JOURNEY SCRAPBOOK

Technical Training - Week 7
03-Oct-2023 to 05-Oct-2023 (Days 23 – 25)

YUVA SAHITH VARMA SANGARAJU

BATCH - 5

Day 23 – Azure ML

- Teachable machine
- Azure ML workspace and studio
- Creating and configuring compute cluster
- Train and test data
- ML types, algorithms and applications
- Running a sample notebook
- Creating a pipeline using the titanic dataset
- Creating data assets using raw GitHub content

Azure Machine Learning ...

Create a machine learning workspace

Resource group * (New) rg_idashell
Create new

Workspace details

Configure your basic workspace settings like its storage connection, authentication, container, and more. Learn more

Name * idaml

Region * East US

Storage account * (new) idaml1659071128
Create new

Key vault * (new) idaml9873364299
Create new

Application insights * (new) idaml6853660130
Create new

Container registry * None
Create new

Review + create | < Previous | Next : Networking

All workspaces
Home
Model catalog PREVIEW
Authoring
Notebooks
Automated ML
Designer
Prompt flow PREVIEW
Assets
Data
Jobs
Components
Pipelines
Environments
Models

idaml

New
Customize view

Generative AI with Prompt flow PREVIEW ...

View prompt flow

QnA with Your Own Data Using ...

Q&A with GPT3.5 using domain knowledge from Faiss index to make the answer more grounded

Start Clone

Bring Your Own Data QnA

Create flows for Q&A with GPT3.5 using data from your own indexed files to make the answer more grounded for entreprise chat scenarios.

Start Clone

Ask Wikipedia

Q&A with GPT3.5 using information from Wikipedia to make your answers more grounded.

Start Clone

Chat wi

ChatGPT-
Wikipedia

Sta

Generative AI models PREVIEW ...

View all

openai-whisper-large
Speech recognition

databricks-dolly-v2-12b
Text generation

gpt-4-32k
Chat completions

gpt-4
Chat comp

Create compute cluster

1. Virtual Machine
2. Advanced Settings

Select virtual machine

Select the virtual machine size you would like to use for your compute cluster.

Location *

East US

Virtual machine tier

- ( ) Dedicated
- (•) Low priority

Virtual machine type

- (•) CPU
- ( ) GPU

Virtual machine size

- ( ) Select from recommended options
- (•) Select from all options

Add filter

Search by VM name...

Showing 192 VM sizes

Back | Next | Cancel

Create compute cluster

Virtual Machine

2 Advanced Settings

Configure Settings

Configure compute cluster settings for your selected virtual machine size.

Name	Category	Cores	Available quota	RAM	Storage	Cost/Node
Standard_DS3_v2	General purpose	4	12 cores	14 GB	28 GB	$0.06/hr

Compute name *

idaclus

Minimum number of nodes *

0

Maximum number of nodes *

1

Idle seconds before scale down *

120

Enable SSH access

Back | Create | Download a template for automation. | Cancel

tough_puppy_06n5f0vq

Running

Overview | Metrics | Images | Child jobs | Outputs + logs | Code | Explanations (preview) | Fairness (preview) | Monitoring

Refresh | Debug and monitor | Resubmit | Register model | Cancel | Delete | Compare (preview)

Properties

Status
Running

Created on
Oct 3, 2023 11:10 AM

Start time
Oct 3, 2023 11:14 AM

Name
6fc076b6-421a-4fbd-9789-b001d2cc3c7f

Script name
Users/Shellunext_1693422046249/azureml-getting-started/azureml-getting-started-studio.ipynb

Created by
Shellunext unextIDA149

Tags

No tags

Metrics

No data

Description

Click edit icon to add a description

Create data asset

- Data type
- Data source
- Web URL
- 4 Review

Review

Review the settings for your data asset and make any changes as needed.

Data type

Name
mlds

Description
--

Type
file

Data source

Type
WebURL

Web URL

Web URL
https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv

Skip data validation
false

Back | Create

Undo | Redo | Validate | Show lineage | ...

Search by name, tags and description

Tags : **All** | Add filter

Data | Component

95

Machine Learning Algorithms (19)

Regression (6)

Boosted Decision Tree Regression
Microsoft
Creates a regression model using the Boosted Decision Tree algorithm. [Learn More](https://aka....
azureml.Designer:true ... 1/10/2023

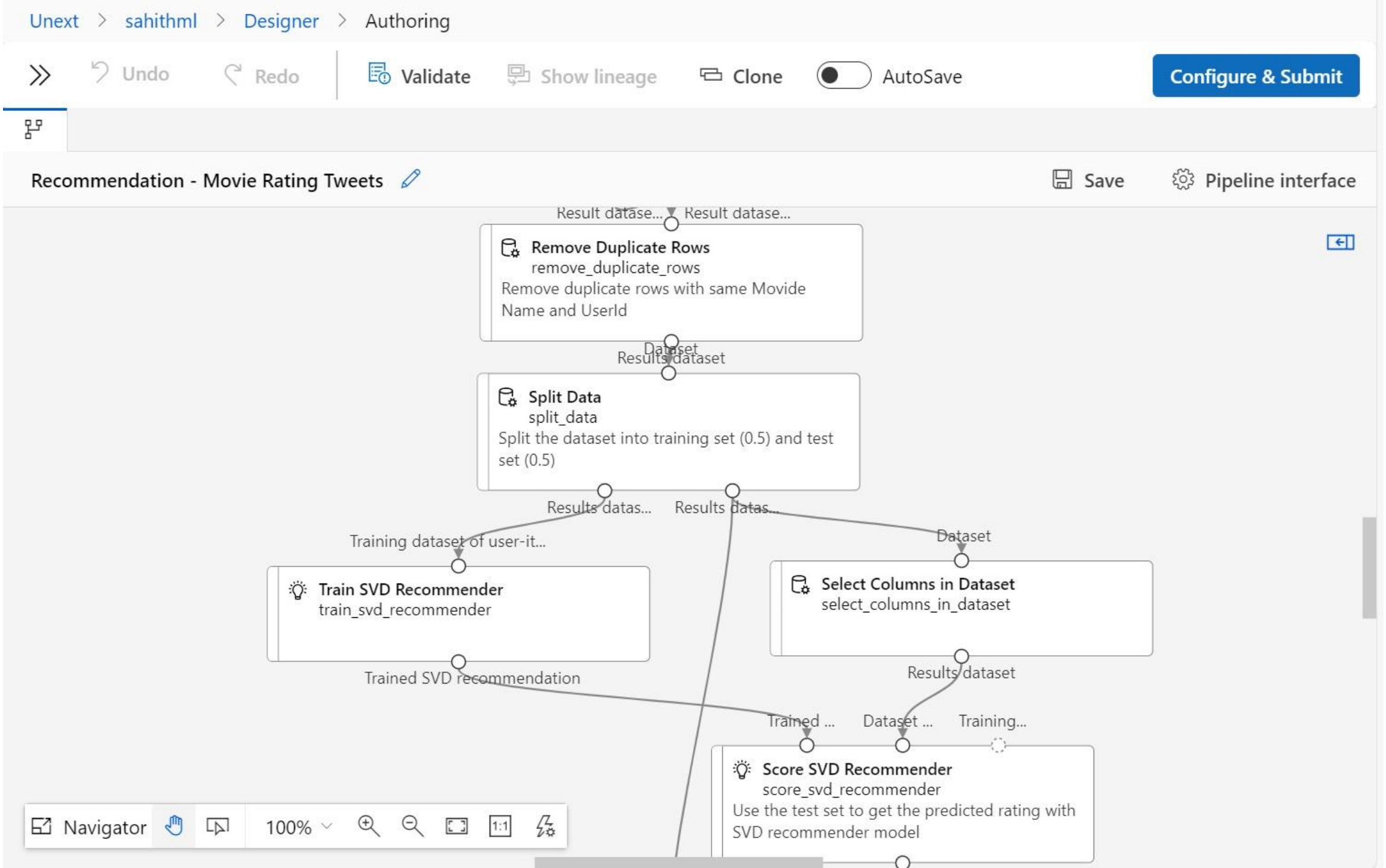
Decision Forest Regression
Microsoft
Creates a regression model using the decision forest algorithm. [Learn More](https://aka.ms/aml...
azureml.Designer:true ... 1/10/2023

