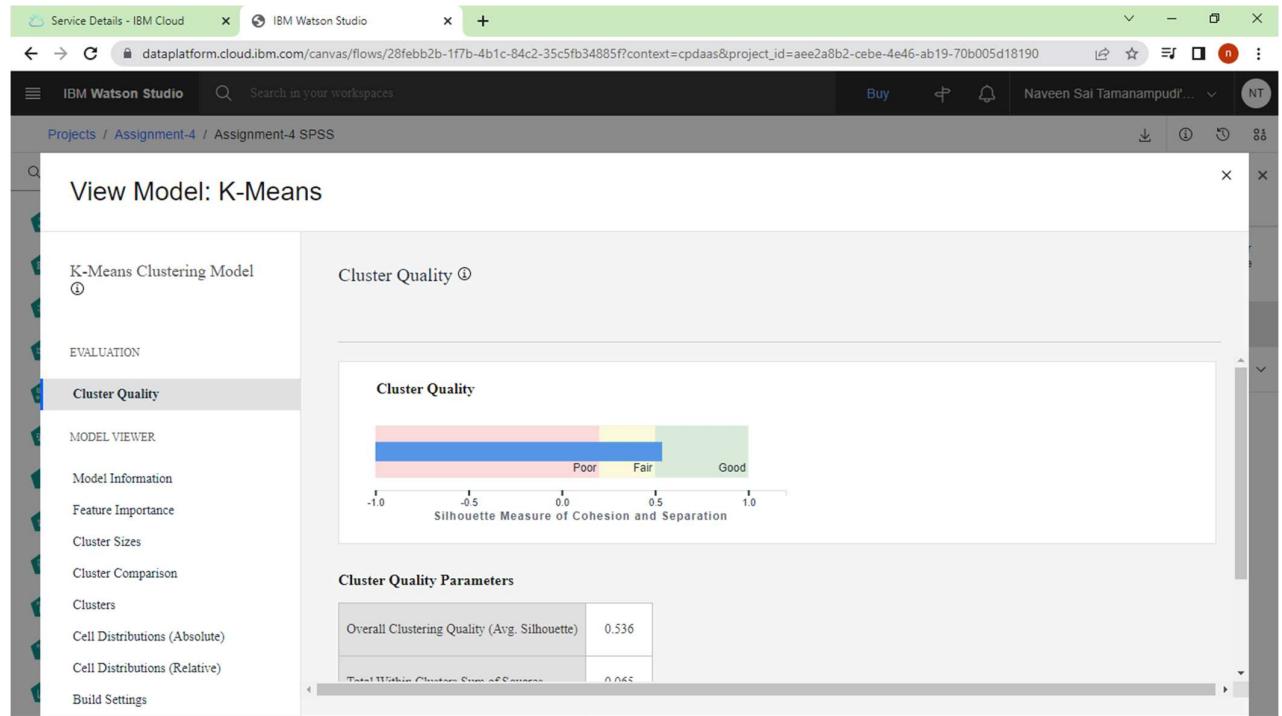
Creating a Partition Node with 70:30 split:

The screenshot shows the IBM Watson Studio interface with a flow canvas. A 'Partition' node is selected, indicated by a dashed border. On the right, the 'Partition' configuration panel is open, showing 'Training Partition(%)' set to 70 and 'Testing Partition(%)' set to 30. There are checkboxes for 'Create validation partition' and 'Repeatable partition assignment', with the latter checked. A 'Seed' value of 1234567 is specified. At the bottom, there are 'Cancel' and 'Save' buttons.

Creating an K-Means Clustering Model:



Output:



Service Details - IBM Cloud IBM Watson Studio

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model

EVALUATION

Cluster Quality

MODEL VIEWER.

Model Information

Feature Importance

Cluster Sizes

Cluster Comparison

Clusters

Cell Distributions (Absolute)

Cell Distributions (Relative)

Build Settings

Cluster Quality

Silhouette Measure of Cohesion and Separation

| Overall Clustering Quality (Avg. Silhouette) | 0.536 |
|--|-------|
| Total Within Clusters Sum of Squares | 0.065 |
| Average Within Cluster Sum of Squares | 0.013 |
| Average SSB (Between ss) | 0.166 |

Service Details - IBM Cloud IBM Watson Studio

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model

EVALUATION

Cluster Quality

MODEL VIEWER.

Model Information

Feature Importance

Cluster Sizes

Cluster Comparison

Clusters

Cell Distributions (Absolute)

Cell Distributions (Relative)

Build Settings

Model Information

| Algorithm | K-Means |
|--------------------|---------------------------|
| Model Class | Center Based |
| Number of Features | 3 |
| Distance Measure | Euclidean |
| Number of Clusters | 5 |
| | Cluster 1 13 (9.77%) |
| | Cluster 2 52 / 20 (84.0%) |

Service Details - IBM Cloud IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/28feb2b-1f7b-4b1c-84c2-35c5fb34885f?context=cpdaas&project_id=aee2a8b2-cebe-4e46-ab19-70b005d18190

IBM Watson Studio Search in your workspaces Buy Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model ①

EVALUATION

Cluster Quality

MODEL VIEWER

Model Information

Feature Importance

Cluster Sizes

Cluster Comparison

Clusters

Cell Distributions (Absolute)

Cell Distributions (Relative)

Build Settings

Model Information ①

| Number of Clusters | |
|--------------------------------------|-------------|
| Cluster 1 | 13 (9.77%) |
| Cluster 2 | 53 (39.85%) |
| Cluster 3 | 19 (14.29%) |
| Cluster 4 | 19 (14.29%) |
| Cluster 5 | 29 (21.8%) |
| Ratio of sizes (Largest to smallest) | 4.077 |

Service Details - IBM Cloud IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/28feb2b-1f7b-4b1c-84c2-35c5fb34885f?context=cpdaas&project_id=aee2a8b2-cebe-4e46-ab19-70b005d18190

IBM Watson Studio Search in your workspaces Buy Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model ①

