



15th–19th November

EU Connect 2021

The Clinical Data
Science Conference



THE
POWER
TO KNOW.



RWE WORKSHOP- FRIDAY NOV 19 HANDS-ON SCRIPT

Contents

Agenda	2
Introduction	3
PART 1 – Define, Build and Analyze a Cohort	4
Access SAS Health – Cohort Builder	4
Creation of Cohort with Stringent Eligibility Criteria.....	5
Analyze the Cohort by Running Add-in Templated Analyses.....	17
PART 2 – Identify Variable Importance via Data Driven (ML) & SME Approach	21
Identify the variable importance using ML (SAS Studio) and SME.....	21
Identify the variable importance using ML (R Studio) and SME	24
Identify the variable importance using ML (Python) and SME	26
PART 3 – Visualize and Interpret the Data.....	28
Visualize Baseline Characteristics for all cohorts	28
Test Scientific Robustness	32
PART 4 – Build Final Model(s)	41
Build Models using model pipelines (SAS, R, Python models)	41

Agenda



AGENDA

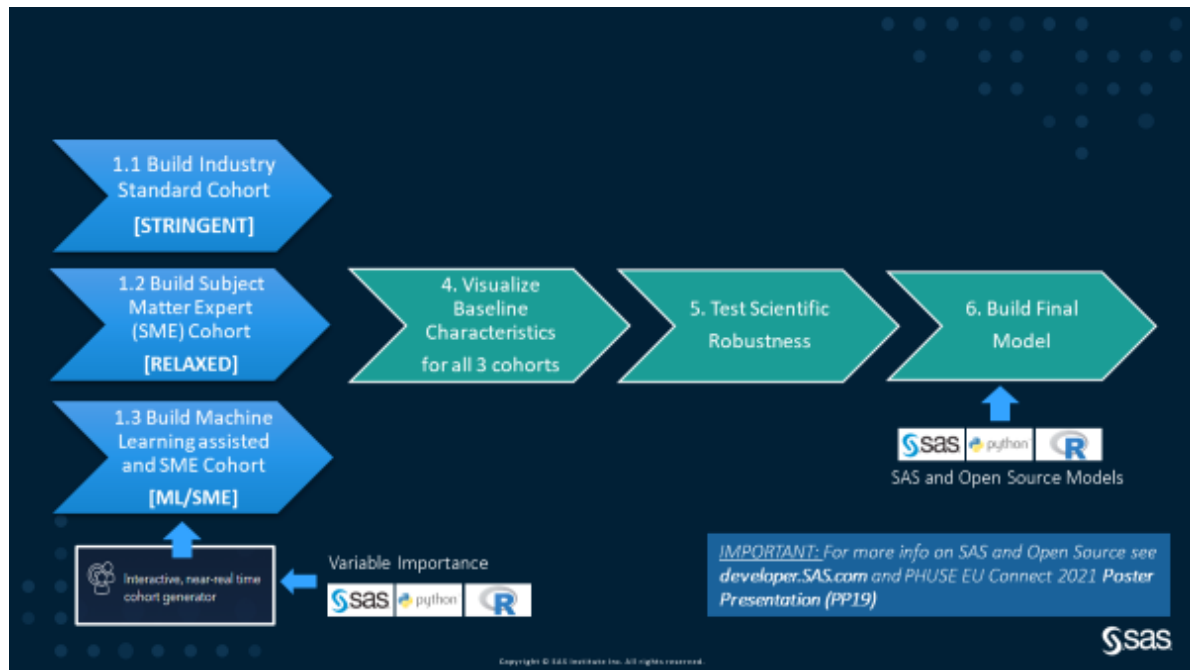
- 01** Welcome, Intro and Agenda Overview
2.45-2.55 pm CET / 8.45-8.55 am EST
- 02** RWE generation in Clinical Development
2.55-3.25 pm CET / 8.55-9.25 am EST
- 5 min break
- 03** Hands-On (break out) or Live Demo (main room)
3.30-4.30 pm CET / 9.30-10.30 am EST
- 04** Q&A, Quiz and Close Out
4.30-4.45 pm CET / 10.30-10.45 am EST



Copyright © SAS Institute Inc. All rights reserved.

Introduction

In the following sections, we will demonstrate the below flow of activities as explained in the introduction (PHUSE EU Connect 2021 - RWE Workshop).



Scope for the hands-on session: In the interest of time, we will exclude 1.2 as well as building the actual cohort in 1.3.

The following protocol (inclusion/exclusion criteria) is used for this hands-on workshop.

The screenshot shows the ClinicalTrials.gov protocol page for the PRECISE trial. The title is 'The PRECISE Protocol: Prospective Randomized Trial of the Optimal Evaluation of Cardiac Symptoms and Revascularization (PRECISE)'. The sponsor is HeartFlow, Inc. The collaborators are Duke Clinical Research Institute and Cardiovascular Research Foundation, New York. The information is provided by the sponsor, HeartFlow, Inc. The study description states: 'The study will be a prospective, pragmatic, randomized clinical trial of the comparative effectiveness of diagnostic evaluation strategies for stable CAD, to be performed in outpatient settings, including primary care and cardiology practices.' The conditions or diseases are Coronary Artery Disease. The diagnostic tests are CCTA with selective IFFs. The phase is Not Applicable. The ClinicalTrials.gov identifier is NCT01702344. The recruitment status is Recruiting. The first posted date is October 11, 2009. The last update posted is February 20, 2019. The SAS logo is in the bottom right corner.

PART 1 – Define, Build and Analyze a Cohort

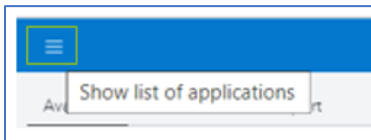


Access SAS Health – Cohort Builder

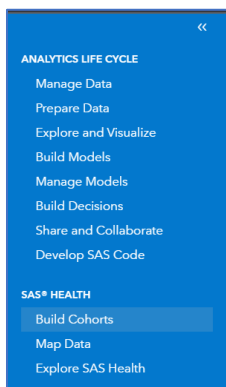
Step 1. Navigate to the following URL and use your Username & Password (*provided upfront*) to access the SAS Viya environment

<https://sashealth1.eastus.cloudapp.azure.com/SASDrive>

Step 2. Navigate to the **menu panel** on the left-hand side by selecting the 'Show List of Applications Icon' (=left top corner).

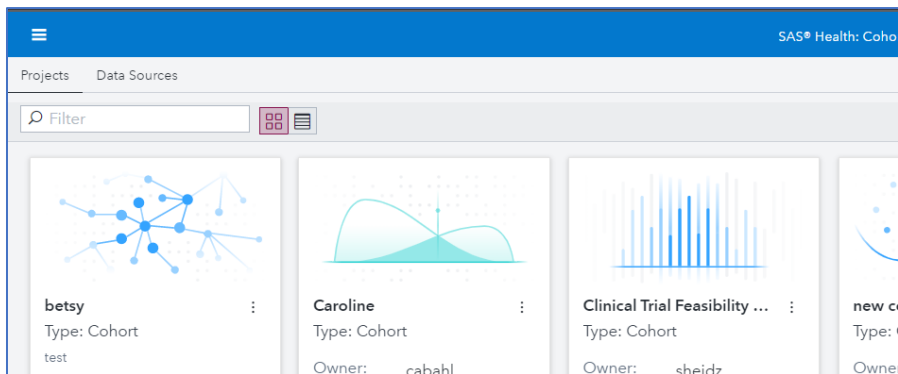


Step 3. In the menu panel on the left-hand side choose '**Build Cohorts**'.

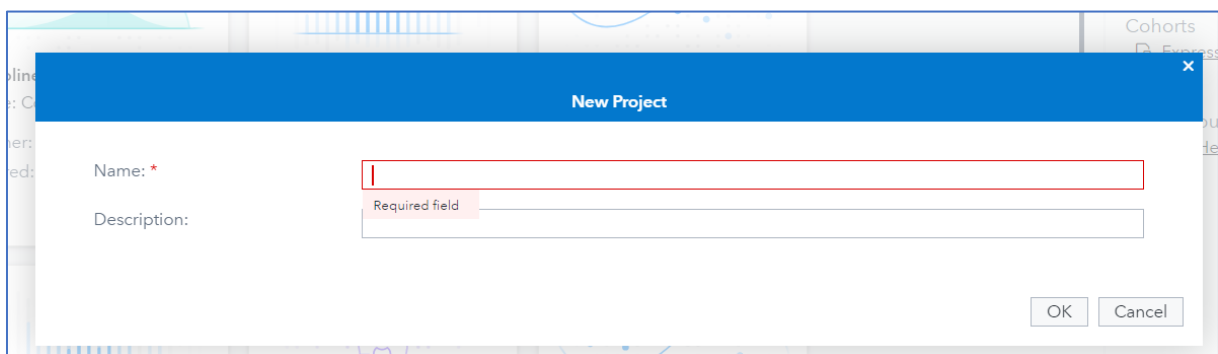
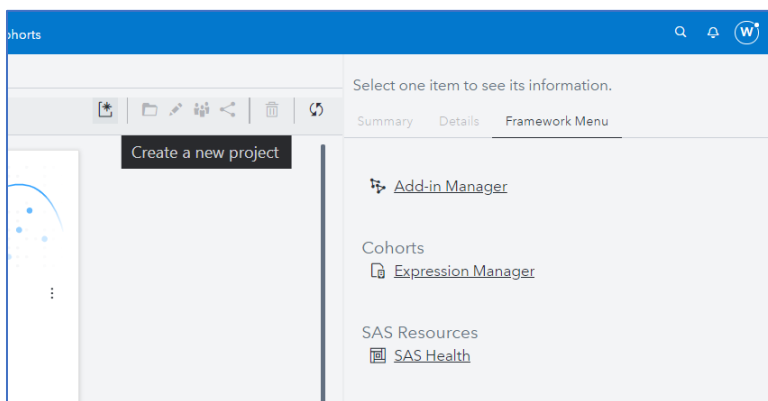


Creation of Cohort with Stringent Eligibility Criteria

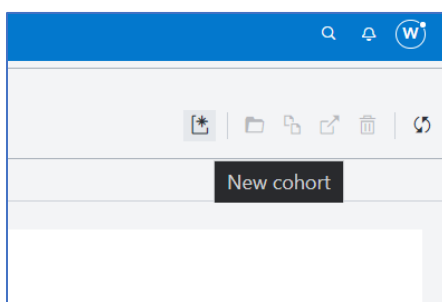
Step 1. Navigate to the **Projects** Tab.



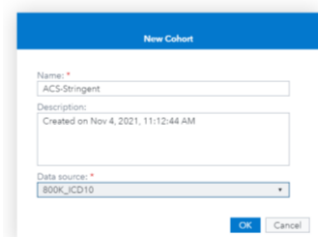
Step 2. Create a new project by clicking “**Create a new project**” and type in a project name “**SASRWE_[your initial]**”, and then click “**OK**”.



Step 3. Create a new cohort by clicking “**New cohort**” on the top right-hand side.

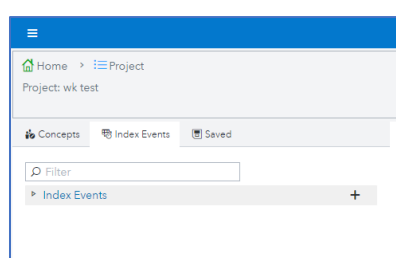


Step 4. Type in a cohort name **“ACS-Stringent_[your initial]”** and select the **“800K_ICD10”** data from the **Data source** pull-down list, and then click **“OK”**.



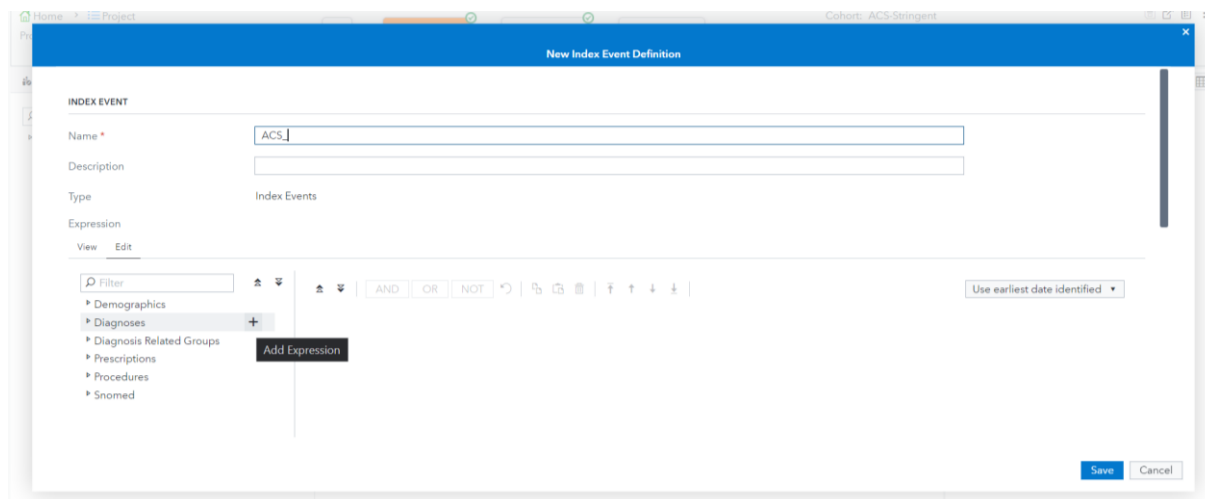
A dialog box titled "New Cohort" with a blue header. It contains three input fields: "Name:" with the text "ACS-Stringent", "Description:" with the text "Created on Nov 4, 2021, 11:12:44 AM", and "Data source:" with a dropdown menu showing "800K_ICD10". At the bottom right are "OK" and "Cancel" buttons.

Step 5. Navigate to the **Index Events** Tab on the left-hand side. Create a new index event by clicking **“+”** icon next to Index Events on the first row.



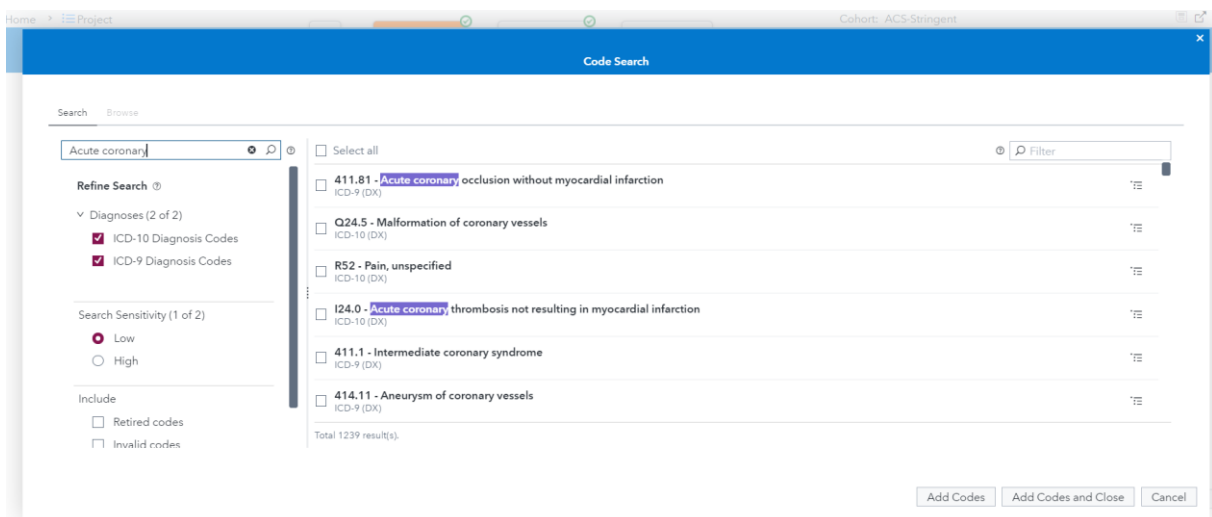
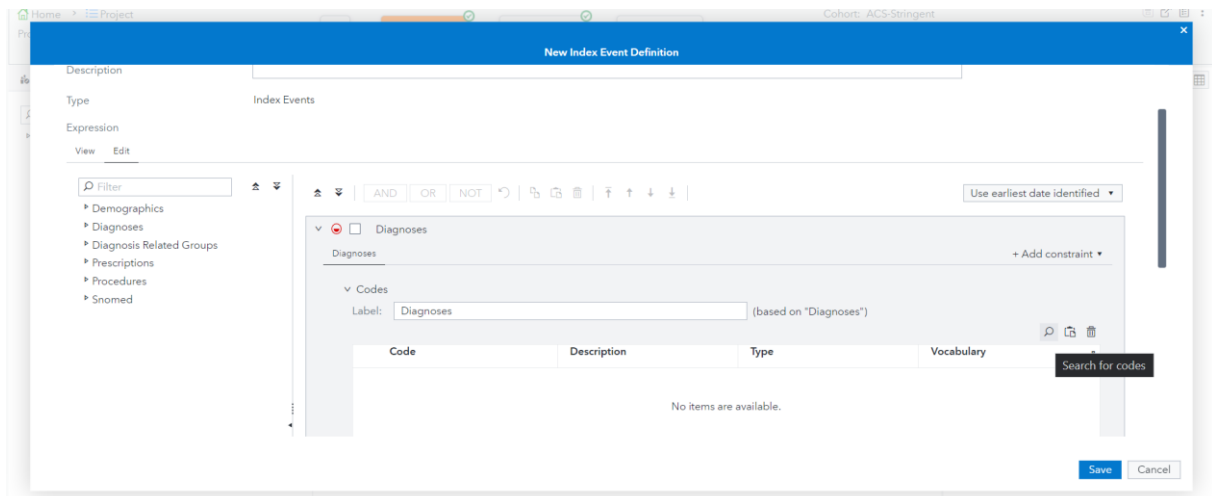
A screenshot of the software interface showing the "Index Events" tab. The tab is highlighted in blue. Below the tab name is a filter input field and a "+" icon to the right.

Step 6. Type in **“ACS_[your initial]”** for the name of the index event, and click the concept **“Diagnoses”** to create an expression to define this index event by using diagnosis codes.

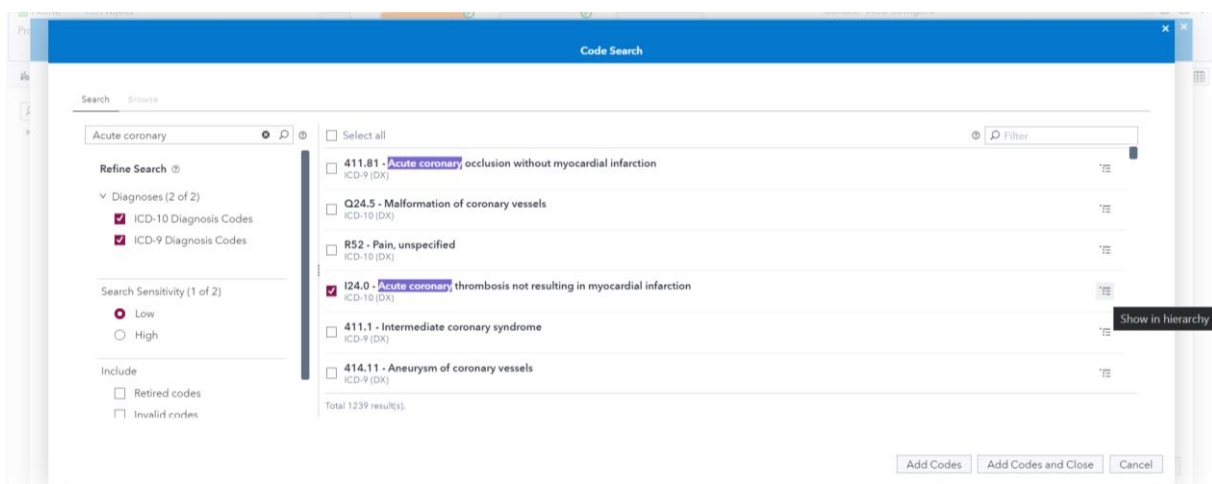


A screenshot of the "New Index Event Definition" dialog box. The "Name" field contains "ACS_". The "Type" is set to "Index Events". The "Expression" section shows a tree view with "Diagnoses" selected. A black "Add Expression" button is visible over the "Diagnoses" item. The "Filter" field is empty. The "Use earliest date identified" dropdown is set to "Use earliest date identified". "Save" and "Cancel" buttons are at the bottom right.