Fast Forest Quantile Regression

Pipeline-Created-on-10-03-2023

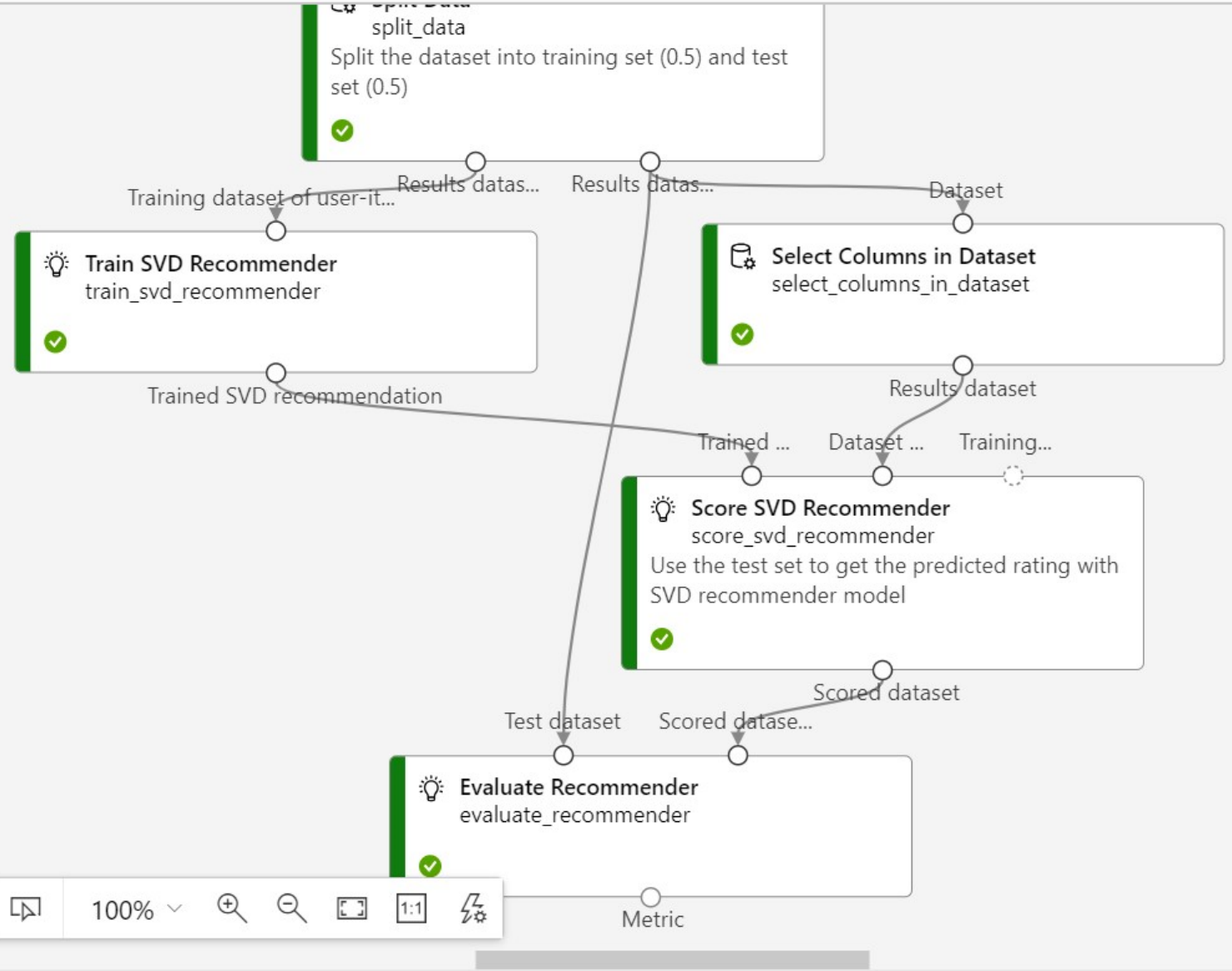
Save | Pipeline interface

Navigator | 100%


Unext
sahithml
Designer
Authoring
Undo
Redo
Validate
Show lineage
Clone
AutoSave
Configure & Submit
Recommendation - Movie Rating Tweets
Save
Pipeline interface
Result datase...
Result datase...
Remove Duplicate Rows
remove_duplicate_rows
Remove duplicate rows with same Movide Name and UserId
Dataset
Results dataset
Split Data
split_data
Split the dataset into training set (0.5) and test set (0.5)
Results datas...
Results datas...
Training dataset of user-it...
Dataset
Train SVD Recommender
train_svd_recommender
Select Columns in Dataset
select_columns_in_dataset
Trained SVD recommendation
Results dataset
Trained ...
Dataset ...
Training...
Score SVD Recommender
score_svd_recommender
Use the test set to get the predicted rating with SVD recommender model
Navigator
100%

Refresh | Clone | Resubmit | Publish | Schedule | Show lineage | ... | Cancel

Recommendation - Movie Rating Tweets | Completed | Share | Job overview



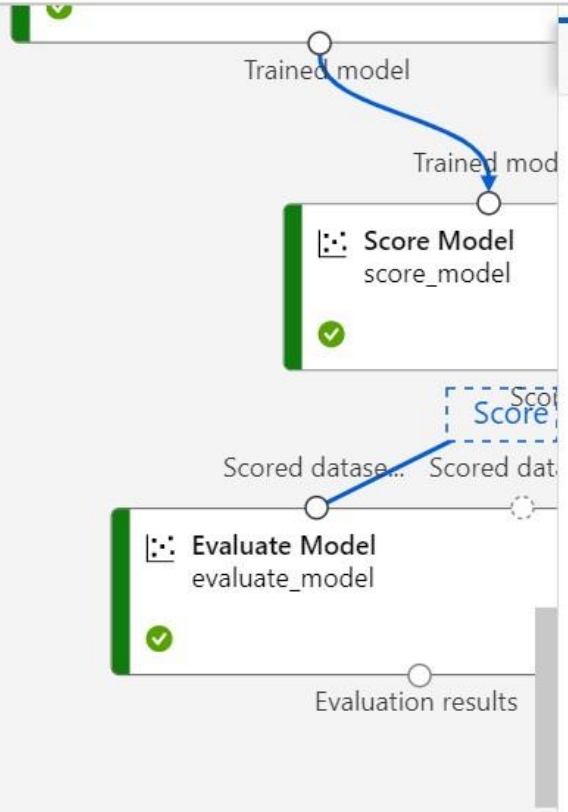
Navigator | 100%

Day 24 – Azure ML Studio

- Implementing pipeline components such as Clean, Split, Train, Score and Evaluate Model
- Running the pipeline as a job
- Hyperparameter optimization – Grid and Random sweeps
- Creating an automated ML with a new experiment
- Implementing the Azure ML Hands-On assignment
- Loading customer dataset from blob storage for preprocessing
- Model development by choosing appropriate algorithm and tuning hyperparameters
- Azure ML steps and importance of some components

Refresh | Clone | Resubmit | Publish | Schedule | Show lineage | Create inference pipeline | Delete | Cancel

TitanicPipeline | Completed | Share | Job overview



Navigator | 100%

Scored_dataset

Rows 267 | Columns 13

PassengerId	Survived	Pclass	Name
756	1	2	Hamalainen, Master. Viljo
204	0	3	Youseff, Mr. Gerious
563	0	2	Norman, Mr. Robert Douglas
54	1	2	Faunthorpe, Mrs. Lizzie (Elizabeth Anne Wilkinson)

To view, select a column in the table

Scored_dataset

Rows 267 Columns 13

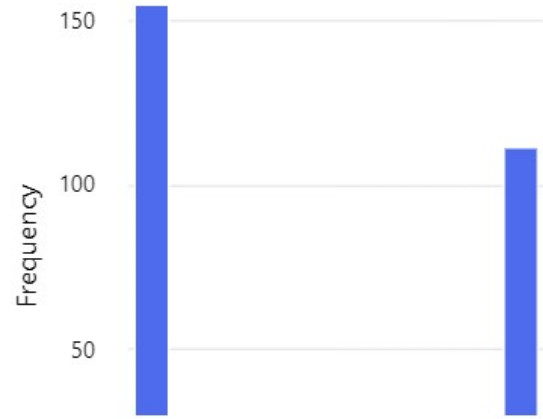
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
756	1	2	Hamalainen, Master. Viljo	male	0.67	1	1	250649
204	0	3	Youseff, Mr. Gerious	male	45.5	0	0	2628
563	0	2	Norman, Mr. Robert Douglas	male	28	0	0	218629
54	1	2	Faunthorpe, Mrs. Lizzie (Elizabeth Anne Wilkinson)	female	29	1	0	2926
552	0	2	Sharp, Mr. Percival James R	male	27	0	0	244358
803	1	1	Carter, Master. William Thornton II	male	11	1	2	113760

Survived

Statistics

Mean	0.4195
Median	0
Min	0
Max	1
Standard deviation	0.4944
Unique values	2
Missing values	0
Feature type	Numeric Label

Visualizations



plucky_stick_99fnbvfw

Not started

Overview | Data guardrails | Models | Outputs + logs | Child jobs

Refresh | Edit and submit (preview) | Register model | Cancel | Delete | Compare (preview)

Properties

Status
Not started

Created on
Oct 4, 2023 11:16 AM

Start time
--

Compute target
idaclus

Name
AutoML_3ecf1da9-cf74-4296-90cd-e9fe768de691

Script name
--

Created by

Inputs

Input name: training_data
Dataset: titanic_ds:1

Best model summary

No data

Run summary

Task type
Regression View configuration settings

Featurization
Auto

Primary metric

dataset

Container

Search

Overview
Diagnose and solve problems
Access Control (IAM)

Settings

Shared access tokens
Access policy
Properties
Metadata

Upload | Change access level | Refresh | Delete | Change tier | Acquire lease | Break lease | View snapshots

Authentication method: Access key (Switch to Azure AD User Account)
Location: dataset

Search blobs by prefix (case-sensitive) | Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Blob type
customer_data.csv	10/4/2023, 2:00:11 PM	Hot (Inferred)		Block blob

Schema preview

Select the column(s) you want to import

Parsing options

	Column name	Type
☐	Path	String
☑	CustomerID	Integer
☑	Age	Integer
☑	AnnualIncome	Decimal (dot '.')
☑	SpendingScore	Integer