EVALUATION

Cluster Quality

MODEL VIEWER

Model Information

Feature Importance

Cluster Sizes

Cluster Comparison

Clusters

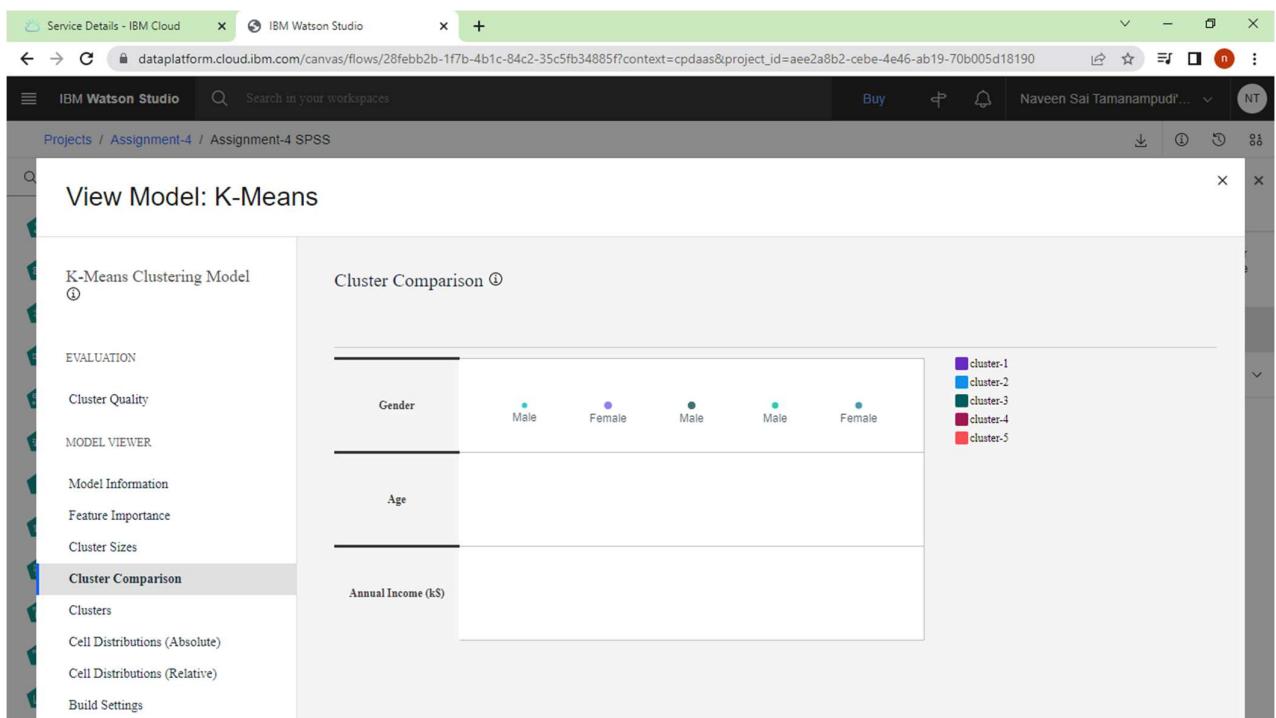
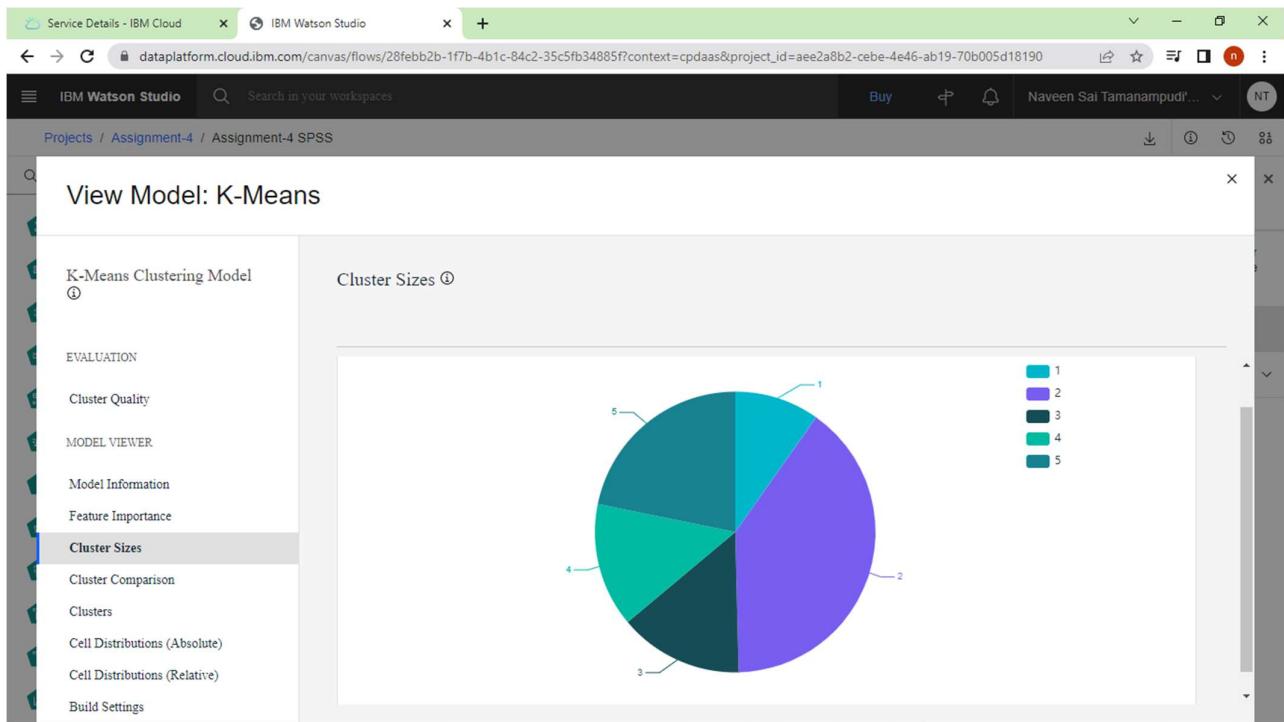
Cell Distributions (Absolute)

Cell Distributions (Relative)