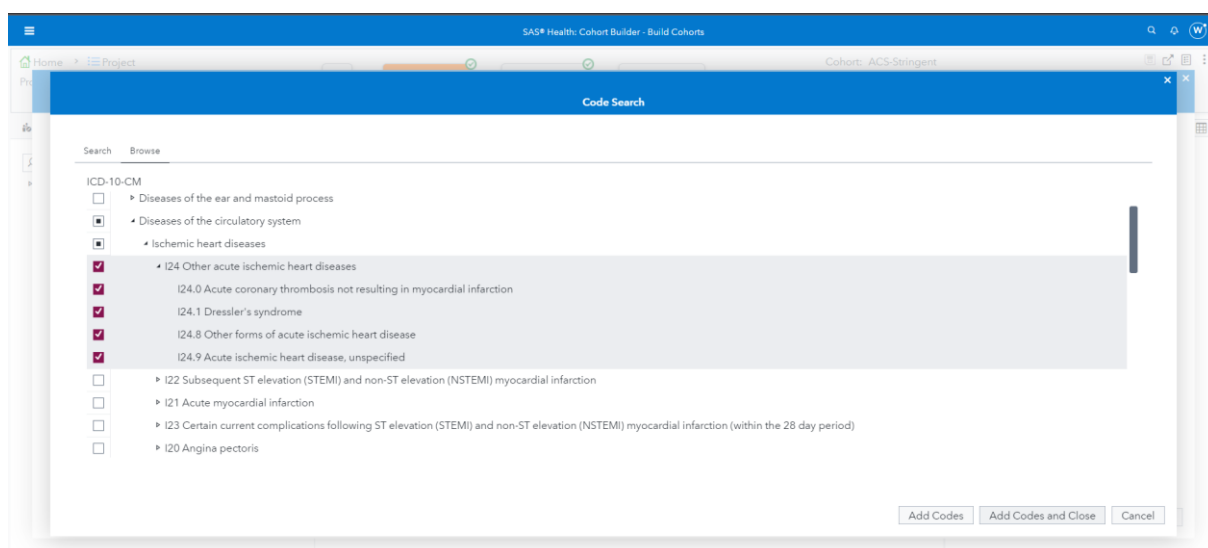
Step 7. Click “**Search for codes**” (the magnifying glass icon) on the right-hand side, and then enter “**Acute coronary**” in the text box on the top left corner (then **hit Enter**) to search for relevant diagnosis code expression.



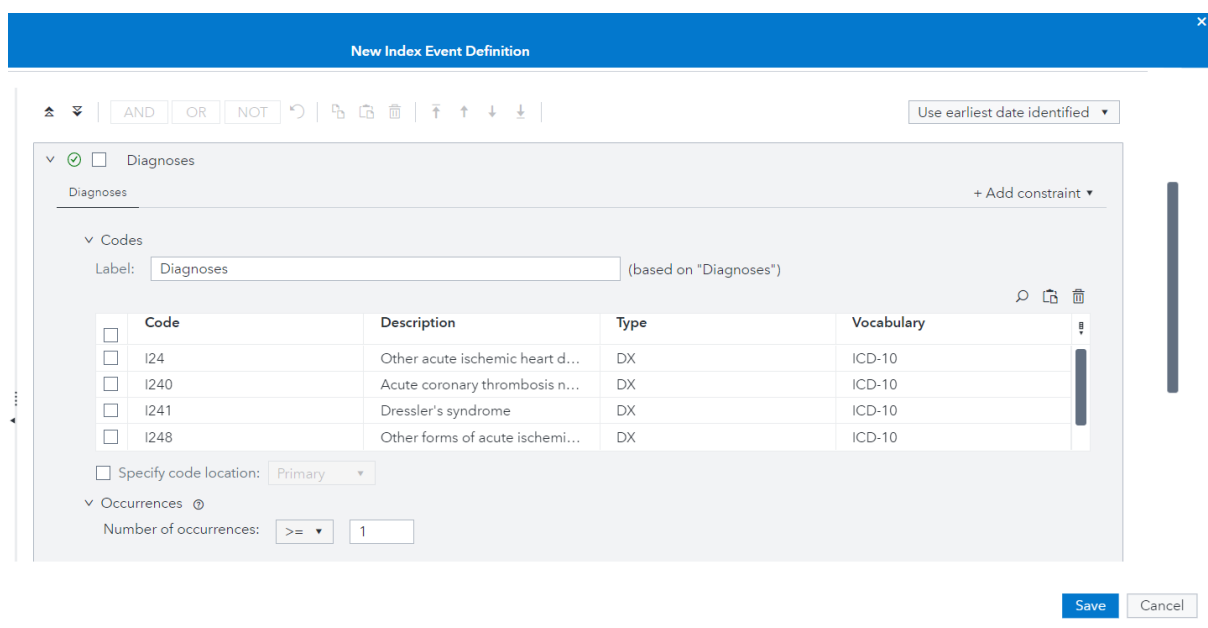
Step 8. Select “**I24.0**”, and then click “**Show in hierarchy**” icon at the far right-hand side for the same record.



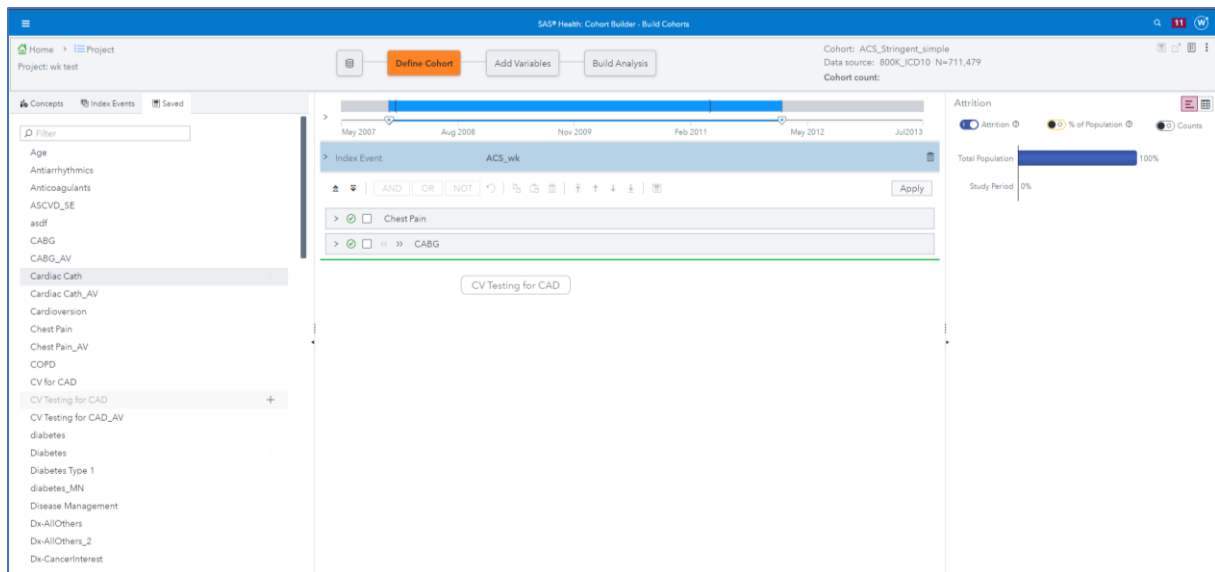
Step 9. The full hierarchy will be visible including the initial selection you made. After making the final selection of all “I24.x”, click “Add Codes and Close”



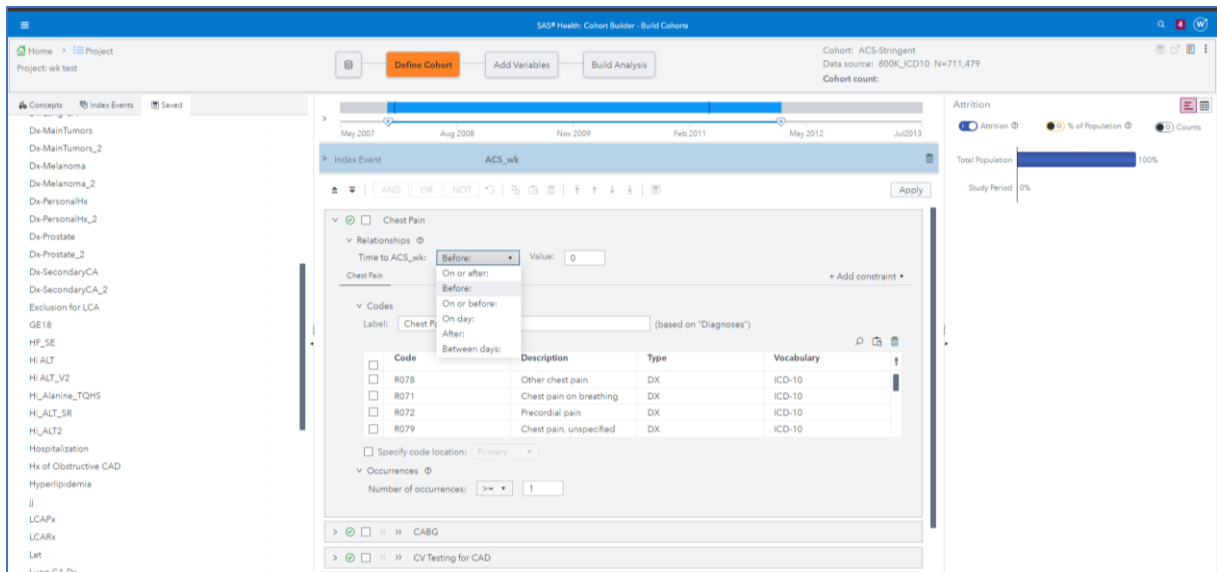
Step 10. Click “Save” on the bottom right-hand side.



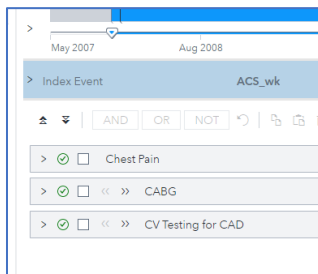
Step 11. Select the index event **ACS_[your initial]** you just created from the left-hand side, and then drag and drop it in the center of canvas. (Note: You can drop it once you notice a green line appearing). The index event now is defined for this cohort.



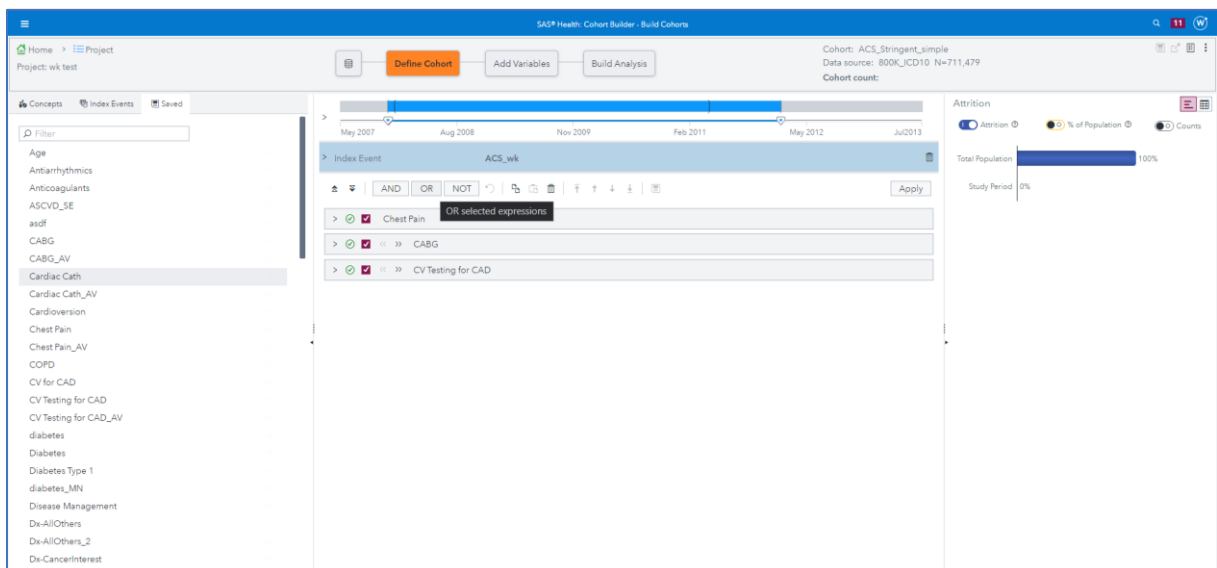
*NOTE: Please be informed that you can add more inclusion/exclusion criteria here, such as exclusion criteria, “Hx of Constructive CAD”, “Cardiac Cath”. For performance reason, we suggest only using “**Chest Pain**” and “**CABG**”. Please remember we are performing this hands-on session with approximately 40 participants and we have limited time for hands-on. Thank you for your understanding.*



Repeat the same steps for other 2 variables.



Step 15. Tick the checkbox along the three concepts you edited and click the “OR” Boolean operator on the top to make it a compound expression.

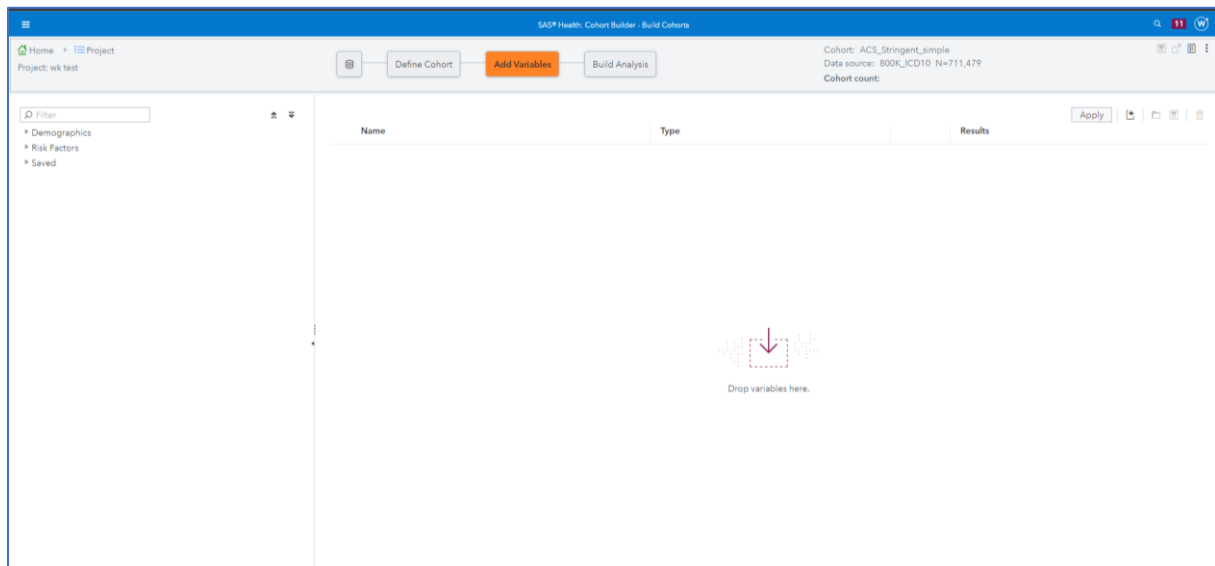


Step 16. Tick the checkbox along the compound expression and click the “**NOT**” Boolean operator to make it as the **exclusion criteria**.

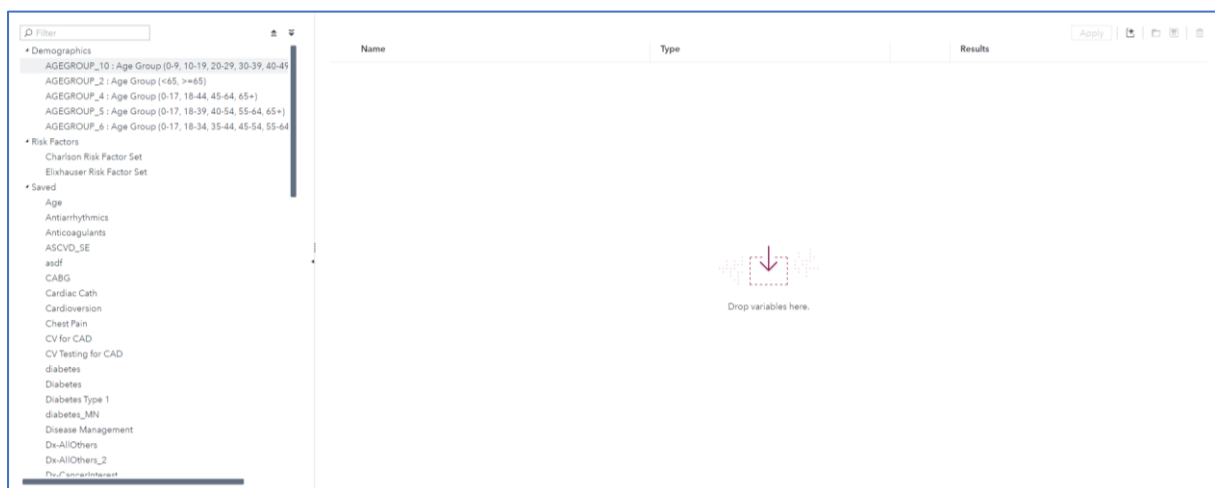
The screenshot displays the SAS® Health Cohort Builder - Build Cohorts interface. The top navigation bar includes 'Home', 'Project', 'Define Cohort', 'Add Variables', and 'Build Analysis'. The project name is 'wik test'. The cohort name is 'ACS_Stringent_simple' with a data source of '800K_ICD10 N=711,479' and a cohort count of 0. The left sidebar shows a list of concepts, including 'Age', 'Antiarhythmics', 'Anticoagulants', 'ASCVD_SE', 'asdf', 'CABG', 'CABG_AV', 'Cardiac Cath', 'Cardiac Cath_AV', 'Cardioversion', 'Chest Pain', 'Chest Pain_AV', 'COPD', 'CV for CAD', 'CV Testing for CAD', 'CV Testing for CAD_AV', 'diabetes', 'Diabetes', 'Diabetes Type 1', 'diabetes_MN', 'Disease Management', 'Dx-AllOthers', 'Dx-AllOthers_2', and 'Dx-CancerInterest'. The main panel shows a timeline from May 2007 to Jul 2013. The 'Index Event' section is active, showing a list of events: 'ACS_wk', 'Chest Pain', 'CABG', and 'CV Testing for CAD'. The 'ACS_wk' event is selected, and its expression '(Chest Pain OR CABG OR CV Testing for CAD)' is shown. The 'NOT' Boolean operator is selected, and a tooltip 'NOT selected expressions' is visible. The 'Apply' button is present. The right panel shows 'Attrition' settings with 'Total Population' at 100% and 'Study Period' at 0%.

*NOTE: Please **DO NOT** hit the APPLY button at this stage. We will first add the variables for further analyses in the next step before we generate the cohort. Please remember we are performing this hands-on session with approximately 40 participants and we have limited time for hands-on. Thank you for your understanding.*

Step 17. Next, click **Add Variables** on the top to show the list of variables. This is the place you can define the variables you plan to use for further analyses.

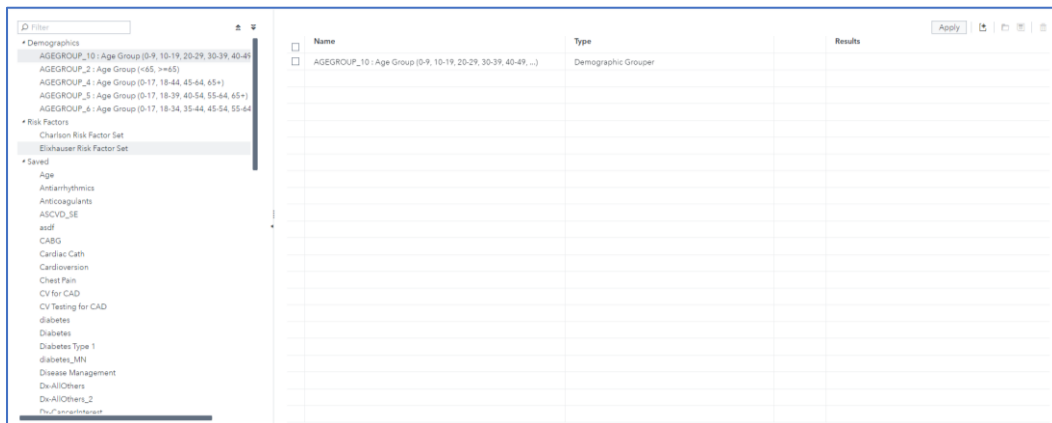


Step 18. Navigate to the left-hand side to find all the predefined and previously saved variables. Click the arrow icon on the left of **Demographics**, and then click **“AGEGROUP_10”** to add it

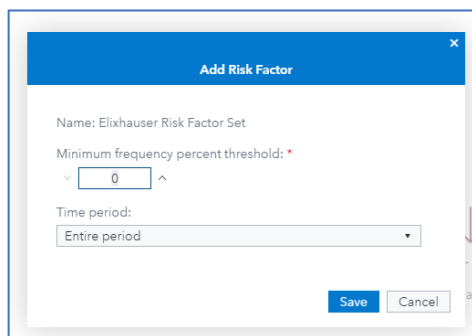


NOTE: The way of creating a new variable is similar as the steps you performed for the index event concept creation earlier. In the interest of time, we only reuse previously defined variables in this step.

