

SAS FANS

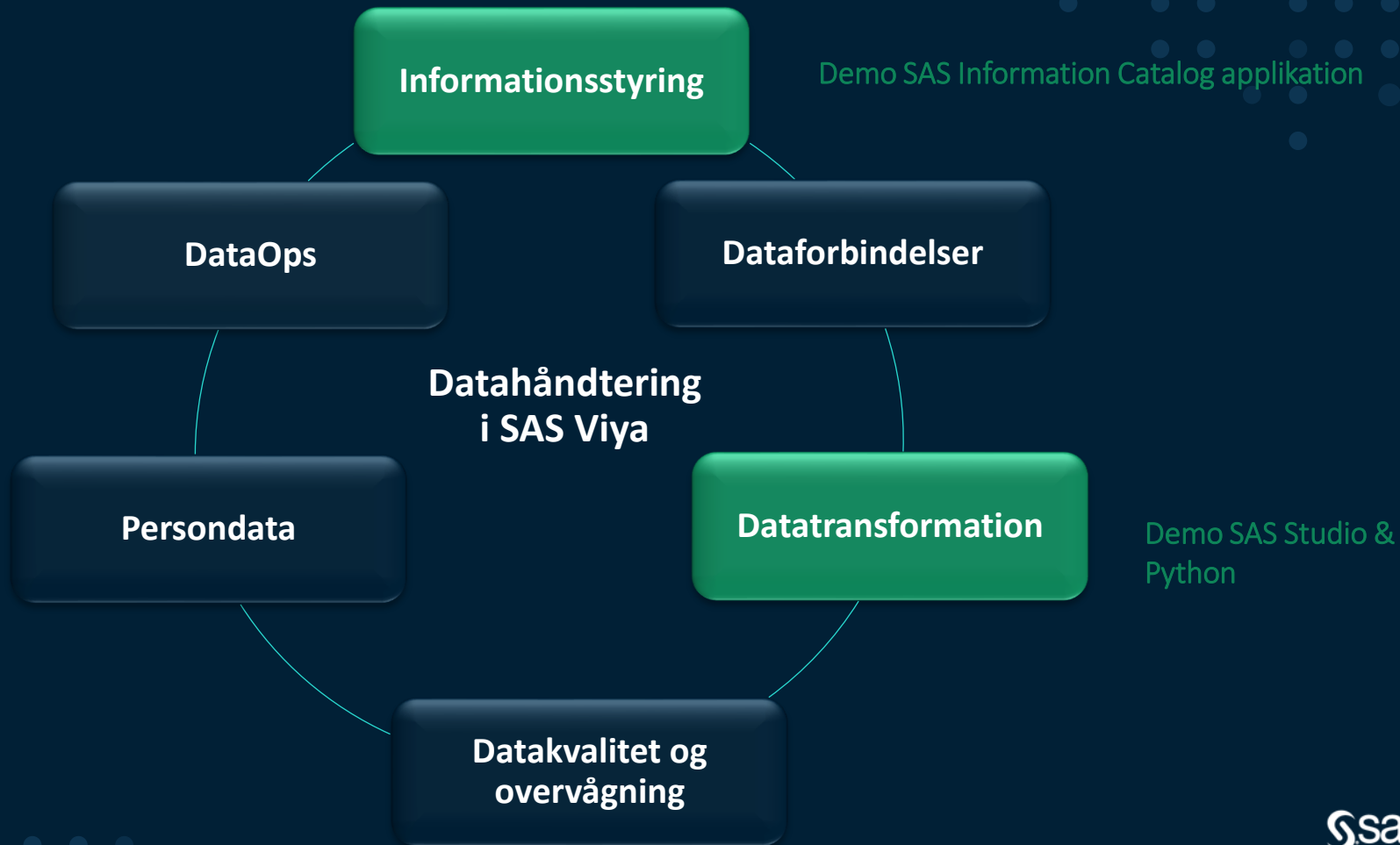
Data Management

Cecily Hoffritz

Data engineering advisory, SAS Nordics

2. december 2021

30 years employed @ 



SAS Information Governance

Styring over et broget datalandskab med et informationskatalog

The screenshot displays the SAS Information Catalog interface. At the top, it says "SAS Information Catalog - Discover Information Assets" and "47 assets cataloged". A search bar asks "What assets are you looking for?". Below the search bar, there are sections for "WELCOME" and "COLLECTIONS". The "COLLECTIONS" section shows a table of assets with columns for Name, Status, Library, Date Modified, and Modified By. A dropdown menu is open over the table, showing "Asset types: (1 of 3)" with options: "Datasets", "Studio flows", and "SAS analytics objects" (which is selected).