Build Settings

Feature Importance ①

| Feature | Importance |
|---------------------|------------|
| Gender | 1.00 |
| Age | 0.89 |
| Annual Income (k\$) | 0.56 |



Service Details - IBM Cloud IBM Watson Studio

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model

- EVALUATION
- Cluster Quality
- MODEL VIEWER
- Model Information
- Feature Importance
- Cluster Sizes
- Cluster Comparison
- Clusters**
- Cell Distributions (Absolute)
- Cell Distributions (Relative)
- Build Settings

Clusters

| Cluster | cluster_1 | cluster_2 | cluster_3 | cluster_4 | cluster_5 |
|---------|--------------------------|----------------------------|--------------------------|--------------------------|----------------------------|
| Size | | | | | |
| Inputs | Gender Male (100.00%) | Gender Female (100.00%) | Gender Male (100.00%) | Gender Male (100.00%) | Gender Female (100.00%) |
| | Age 27.12 | Age 45.59 | Age 36.24 | Age 56.83 | Age 27.23 |
| | Annual Income | Annual Income | Annual Income | Annual Income | Annual Income |

Service Details - IBM Cloud IBM Watson Studio

Projects / Assignment-4 / Assignment-4 SPSS

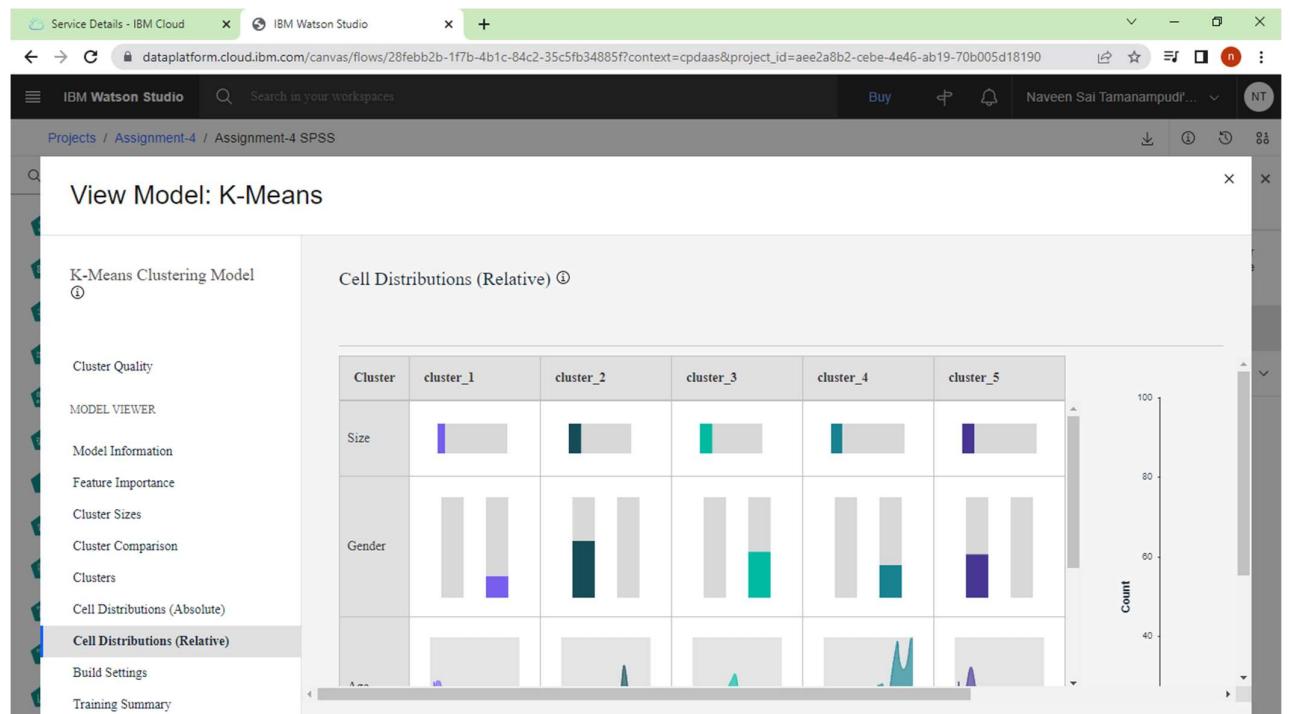
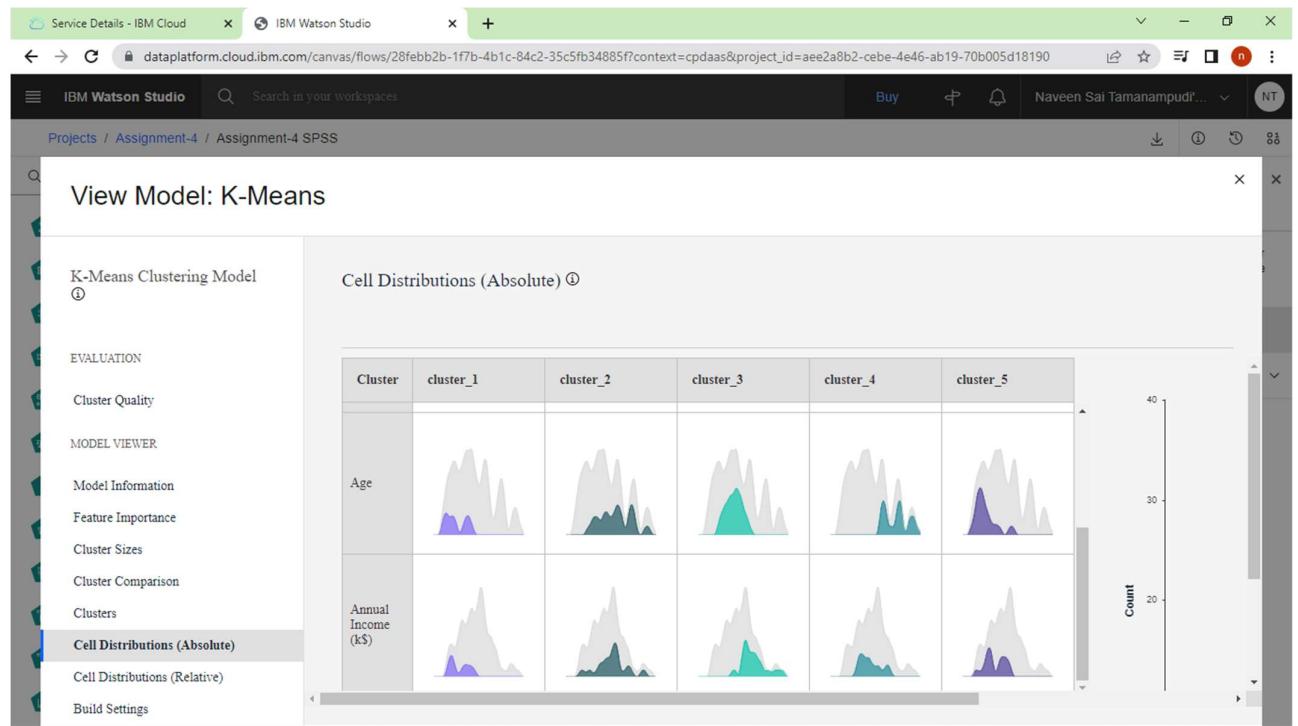
View Model: K-Means