Step 19. Click the arrow icon on the left of **Risk Factors** and select “**Elixhauser Risk Factor Set**”.
Drag and drop it on the canvas



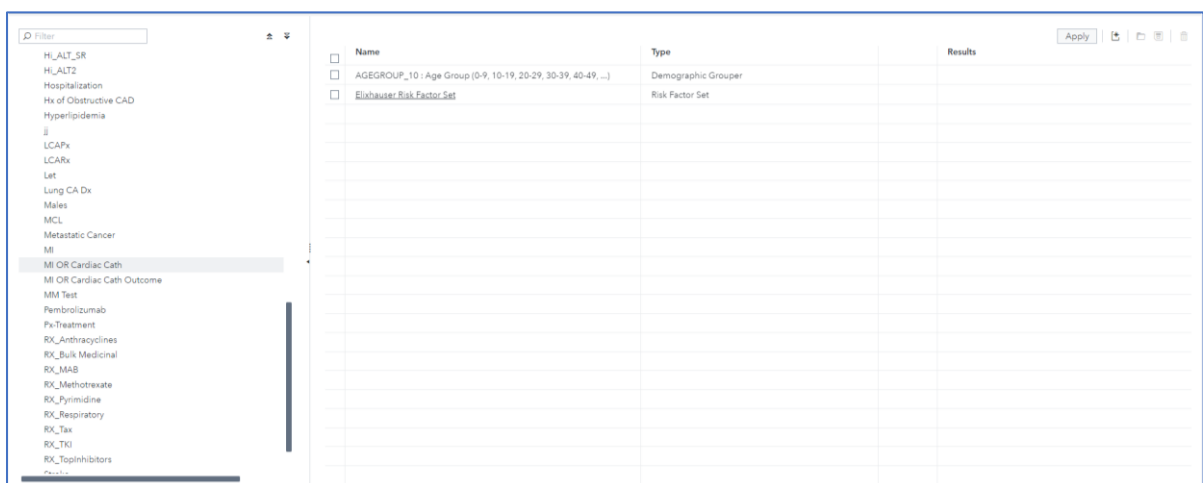
Afterwards hit the **Save** button to add (no need to make any changes).



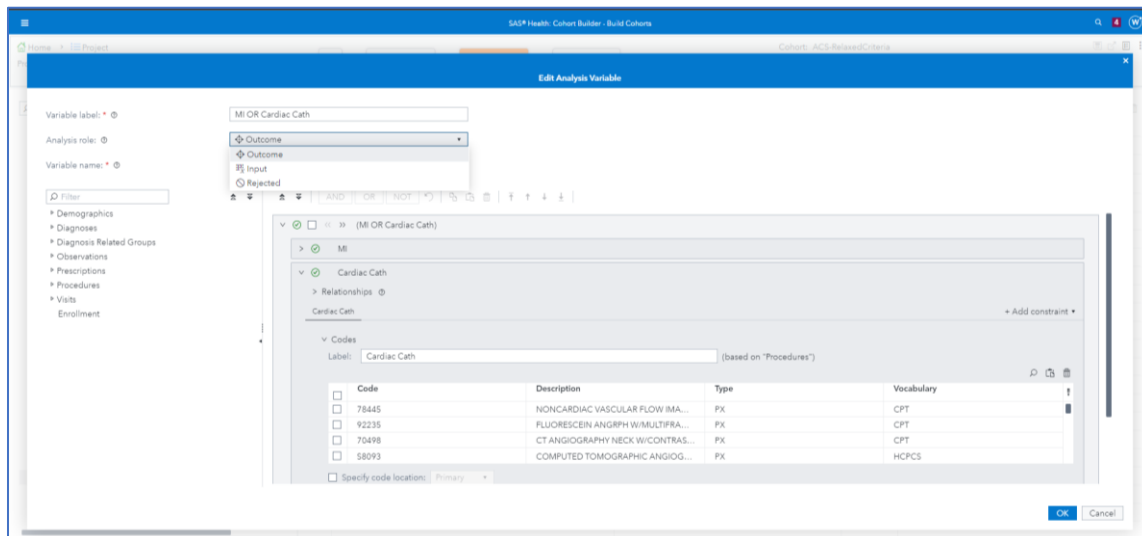
Step 20. Click the arrow icon on the left of **Saved**.

Find and add the following (how: *drag and drop on the canvas*):

- “**MI OR Cardiac Cath**”
- “**Stroke**”

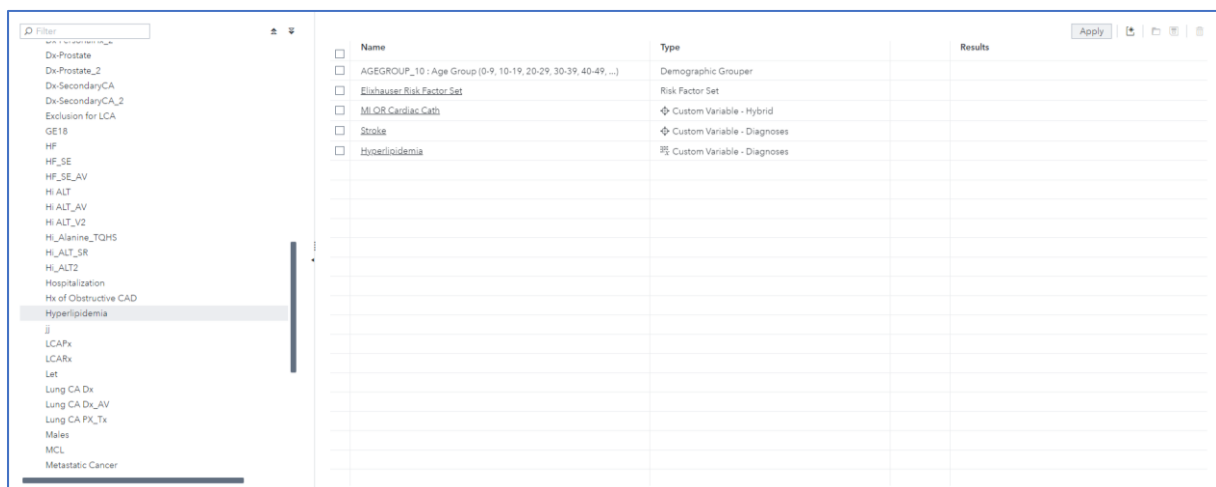


Step 21. Click the variable “**MI OR Cardiac Cath**” in the right-hand side pane. Select “**Outcome**” as **Analysis role**. Hit afterwards “**OK**”



Repeat the same action for the variable “**Stroke**”.

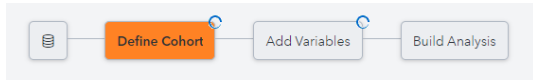
Step 22. Navigate back to **Saved** tab and drag/drop another variable “**Hyperlipidemia**” to enrich the cohort. Keep it as a “**Input**” variable.



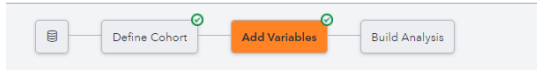
NOTE: Please be informed that you can add more analysis variables here. For performance reason, we suggest only using the variables specified above. Please remember we are performing this hands-on session with approximately 40 participants and we have limited time for hands-on. Thank you for your understanding.

Click **“Apply”** button.

Please note that this can take a few minutes. Especially when all hands-on participants execute around same time...



Once the cohort is created, a green check mark will appear next to ‘Define Cohort’ & “Add Variables”

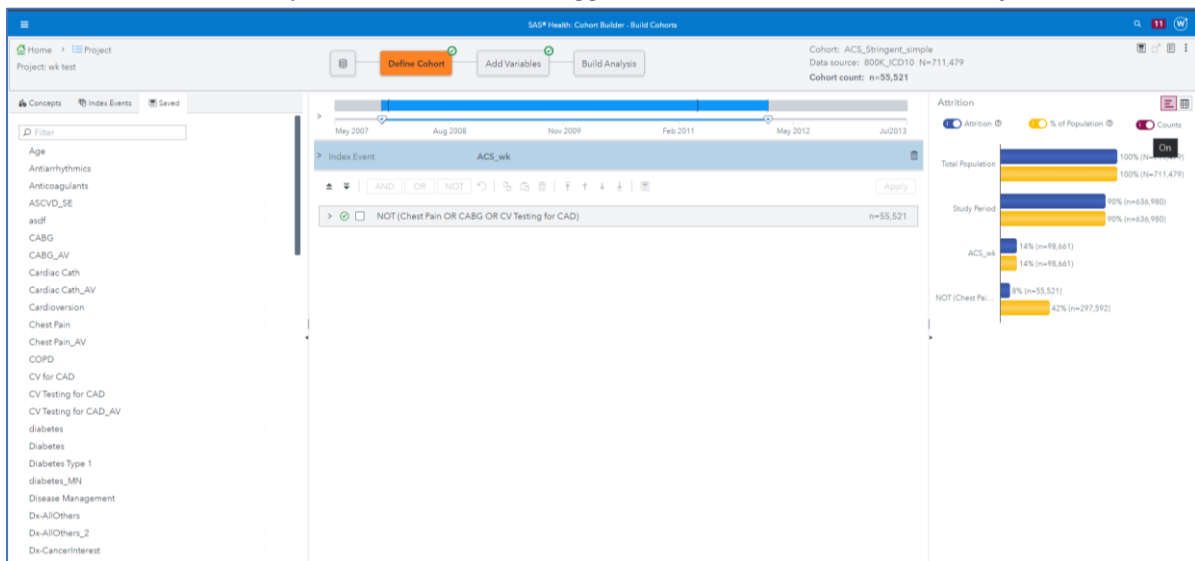


The number of counts or summary statistics will be displayed for each variable.

Click **“Save cohort”** (the floppy disk icon) on the top right-hand side to save it.

Name	Type	Results
☐ AGEGROUP_10 : Age Group (0-9, 10-19, 20-29, 30-39, 40-49, ...)	Demographic Grouping	
☐ RiskFactorSet	Risk Factor Set	Median: 15, Mean: 16.2, Min: -8, Max: 63
☐ MI OR Cardiac Cath	Custom Variable - Hybrid	Count = 35,890
☐ Stroke	Custom Variable - Diagnoses	Count = 6,807
☐ Hypertension	Custom Variable - Diagnoses	Count = 35,949

Step 23. Navigate back to **“Define Cohort”** tab and review the **Attrition** pane on the right-hand side to check the number of patients. Turn on the toggle buttons **“Counts”** and **“% of Population”**

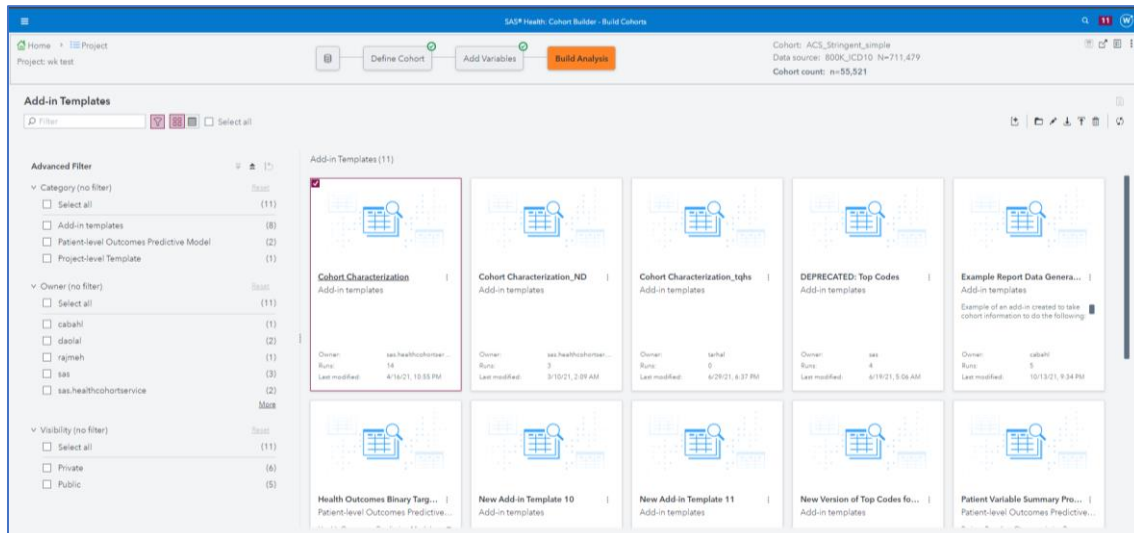


When you hit the **Table icon** you will get a tabular format (instead of **Chart**) for Attrition.

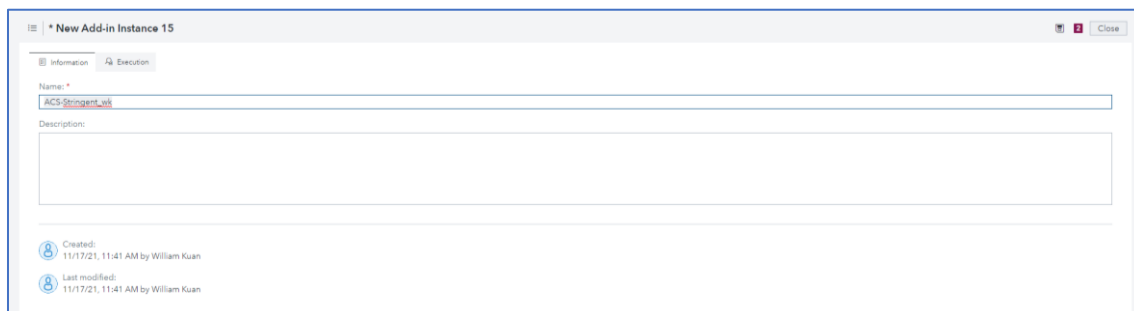
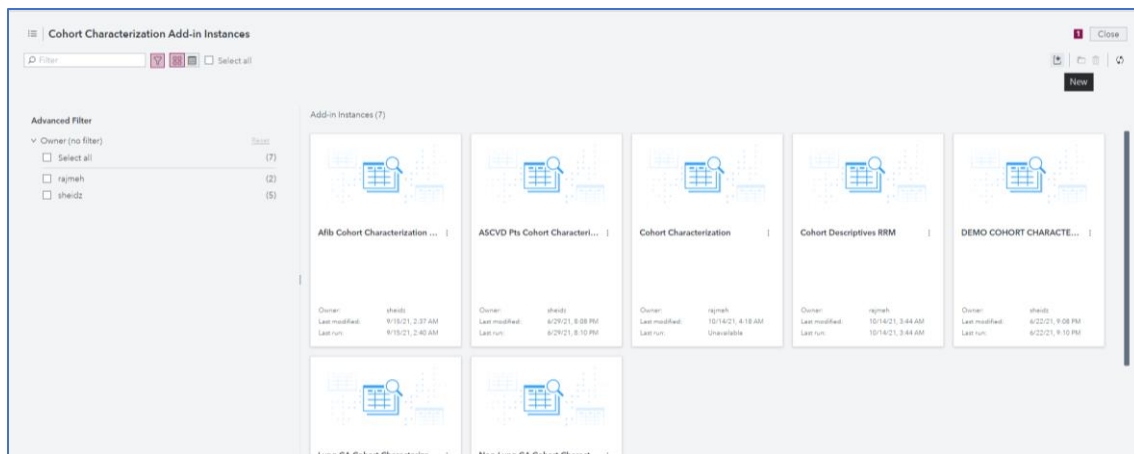
Attrition		
Total Populat...	N=711,479 / 10...	N=711,479 / 10...
Study Period	n=636,980 / 90%	n=636,980 / 90%
ACS_wk	n=98,661 / 14%	n=98,661 / 14%
NOT (Chest ...	n=55,521 / 8%	n=297,592 / 42%

Analyze the Cohort by Running Add-in Templated Analyses

Step 1. Next, click **Build Analysis** on the top to go to the list of **add-in templates** we can use to gain immediate insights from our cohort. Point to and double click the add-in template **“Cohort Characterization”**.

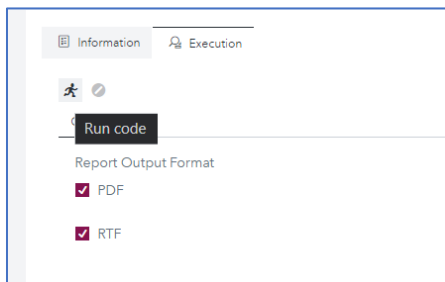


Step 2. Create a new instance by clicking **“New”** on the top right-hand side. Type in **“ACS-Stringent_[your initial]”** for the name of this instance.

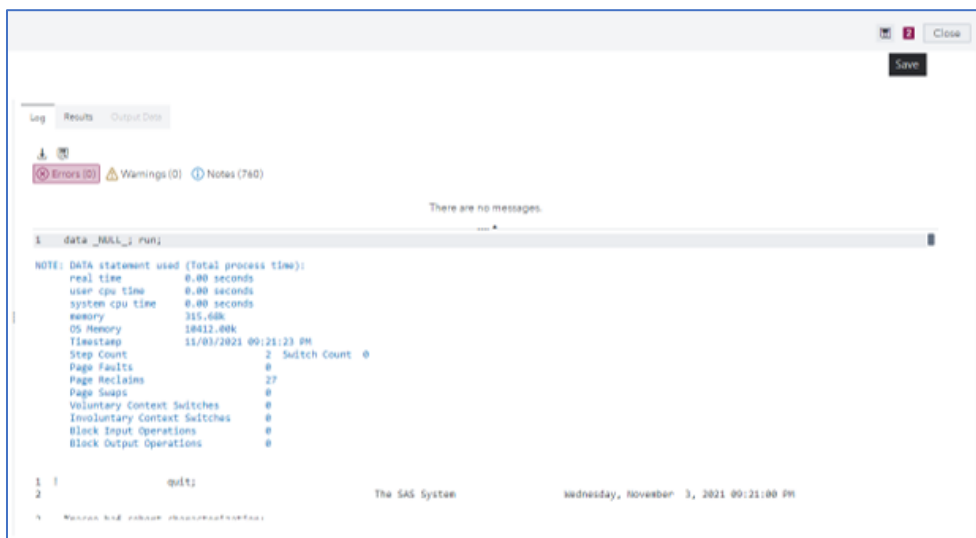


Step 3. Click the **Execution** tab on the top. This is the place you can execute the pre-registered SAS code for this descriptive analysis.

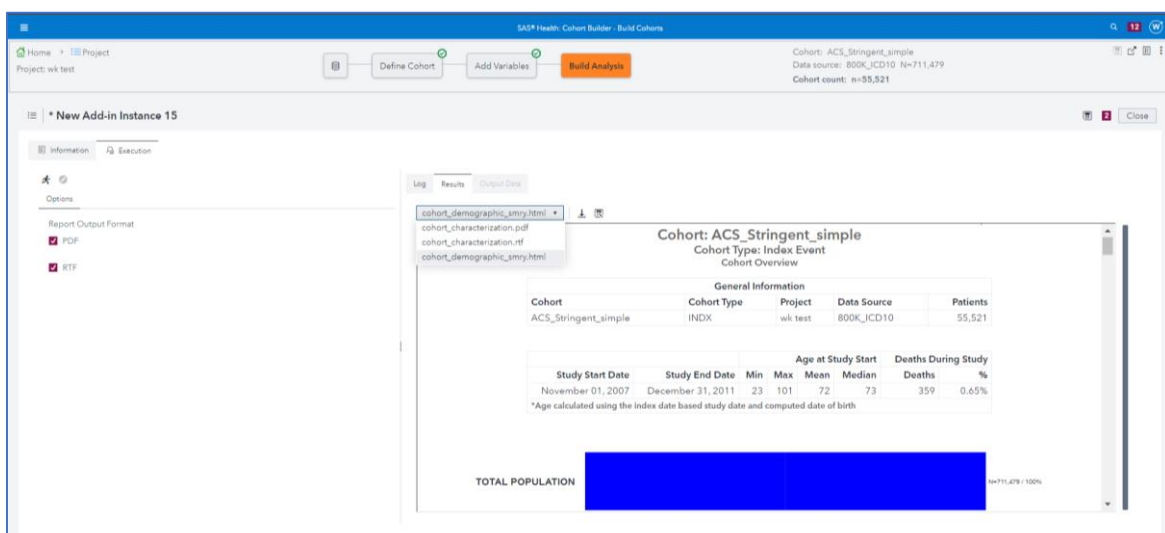
Select both **PDF** and **RTF**. Run the analysis by clicking **“Run code”** (the running man icon).