	Name	Status	Library	Date Modified	Modified By
<input type="checkbox"/>	worldcities	○	CATAL...	Oct 19, 2020 3:35 PM	..
<input type="checkbox"/>	watercluster	●	CATAL...	Sep 2, 2020 3:14 PM	..
<input type="checkbox"/>	DB_SPEC-CHA...	●	CATAL...	Nov 6, 2020 3:14 PM	sas Unix Service Acct
<input type="checkbox"/>	class_label	○	CATAL...	Jul 9, 2019 4:49 PM	..
<input type="checkbox"/>	CLASS	⚠	NZLIB		..
<input type="checkbox"/>	NETEZZA所有...	○	NZLIB		..
<input type="checkbox"/>	CARS	■	NZLIB		..
<input type="checkbox"/>	newcars	○	PCFILE		..

Et informationskatalog gør det muligt

- At søge efter data og undersøge deres analyseparathed.
- At søge efter analyseaktiver med henblik på genbrug og læring

Hvorfor et informationskatalog?

- Tidsbesparende - afkorter processen fra data til indsigt
- Øget datakvalitet og forbedret dataanalyse
- Øget samarbejde mellem teams – bedre data er et fællesanliggende
- Bedre styring på data- og analyseaktiverne
- Øget tillid til resultaterne – man kan stole på data

Informationsstyring med SAS Information Governance

Er data analyseparate?

WATER_CLUSTER

Samples

Columns

21

Rows

46.7 K

Size

9.6 MB

Completeness:



Status:

Approved



Actions

Overview

Column Analysis

Sample Data

Date analyzed: 27 Jan 2021 4:25 pm

Descriptive Measures

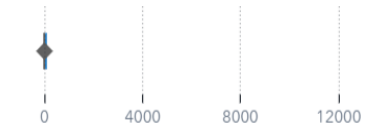
Metadata Measures

Data Quality Measures

#	Name	Maximum	Standard Deviation	Missing	Blanks	Outliers
11	⊕ Long	29.976798	0.053542	0	--	Yes
12	⊕ Property_type	0	0	0	--	No
13	⊕ Meter_Location	--	--	0	0	--
14	⊕ Clli	--	--	0	0	--
15	⊕ DMA	2	0.402329	0	--	Yes
16	⊕ Weekday	7	1.997622	0	--	No
17	⊕ Weekend	1	0.451387	0	--	No
18	⊕ Daily_W_C_M3	11910	230.931	0	--	Yes
19	⊕ Week	52	15.05769	0	--	No
20	⊕ US Holiday	--	--	45,056	0	--

Column Properties

Frequency distribution:



Quantiles

Minimum: 0

25%: 0.385

50%: 0.707

75%: 1.173

Maximum: 11910

Column name: ⊕ Daily_W_C_M3

Label: --

Type: double

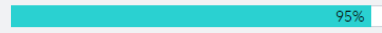
Format: --

Informationsstyring med SAS Information Governance

Er data analyseparate?

WATER_CLUSTER
Samples

Columns **21** Rows **46.7 K** Size **9.6 MB**

Completeness: 

Status: Approved

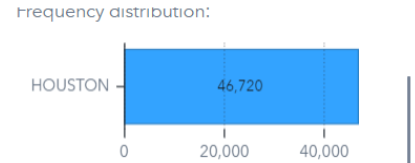
☆ **Actions**

Overview Column Analysis Sample Data

Date analyzed: 27 Jan 2021 4:25 pm

Descriptive Measures **Metadata Measures** Data Quality Measures

#	Name	Label	Type	Actual Type	Logical Type	Format	Le #
1	Year		double	--	Unary		
2	Month	Month	double	--	Nominal		
3	Day		double	--	Interval		
4	Date		double	--	Interval	MMDDYY	
5	Serial		double	--	Interval	BEST	
6	Property		double	--	Interval	BEST	
7	Address		char	String	Nominal	\$CHAR	
8	City		char	String	Unary	\$CHAR	
9	Zip		double	--	Interval	BEST	
10	Lat		double	--	Interval	BEST	
11	Long		double	--	Interval	BEST	




Column name: City
Label: --
Type: char
Actual type: String
Format: \$CHAR
Length: 7
Minimum length: 7
Maximum length: 7
Primary key candidate: No
Logical type: Unary
Semantic type: CITY
Information privacy: Candidate

Informationsstyring med SAS Information Governance

Er data analyseparate?