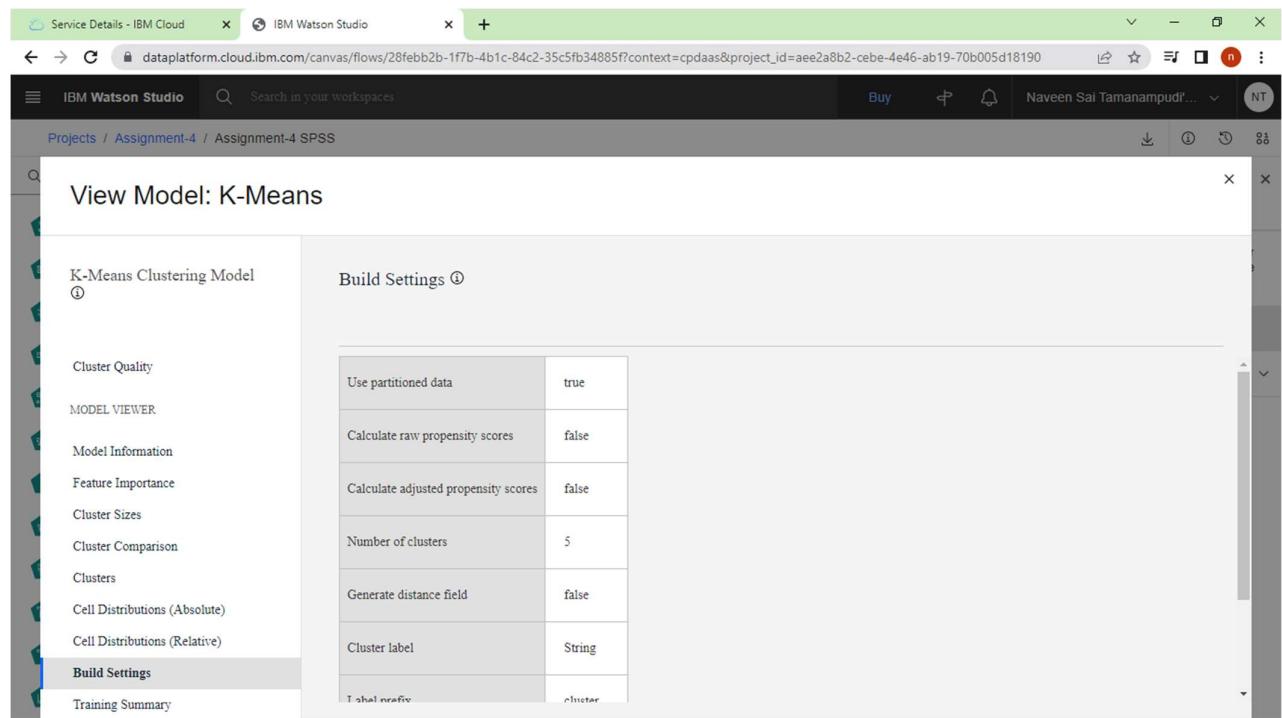
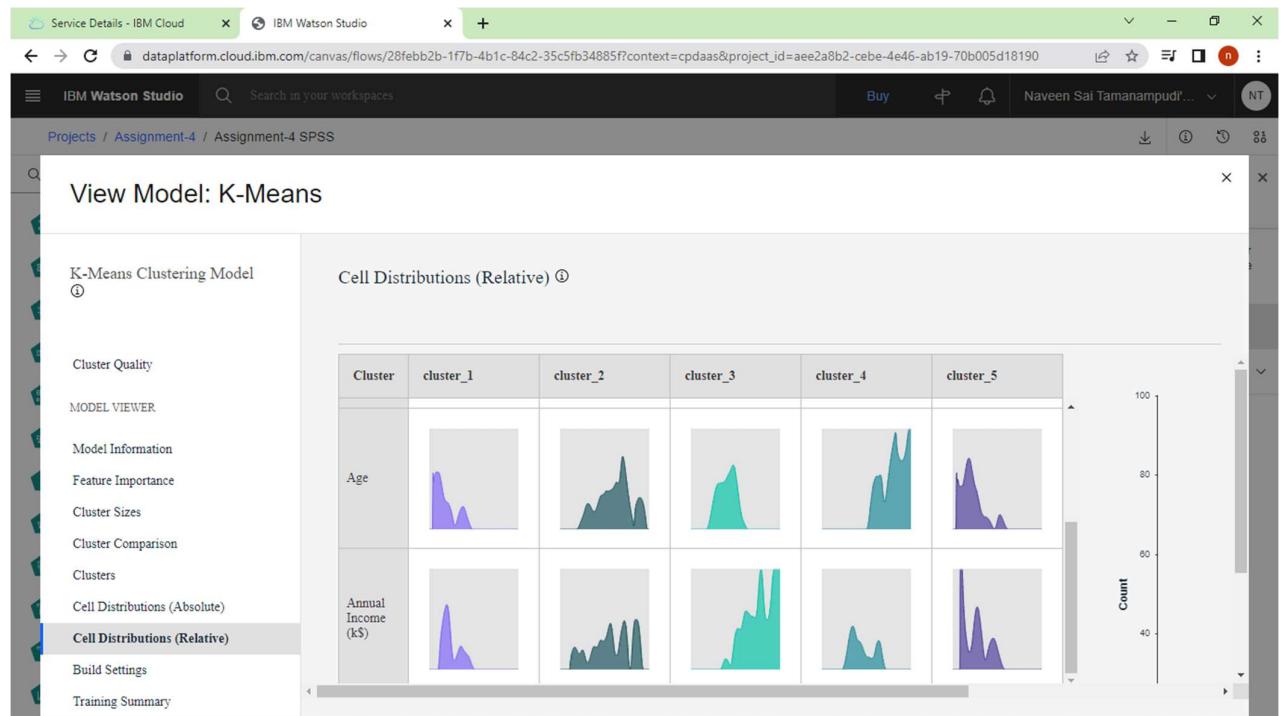
K-Means Clustering Model