Step 4. Once it is completed, click the **“Save”** button on the right-hand side, and navigate to the **Log** tab to review the SAS Log file (code).



Step 5. Navigate to **Results** tab to see the output of this analysis. You can adjust the size of window by **dragging** the line in the center with four dots ()



Step 6. Navigate back to **Information** tab. The platform provides full transparency / governance on when this specific Add-In was created, last modified as well as the most recent run.

☰

ACS-Stringent_wk

📄 Information

🚦 Execution

Name: *

ACS-Stringent_wk

Description:

👤

Created:
11/17/21, 11:41 AM by William Kuan

👤

Last modified:
11/17/21, 11:52 AM by William Kuan

👤

Last run:
11/17/21, 11:48 AM by William Kuan

*NOTE: Other related analysis templated that would be used for this use case are Logistic Regression Model and Cohort Comparison (to compare cohorts with relaxed vs. stringent criteria). See screen shots below. The process to be followed would be similar. In the interest of time, we **DO NOT** run these analyses. Thank you for your understanding.*

[Logistic Regression Model Add-in]

Information

Execution

Data

Study Definition

Incidence Rates

Build Model

Export

Prediction Model Setting

Target outcome variable (maximum of two variables):

MI OR Cardiac Cath

Target variable value:

1

Home

Project

Project: wk test

Define Cohort

Add Variables

Build Analysis

Cohort: ACS_RetainedCohort

Data source: 0009_CD10 N=71,479

Cohort count: n=65,037

12

New Add-in Instance 4

Information

Execution

Data

Study Definition

Incidence Rates

Build Model

Export

Prediction Model Setting

Target outcome variable (maximum of two variables)

MI OR Cerebral Cereb

Target variable value:

1

Input variables

Selected items (3)

MI OR Cerebral Cereb

AGEGROUP_10 (8)

Antiarhythmics

Anticoagulants

IX_Bulk Meds

AV_E0001_INTX_E001

Log

Results

Output Data

AV_E0001_FUG_cohortprofile.html

AV_E0001_FUG_cohortprofile.html

AV_E0001_FUG_cohortprofile.html

AV_E0001_FUG_LRMModel.html

AV_E0001_FUG_LRMModel.html

AV_E0001_FUG_profile.html

AV_E0001_FUG_profile.html

AV_E0001_FUG_profile.html

New Add-in Instance 4_listing.htm

New Add-in Instance 4_listing.pdf

New Add-in Instance 4_listing.rtf

Cohort population characteristics

Characteristic	Level	MI OR Cerebral Cereb		Total	Statistic (P-value)
		Outcome=Yes	Outcome=No		
Ethnicity	Non-Hispanic	1608 (12.5)	1688 (12.36)	1695 (14.4)	3.80 (0.0001)
	Hispanic	1007 (7.8)	2080 (17.1)	3087 (14.3)	5.70 (0.0000)
	East North Cen	2058 (16.0)	4580 (37.1)	6638 (29.0)	
	East South Cen	1190 (9.2)	2080 (17.1)	3087 (14.3)	
	West North Cen	1191 (9.2)	3087 (25.0)	4278 (19.0)	
	South Atlantic	948 (7.5)	1688 (13.6)	2636 (11.6)	
	West South Cen	688 (5.5)	1007 (8.3)	1695 (7.4)	
	Mid South	1007 (7.8)	2080 (17.1)	3087 (14.3)	
	North Atlantic	3480 (27.4)	6194 (51.4)	9674 (43.4)	
	West North Cen	1194 (9.2)	2080 (17.1)	3274 (14.6)	
Gender Code	Male	1770 (13.8)	3070 (25.0)	4840 (21.5)	
	Female	1090 (8.5)	2190 (18.1)	3280 (14.6)	5.20 (0.0000)
Race Code	Hispanic	1007 (7.8)	2080 (17.1)	3087 (14.3)	5.20 (0.0000)
	Non-Hispanic	1608 (12.5)	1688 (13.6)	1695 (14.4)	3.80 (0.0001)
Age Group	18-24	1007 (7.8)	2080 (17.1)	3087 (14.3)	5.20 (0.0000)
	25-34	1007 (7.8)	2080 (17.1)	3087 (14.3)	5.20 (0.0000)
Sex	Male	1770 (13.8)	3070 (25.0)	4840 (21.5)	
	Female	1090 (8.5)	2190 (18.1)	3280 (14.6)	5.20 (0.0000)
Date	2010-2011	1007 (7.8)	2080 (17.1)	3087 (14.3)	5.20 (0.0000)
	2012-2013	1007 (7.8)	2080 (17.1)	3087 (14.3)	5.20 (0.0000)

[Cohort Comparison Add-in]

Manage Cohorts

Manage Project Add-ins

New Add-in Instance 5

Close

Information

Execution

Options

Log

Results

Output Data

Cohorts

Available Items (4):

ACS-RelaxedCriteria
ACS-RelaxedCriteria-44acd1b7-0104-4cb2-b54b-7cd4842...

ACS-Stringent
ACS-Stringent-80075e8f-cdc-4a7c-b16-671fdb58273

PHUSE_RWE_001
PHUSE_RWE_001-03432366-3113-4bb0-9166-976090cc48b

PHUSE_RWE_002
PHUSE_RWE_002-89774d1-ba34-4e68-92d7-1e0b86d319a

Selected Items (0):

No items

Add

No results were found. Submit the code to obtain results.

Manage Cohorts

Manage Project Add-ins

Rvs5_wk

Close

Information

Execution

Options

Log

Results

Output Data

Cohorts

Available Items (2):

PHUSE_RWE_001
PHUSE_RWE_001-03432366-3113-4bb0-9166-976090cc48b

PHUSE_RWE_002
PHUSE_RWE_002-89774d1-ba34-4e68-92d7-1e0b86d319a

Selected Items (2):

ACS-RelaxedCriteria
ACS-RelaxedCriteria-44acd1b7-0104-4cb2-b54b-7cd4842...

ACS-Stringent
ACS-Stringent-80075e8f-cdc-4a7c-b16-671fdb58273

New Add-in Instance 5_listing.htm

Download

Print

Multiple Cohorts: Population Characteristics

Characteristic	Level	Cohort name			
		ACS-RelaxedCriteria		ACS-Stringent	
		N	%	N	%
AGEGROUP_10	20-29	245	0.59	212	0.64
	30-39	755	1.68	554	1.66
	40-49	1586	3.52	1144	3.43
	50-59	2609	5.79	1936	5.81
	60-69	10391	23.07	7915	23.76
	70-79	16211	35.99	11980	35.96
	80-89	10536	23.39	7671	23.03
	90+	2684	5.96	1899	5.70
AV_ANTIARRHY_FLG	0	44908	99.71	33213	99.71
	1	129	0.29	98	0.29
AV_ANTIAGOAG_FLG	0	43558	96.72	32234	96.77
	1	1479	3.28	1077	3.23

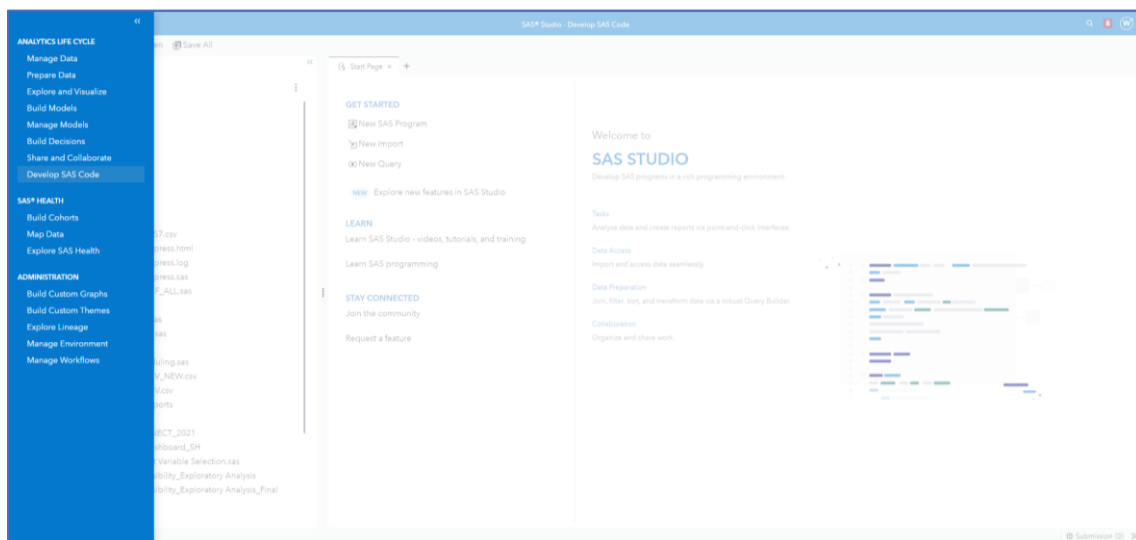
PART 2 – Identify Variable Importance via Data Driven (ML) & SME Approach



In the previous example, the inclusion/exclusion criteria for the cohort are selected by subject-matter experts. What if we take another approach, a data-driven approach, to select the eligibility criteria related to the outcome of interest by the support of a machine learning technique and SME? How will that impact the cohort?

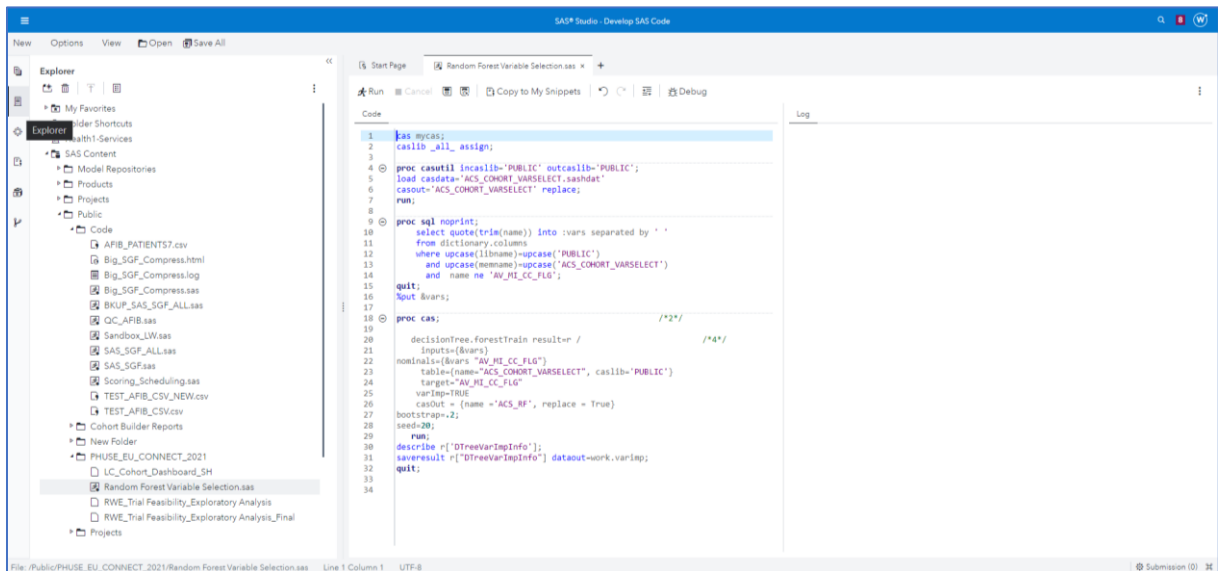
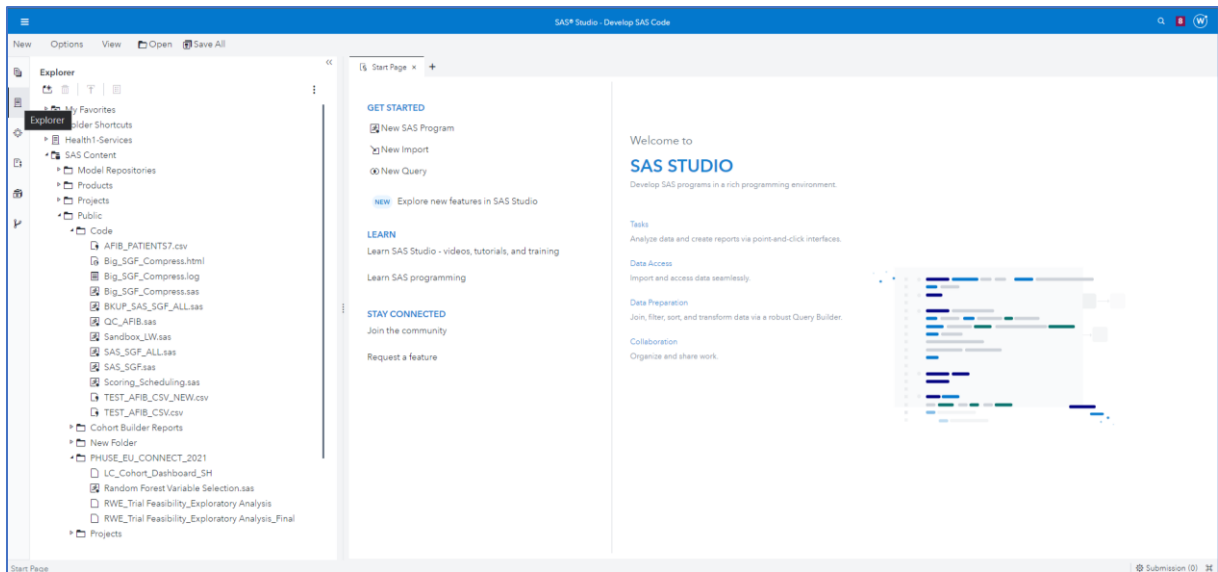
Identify the variable importance using ML (SAS Studio) and SME

Step 1. Navigate to “Develop SAS Code” to launch SAS Studio.

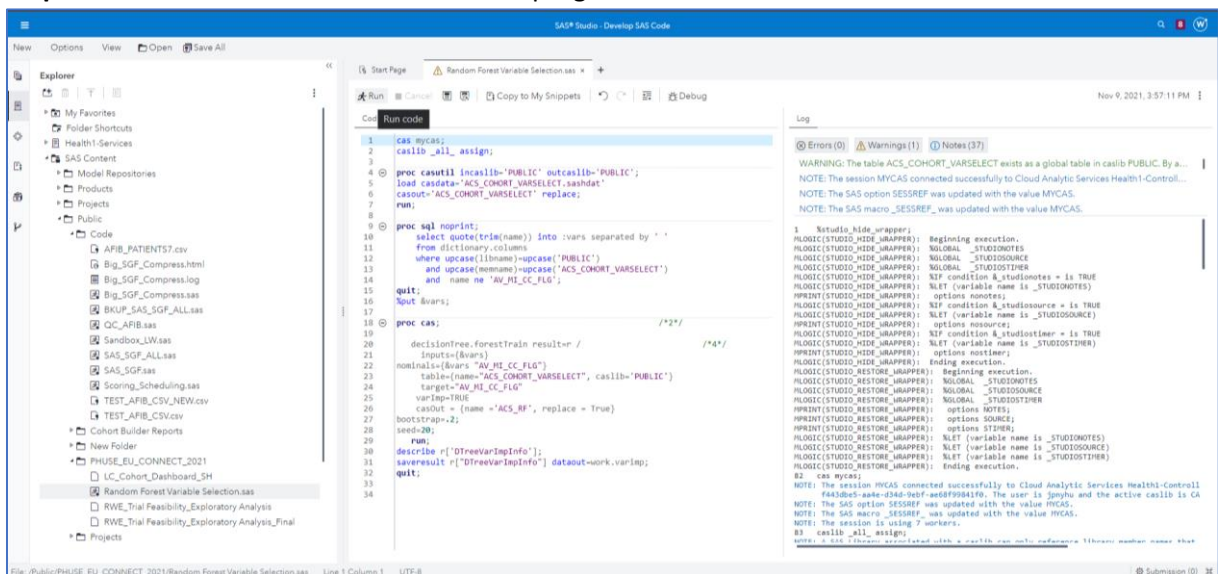


Step 2. Click “Explorer” on the left-hand side, and then double click the SAS program “Random Forest Variable Selection.sas” (SAS Content / Public / PHUSE_EU_CONNECT_2021) to open.

This is the program you will use to run a random forest algorithm to **identify the variable importance**.



Step 3. Click “Run code” to execute the SAS program.



Step 4. Once it is completed, navigate to “**Libraries**” on the left-hand side, and double click the result dataset “**VARIMP**” under the WORK folder.

```

1 cas mycas;
2 caslib all_assign;
3
4 proc casutil incaslib="PUBLIC" outcaslib="PUBLIC";
5 load casdata="ACS_COHORT_VARSELECT.sashdat"
6 casout="ACS_COHORT_VARSELECT" replace;
7 run;
8
9 proc sql noprint;
10 select quote(trim(name)) into :vars separated by ' '
11 from dictionary.columns
12 where upcase(libname)="PUBLIC"
13 and upcase(memname)="ACS_COHORT_VARSELECT"
14 and name ne "AV_ML_CC_FLG";
15 quit;
16 %put &vars;
17
18 proc cas; /*2*/
19
20 decisionTree.ForestTrain result= / /*4*/
21 inputs=&vars
22 nominal=&vars "AV_ML_CC_FLG"
23 table=(name="ACS_COHORT_VARSELECT", caslib="PUBLIC")
24 target="AV_ML_CC_FLG"
25 varImp=TRUE
26 casout = (name = "ACS_RF", replace = true)
27 bootstrap=2;
28 seed=20;
29 run;
30 describe r["DTreeVarImpInfo"];
31 save result r["DTreeVarImpInfo"] dataout=work.varimp;
32 quit;
33
34

```

Variable	Importance	Std
ELIX_RISK_GROUP	8.4157660524	9.0320063189
ELIX_ARRHY	7.1973432381	10.9118209008
ELIX_COPD	6.8698537942	10.127913702
ELIX_RISK_SCORE	5.3759545696	2.7208401029
AV_STROKE_FLG	5.279686092	5.8244019389
ELIX_RF	1.9877886061	5.2273297216
ELIX_VD	1.9815165697	3.249297233
ELIX_OTHND	1.931452558	5.6863096787
ELIX_DEP	1.75798055	3.6189859827
ELIX_DA	1.5879503153	3.294495888
ELIX_LD	1.5700541582	5.6572364723
ELIX_BLA	1.5218822009	4.405726285
ELIX_PVD	1.4656543314	4.6914810912
ELIX_COAG	1.4391626981	3.5045203398
AV_ANTICOAG_FLG	1.1302807709	2.2671818946
AV_E0001_FLG	1.1220891113	1.551854156
ELIX_WL	1.0235829467	3.0800142376
ELIX_RHEUM_A	0.9633786293	3.1478803998
ELIX_TUMOR	0.7559743623	1.8998993338
ELIX_PCD	0.6666796943	2.7322887178

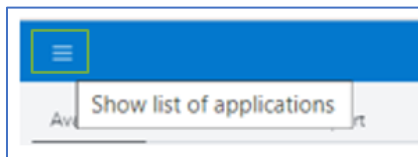
The Subject Matter Expert (SME) will now investigate this variable importance list (created by the Random Forest ML algorithm). Afterwards he/she will create a 3rd Cohort based on the combination of this output and his/her domain expertise.

Identify the variable importance using ML (R Studio) and SME

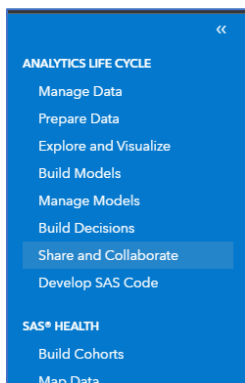


This is the part to demonstrate how you can run the ML method we used in SAS by using R code instead. An open-source user can call SAS CAS actions and leverage the power and performance scalability of SAS advanced analytics platform as well.

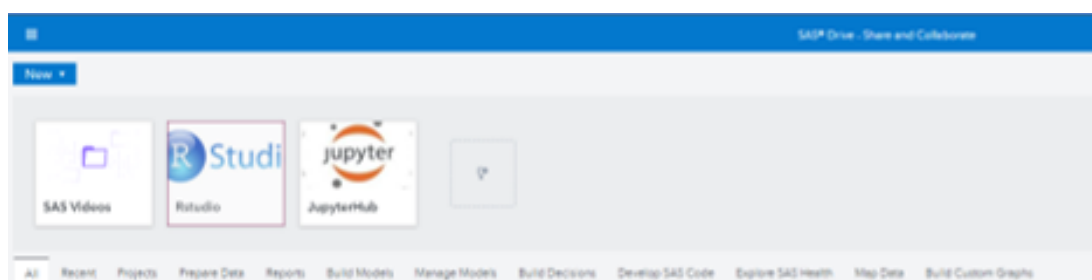
Step 1. Navigate to the **menu panel** on the left-hand side by selecting the ‘Show List of Applications Icon’ (=left top corner).