WATER_CLUSTER
Samples

Columns **21** Rows **46.7 K** Size **9.6 MB**

Completeness:  95%









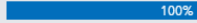
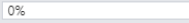



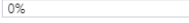








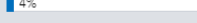
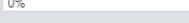


Status: Approved

☆ Actions

Overview Column Analysis Sample Data

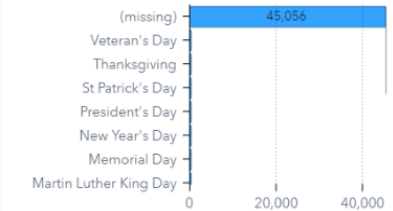
Date analyzed: 27 Jan 2021 4:25 pm


Descriptive Measures Metadata Measures **Data Quality Measures**

#	Name	Completeness	Uniqueness	Most Common Value	Least Common Value	Pattern Count	Semantic
9	Zip	 100%	 0%	77094	77020 (2 more)	--	POSTAL C
10	Lat	 100%	 0%	-95.408872 (1 more)	-95.71878 (19+ more)	--	LATITUDE
11	Long	 100%	 0%	29.785747 (1 more)	29.651167 (19+ more)	--	LONGITU
12	Property_type	 100%	 0%	0	0	--	GENERIC
13	Meter_Location	 100%	 0%	external	internal	1	
14	Clli	 100%	 0%	HSTNTXJA	HSTNTXHO	1	
15	DMA	 100%	 0%	2	1	--	
16	Weekday	 100%	 0%	4 (1 more)	1	--	
17	Weekend	 100%	 0%	0	1	--	
18	Daily_W_C_M3	 100%	 11%	0	0.023 (19+ more)	--	
19	Week	 100%	 0%	51 (19+ more)	52	--	DATE
20	US Holiday	 4%	 0%	Veteran's Day (11 m...	Independence Day	13	
21	CLUSTER	 100%	 0%	4	1	--	

Column Properties

Frequency distribution:



Column name:  US Holiday

Label: --

Type: char

Actual type: String

Format: \$CHAR

Length: 27

Minimum length: 8

Maximum length: 27

Primary key candidate: No

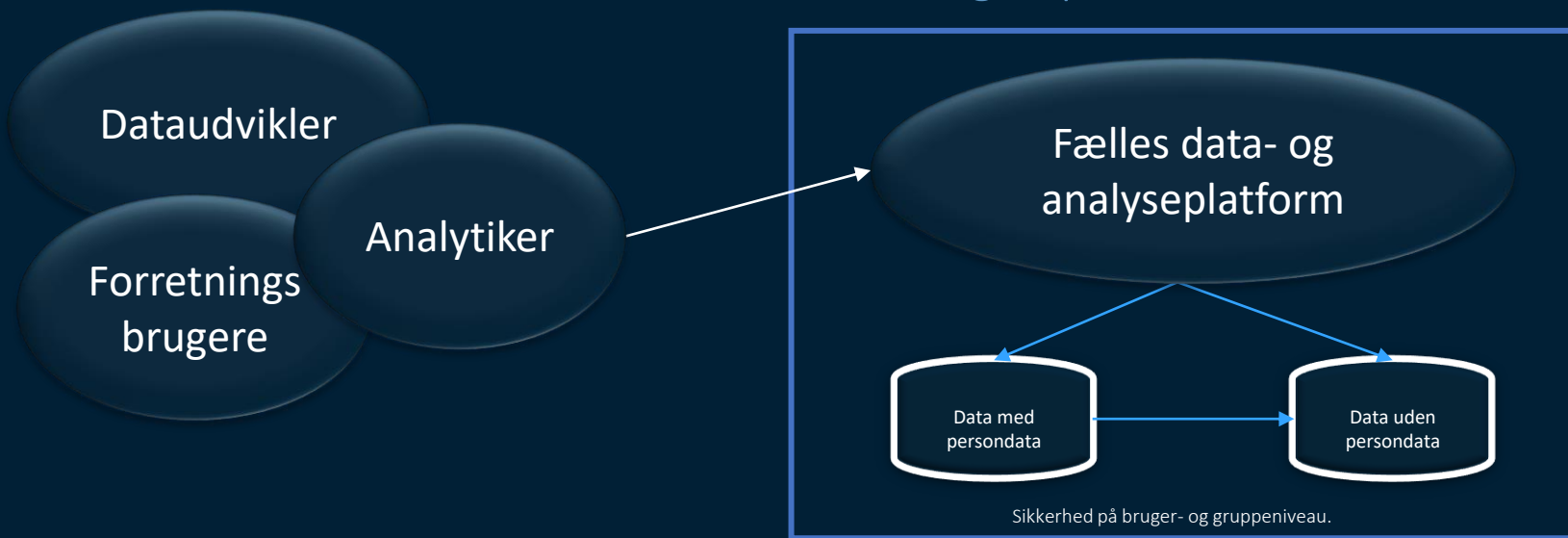
Logical type: Nominal

Semantic type: --

Information privacy: --

Informationsstyring med SAS Information Governance

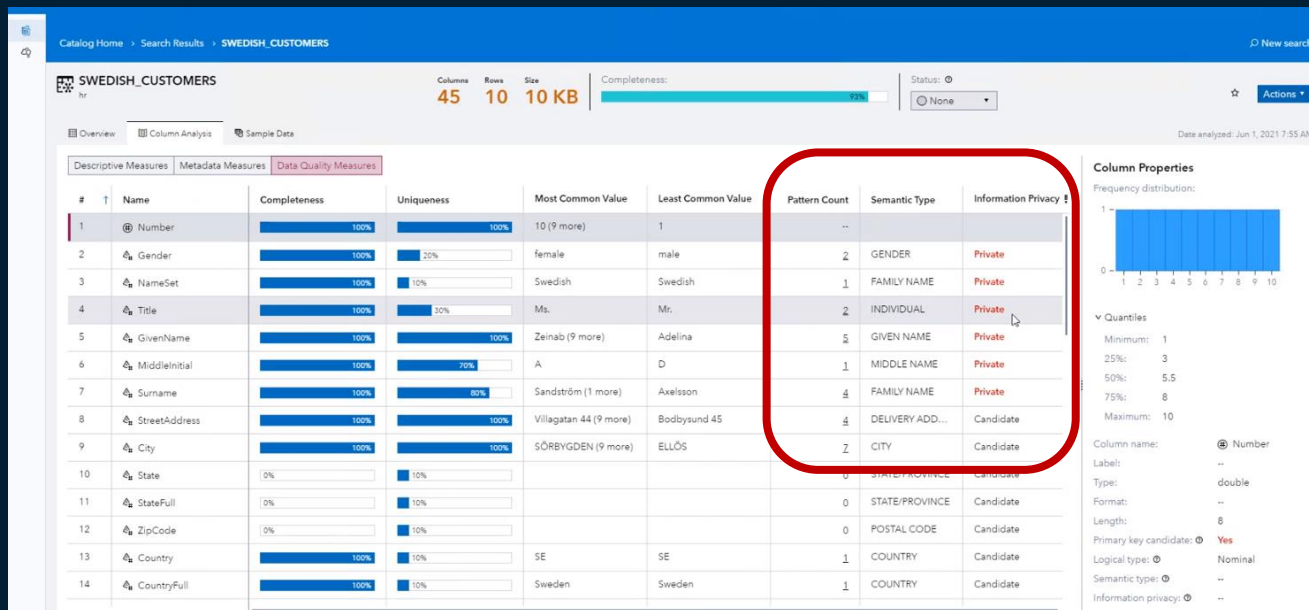
Use case: håndtering af persondata



- Sporbarhed gennem hele processen, fra data til analyse.
- Logning af aktivitet på data - hvilke data kigger man specifikt på og hvornår.

Informationstyring med SAS Information Governance

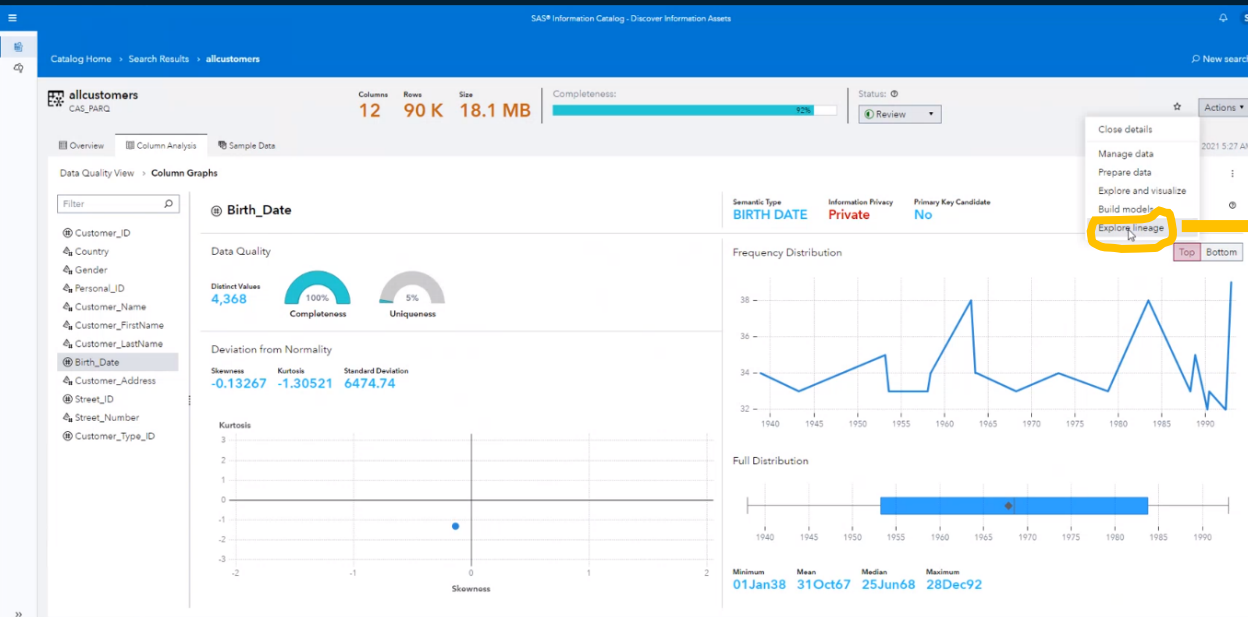
Håndtering af persondata med SAS



Trin 1- Overblik:
Få et straks overblik over evt. persondata gennem den automatiske persondata-scanning