- EVALUATION
- Cluster Quality
- MODEL VIEWER
- Model Information
- Feature Importance
- Cluster Sizes
- Cluster Comparison
- Clusters**
- Cell Distributions (Absolute)**
- Cell Distributions (Relative)
- Build Settings

Cell Distributions (Absolute)

| Cluster | cluster_1 | cluster_2 | cluster_3 | cluster_4 | cluster_5 |
|---------|-----------|-----------|-----------|-----------|-----------|
| Size | | | | | |
| Gender | | | | | |





Service Details - IBM Cloud IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/28febb2b-1f7b-4b1c-84c2-35c5fb34885f?context=cpdaas&project_id=aee2a8b2-cebe-4e46-ab19-70b005d18190

IBM Watson Studio Search in your workspaces Buy Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model ①

- Cluster Quality
- MODEL VIEWER
- Model Information
- Feature Importance
- Cluster Sizes
- Cluster Comparison
- Clusters
- Cell Distributions (Absolute)
- Cell Distributions (Relative)
- Build Settings**
- Training Summary

Build Settings ①

| | |
|--------------------------------------|---------|
| Calculate adjusted propensity scores | false |
| Number of clusters | 5 |
| Generate distance field | false |
| Cluster label | String |
| Label prefix | cluster |
| Optimize | Memory |
| Mode | Simple |

Service Details - IBM Cloud IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/28febb2b-1f7b-4b1c-84c2-35c5fb34885f?context=cpdaas&project_id=aee2a8b2-cebe-4e46-ab19-70b005d18190

IBM Watson Studio Search in your workspaces Buy Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

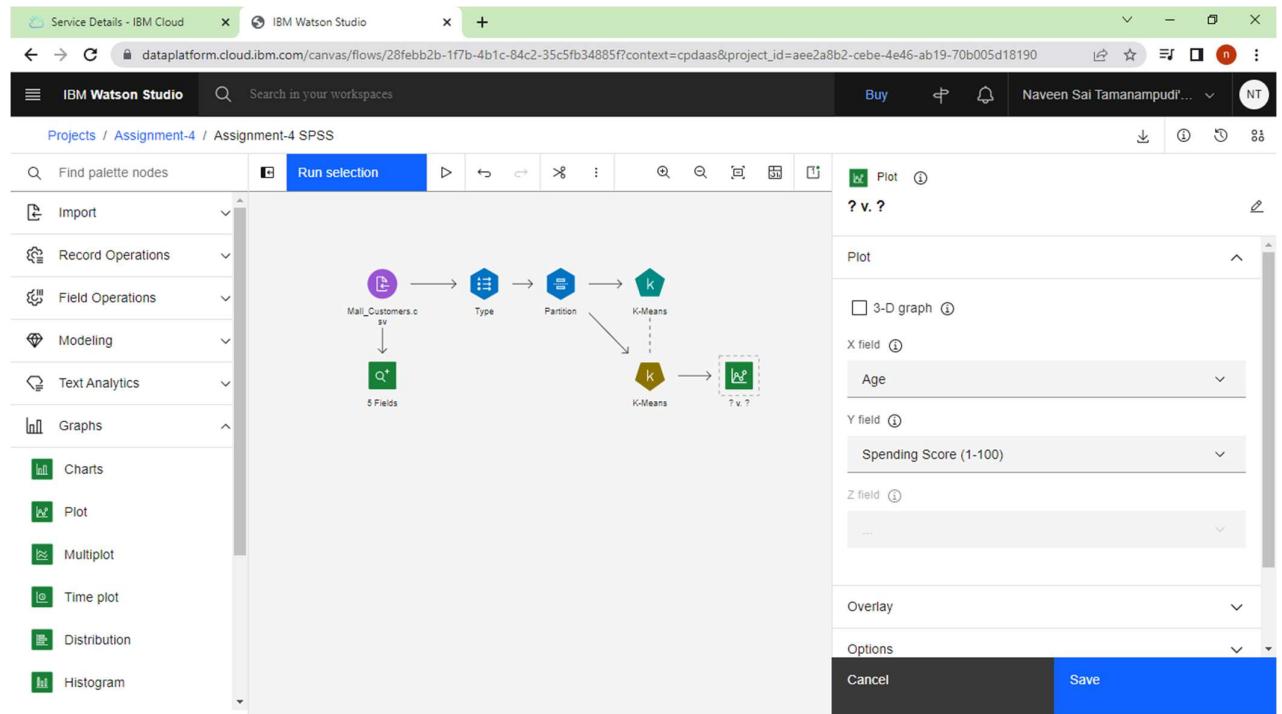
K-Means Clustering Model ①

- Cluster Quality
- MODEL VIEWER
- Model Information
- Feature Importance
- Cluster Sizes
- Cluster Comparison
- Clusters
- Cell Distributions (Absolute)
- Cell Distributions (Relative)
- Build Settings
- Training Summary**