Save Cancel

Undo | Redo | Validate | Show lineage | Clone | ... | Configure & Submit

Search by name, tags and description

Tags : **All** | Add filter

Data | Component

95

Enables entering and editing small datasets by typing values. [Learn More](https://aka.ms/aml/en...

azureml.Designer:true ... 1/10/2023

Export Data
Microsoft
Writes a dataset to cloud-based storage in Azure, such as Azure blob storage, Azure Data Lake Stora...

azureml.Designer:true ... 1/10/2023

Import Data
Microsoft
Load data from web URLs or from various cloud-based storage in Azure, such as Azure SQL ...

azureml.Designer:true ... 1/10/2023

Recommendation (5)

Cust_HandsOn

Save | Pipeline interface

Import Data

Data source *

URL via HTTP

Data source URL *

https://idamlws8340933661.blob.core.windows.net/dataset/customer_data.csv

Validated

Preview schema

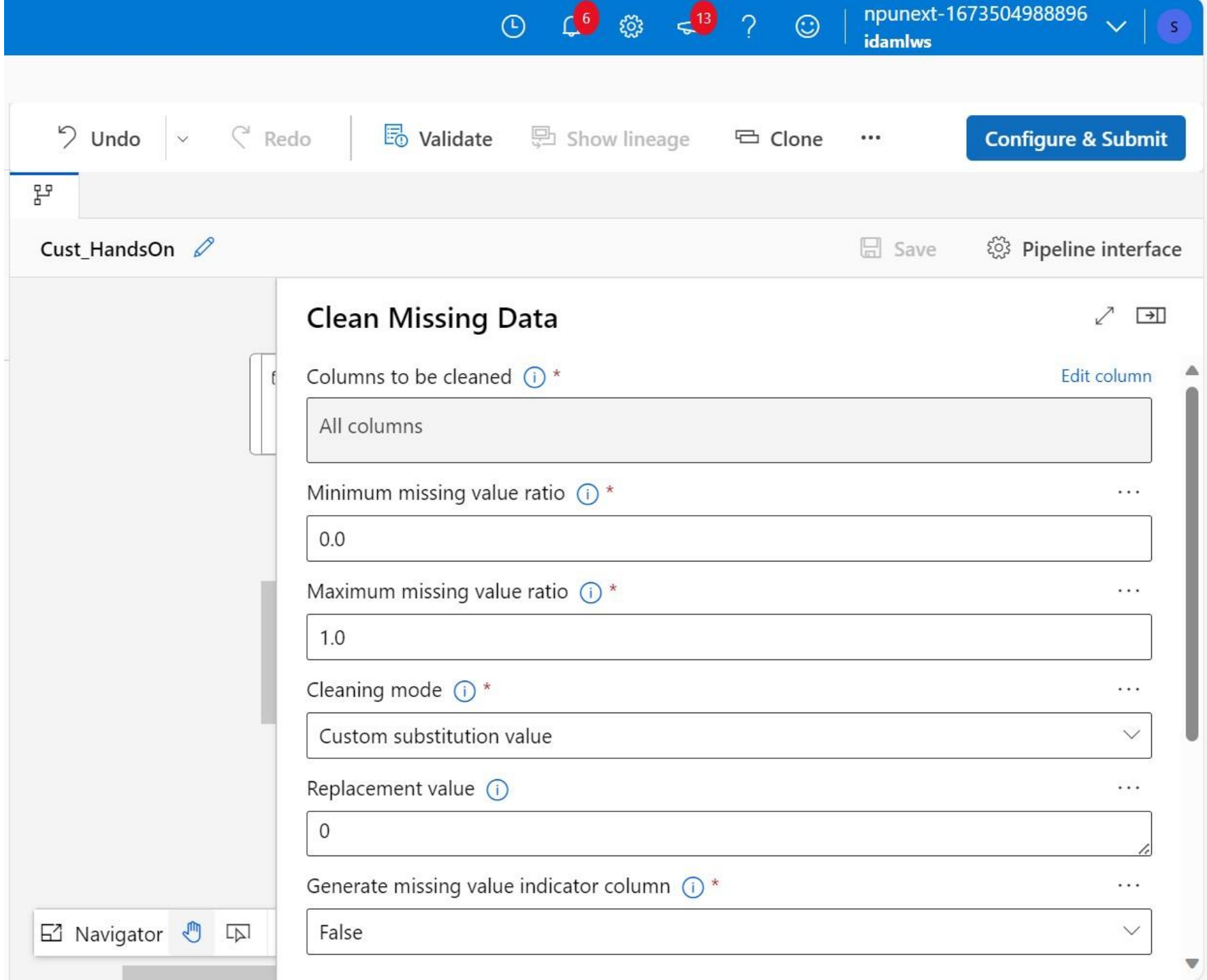
Output settings

Input settings

Run settings

Node information

Navigator


6
13
npunext-1673504988896
idamlws
S
Undo
Redo
Validate
Show lineage
Clone
Configure & Submit
Cust_HandsOn
Save
Pipeline interface
Clean Missing Data
Columns to be cleaned *
Edit column
All columns
Minimum missing value ratio *
0.0
Maximum missing value ratio *
1.0
Cleaning mode *
Custom substitution value
Replacement value
0
Generate missing value indicator column *
False
Navigator