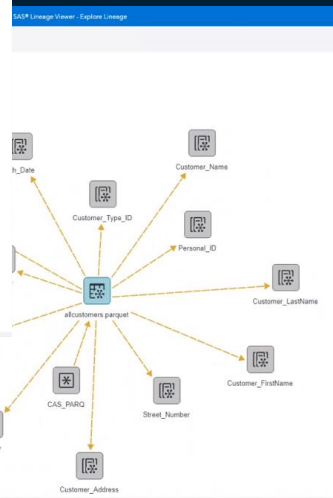
Informationstyring med SAS Information Governance

Håndtering af persondata med SAS



Trin 2 - Sporbarhed:
Skab overblik over de
sammenhænge, hvor data
indgår.

- Close details
- Manage data
- Prepare data
- Explore and visualize
- Build model
- Explore lineage



Informationsstyring med SAS Information Governance

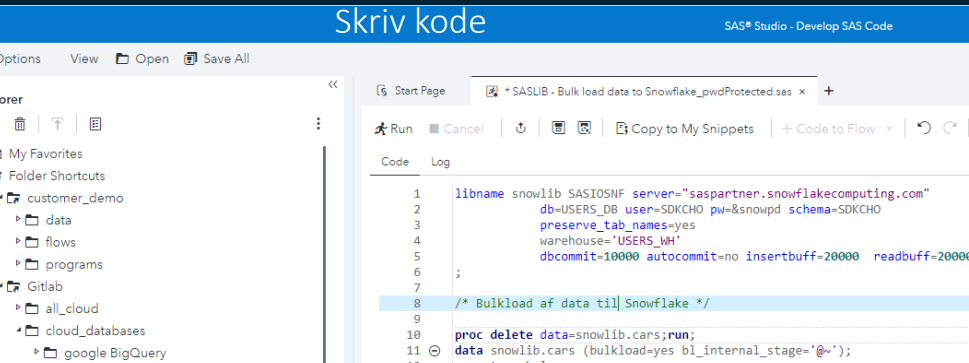
Håndtering af persondata med SAS

The screenshot displays the SAS Studio interface for developing SAS code. On the left, a 'Steps' pane lists various actions like 'Find_PD', 'Data (Input and Output)', and 'SAS Program'. The main workspace shows a flow diagram with a step named 'Find_PD' connected to a data source 'PARQUET_CU STOMER'. Below the flow, the configuration for the 'Find_PD' step is visible, including tabs for 'Search Options', 'Anonymization', and 'Node'. The 'Anonymization' tab is active, showing options for 'X_replace', 'Encrypt', and 'Mask'. A dropdown menu for 'Select method for anonymization' is open, and a list for 'Select columns to anonymize' is also visible. Two blue arrows point to the 'Find_PD' step icon and the 'Anonymization' configuration area.

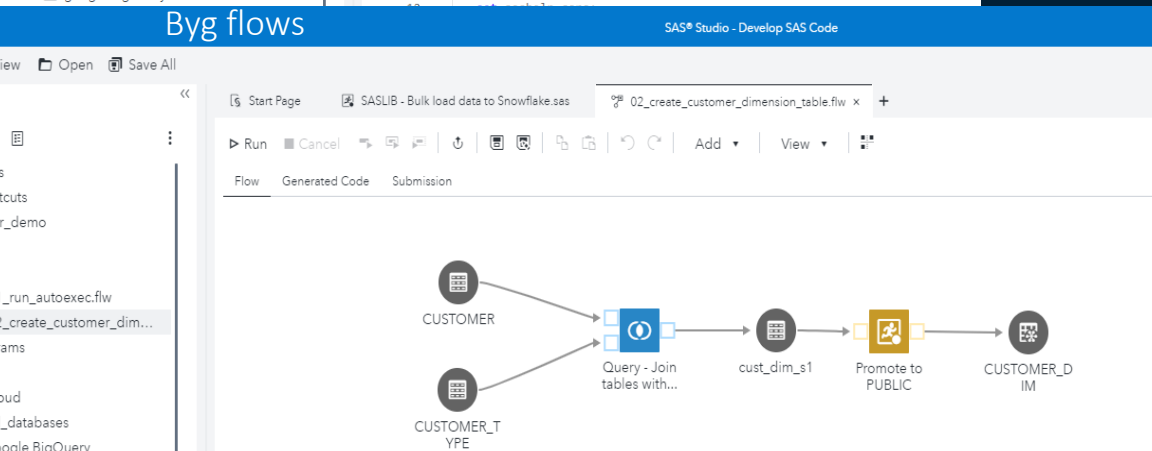
Trin 3 - Beskyttelse:
Skab anonymiserede data til analysen.

Datatransformation

Gør data klar til analyse med SAS Studio



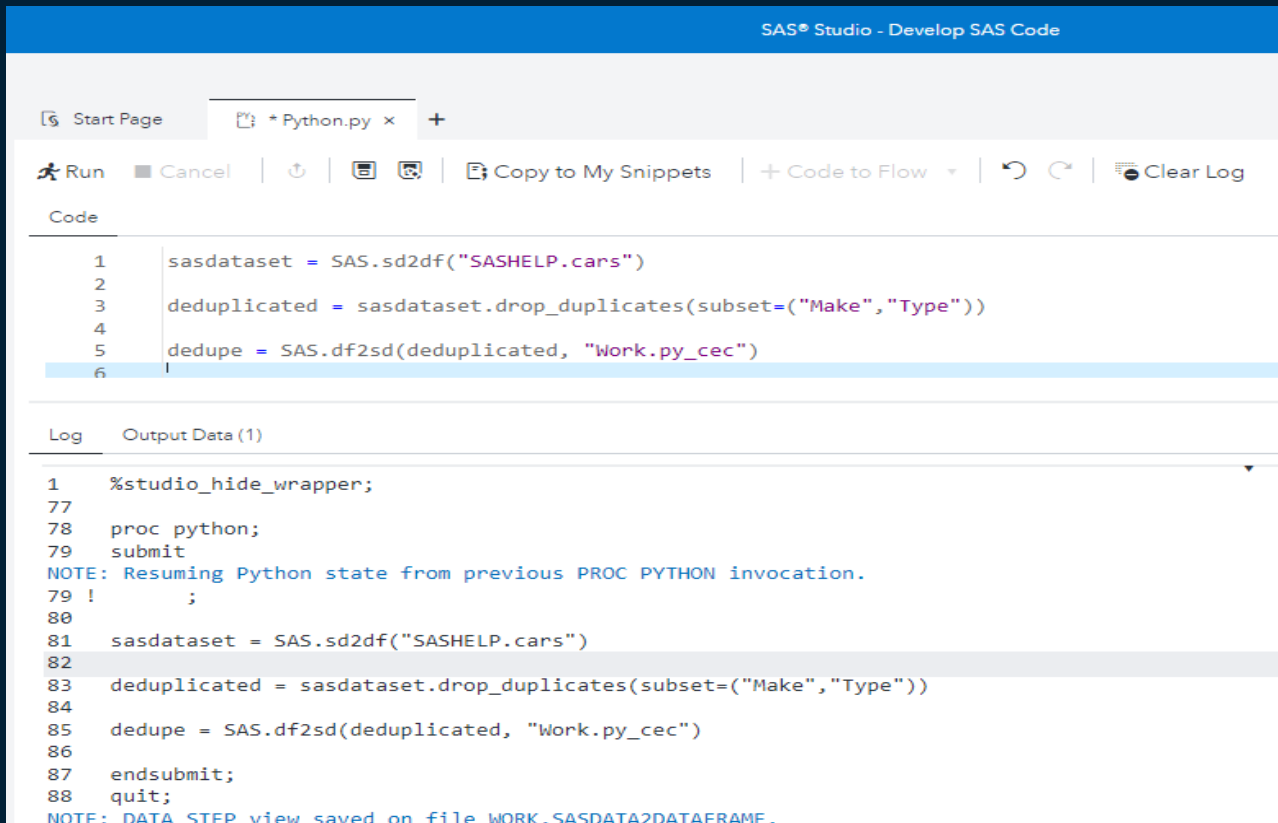
- SAS Studio er én applikation med et væld af funktionalitet til datatransformation, dataintegration og analyse/statistik for alle typer af brugere



Agilitet
Samarbejde
Integration

SAS Studio & Python

Kør Python kode i Python.py editor i SAS Studio



The screenshot displays the SAS Studio interface for developing SAS code. The title bar reads "SAS® Studio - Develop SAS Code". The editor window shows a Python script with the following code:

```
1 sasdataset = SAS.sd2df("SASHELP.cars")
2
3 deduplicated = sasdataset.drop_duplicates(subset=("Make","Type"))
4
5 dedupe = SAS.df2sd(deduplicated, "Work.py_cec")
6
```

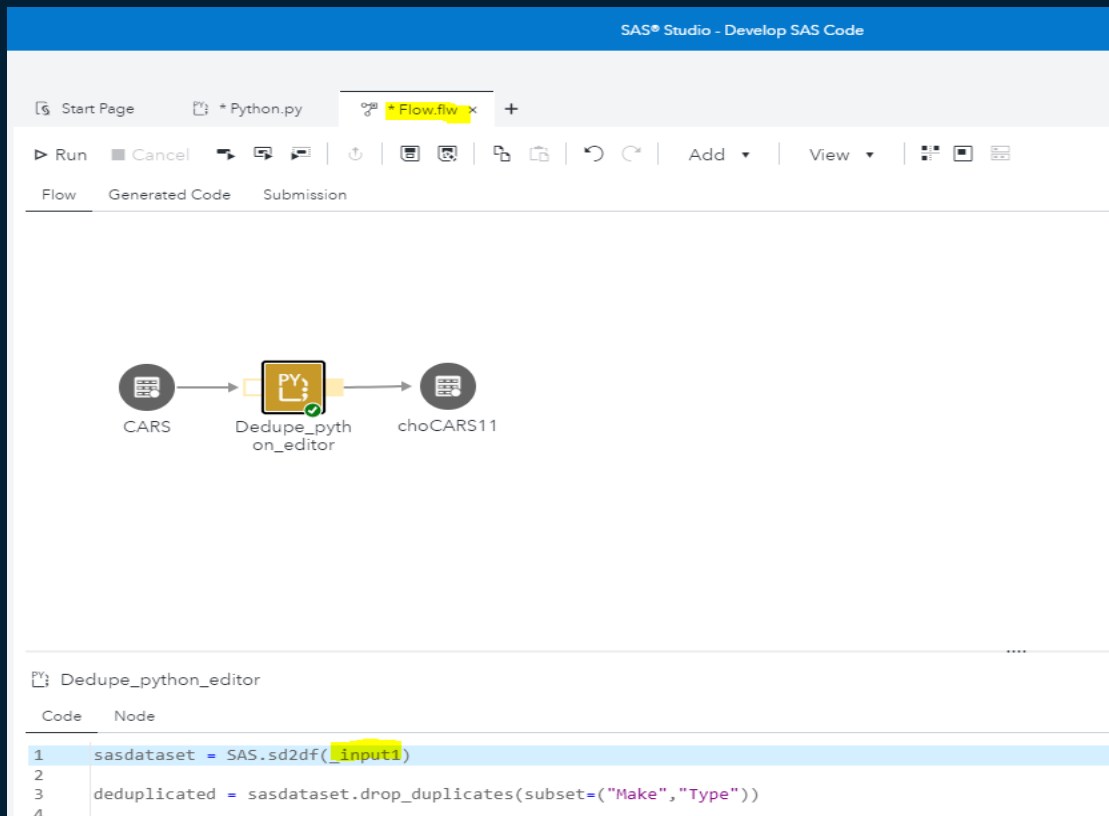
Below the code editor, the Log window is visible, showing the execution output:

```
Log Output Data (1)
1 %studio_hide_wrapper;
77
78 proc python;
79 submit
NOTE: Resuming Python state from previous PROC PYTHON invocation.
79 ! ;
80
81 sasdataset = SAS.sd2df("SASHELP.cars")
82
83 deduplicated = sasdataset.drop_duplicates(subset=("Make","Type"))
84
85 dedupe = SAS.df2sd(deduplicated, "Work.py_cec")
86
87 endsubmit;
88 quit;
```

At the bottom of the log, a note indicates: "NOTE: DATA STEP view saved on file WORK.SASDATA2DATAFRAME."

SAS Studio & Python

Integrer Python i en SAS Studio data pipeline



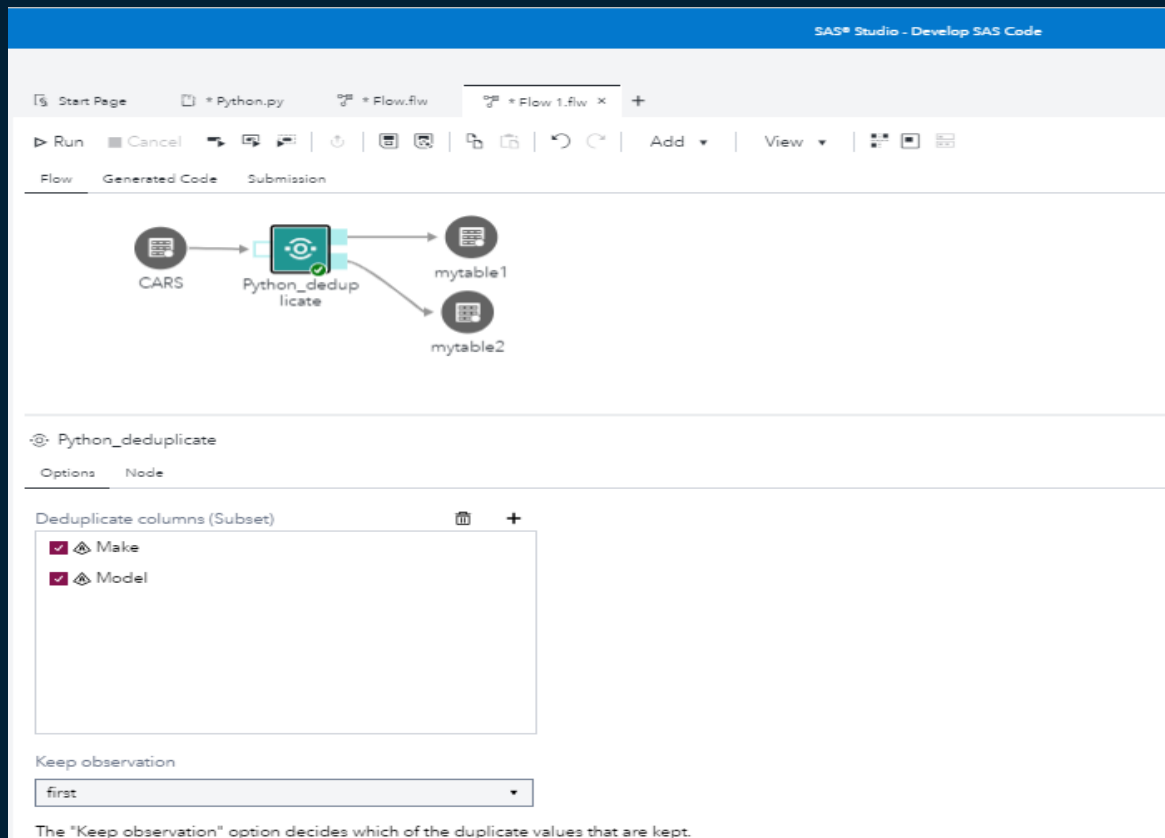
The screenshot displays the SAS Studio interface for developing SAS code. The top bar indicates "SAS® Studio - Develop SAS Code". The breadcrumb navigation shows "Start Page" > "Python.py" > "Flow.flw". The toolbar includes "Run", "Cancel", and various workflow management icons. Below the toolbar, the workflow is visible, showing a sequence of nodes: "CARS" (input), "Dedupe_python_editor" (Python node), and "choCARS11" (output). The Python node is highlighted with a yellow border and a green checkmark. Below the workflow, the code editor for the "Dedupe_python_editor" node is shown, containing the following Python code:

```
Code Node
1 sasdataset = SAS.sd2df(_input1)
2
3 deduplicated = sasdataset.drop_duplicates(subset=("Make", "Type"))
4
```

SAS Studio & Python

Gør Python super nemt at anvende for SAS Studio brugere

Kodefrit



The screenshot displays the SAS Studio interface for developing SAS code. The top navigation bar includes "Start Page", "Python.py", "Flow.fiw", and "Flow 1.fiw". The main workspace shows a workflow diagram with a "CARS" data source connected to a "Python_deduplicate" node, which then outputs to "mytable1" and "mytable2". Below the diagram, the "Python_deduplicate" node configuration is shown, including a list of deduplicated columns ("Make" and "Model") and a "Keep observation" dropdown set to "first".

SAS® Studio - Develop SAS Code

Start Page Python.py Flow.fiw Flow 1.fiw

Run Cancel [Icons] Add View

Flow Generated Code Submission

CARS Python_deduplicate mytable1 mytable2

Python_deduplicate

Options Node

Deduplicate columns (Subset)

- Make
- Model

Keep observation

first

The "Keep observation" option decides which of the duplicate values that are kept.

Migrering af data til UTF-8 for SAS Viya

- Migrér til UTF-8 encoding links:

- [Determine the Encoding of a Data Set.](#)
- [Migrating Data from WLATIN1 to UTF-8.](#)
- [Determine Storage Size Requirements.](#)
- [Use CEDA to Read Data.](#)
- [Determine Whether the CVP Engine Is Needed to Read Your Data without Truncation.](#)
- [Convert Indexes and Integrity Constraints to UTF-8.](#)
- [Convert Format Catalogs to UTF-8.](#)
- [Read External Files.](#)

[Læs mere om SAS og UTF8](#)

[Læs mere om session encoding og dataafkortning](#)

Tak
Cecily Hoffritz | LinkedIn