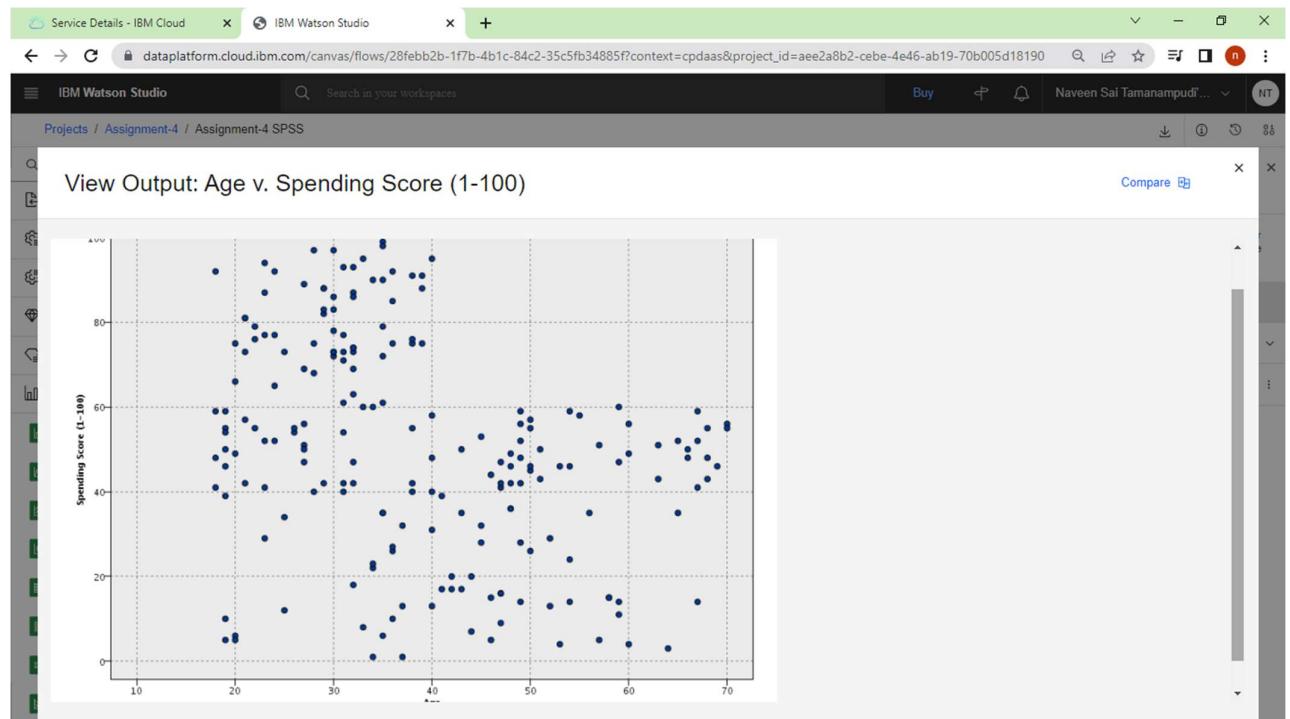
Training Summary ①

| | |
|------------------------------|------------------------------|
| Algorithm | K-means |
| Model type | Clustering |
| Date built | Mon Apr 25 16:59:37 UTC 2022 |
| Elapsed time for model build | 0 hours, 0 mins, 0 secs |

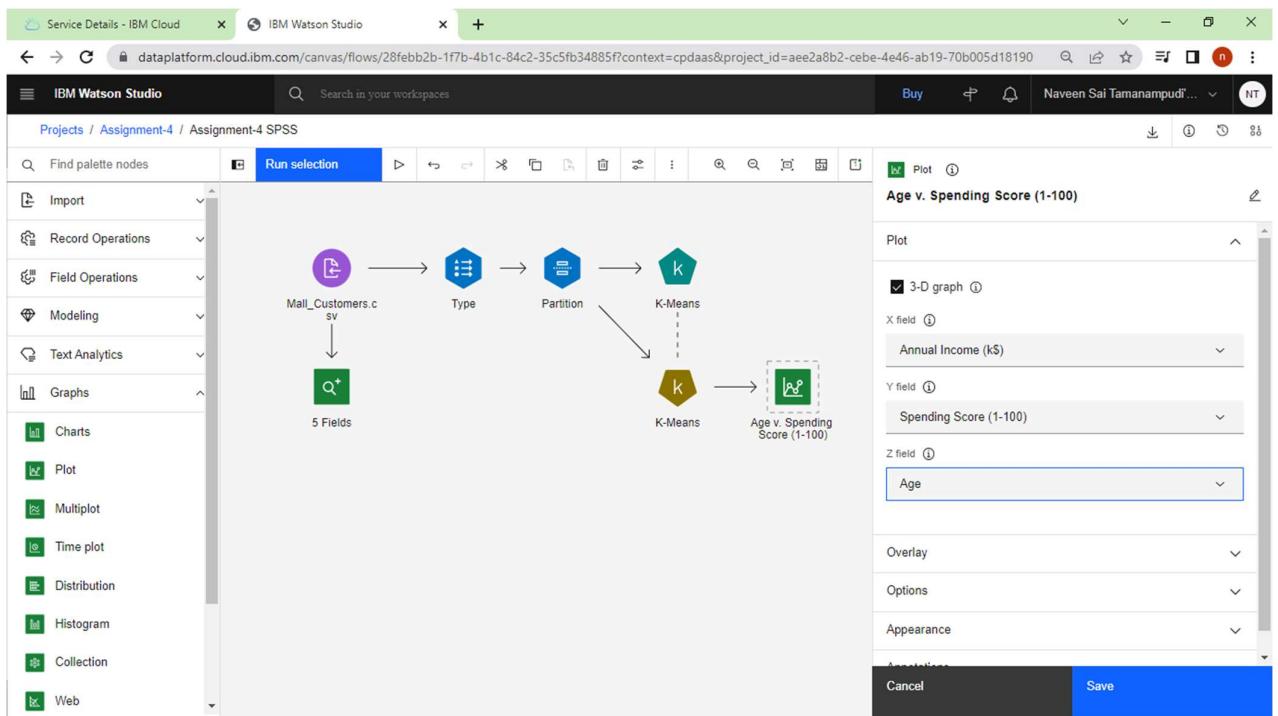
Creating a Plot Node with Age vs Spending Score(1-100):