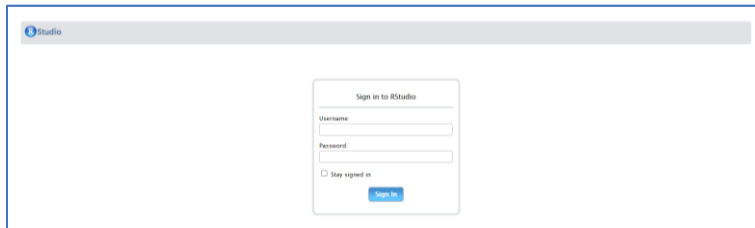
Step 2. In the menu panel on the left-hand side choose ‘Share and Collaborate’.



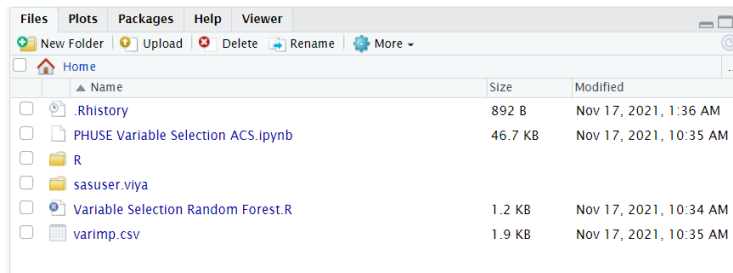
Step 3. Click “Shared” folder on the left-hand side, and then double click “Rstudio” to launch the application. (Note: You can also access it from Quick Access bar on the top if you pinned the application on it already)



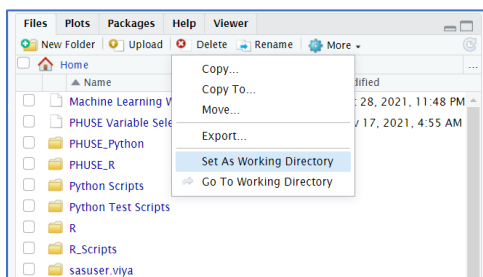
Step 4. Enter your Username and Password to sign in to RStudio



Step 5. Open the **Variable Selection Random Forest.R** file which has been pre-loaded into your directory environment.



Step 6 (optional). Click “More” button and set current directory as working directory.



Step 7. Provide your **username** and **password** in CAS api call. The CAS() call creates a CAS connection object.

```
1 library(swat)
2 #library(getPass)
3 library(tidyr)
4
5 Sys.setenv(CAS_CLIENT_SSL_CA_LIST="/opt/sas/viya/config/etc/SASSecurityCertificateFramework/cacerts/
6 session<-CAS("Health1-Controller.azure.com",5570,username="***",password="****")
7
```

Step 8. Run the R code to execute the program.

1/ The variable importance (ranking) can be observed in the Console by entering “**results\$DTreeVarImpInfo**” to see CAS data frame object.

2/ The variable importance table is also exported as “**varimp.csv**” file to your working directory.

The Subject Matter Expert (SME) will now investigate this variable importance list (created by the Random Forest ML algorithm). Afterwards he/she will create a 3rd Cohort based on the combination of this output and his/her domain expertise.

The way of creating this 3rd Cohort is similar as the steps you performed in Part 1.

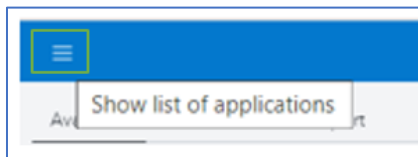
In the interest of time, we skip this specific step and proceed with the rest of the hands-on script.

Identify the variable importance using ML (Python) and SME

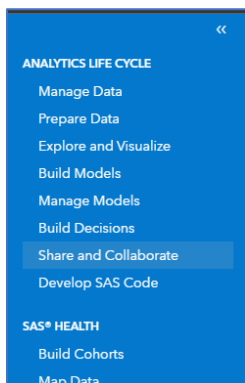


This is the part to demonstrate how you can run the ML method we used in SAS by using R code instead. An open-source user can call SAS actions and leverage the power and performance scalability of SAS advanced analytics platform as well.

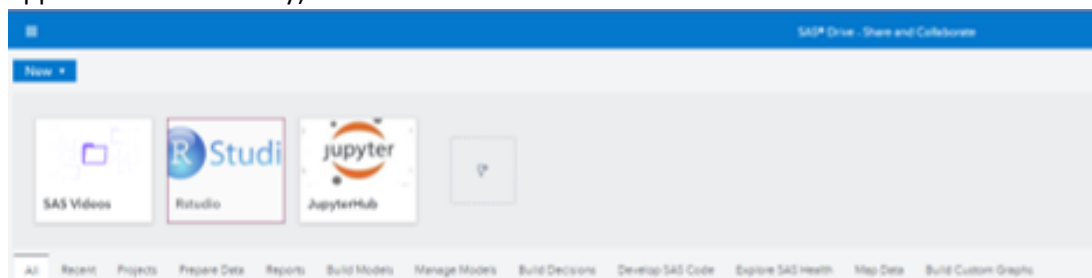
Step 1. Navigate to the **menu panel** on the left-hand side by selecting the ‘Show List of Applications Icon’ (=left top corner).



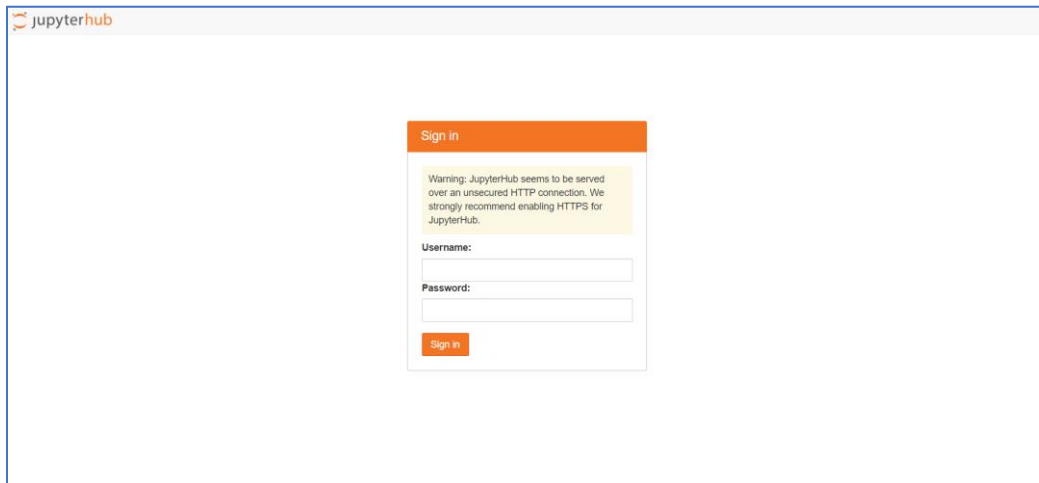
Step 2. In the menu panel on the left-hand side choose ‘Share and Collaborate’.



Step 3. Click “Shared” folder on the left-hand side, and then double click “**JupyterHub**” to launch the application. (Note: You can also access it from Quick Access bar on the top if you pinned the application on it already)



Step 4. Enter your **username** and **password** to log in to jupyterhub

The image shows the JupyterHub login interface. At the top left is the 'jupyterhub' logo. In the center, there is a 'Sign in' box. Inside this box, there is a warning message: 'Warning: JupyterHub seems to be served over an unsecured HTTP connection. We strongly recommend enabling HTTPS for JupyterHub.' Below the warning, there are input fields for 'Username:' and 'Password:'. At the bottom of the box is a 'Sign in' button.

Step 5. Open the “PHUSE Variable Selection ACS.ipynb” notebook

The image shows a list of notebooks in JupyterHub. The first notebook is 'PHUSE Variable Selection ACS.ipynb'. To its right, it says '12 days ago' and '46.7 kB'.

Step 6. Enter your **username** and **password** in the `swat.cas()` api function

The image shows a Jupyter notebook cell. The title is 'Variable Selection Example Using SWAT'. Below the title, it says 'In this example we further reduce the variable in ACS_COHORT_VARSELECT table and feed selected variable into a neural network model'. Below that, it says 'Connect to CAS'. The code cell contains the following code:

```
In [2]: import swat
import matplotlib.pyplot as plt
conn=swat.CAS("Health1-Controller.azure.com",5570,username="****",password="*****")
```

Step 7. Execute the notebook cells to get the variable importance table.

A csv file is also saved into your jupyterhub root directory.

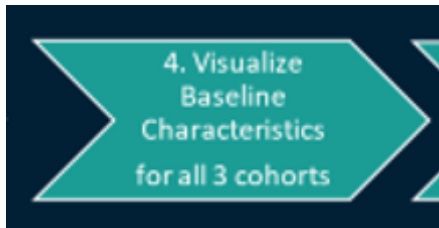
The Subject Matter Expert (SME) will now investigate this variable importance list (created by the Random Forest ML algorithm). Afterwards he/she will create a 3rd Cohort based on the combination of this output and his/her domain expertise.

The way of creating this 3rd Cohort is similar as the steps you performed in Part 1.

In the interest of time, we skip this specific step and proceed with the rest of the hands-on script.

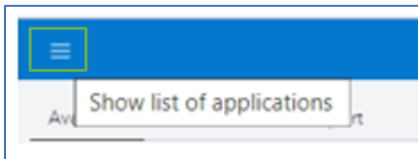
PART 3 – Visualize and Interpret the Data

Visualize Baseline Characteristics for all cohorts

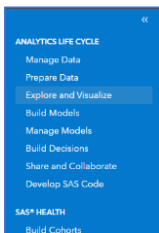


Next, let's explore **potential shifts in baseline characteristics** between the 3 cohorts of interest.

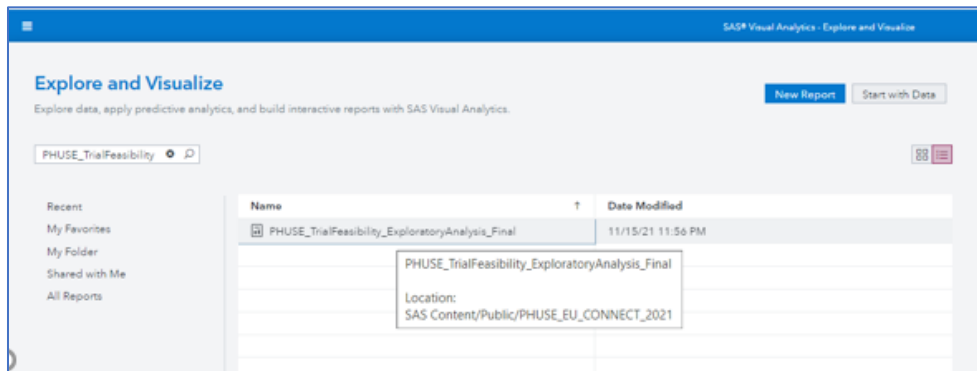
Step 1. Navigate to the **menu panel** on the left-hand side by selecting the 'Show List of Applications Icon' (=left top corner).



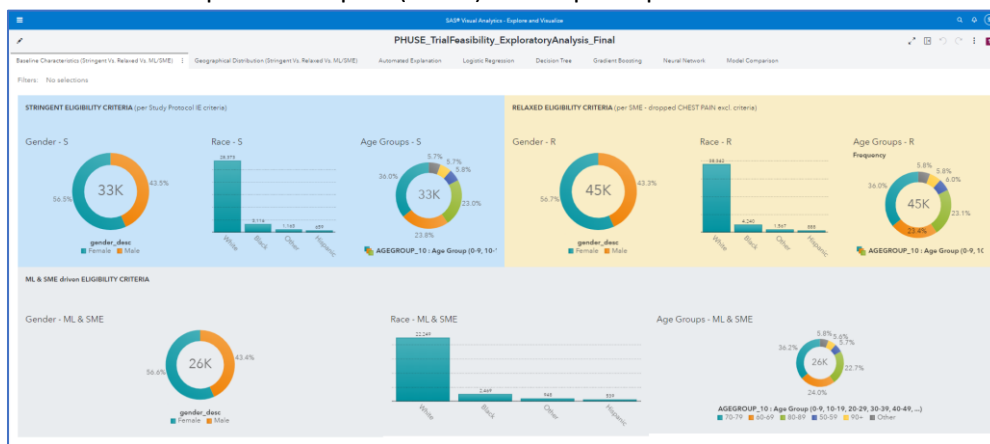
Step 2. In the menu panel on the left-hand side choose '**Explore and Visualize**'.



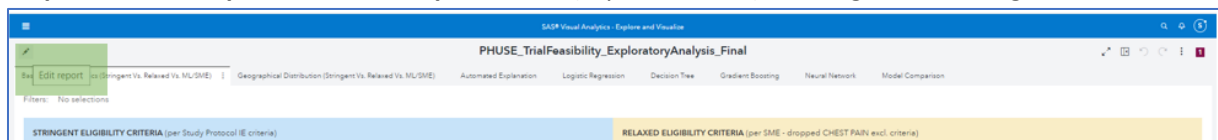
Step 3. Enter “PHUSE_TrialFeasibility_ExploratoryAnalysis_Final” in the search box on the left-hand side.



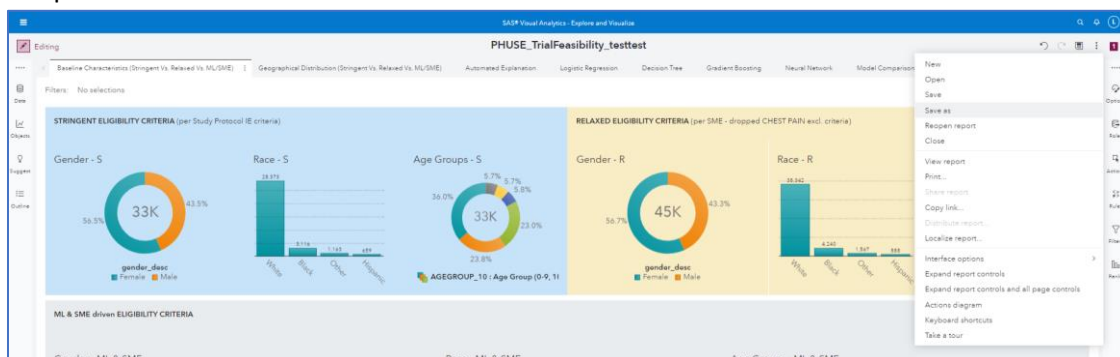
Double click the pre-built report (name). The report opens.



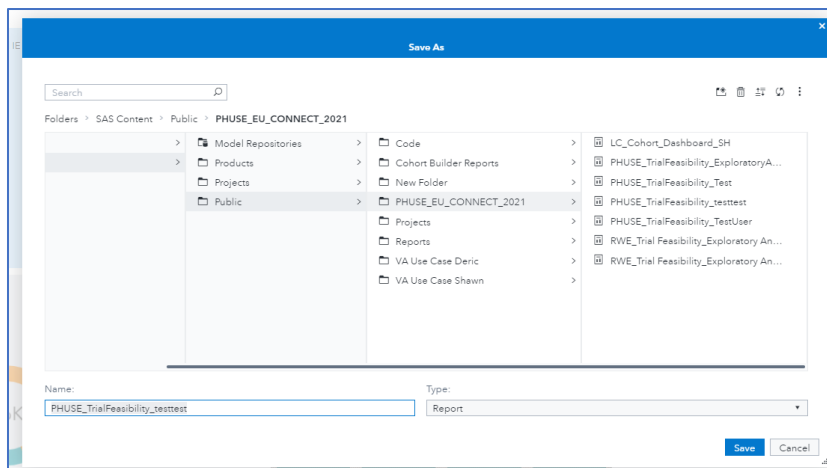
Step 4. Edit the report. Hit first the **pencil icon** (top left corner) to navigate to editing mode.



Once in editing mode, navigate to the top right corner, hit the **menu icon (3 dots)** and select the **Save as** option.



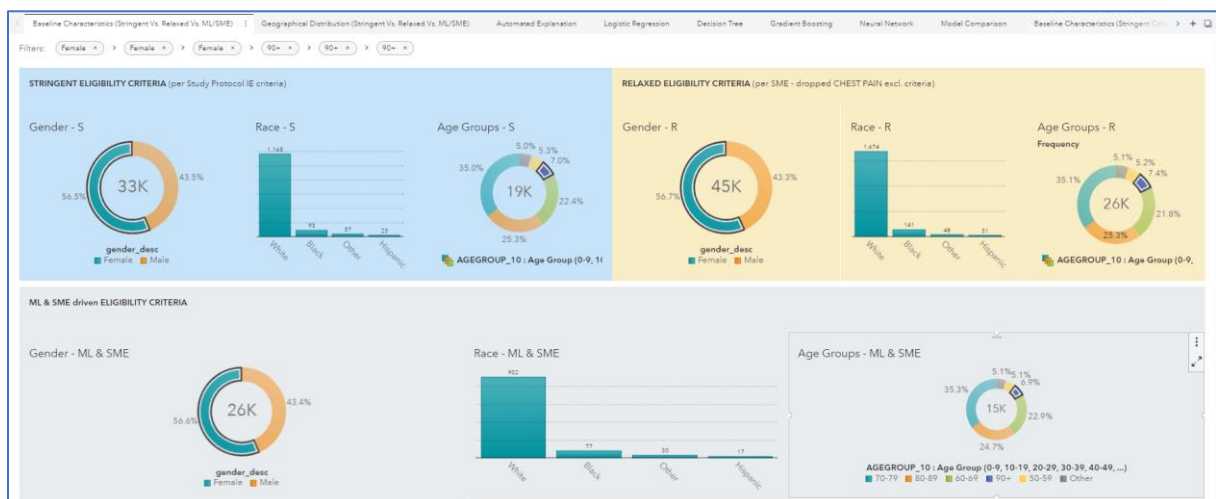
Save copy of the report under same folder (SAS Content/Public/PHUSE_EU_Connect_2021). Save as “PHUSE_TrialFeasibility_[UserInitials]”.



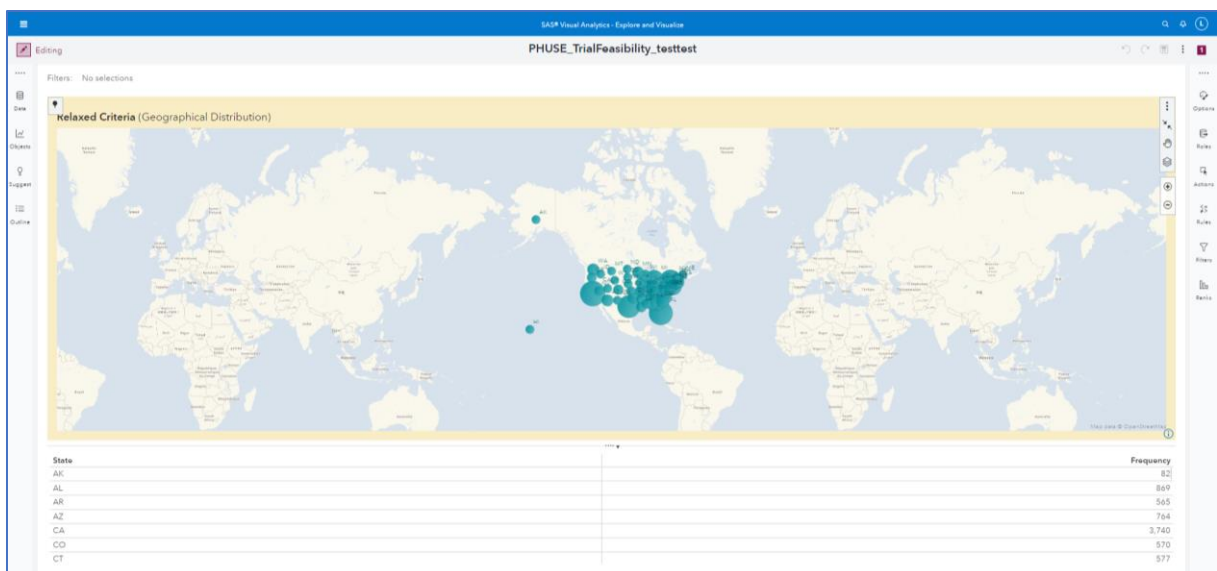
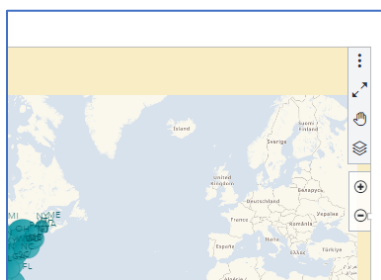
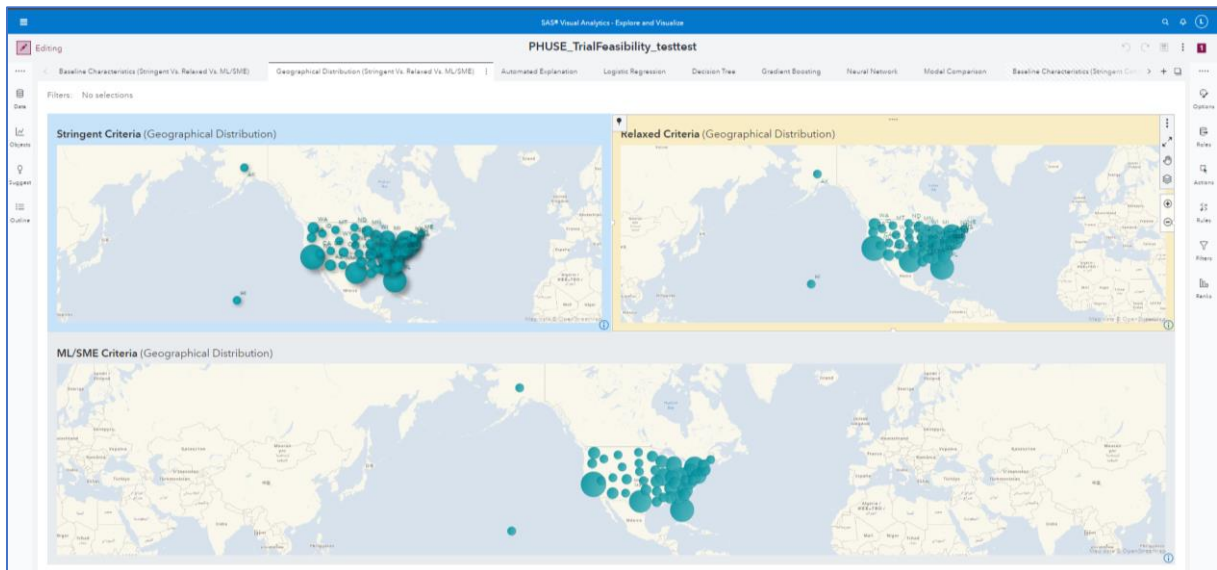
You can stay in Editing mode. If you want however to exit editing mode, hit the pencil icon (top left corner) again.

Step 5. In the “Baseline Characteristics” tab, we can see the comparison of baseline characteristics between our 3 cohorts (Stringent, Relaxed and ML/SME). Data was automatically loaded from Cohort Builder into SAS Viya In Memory (CAS engine).

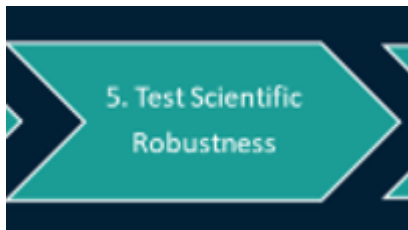
Click around the pie chart(s) or bar chart(s) to see how the data and visual representation is automatically changing accordingly. In the example below we filtered on “Female” and “90+” Age group for all 3 cohorts. You can easily remove the filters again by removing these on top of the screen.



Step 6. The next Characteristics tab represents the **Geographical distribution** of all 3 cohorts. You can zoom in/zoom out; hover over to see numbers or get a tabular view with all figures (by maximizing the screen using the arrow icon on the right-hand side, click it again to restore).



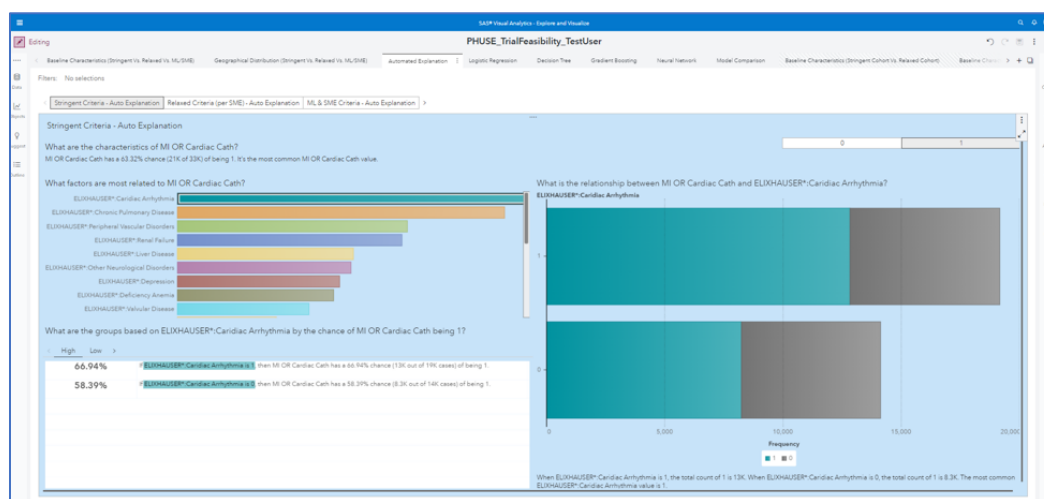
Test Scientific Robustness



Step 7. The next tab is an ***Automated Explanation*** overview for all 3 cohorts of interest.

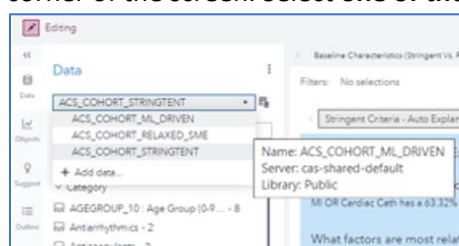
Explore relationship between variables: Using machine learning and natural language generation (NLG) we are ‘out of the box’ providing information on which factors (variables) are most important in predicting the outcome “**MI or Cardiac Cath**”.

Please navigate through the 3 tabs (*underneath automated explanation*) to get a better understanding of the data at hand. Select filter (0 or 1) and individual factors to explore the interactivity of the report.

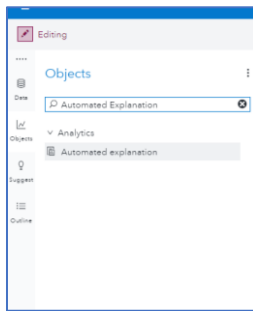


Step 8. Let's now create your own Automated Analysis.

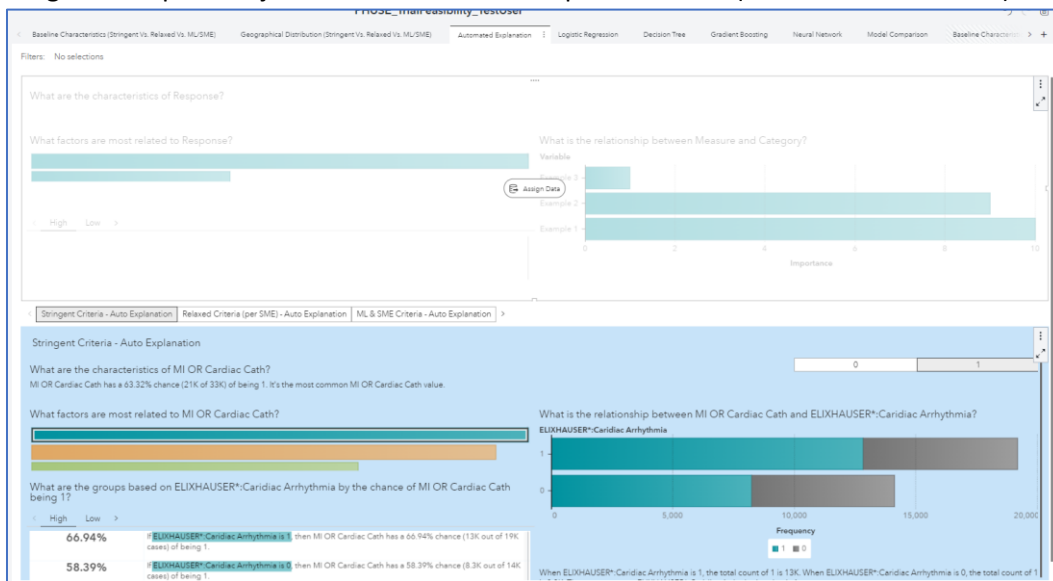
Navigate to the **data** icon on the left-hand side. Note that you can **pin/unpin the pane** at bottom left corner of the screen. Select **one of the 3** data sources (Stringent, Relaxed, ML_driven).



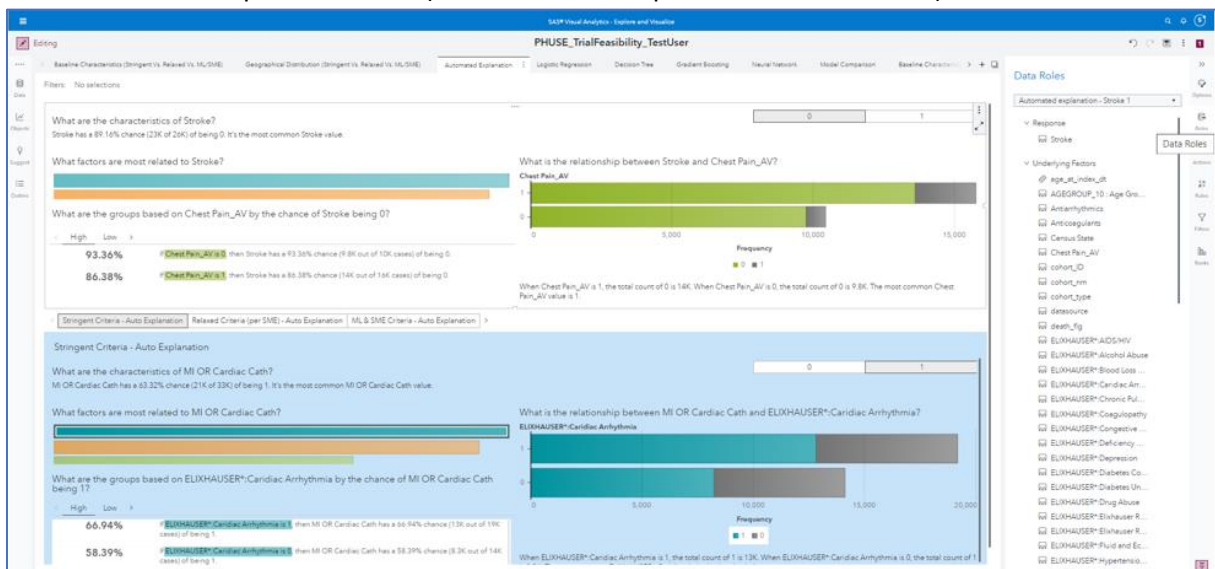
Now that you have selected the proper data source please navigate to the **Objects** library (left hand side, below Data). Enter **“Automated Explanation”** in the Filter text box.



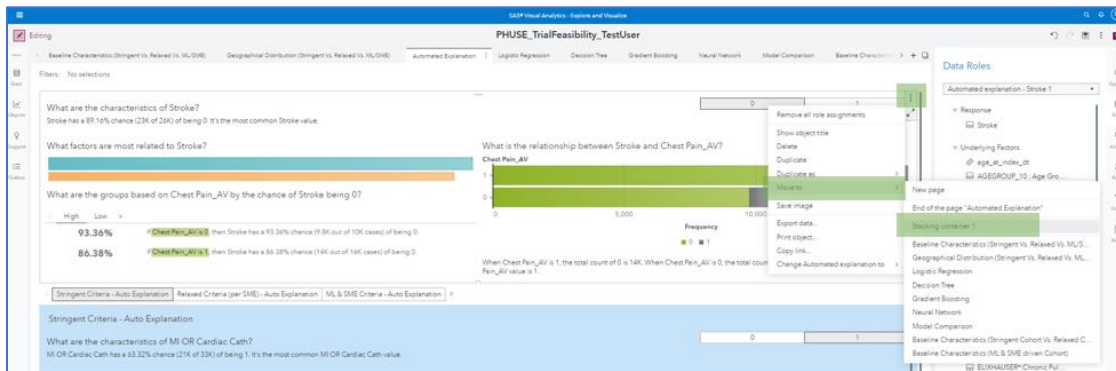
Drag and Drop the object on the Automated Explanation Tab (see screenshot below).



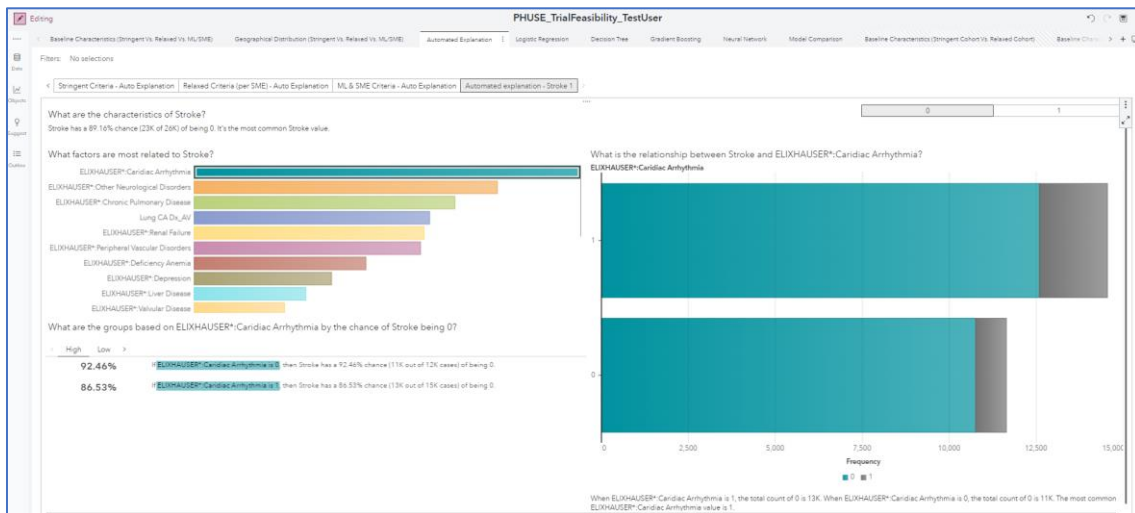
On the right-hand side open the **Data Roles** option. Select **Stroke** or **HF_SE** or **Elixhauser Congestive Heart Failure** as Response variable (in the below example **Stroke** was selected).



Hit the **3 dots** besides the new automated explanation report, select **move to**, select **Stacking Container 1**.



As a result, your new Automated explanation report is now **part of the container** (together with the already existing 3 reports).



Finally, you can remove variables that you feel are duplicate or not relevant. This can be done directly in the graph (right mouse click/remove) - or in the Underlying Factors menu (on the right-hand side).

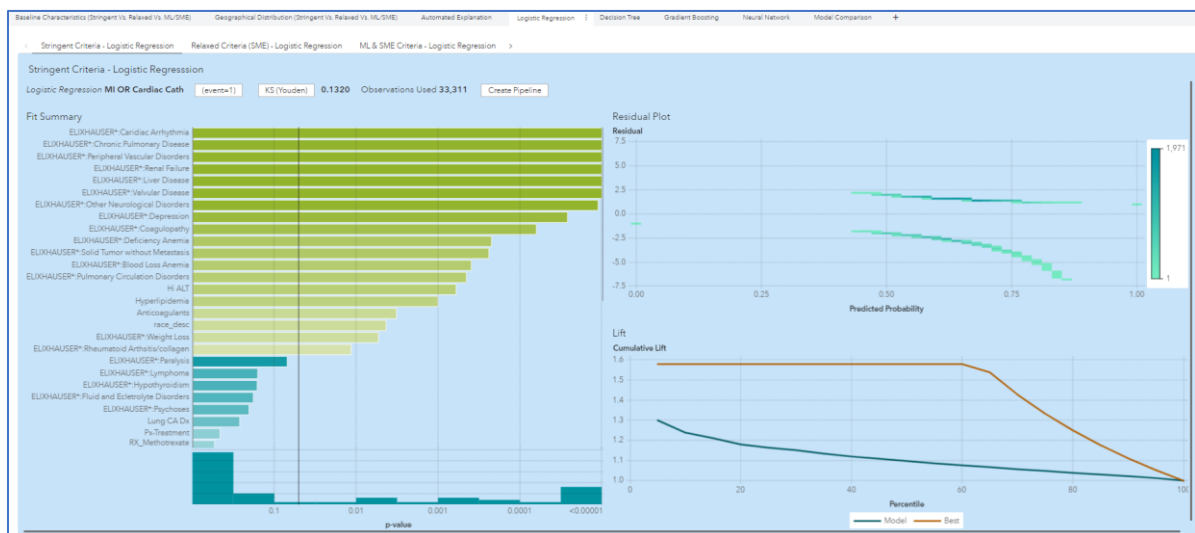
For example, please remove some of the below variables.

- ☐ AGEGROUP_10 : Age Group (0-9, 10-19, 20-29, 30-39, 40-49, ...) - 8
- ☐ Census State - 10
- ☐ Chest Pain_AIV - 2
- ☐ cohort_ID - 1
- ☐ cohort_nm - 1
- ☐ cohort_type - 1
- ☐ datasource - 1
- ☐ ELIXHAUSER*:Congestive Heart Failure - 2
- ☐ ELIXHAUSER*:Elixhauser Risk Factor Set Group - 4
- ☐ ELIXHAUSER*:Elixhauser Risk Factor Set Score - 62
- ☐ Gender Code - 2
- ☐ HF_SE - 2
- ☐ MI OR Cardiac Cath Outcome - 2
- ☐ project_ID - 1
- ☐ project_nm - 1
- ☐ Race Code - 4
- ☐ State - 52
- ☐ Unique Member ID - 26K
- ☐ ZIP Code - 1
- ☐ Year of Birth
- ☐ age_at_death
- ☐ age_at_study_end
- ☐ age_at_study_start

Step 9. Next, let's **explore different ML models** on all 3 cohorts.

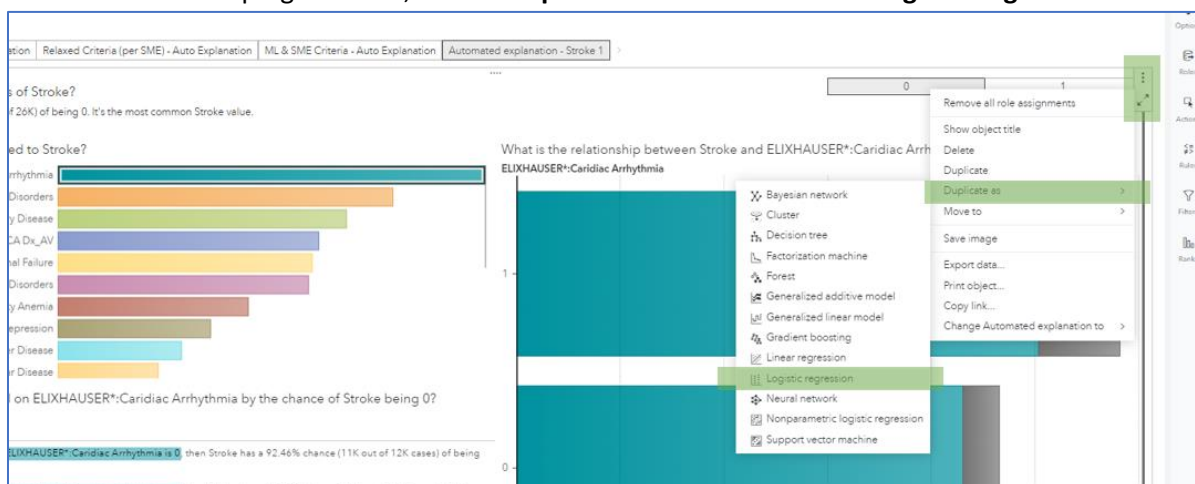
Is the golden standards / industry standard approach (*logistic regression*) the best or is there another (ML) approach we also need to consider? In addition, we also want to **test the impact of the eligibility criteria** between all 3 cohorts on the **statistical outcome**.

Please navigate through the 3 tabs (*underneath logistic regression*) to get a better understanding of the logistic regression analysis for Stringent, Relaxed and ML-SME Cohorts.

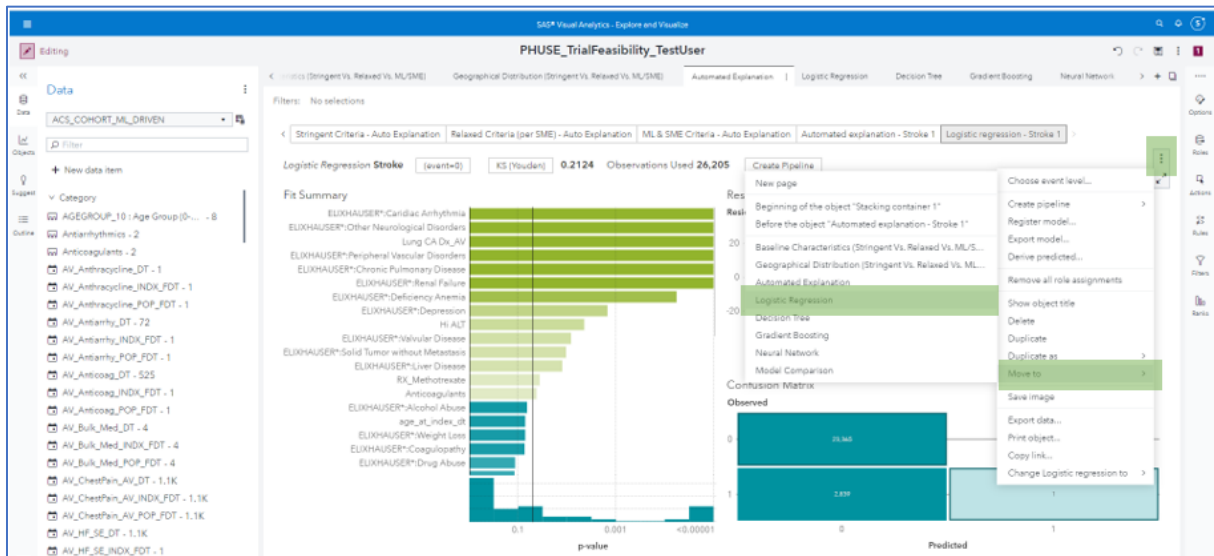


Let's create now your own ML models. Instead of creating the ML models from scratch we will leverage the work we already completed on the automated analysis (previous step).

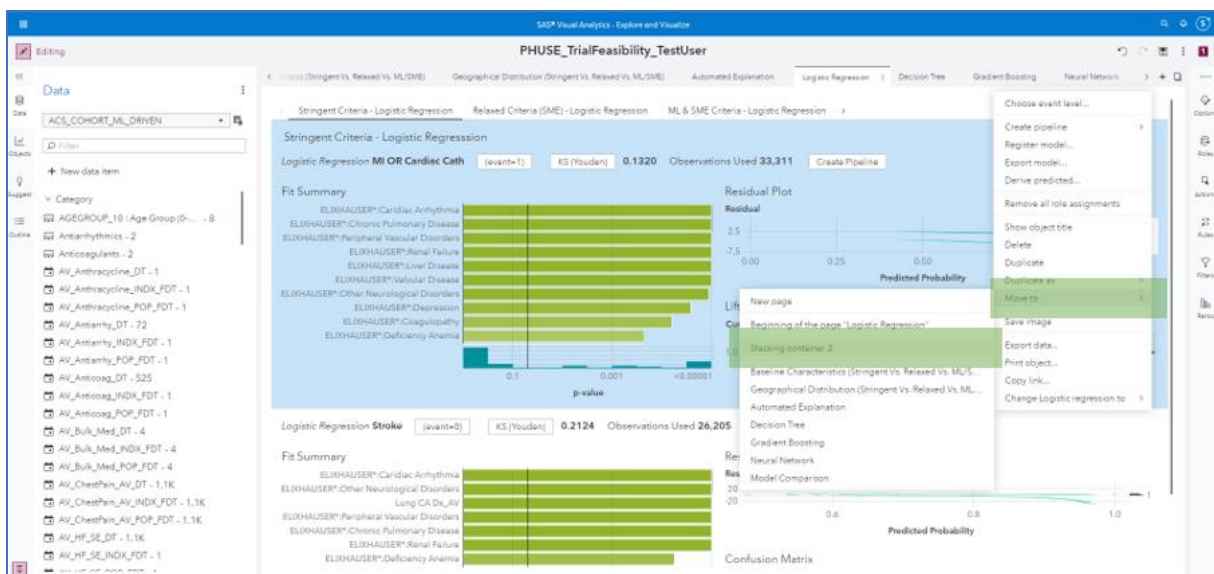
Navigate to the Automated Explanation **tab** and newly created Automated Analysis **Report**. Select the 3 dots on the top right corner, select **"Duplicate as"** and afterwards **"Logistic Regression"**.



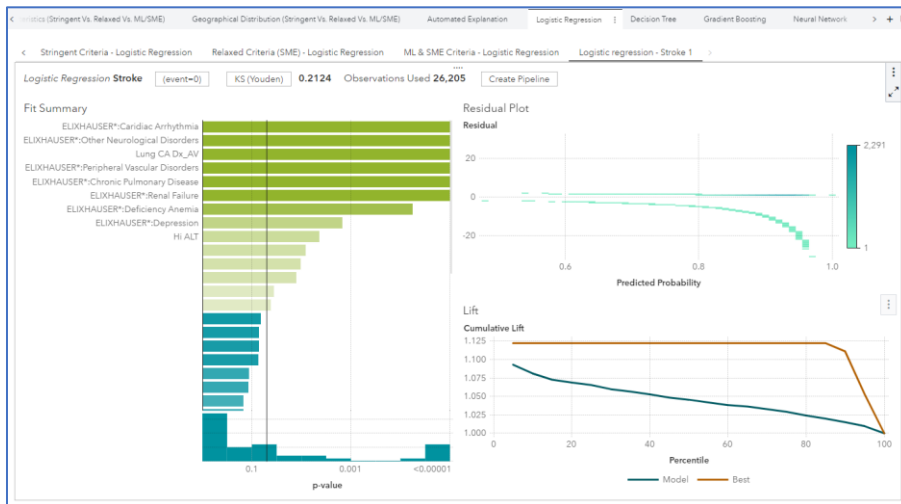
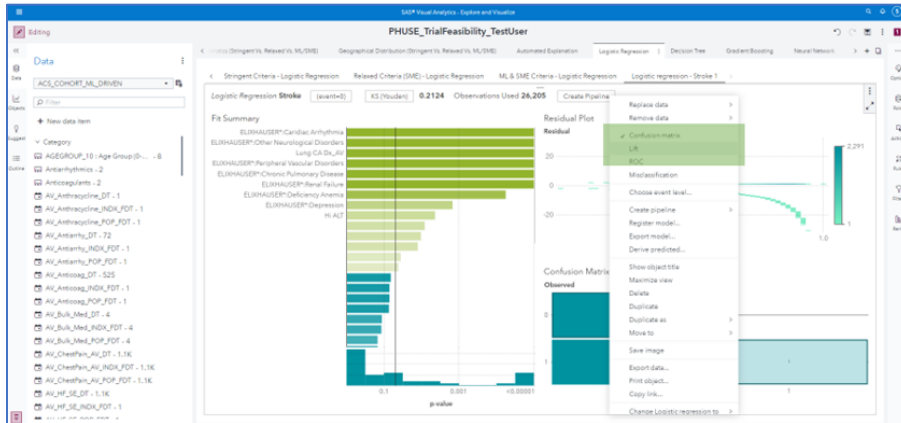
Move this newly created logistic regression report from the Automated Explanation Tab to the Logistic Regression Tab. Select the 3 dots on the top right corner, select **“Move to”** and afterwards **“Logistic Regression”**.



The newly created Logistic Regression analysis is now available under the correct Tab. Please move it still to the same logistic regression container. Select newly created Logistic Regression analysis; hit the 3 dots on the top right corner, select **“Move to”** and afterwards **“Stacking container 2”**.

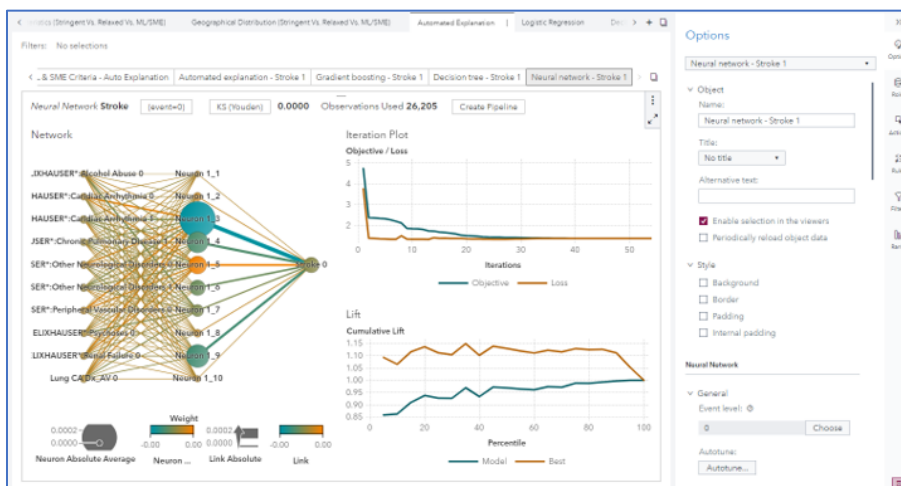


As a final step use **right mouse click** on the **Confusion Matrix**. Select **Lift** instead of Confusion Matrix.



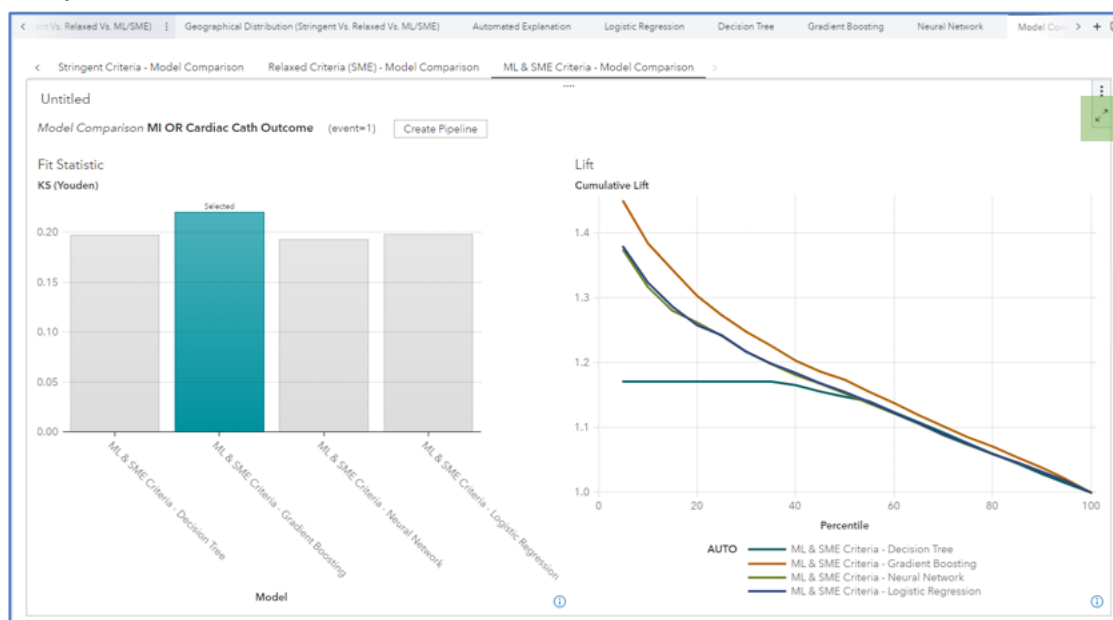
Now repeat the same the same steps for **Gradient Boosting, Decision Tree and Neural Network**. You can **start from the Automated Explanation** (same approach as with Logistic Regression). Feel free to explore the **Options menu** on the right-hand side for each of these objects / reports.

- Duplicate As – Decision Tree
- Duplicate As – Gradient Boosting
- Duplicate As – Neural Network (example below)



Step 10. Finally, you want to **compare all these ML models** to find the best performing/best fit.

First, navigate to the existing **Model Comparison** Tab (below you can find an example from **ML & SME Criteria Cohort – Model Comparison**). You will notice that **Gradient Boosting** is highlighted as the **champion model**.

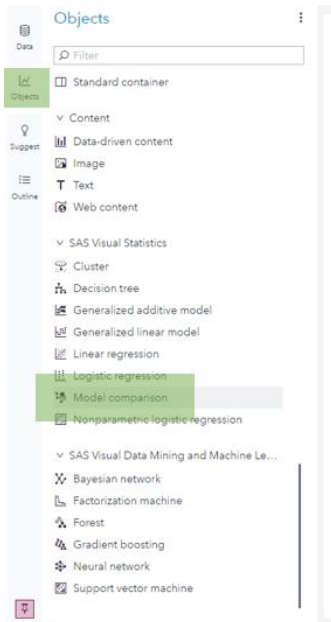


Note: When maximizing reports (using the arrow on the top right-hand corner – see screenshot above) you will for each report get the underlying data (example below). To go back hit again the arrow.



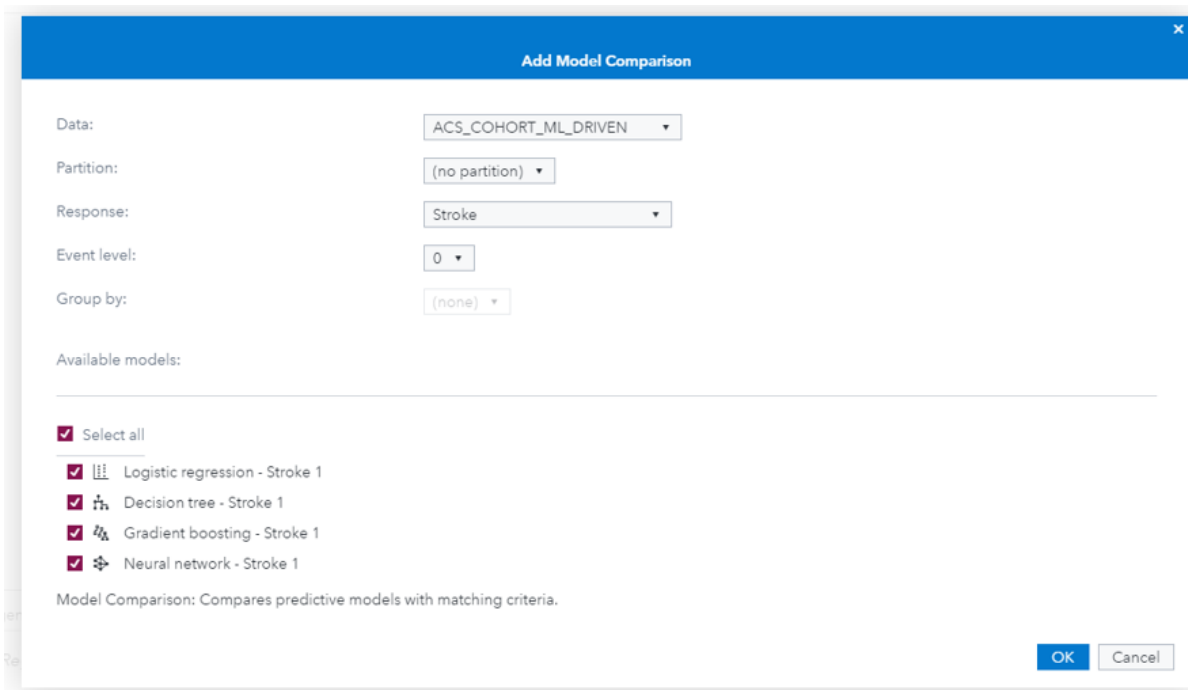
Let's create now your own Model Comparison report.

Navigate to the **Objects menu** (left hand side) and search for Model comparison (or scroll down to the SAS Visual Statistics section). Drag and Drop the **Model Comparison** option on the canvas.

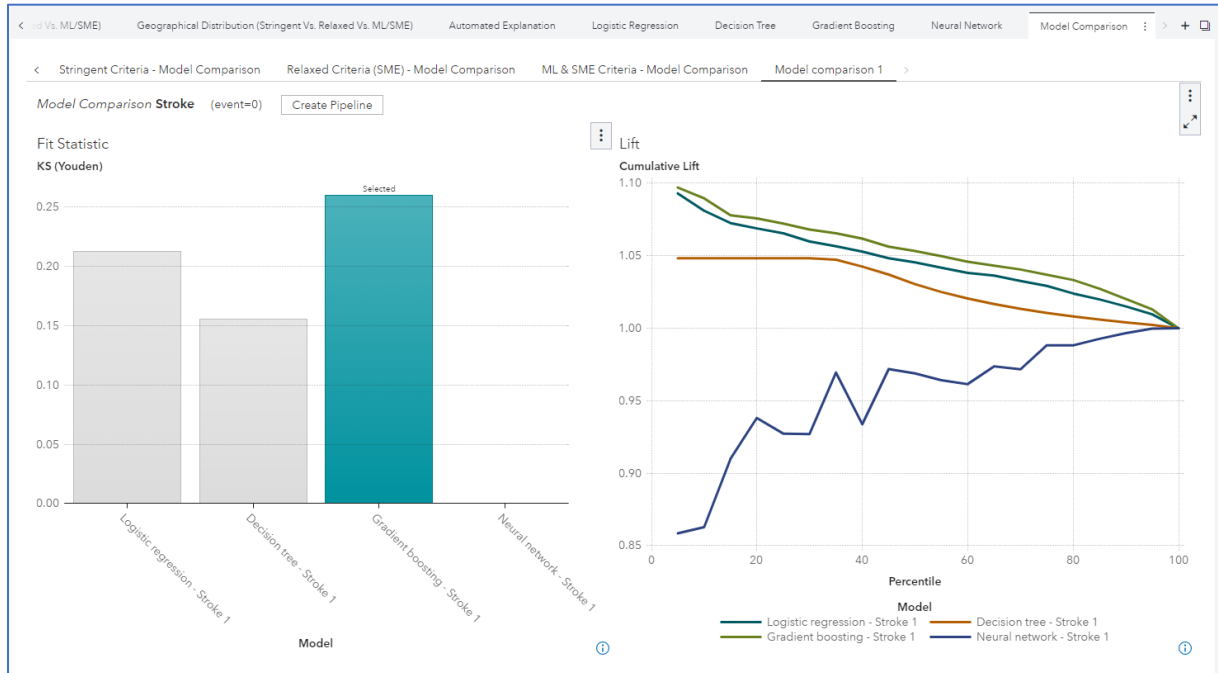


In the next step **select the Data Source** you used to create your ML models previously (in my case it was the **ML_Driven Cohort**).

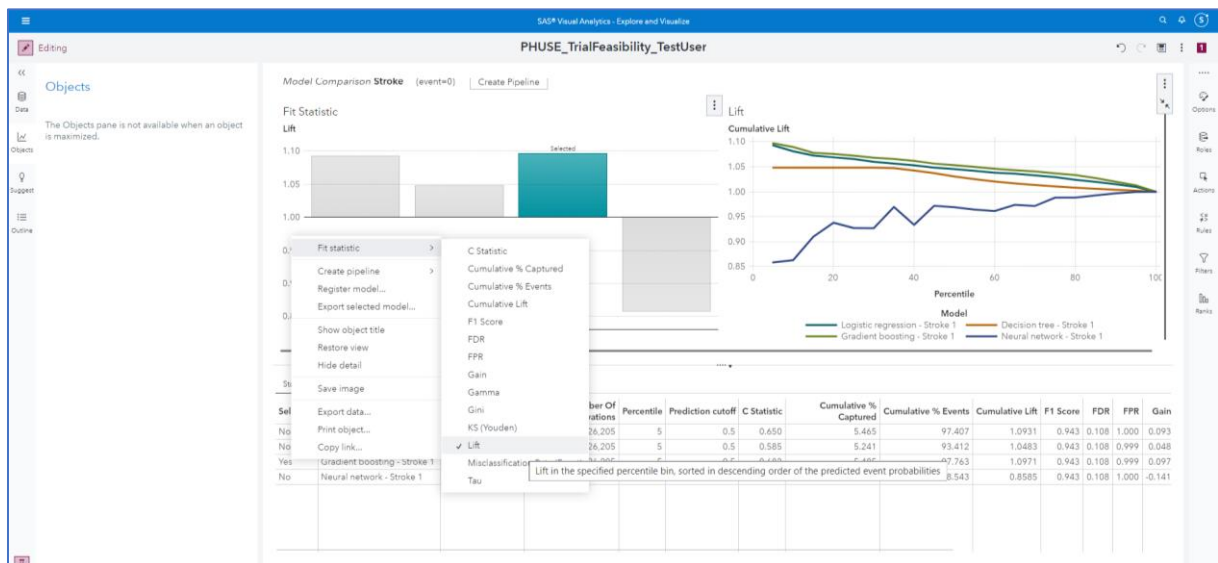
Select the **Response variable** you used for your exercise (in my case it was **Stroke**). Select All available models (Logistic Regression – Stroke 1, Decision Tree – Stroke 1,...) and hit the **OK** button.



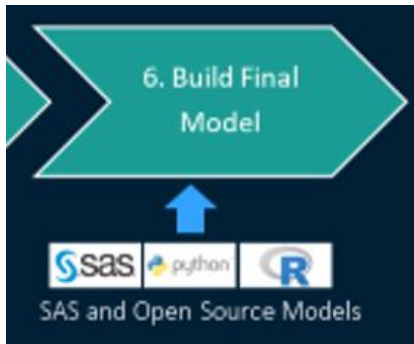
The newly created model comparison is now available.



For each report you can also change the **Fit Statistic method** applied by right clicking a mouse. See below screenshot for options.



PART 4 – Build Final Model(s)

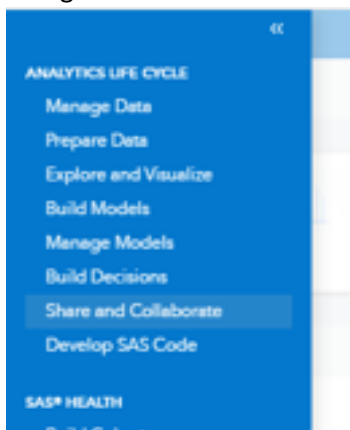


Build Models using model pipelines (SAS, R, Python models)

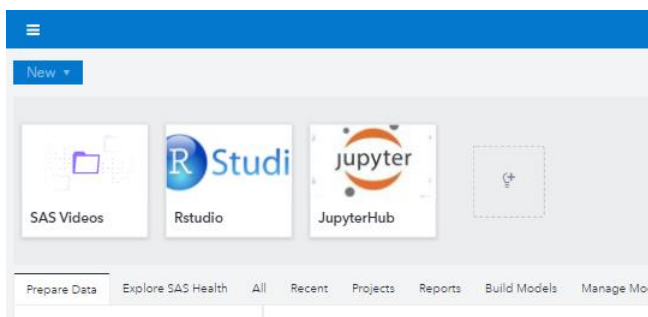
Next, let us compare multiple models built in SAS, R and Python.

Step 1. Open R Studio and Install **RandomForest** package (no need to take action for Python)

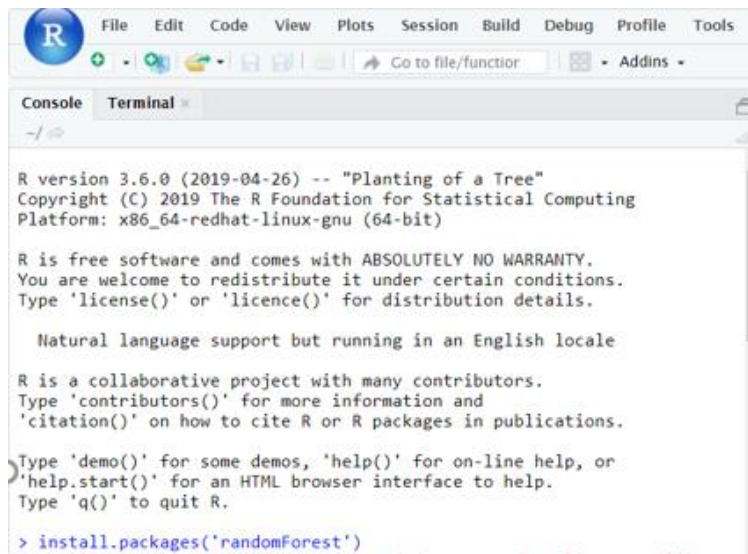
Navigate to the **Share and Collaborate** menu



Select **RStudio** from the Quick Access menu and login with your same credentials.



In R, make sure you **install the randomForest** package using the `install.packages('randomForest')` code. See screenshot below.



```
R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

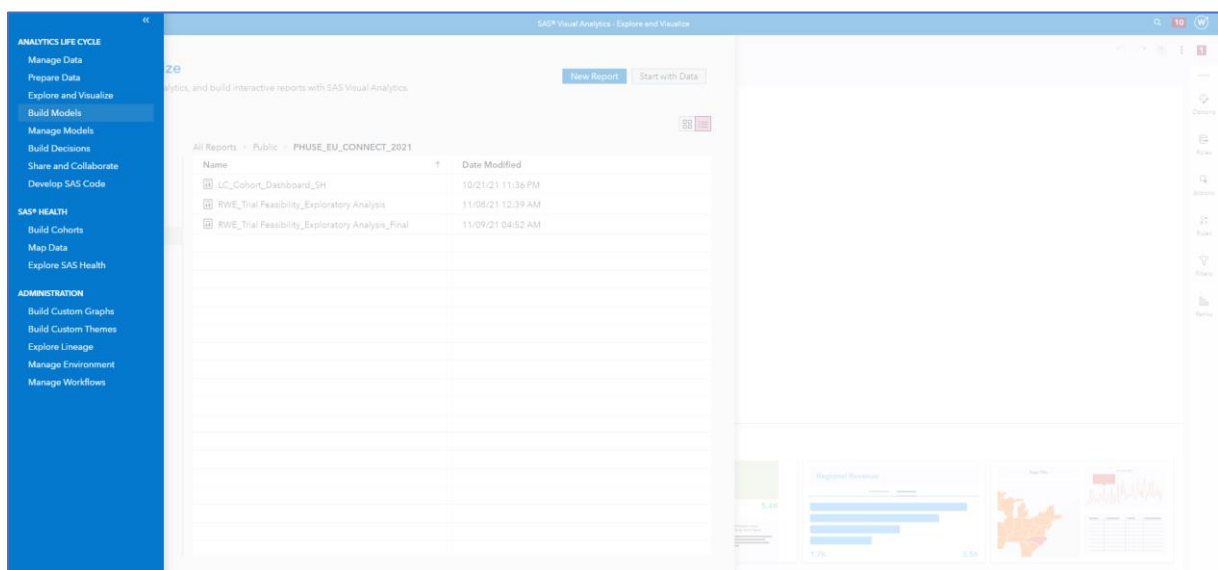
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

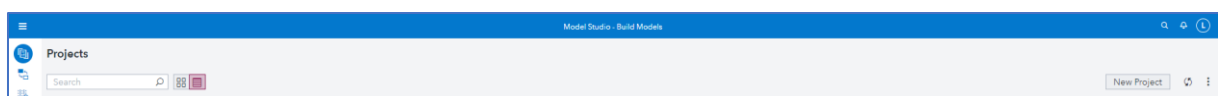
> install.packages('randomForest')
```

Exit R Studio.

Step 2. Navigate to “**Build Models**” to access SAS Model Studio.



Step 3. Hit the New Project button to create a New Model Project.



Step 4. In the New Project window assign a project name.

For the Type of project please keep **Data Mining and Machine Learning**.

Type: *

Data Mining and Machine Learning

Data Mining and Machine Learning

Forecasting

Text Analytics

Data: *

For template option please select **RWE Trial Feasibility PHUSE** and hit the **OK** button.

Browse Templates			
RWE Trial Feasibility PHUSE			
Template Name	Description	Owner	Last Modified
RWE Trial Feasibility PHUSE		Samiul Haque	Nov 17, 2021, 9:17:32 AM

In the data field click on the Browse button and select **ACS_MLTABLE** from available data tables.

New Project

Name: *

Model Pipeline PHUSE (TEST)

Type: *

Data Mining and Machine Learning

Template:

RWE Trial Feasibility PHUSE

Browse

Data: *

Public.ACS_MLTABLE

Browse

Description:

Advanced

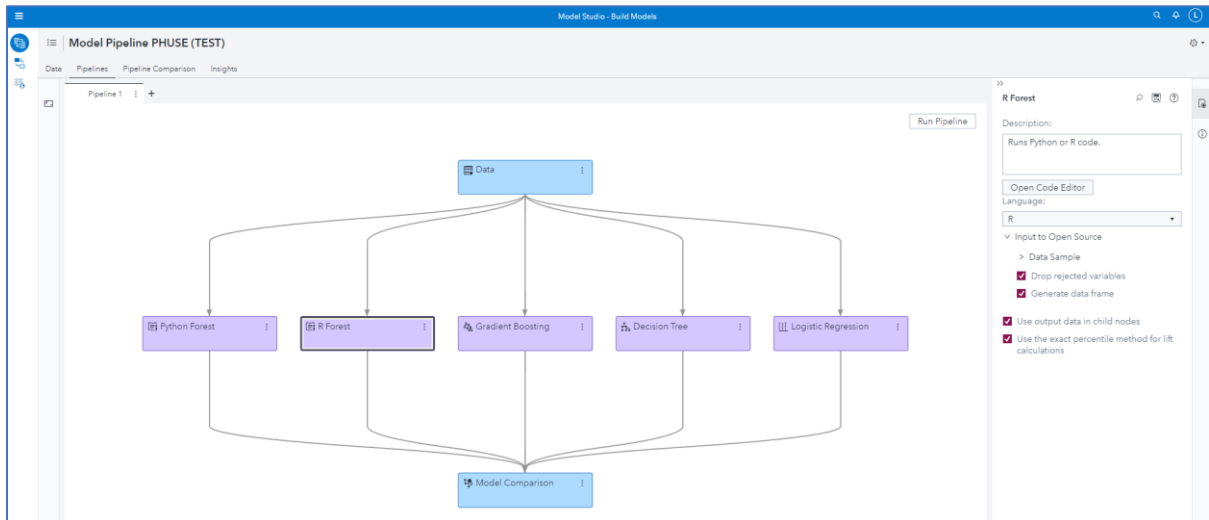
Save

Cancel

Click on **Save**.

Step 5. Navigate to the Pipelines tab to see the model pipeline you created.

Note: This pipeline consists of one **Python** Forest model, one **R** Forest model and 3 **SAS** models (Gradient Boosting, Decision Tree, Logistic Regression).

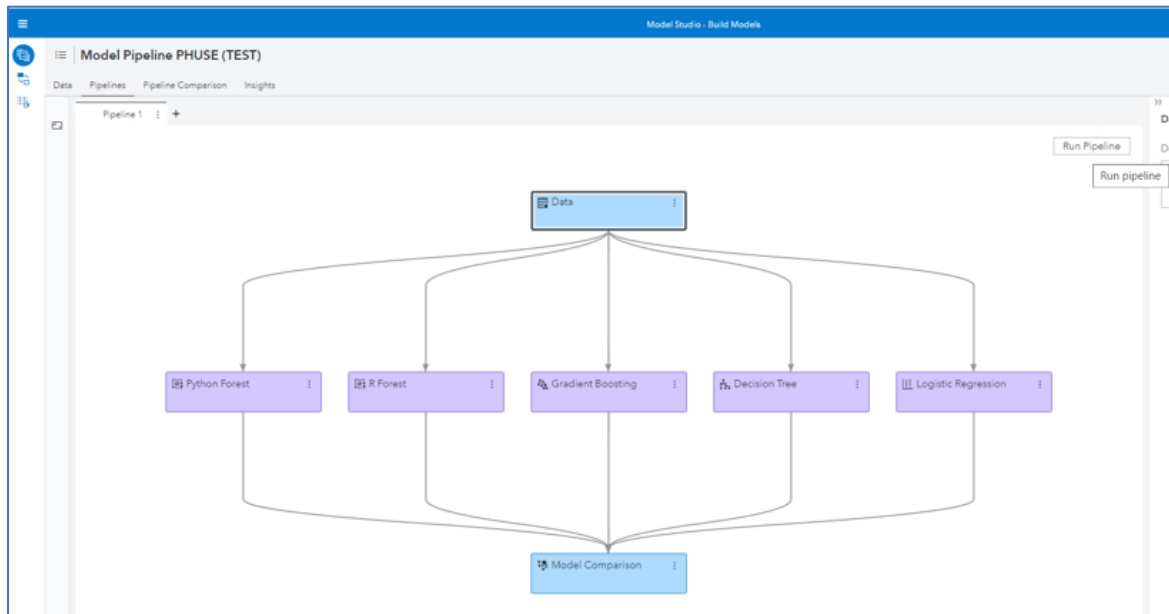


Step 6. You can review the code behind the models.

Click **"R Forest"** model and then click **"Open Code Editor"** on the right-hand side to review the R code. Hit **"Close"** after reviewing the code.

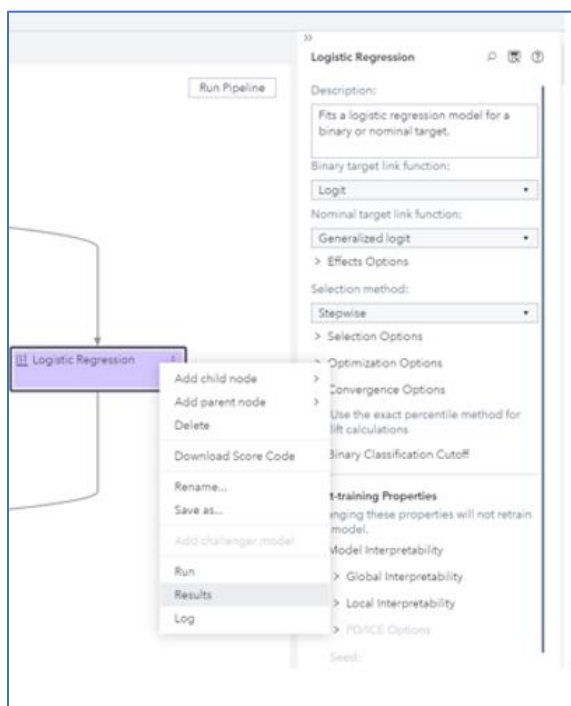
```
1 # Builds Random Forest model with 100 trees
2
3 library(randomForest)
4
5 # Fit RandomForest model w/ training data
6 dm_model <- randomForest(dm_model_formula, ntree=100, data=dm_traindf, importance=TRUE)
7
8 # Save HSE plot to PNG
9 png("rpt_forestHsePlot.png")
10 plot(dm_model, main='randomForest HSE Plot')
11 dev.off()
12
13 # Save VariableImportance to CSV
14 write.csv(importance(dm_model), file="rpt_forestIMP.csv", row.names=TRUE)
15
16 # Score full data
17 dm_scoreddf <- data.frame(predict(dm_model, dm_inputdf, type="prob"))
18 colnames(dm_scoreddf) <- c("P_AV_HI_CC_FL60", "P_AV_HI_CC_FLG1")
```

Step 7. Run the Pipeline (top right corner button).

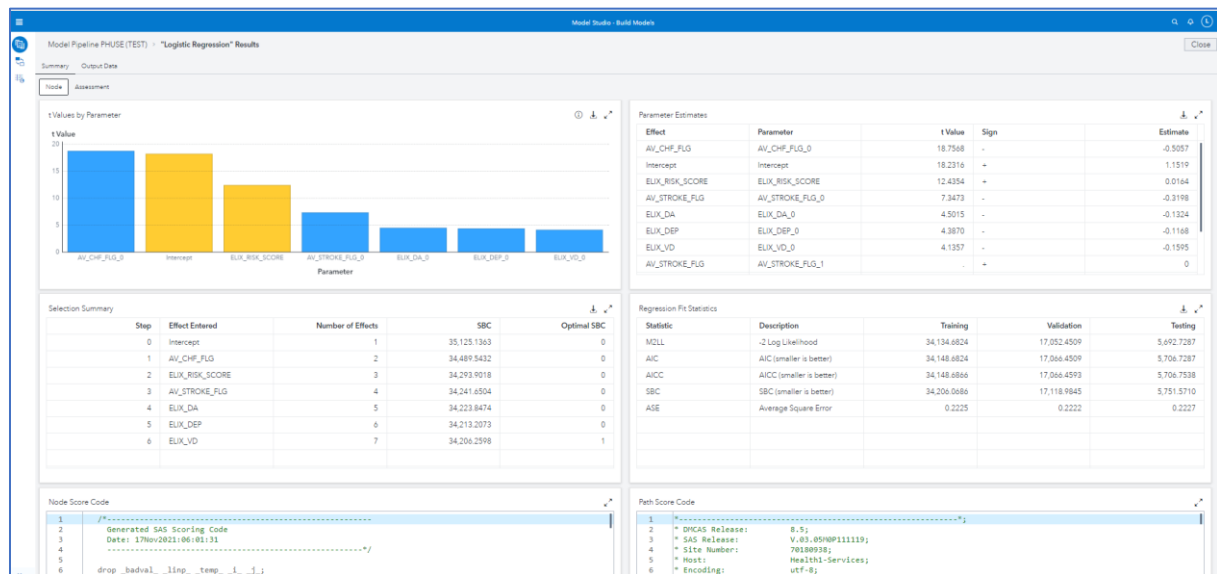


Step 8. You can review the model results from each individual model as well as the model comparison results.

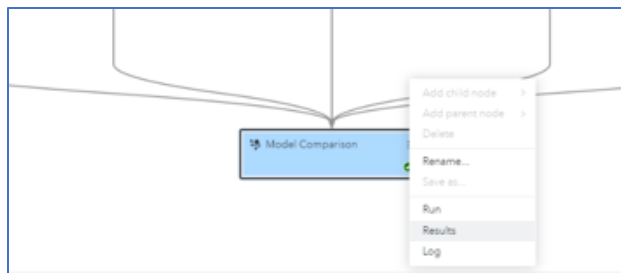
To navigate to a specific model result, please **select the model** and **use the right mouse button** to see the **Results** option (example below for **SAS logistic regression**). You can perform the same action for each individual model.



Results Logistic Regression (see below). Please navigate to the different results sections (nodes tab) as well as all other information (assessment tab). Go back to the pipeline using the Close button on the top right corner.



Step 9. Navigate to the Results of the Model Comparison (SAS, R and Python models) using the same approach (right mouse click – Results).

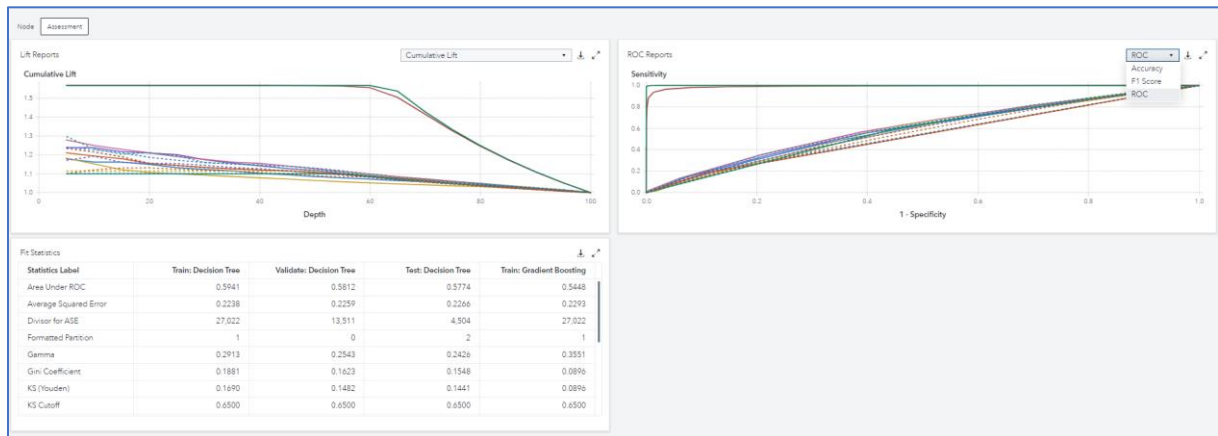


The **Logistic Regression (SAS model)** appears to best performing model.

Champion	Name	Algorithm Name	KS (Youden)	Misclassification Rate
☒	Logistic Regression	Logistic Regression	0.1593	0.3612
☐	Decision Tree	Decision Tree	0.1441	0.3668
☐	R Forest	Open Source Code	0.1368	0.3723
☐	Python Forest	Open Source Code	0.1272	0.3781
☐	Gradient Boosting	Gradient Boosting	0.0757	0.3617

Property Name	Property Value
selectionCriteriaClass	Kolmogorov-Smirnov statistic (KS)
selectionCriteriaInterval	Average squared error
selectionTable	Test
selectionDepth	10
cutoff	0.5000

Please explore the different graphs on the **assessment tab** (example below) to better understand the overall results.



Finally, we will close the workshop with a short Quiz (10 questions).

The TOP 3 participants are receiving some cool gadgets!

Are you Ready?

>> Please go back to the main room if you completed the hands-on session <<