Cust_HandsOn

Save

Pipeline interface

Split Data

Splitting mode *

Split Rows

Fraction of rows in the first output dataset *

0.7

Randomized split *

True

Random seed *

0

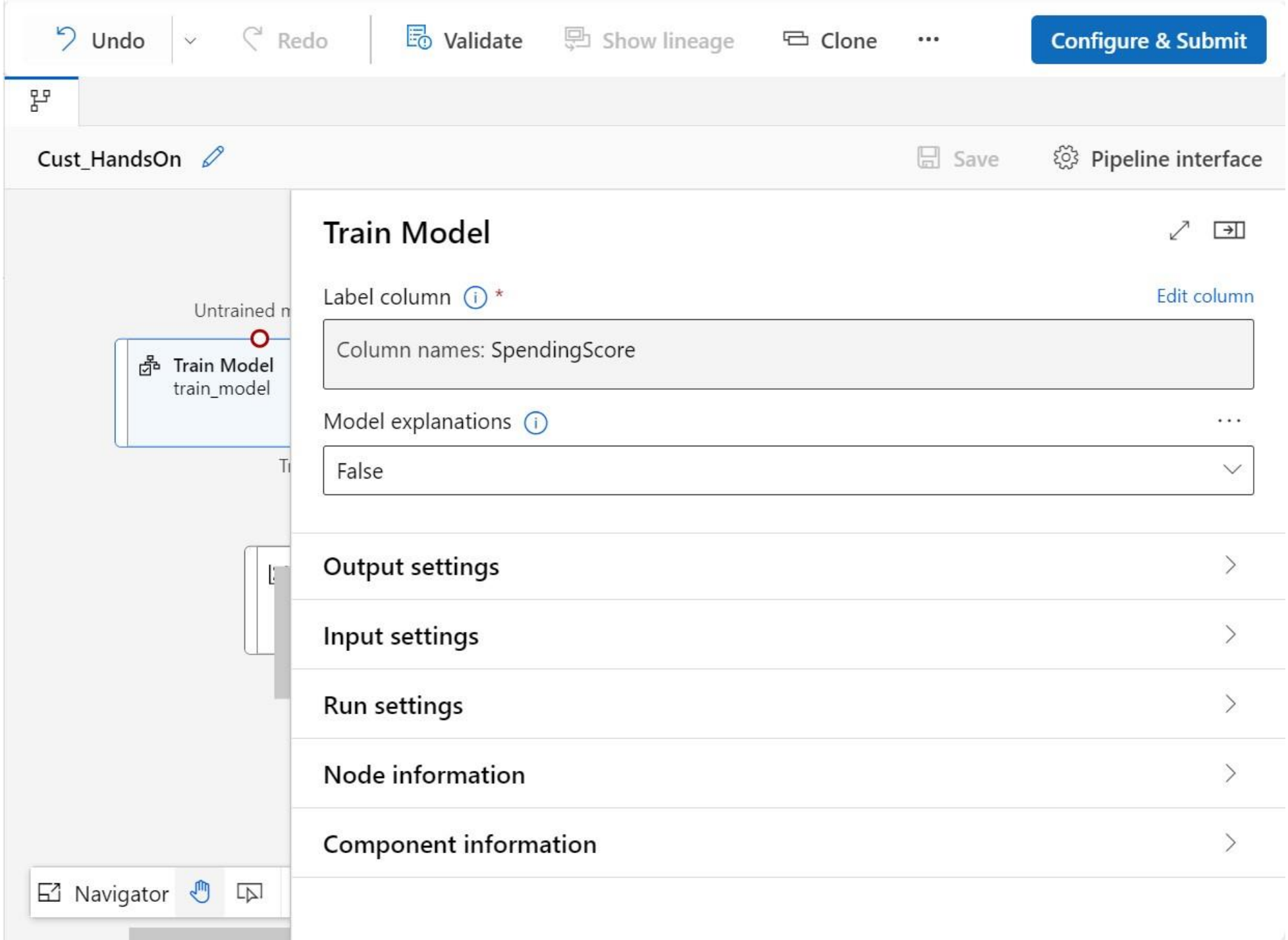
Stratified split *

False

Output settings

Input settings

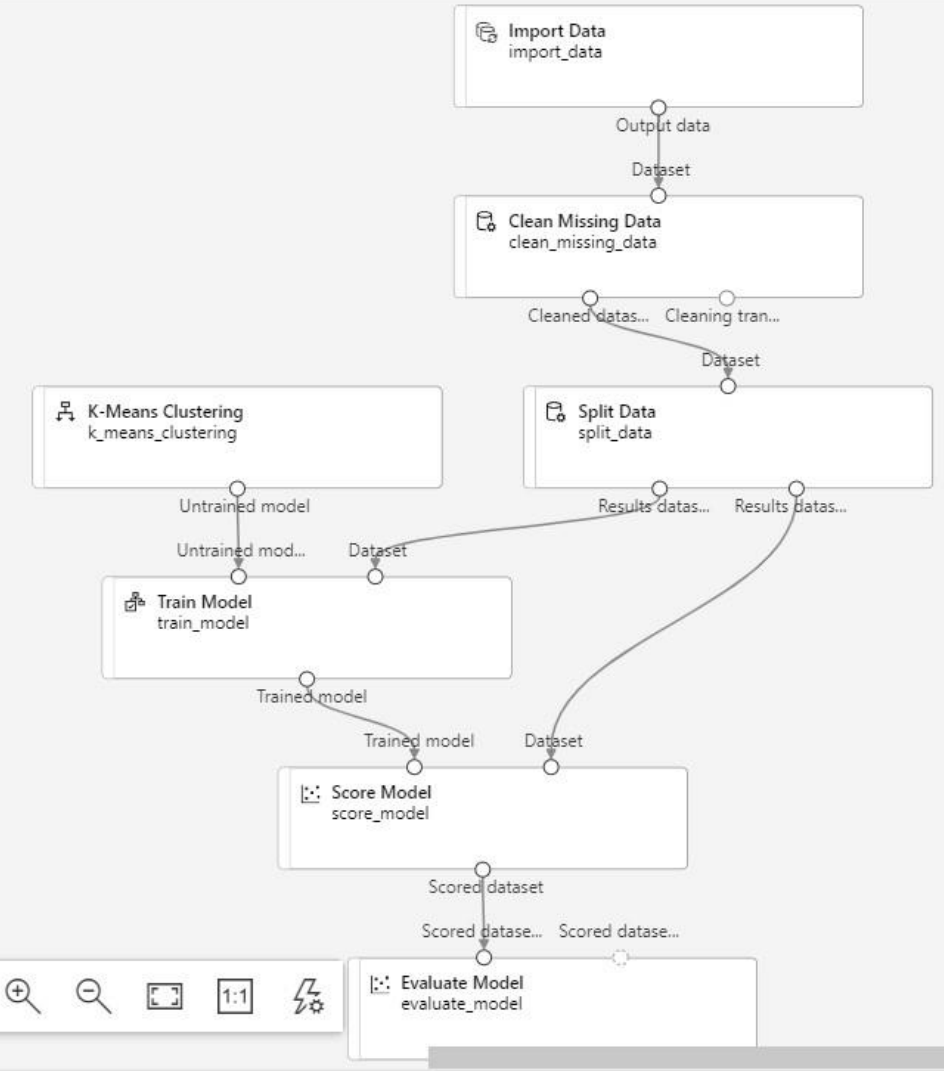
Navigator


Undo
Redo
Validate
Show lineage
Clone
Configure & Submit
Cust_HandsOn
Save
Pipeline interface
Untrained m
Train Model
train_model
Train Model
Label column
Edit column
Column names: SpendingScore
Model explanations
False
Output settings
Input settings
Run settings
Node information
Component information
Navigator

Undo | Redo | Validate | Show lineage | Clone | AutoSave | Configure & Submit

Cust_HandsOn

Save | Pipeline interface



Navigator | 66% | 1:1

Label column

Select a single column | Column names | SpendingScore

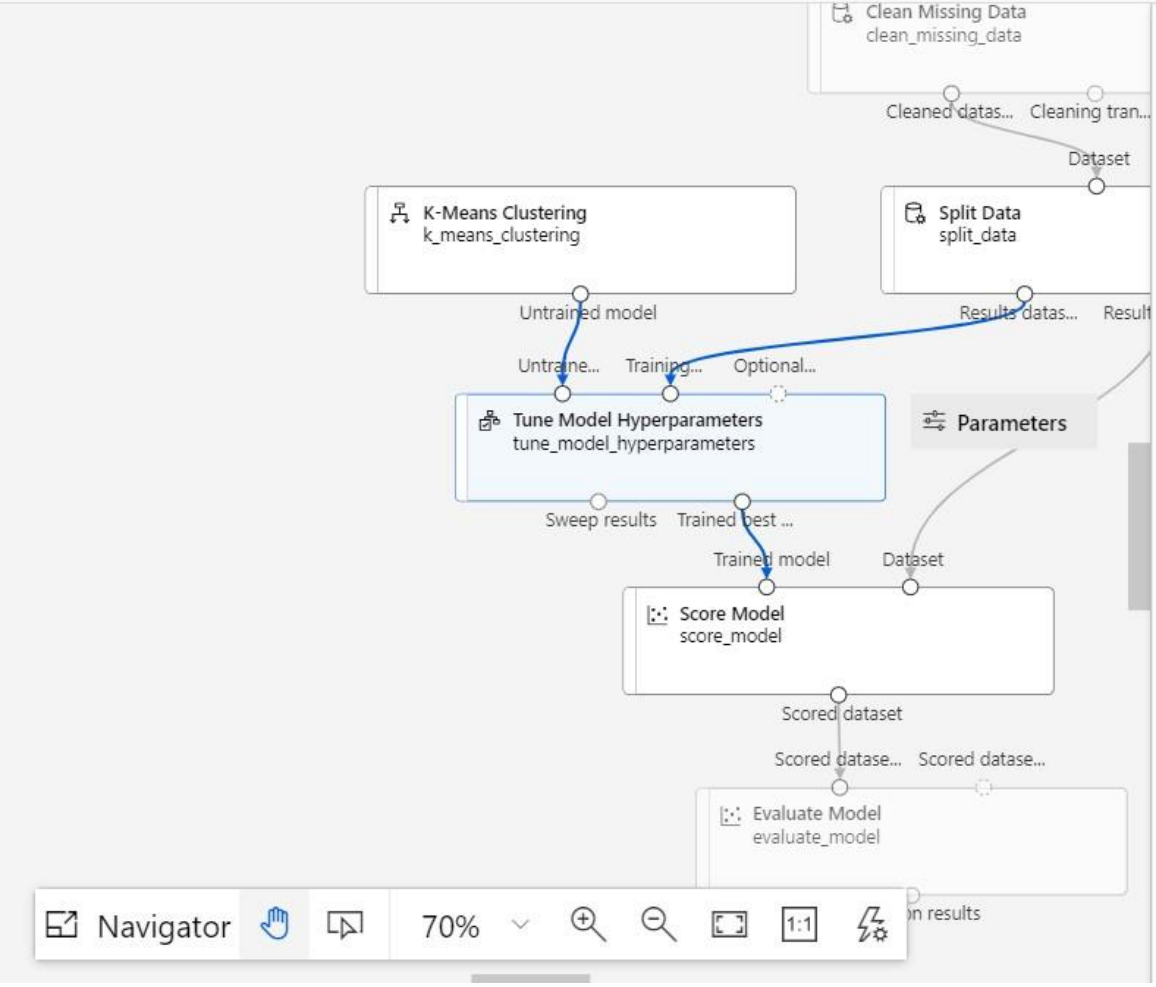
Save | Cancel

Success: Pipeline job has been submitted. View Details

Undo | Redo | Validate | Show lineage | Clone | AutoSave | Configure & Submit

CustHyper_HandsOn

Auto saved | Save | Pipeline interface



Tune Model Hyperparameters

Specify parameter sweeping mode *

Random sweep

Maximum number of runs on random sweep *

5

Random seed *

0

Metric for measuring performance for classification *

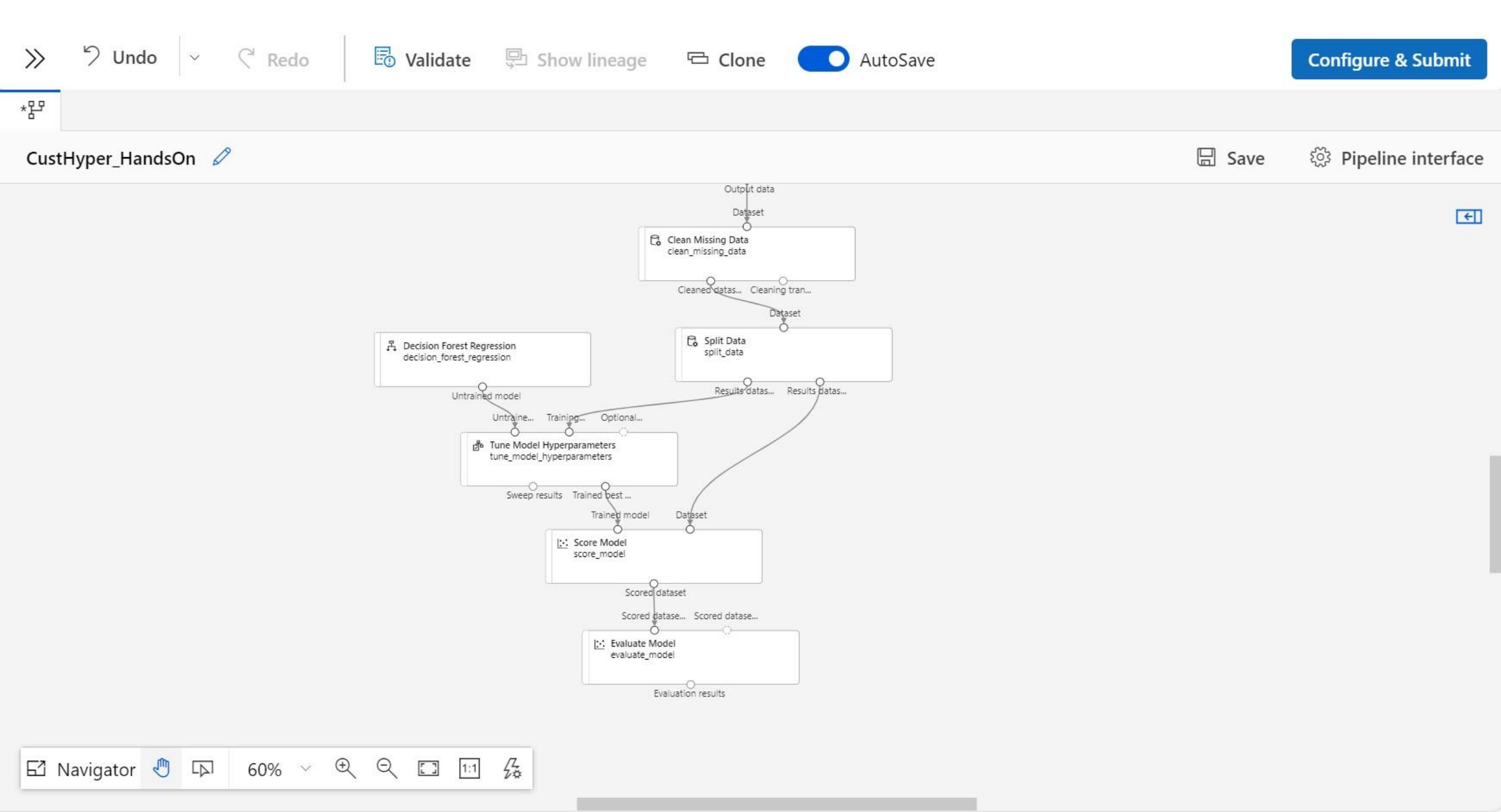
F-score

Metric for measuring performance for regression *

Coefficient of determination

Label column * Edit column

Column names: SpendingScore


Undo
Redo
Validate
Show lineage
Clone
AutoSave
Configure & Submit
CustHyper_HandsOn
Save
Pipeline interface
Output data
Dataset
Clean Missing Data
clean_missing_data
Cleaned datas...
Cleaning tran...
Dataset
Decision Forest Regression
decision_forest_regression
Split Data
split_data
Untrained model
Results datas...
Results datas...
Untraine...
Training...
Optional...
Tune Model Hyperparameters
tune_model_hyperparameters
Sweep results
Trained best ...
Trained model
Dataset
Score Model
score_model
Scored dataset
Scored datase...
Scored datase...
Evaluate Model
evaluate_model
Evaluation results
Navigator
60%
1:1

Graph validation

Refresh

The pipeline looks good,
you can submit or publish now

Set up pipeline job

1. Basics
2. Inputs & outputs
3. Runtime settings
4. **Review + Submit**

Review + Submit

Basics

Job display name
CustHyper_HandsOn

Job description
Pipeline created on 20231004

Experiment
idaexp

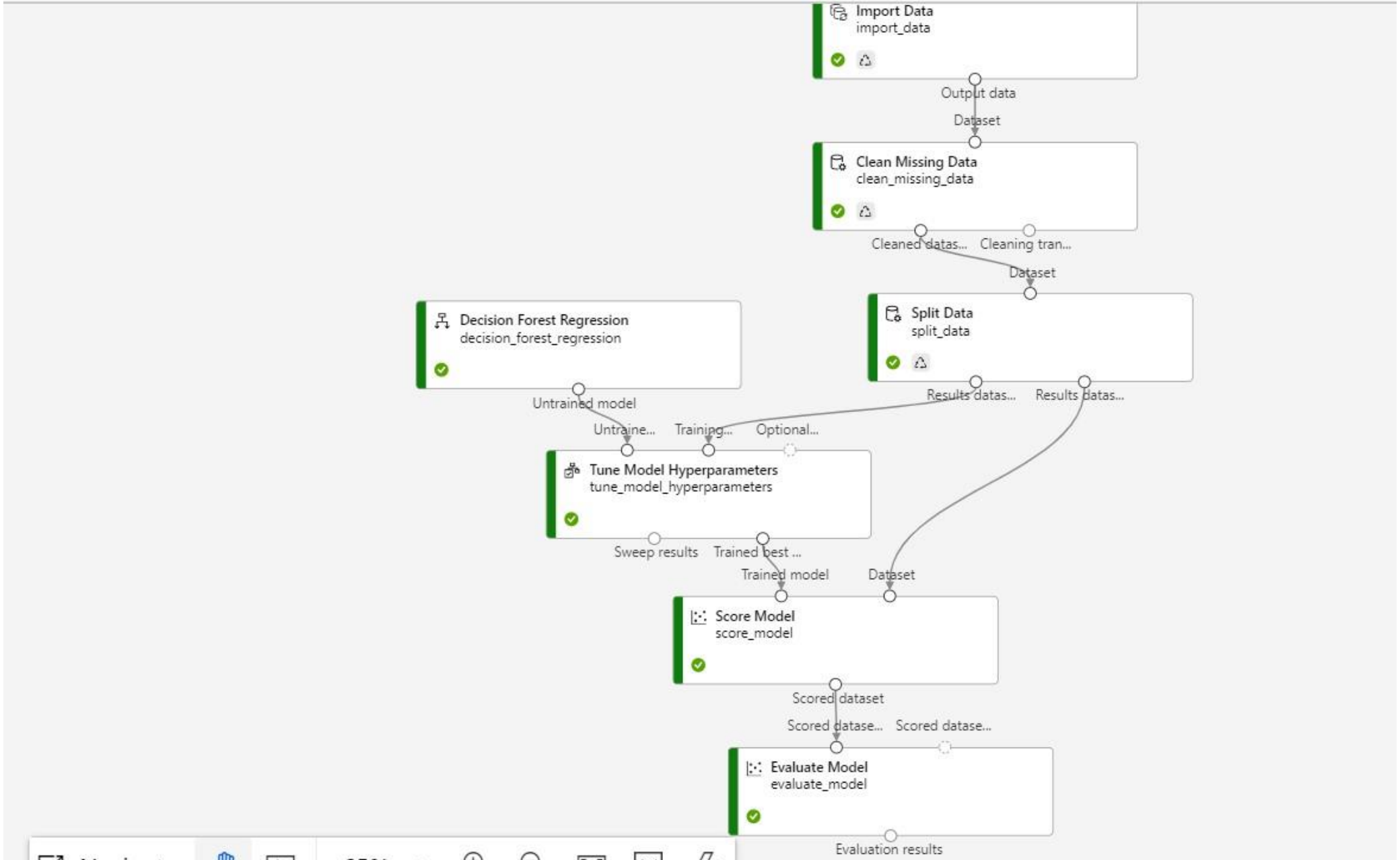
Tags
No tags

Inputs & outputs

Inputs
None

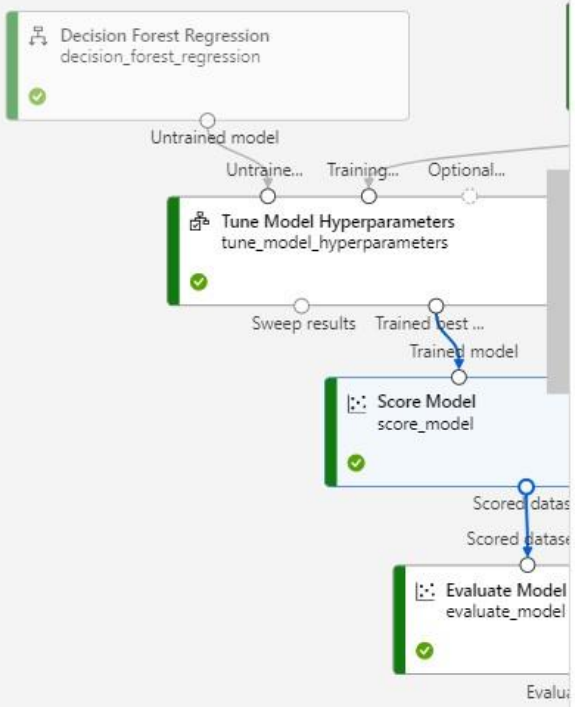
Submit | Back | Next | Close

CustHyper_HandsOn Completed



Refresh | Clone | Resubmit | Publish | Schedule | Show lineage | Create inference pipeline | Delete | Cancel

CustHyper_HandsOn | Completed | Share | Job overview



Navigator | 65%

Scored_dataset

Rows 60 | Columns 5

CustomerID	Age	AnnualIncome	Spending
29	0	84987	0
89	0	146982	0
134	0	43627	0
78	0	0	0
2	0	45194	0
166	0	140427	0
152	0	564965	0
104	0	113795	0
22	0	122765	0
189	0	102036	0
115	0	152883	0

To view, select a column in the table

Scored Labels

0.479863

0.091671

0.046765

0.83052

0.114466

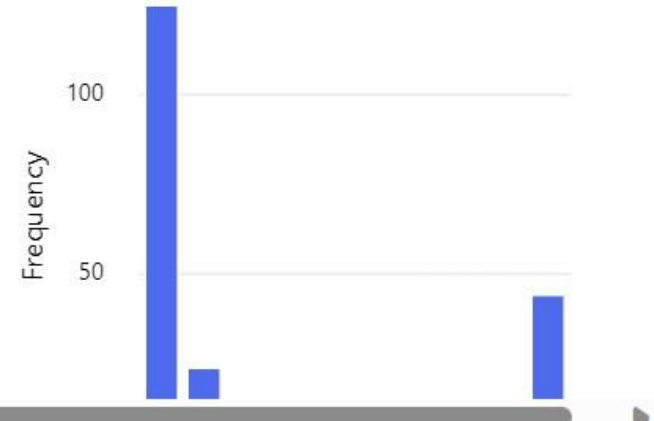
0.345171

Scored Labels

Statistics

Mean	0.342
Median	0.1492
Min	0.029
Max	0.9497
Standard deviation	0.3414
Unique values	266
Missing values	0
Feature type	Numeric Score

Visualizations



Day 25 – Case Study Implementation

- Sales360 Analysis
- Requirements, Planning and Architecture
- Data pipeline, layers and streams
- Azure blob storage to store all the files and tables
- Medallion architecture and using SQL DB as a backup
- Dynamic data loading and transfer using Data factory
- Data cleaning and transformation in Databricks
- Visualization and reporting using Power BI
- CI/CD with Azure DevOps and GitHub
- Connecting all the resources and services
- Documentation of the work done

Create Data Factory

One-click to create data factory with sample pipeline and datasets. Try it

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * : npunext-1680262033736

Resource group * : CaseStudyT3

Create new

Instance details

Name * : Sales360adf

Region * : East US

Version * : V2

Move resources

sales360store

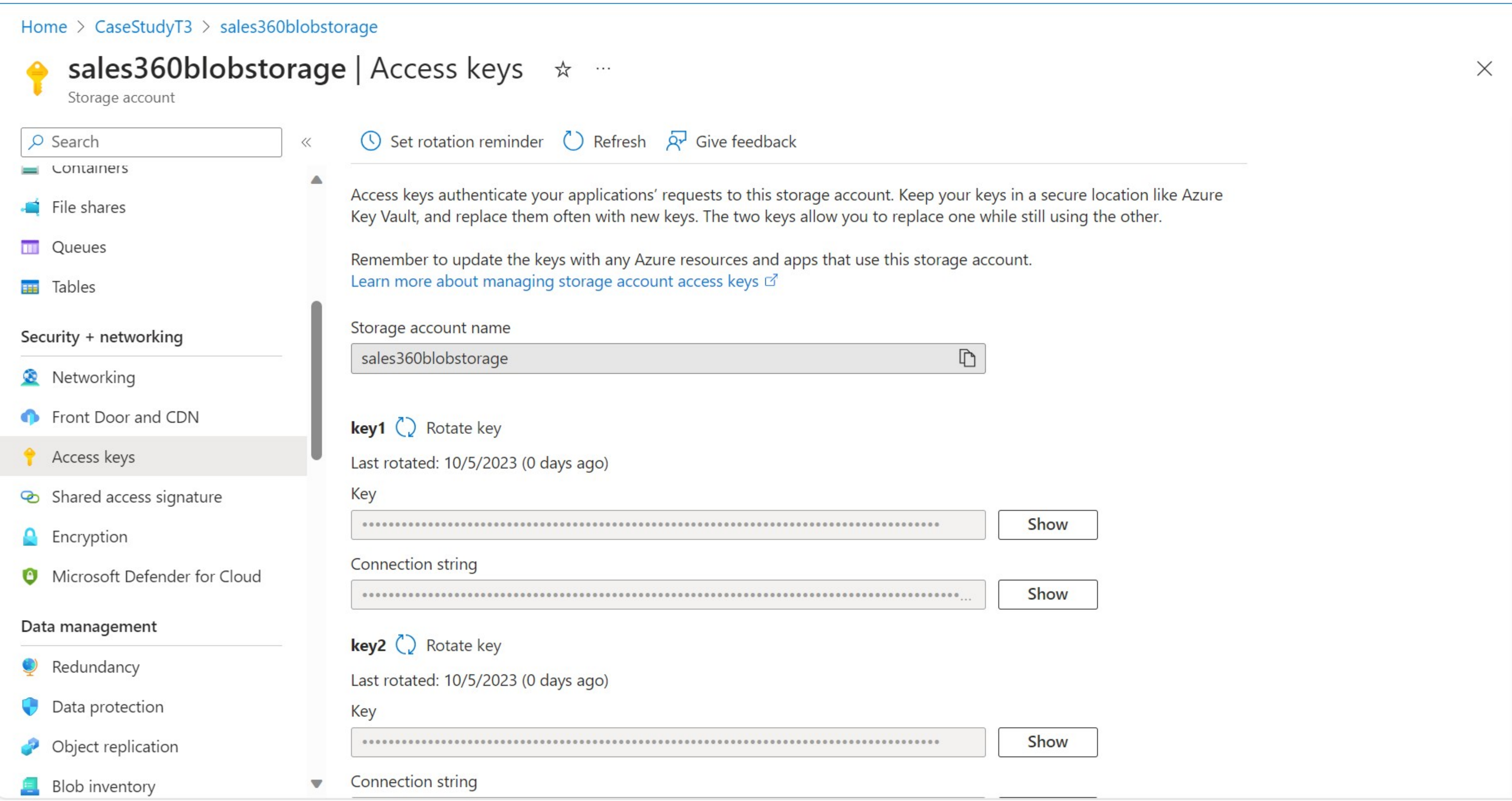
1 Source + target — 2 Resources to move — Review

Selection summary

Source subscription	npunext-1673504988896
Source resource group	salesrg
Target subscription	npunext-1680262033736
Target resource group	CaseStudyT3
Number of resources to move	1

I understand that tools and scripts associated with moved resources will not work until I update them to use new resource IDs

Previous | Move


Home > CaseStudyT3 > sales360blobstorage
sales360blobstorage | Access keys
Storage account
Search
Containers
File shares
Queues
Tables
Security + networking
Networking
Front Door and CDN
Access keys
Shared access signature
Encryption
Microsoft Defender for Cloud
Data management
Redundancy
Data protection
Object replication
Blob inventory
Set rotation reminder
Refresh
Give feedback
Access keys authenticate your applications' requests to this storage account. Keep your keys in a secure location like Azure Key Vault, and replace them often with new keys. The two keys allow you to replace one while still using the other.
Remember to update the keys with any Azure resources and apps that use this storage account.
Learn more about managing storage account access keys
Storage account name
sales360blobstorage
key1
Rotate key
Last rotated: 10/5/2023 (0 days ago)
Key
Show
Connection string
Show
key2
Rotate key
Last rotated: 10/5/2023 (0 days ago)
Key
Show
Connection string

New
Workspace
Recents
Catalog
Workflows
Compute

SQL
SQL Editor
Queries
Dashboards NEW
Alerts
Query History
SQL Warehouses

Data Engineering
Job Runs
Data Ingestion
Delta Live Tables

Machine Learning
Experiments

Sales360DataBricksNotebook

Python | File Edit View Run Help | Last edit was 1 minut... | Provide feedback | Run all | Shellunext's Cluster | Schedule | Share

Cmd 1

```
1  dbutils.fs.mount(source = "wasbs://parent@sales360blobstorage.blob.core.windows.net",mount_point ="/mnt/
   parent",extra_configs = {"fs.azure.account.key.sales360blobstorage.blob.core.windows.net":dbutils.
   secrets.get(scope = "sales360blobscope", key = "sales360blobsecret")})
```

```
True
```

Command took 12.29 seconds -- by shellunext_1693422051344@npunext.onmicrosoft.com at 10/5/2023, 10:17:15 AM on Shellunext's Cluster

Cmd 2

```
1  dbutils.secrets.listScopes()
```

```
[SecretScope(name='sales360blobscope')]
```

Command took 0.75 seconds -- by shellunext_1693422051344@npunext.onmicrosoft.com at 10/5/2023, 10:14:46 AM on Shellunext's Cluster

Cmd 3

```
1  dbutils.fs.mounts()
```

```
[MountInfo(mountPoint='/databricks-datasets', source='databricks-datasets', encryptionType=''),
 MountInfo(mountPoint='/Volumes', source='UnityCatalogVolumes', encryptionType=''),
 MountInfo(mountPoint='/databricks/mlflow-tracking', source='databricks/mlflow-tracking', encryptionType=''),
 MountInfo(mountPoint='/databricks-results', source='databricks-results', encryptionType=''),
 MountInfo(mountPoint='/databricks/mlflow-registry', source='databricks/mlflow-registry', encryptionType=''),
 MountInfo(mountPoint='/Volume', source='DbfsReserved', encryptionType=''),
 MountInfo(mountPoint='/mnt/parent', source='wasbs://parent@sales360blobstorage.blob.core.windows.net', encryptio
nType=''),
```

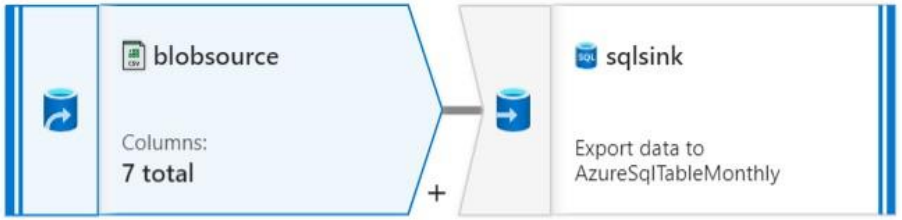
Microsoft Azure | Data Factory ▸ Sales360 | Search factory and documentation | Shellunext_1693422046249@npunext.onmicrosoft.com UNEXT

Your insights matter! Participate in our brief survery about our CDC top-level resource, and help us enhance your experience.

Data Factory | Validate all | Publish all 4 | Preview experience Off

Sales360FetchADF | BlobToSqlDataflow | monthlyprice | AzureSqlTableMonthly | Binary2

Validate | Data flow debug | Debug Settings



Source settings | Source options | Projection | Optimize | Inspect | Data preview

Number of columns **Total** 7

Order	Column	Type

Sales360FetchADF ● | BlobToSqlDataflow ● | monthlyprice ● | AzureSqlTableMonthly ● | Binary2

Validate | Validate copy runtime | Debug | Add trigger | Data flow debug

Activities

Search activities

- Move and transform
  - Copy data
  - Data flow
- Synapse
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning

General | Source | **Sink** | Mapping | Settings | User properties

Sink dataset * : AzureSqlTableMonthly | Open | New | Learn more

Write behavior : (●) Insert ( ) Upsert ( ) Stored procedure

Bulk insert table lock : ( ) Yes (●) No

Table option : (●) None ( ) Auto create table

Pre-copy script :

Write batch timeout : e.g. 00:30:00

Write batch size :

Max concurrent connections :

Disable performance metrics analytics : [ ]

43.204.37.18 - Remote Desktop Connection

Paste | Cut | Copy | Format painter — Clipboard

Get data | Excel workbook | Data hub | SQL Server | Enter data | Dataverse | Recent sources — Data

Transform data | Refresh — Queries

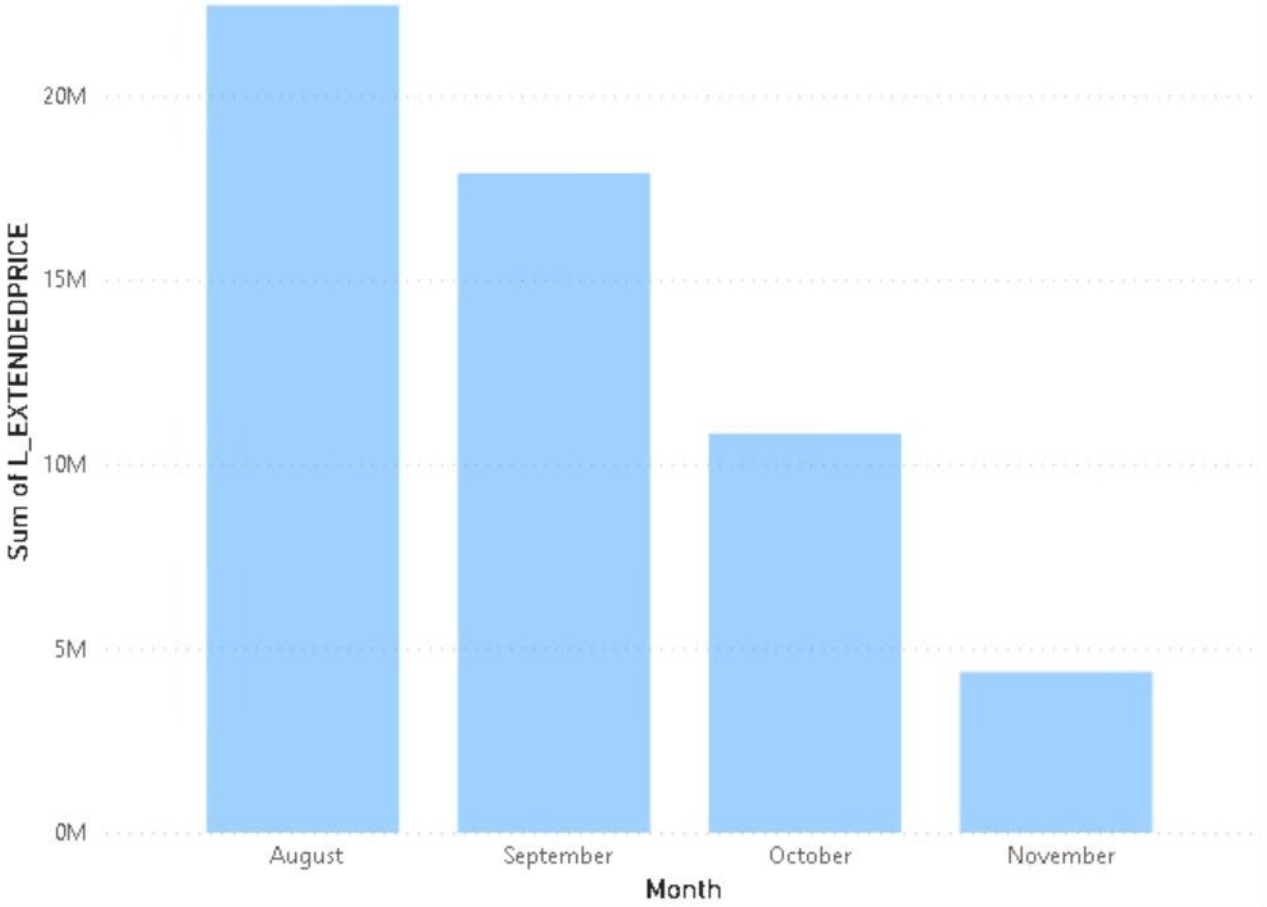
New visual | Text box | More visuals — Insert

New measure | Quick measure — Calculations

Sensitivity — Sensitivity

Publish — Share

Back to report | SUM OF L_EXTENDEDPRICE BY MONTH



Filters

Search

Filters on this visual

L_SHIPDATE
is on or before Sunda...

L_SHIPDATE - Month
is (All)

Sum of L_EXTENDEDP...
is (All)

Add data fields here

Filters on this page

Add data fields here

Filters on all pages

Add data fields here

Visualizations

Build visual

X-axis

L_SHIPDATE
Month

Y-axis

Sum of L_EXTENDEDP...

Legend

Add data fields here

Small multiples

Add data fields here

Fields

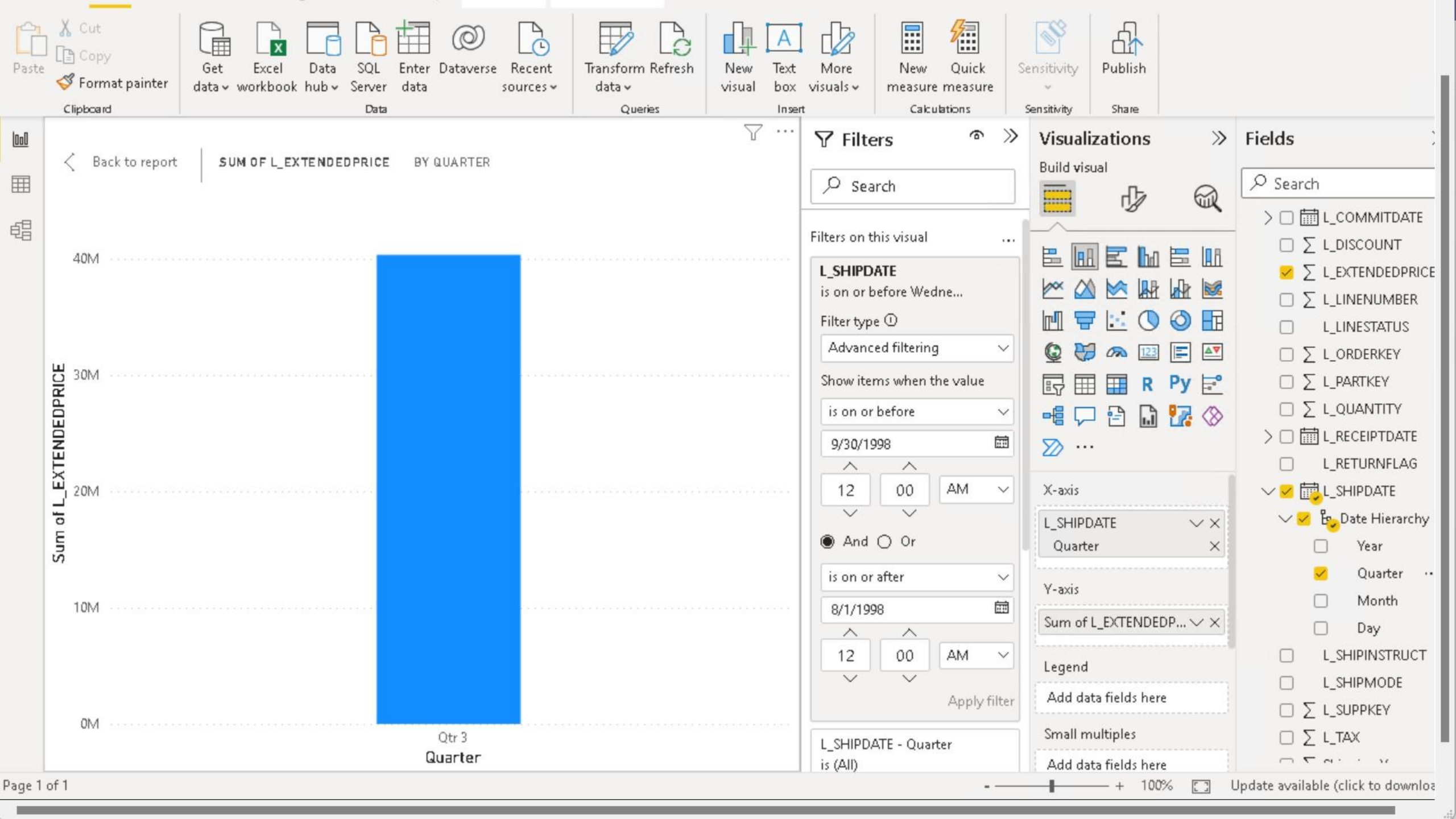
Search

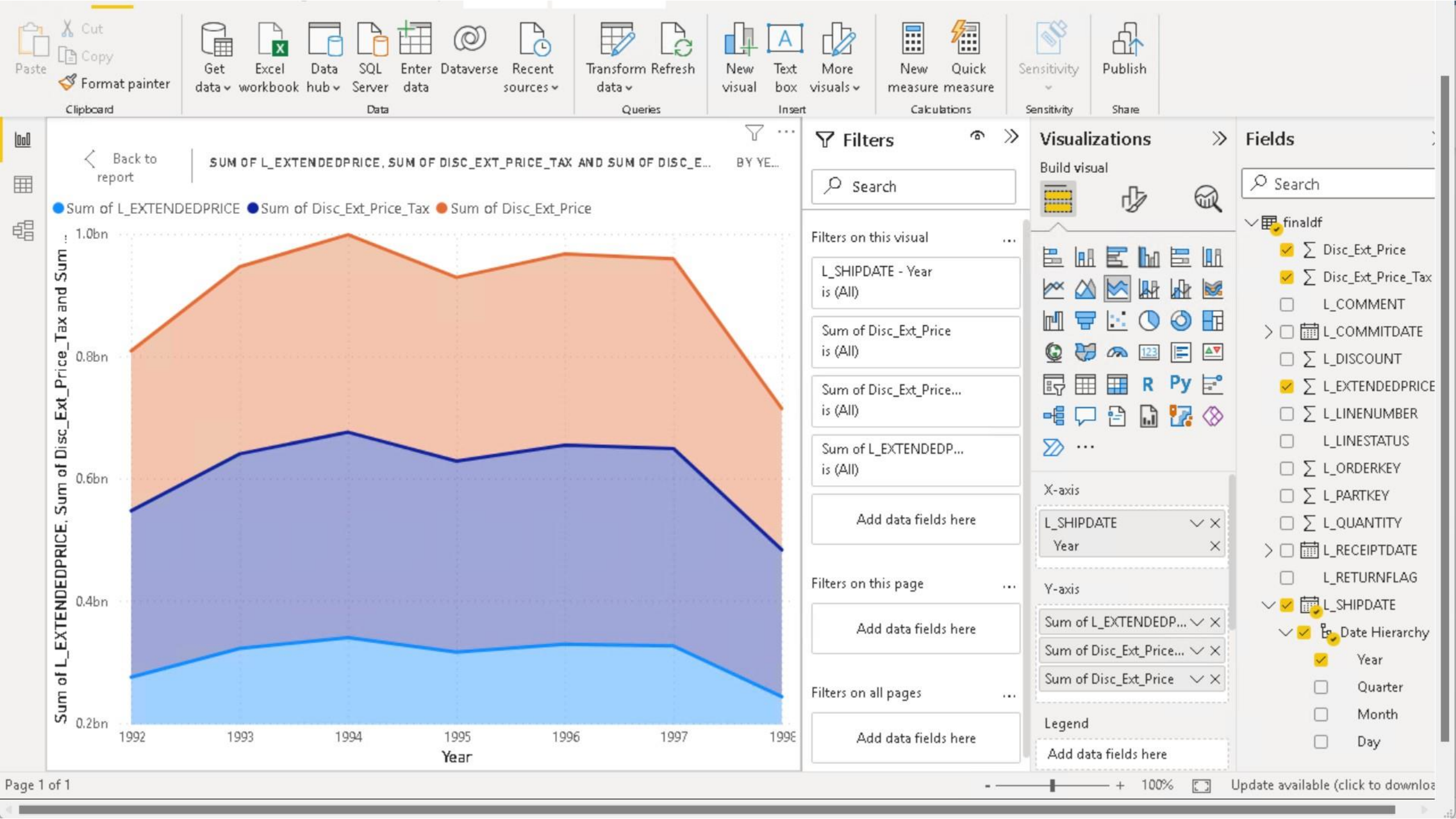
finaldf
- Disc_Ext_Price
- Disc_Ext_Price_Tax
- L_COMMENT
- L_COMMITDATE
- L_DISCOUNT
- L_EXTENDEDPRICE
- L_LINENUMBER
- L_LINESTATUS
- L_ORDERKEY
- L_PARTKEY
- L_QUANTITY
- L_RECEIPTDATE
- L_RETURNFLAG
- L_SHIPDATE
  - Date Hierarchy
    - Year
    - Quarter
    - Month
    - Day

Page 1 of 1

100%

Update available (click to downloa


Paste
Cut
Copy
Format painter
Clipboard
Get data
Excel workbook
Data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
New measure
Quick measure
Calculations
Sensitivity
Sensitivity
Publish
Share
Back to report
SUM OF L_EXTENDEDPRICE
BY QUARTER
40M
30M
20M
10M
0M
Sum of L_EXTENDEDPRICE
Qtr 3
Quarter
Filters
Search
Filters on this visual
L_SHIPDATE
is on or before Wedne...
Filter type
Advanced filtering
Show items when the value
is on or before
9/30/1998
12
00
AM
And
Or
is on or after
8/1/1998
12
00
AM
Apply filter
L_SHIPDATE - Quarter
is (All)
Visualizations
Build visual
X-axis
L_SHIPDATE
Quarter
Y-axis
Sum of L_EXTENDEDP...
Legend
Add data fields here
Small multiples
Add data fields here
Fields
Search
L_COMMITDATE
L_DISCOUNT
L_EXTENDEDPRICE
L_LINENUMBER
L_LINESTATUS
L_ORDERKEY
L_PARTKEY
L_QUANTITY
L_RECEIPTDATE
L_RETURNFLAG
L_SHIPDATE
Date Hierarchy
Year
Quarter
Month
Day
L_SHIPINSTRUCT
L_SHIPMODE
L_SUPPKEY
L_TAX
Page 1 of 1
100%
Update available (click to downloa


Paste
Cut
Copy
Format painter
Clipboard
Get data
Excel workbook
Data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
New measure
Quick measure
Calculations
Sensitivity
Sensitivity
Publish
Share
Back to report
SUM OF L_EXTENDEDPRICE, SUM OF DISC_EXT_PRICE_TAX AND SUM OF DISC_E...
BY YE...
Sum of L_EXTENDEDPRICE
Sum of Disc_Ext_Price_Tax
Sum of Disc_Ext_Price
1.0bn
0.8bn
0.6bn
0.4bn
0.2bn
Sum of L_EXTENDEDPRICE, Sum of Disc_Ext_Price_Tax and Sum ...
1992
1993
1994
1995
1996
1997
1998
Year
Filters
Search
Filters on this visual
L_SHIPDATE - Year
is (All)
Sum of Disc_Ext_Price
is (All)
Sum of Disc_Ext_Price...
is (All)
Sum of L_EXTENDEDP...
is (All)
Add data fields here
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Visualizations
Build visual
X-axis
L_SHIPDATE
Year
Y-axis
Sum of L_EXTENDEDP...
Sum of Disc_Ext_Price...
Sum of Disc_Ext_Price
Legend
Add data fields here
Fields
Search
finaldf
Disc_Ext_Price
Disc_Ext_Price_Tax
L_COMMENT
L_COMMITDATE
L_DISCOUNT
L_EXTENDEDPRICE
L_LINENUMBER
L_LINESTATUS
L_ORDERKEY
L_PARTKEY
L_QUANTITY
L_RECEIPTDATE
L_RETURNFLAG
L_SHIPDATE
Date Hierarchy
Year
Quarter
Month
Day
Page 1 of 1
100%
Update available (click to downloa