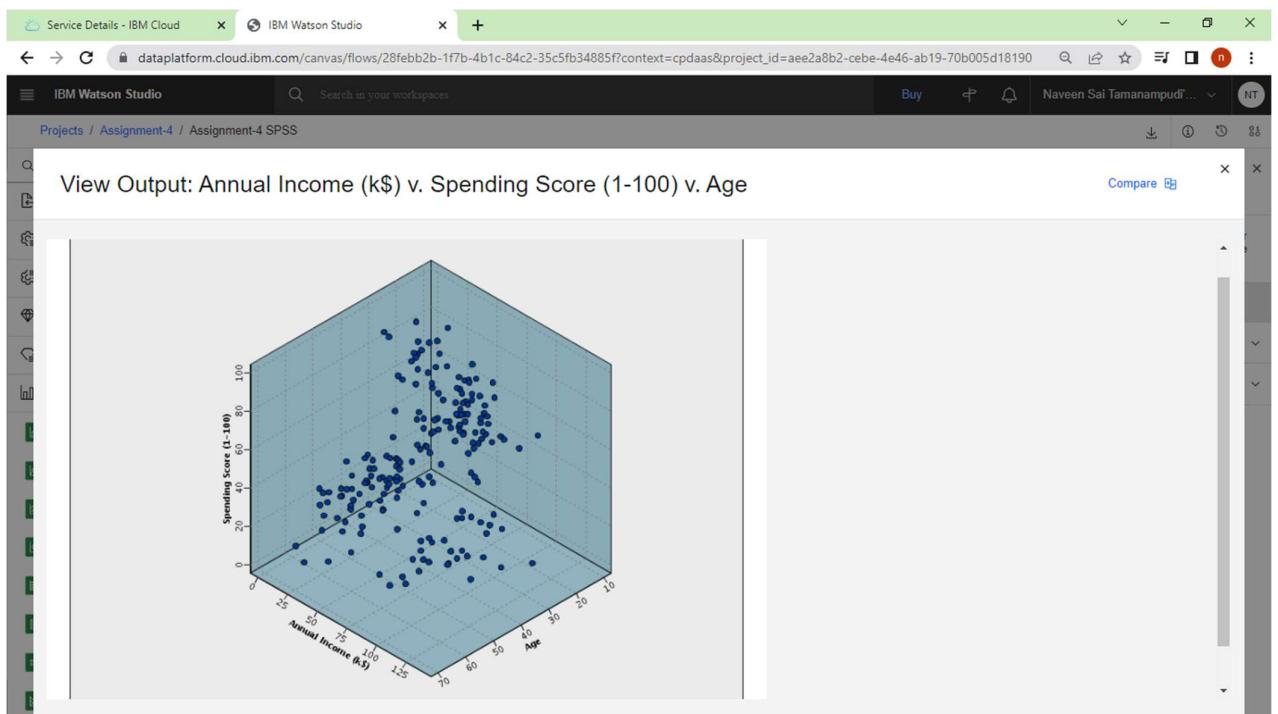
Viewing Plot:



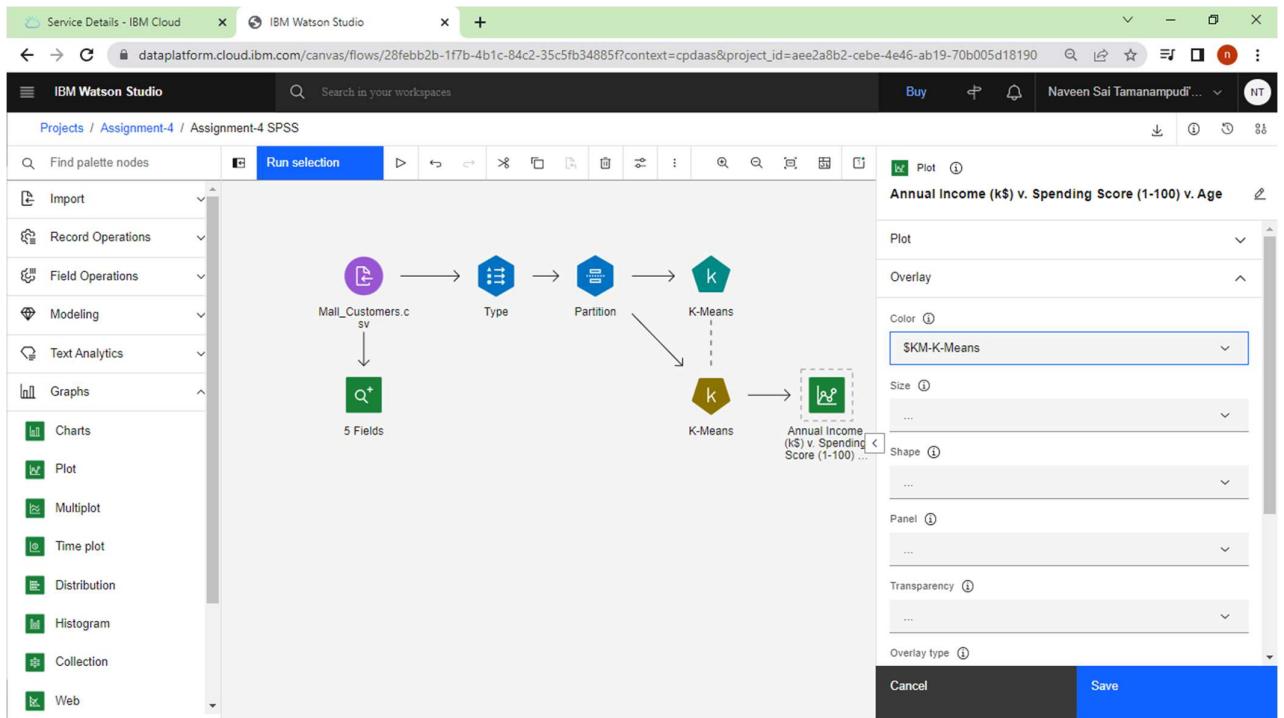
Creating a Plot Node with Annual Income vs Spending Score(1-100) vs Age:



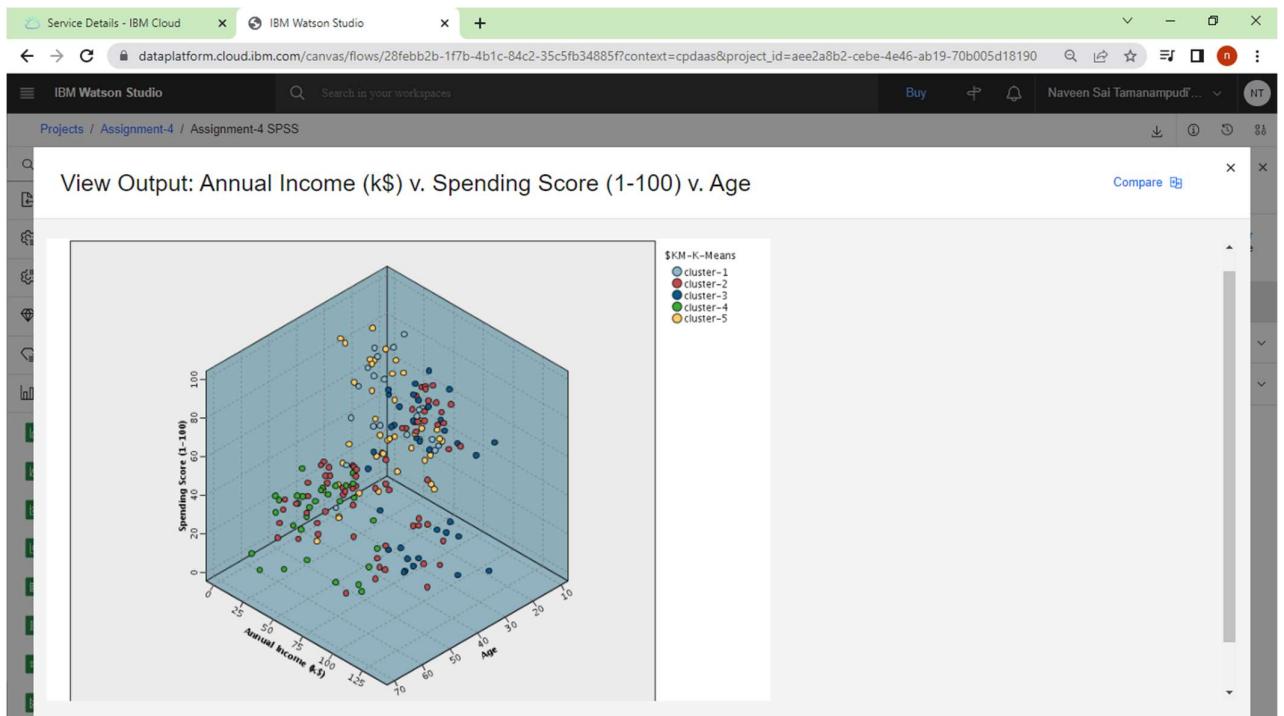
Output:



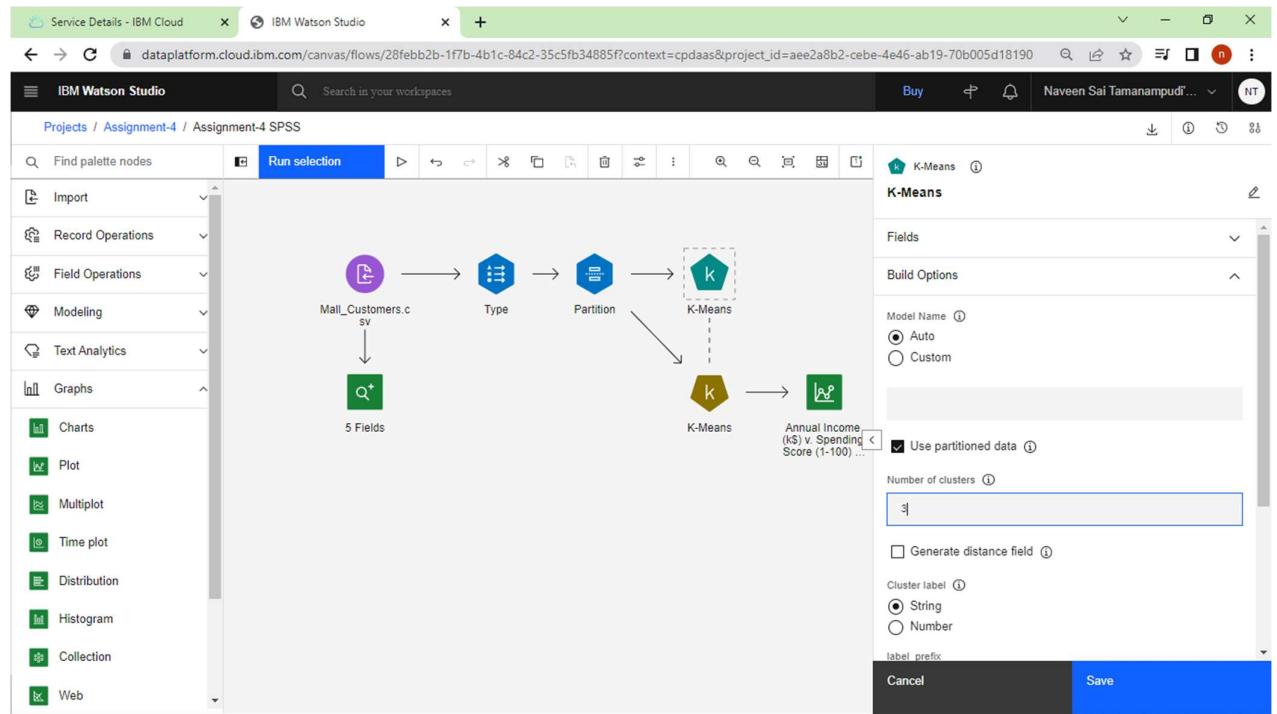
Assigning Colors:



Output:



Changing Number of clusters to 3:



Output:

