16 DEC 2022

BREAST CANCER DATA ANALYSIS PROJECT

BY NATACHA CHETTY

CST 4704

NEW YORK CITY COLLEGE OF TECHNOLOGY

TABLE OF CONTENTS

Abstract

The dataset that was used for this project was retrieved from the Breast Cancer Surveillance Consortium(BCSC). The BCSC has six ongoing breast imaging registries and two historic registries dedicated to researching the delivery and quality of breast cancer screening and related outcomes in the United States. The three CSV files from the BCSC website were combined to create a dataset with 1522340 rows and 13 columns. The dataset's primary focus is on the risk variables connected to breast cancer diagnosis. Knowing the risk factors that individuals should be wary of would be very beneficial given the rising number of breast cancer cases. This can raise awareness among women who are unaware that they may have breast cancer and encourage them to get a checkup. Therefore, this can increase the likelihood of saving lives since the earlier breast cancer is discovered, the better the chance of starting treatment that can result in a full recovery. The risk factors that are included in the dataset are age, ethnicity, family history of breast cancer, age at menarche, age at first birth, breast density, use of hormone replacement therapy, menopausal status, body mass index, history of biopsy, and history of breast cancer. This report will explore the correlation between the different variables.

Introduction

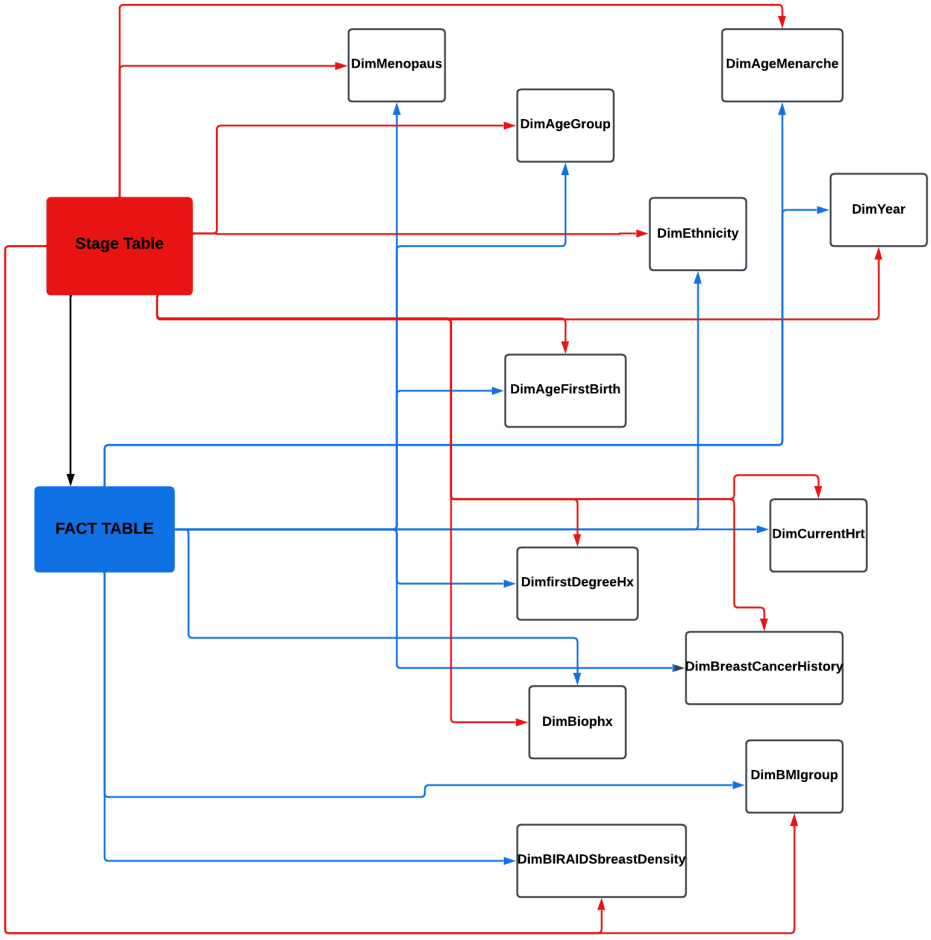
The main objective of this report is to find the most prevalent correlations between certain risk factors in the dataset so as to encourage women who fall in a particular category to get a breast cancer exam.

Those are the key questions/insights that this project is targeting:

1. What is the most common Age Group in the dataset that has breast cancer?
2. Total Number of cases by Ethnicity
3. Total Number of cases by Age Menarche
4. Which Ethnicity has the most breast cancer history?

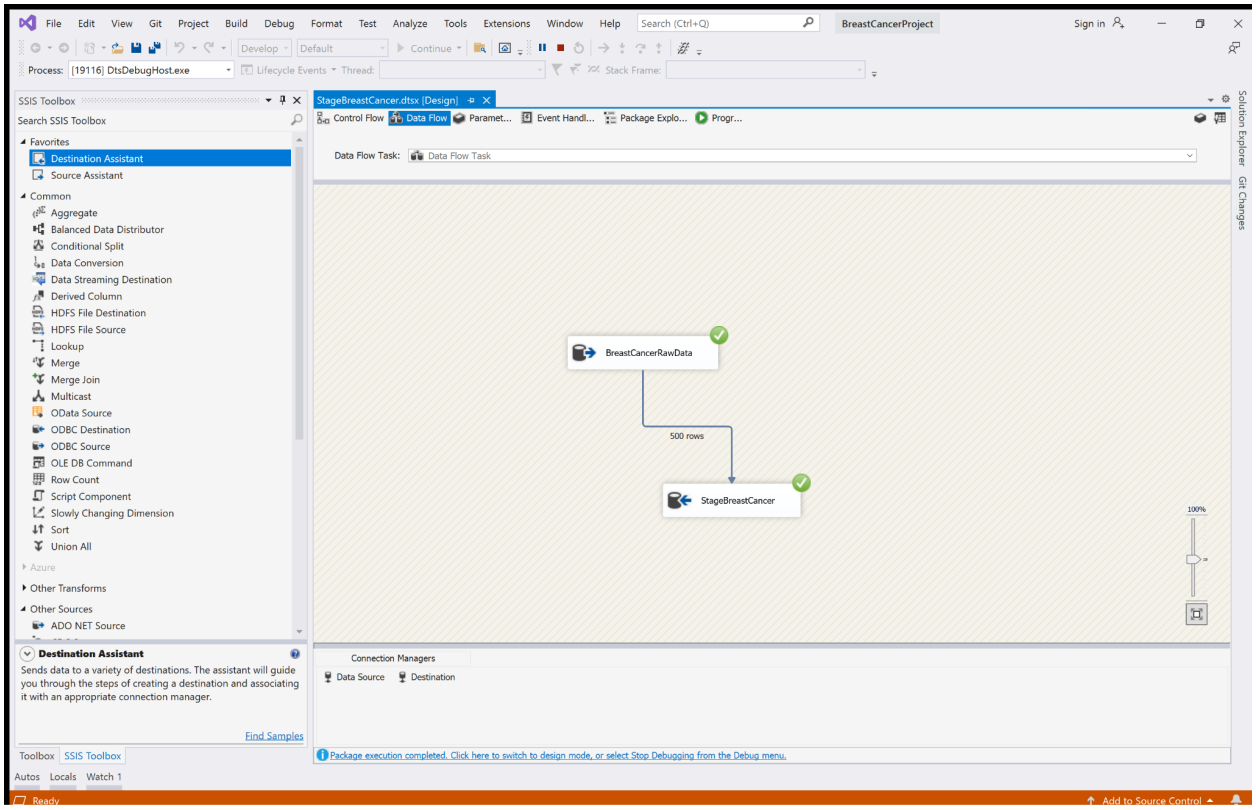
Data Warehouse Architecture and Design

All the risk factors in the dataset represent different dimensions in the data architecture. The best data warehouse architecture for this dataset is a Basic Architecture with Stage area architecture as the only target in the analysis is the number of breast cancer cases.

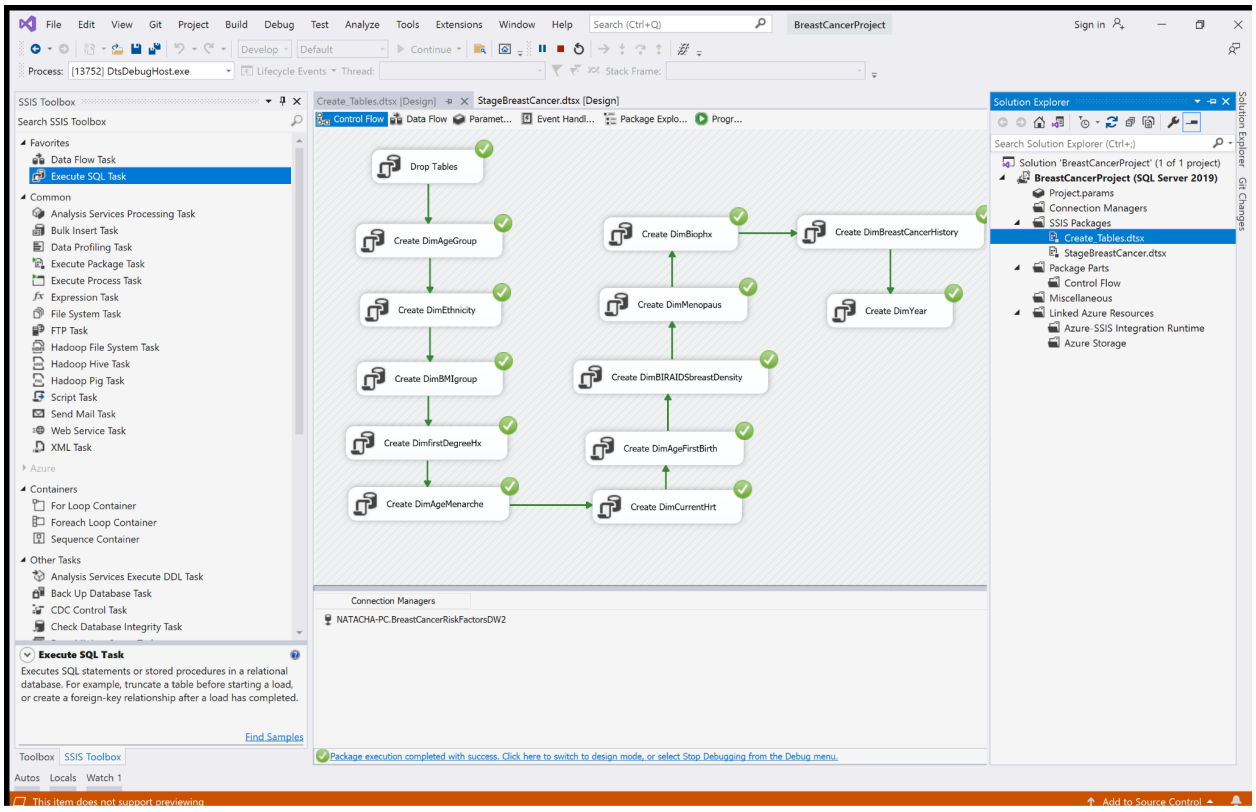


SSIS Design

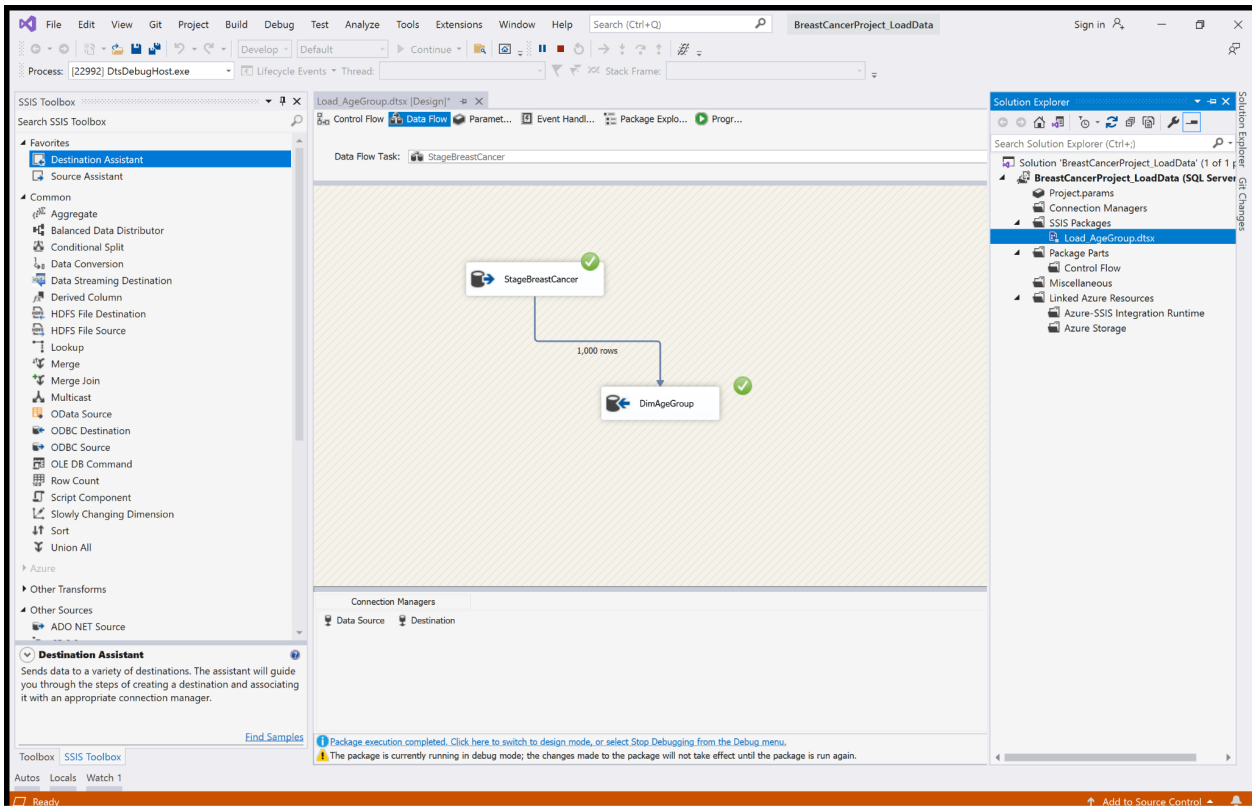
RawData – StageArea:



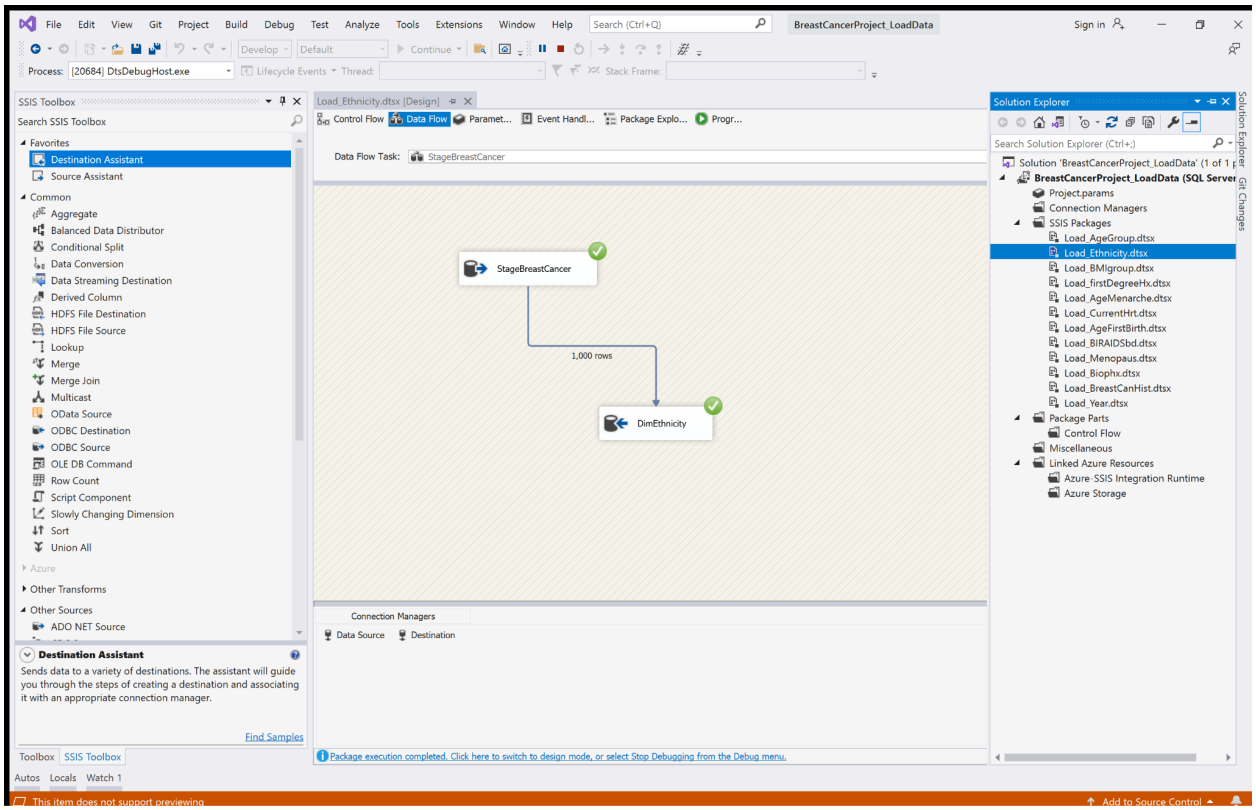
Create Tables:



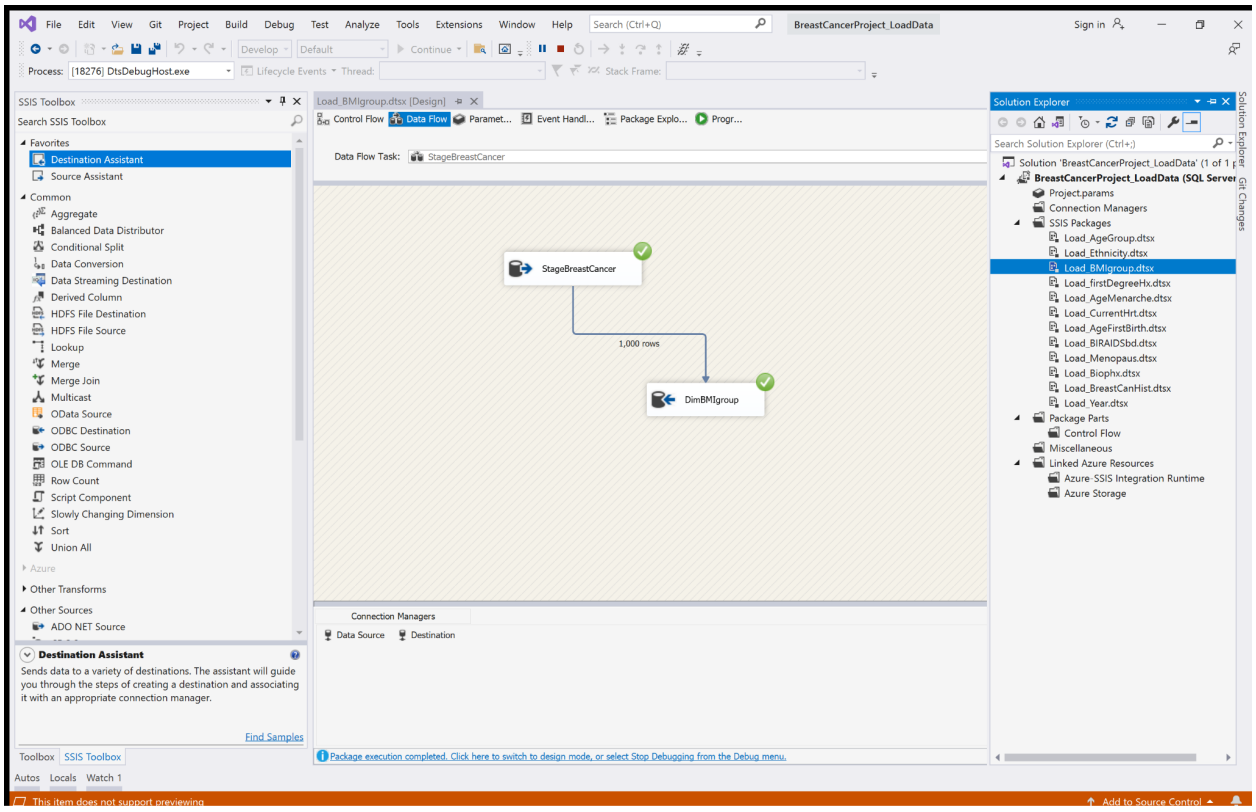
Load DimAgeGroup:



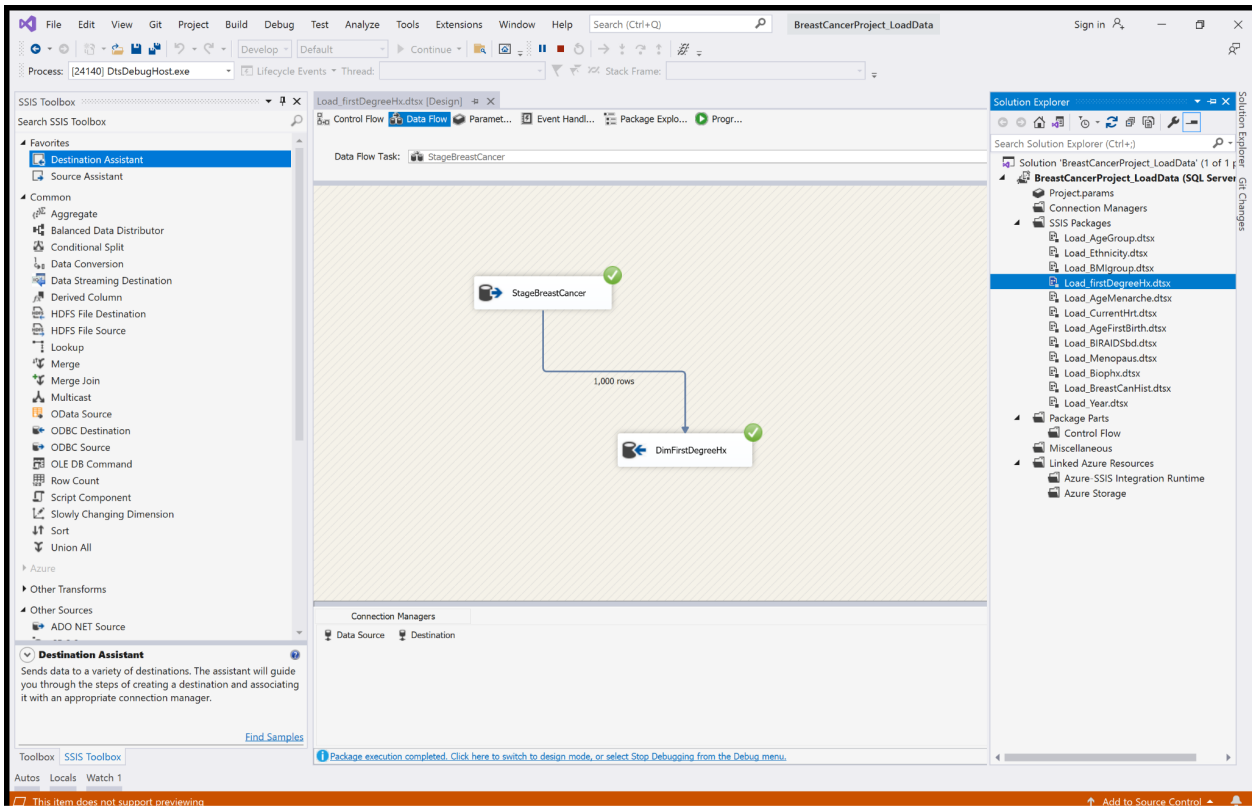
Load DimEthnicity:



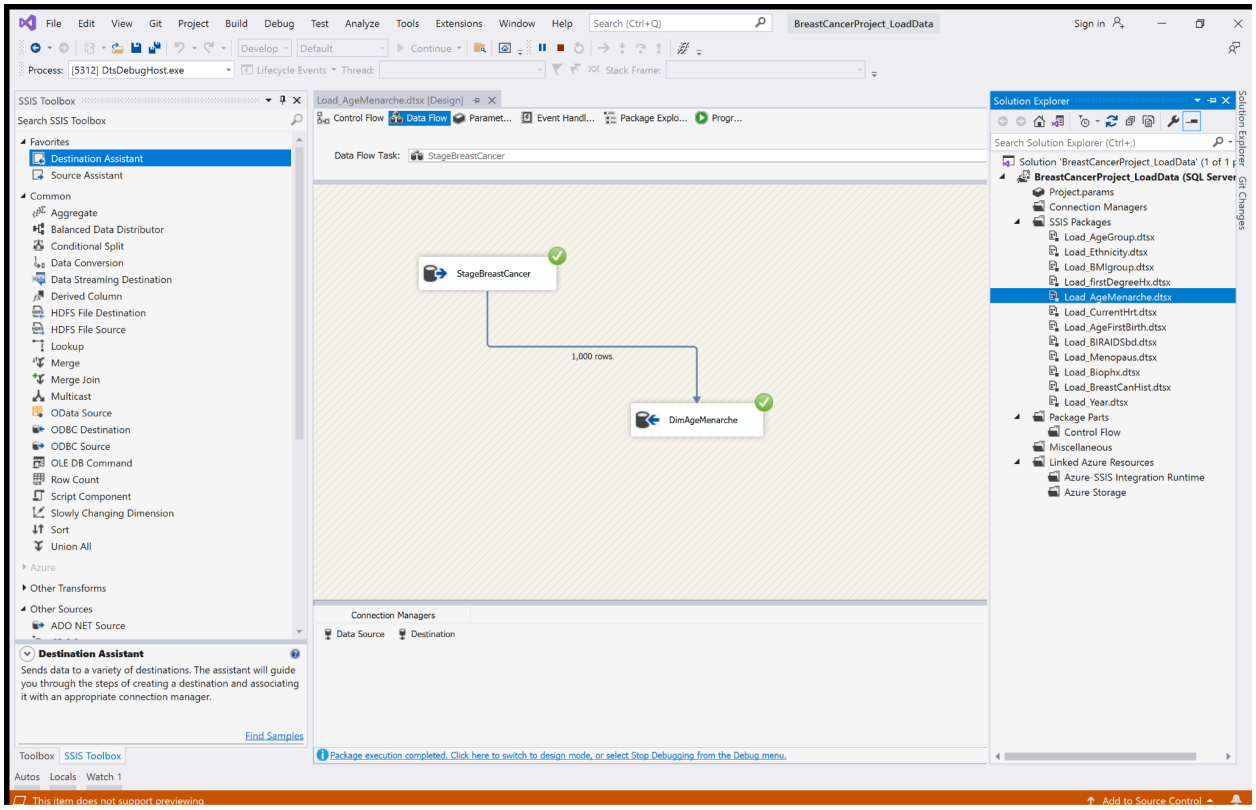
Load DimBMIgroup:



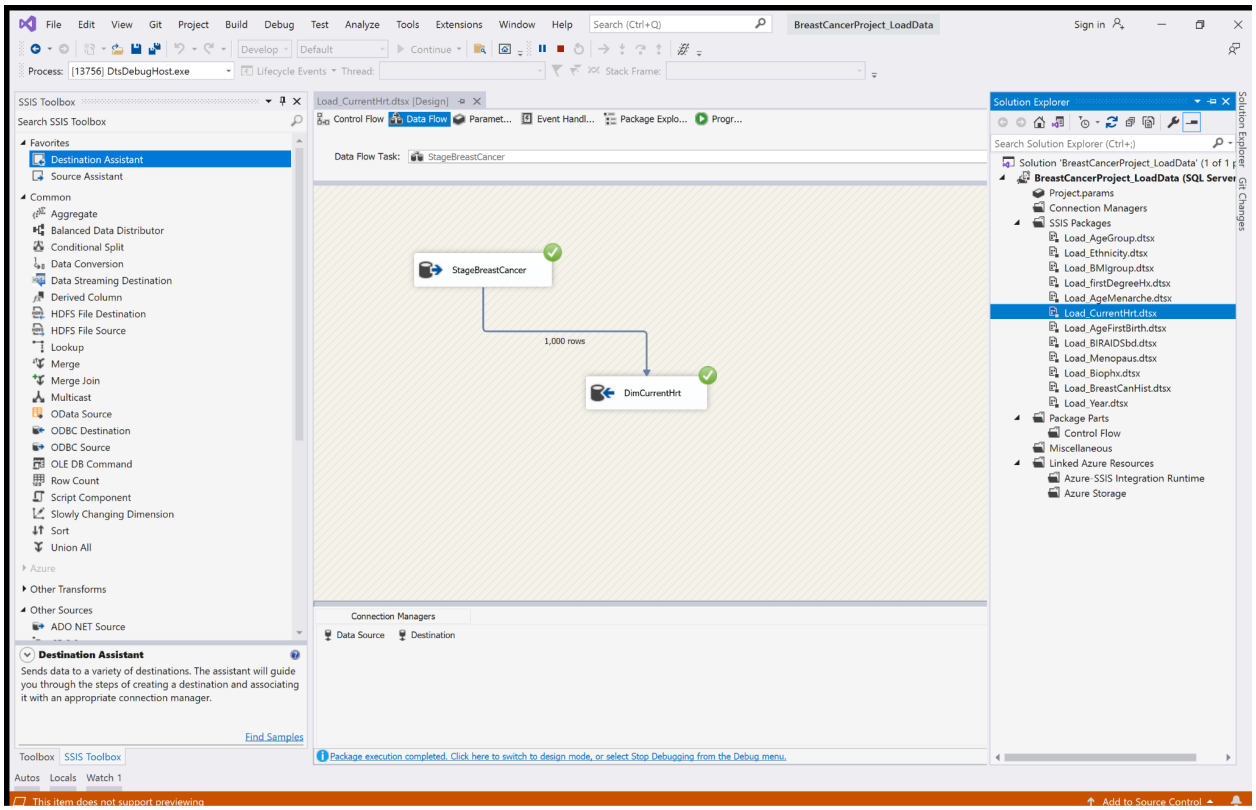
Load DimFirstDegreeHx:



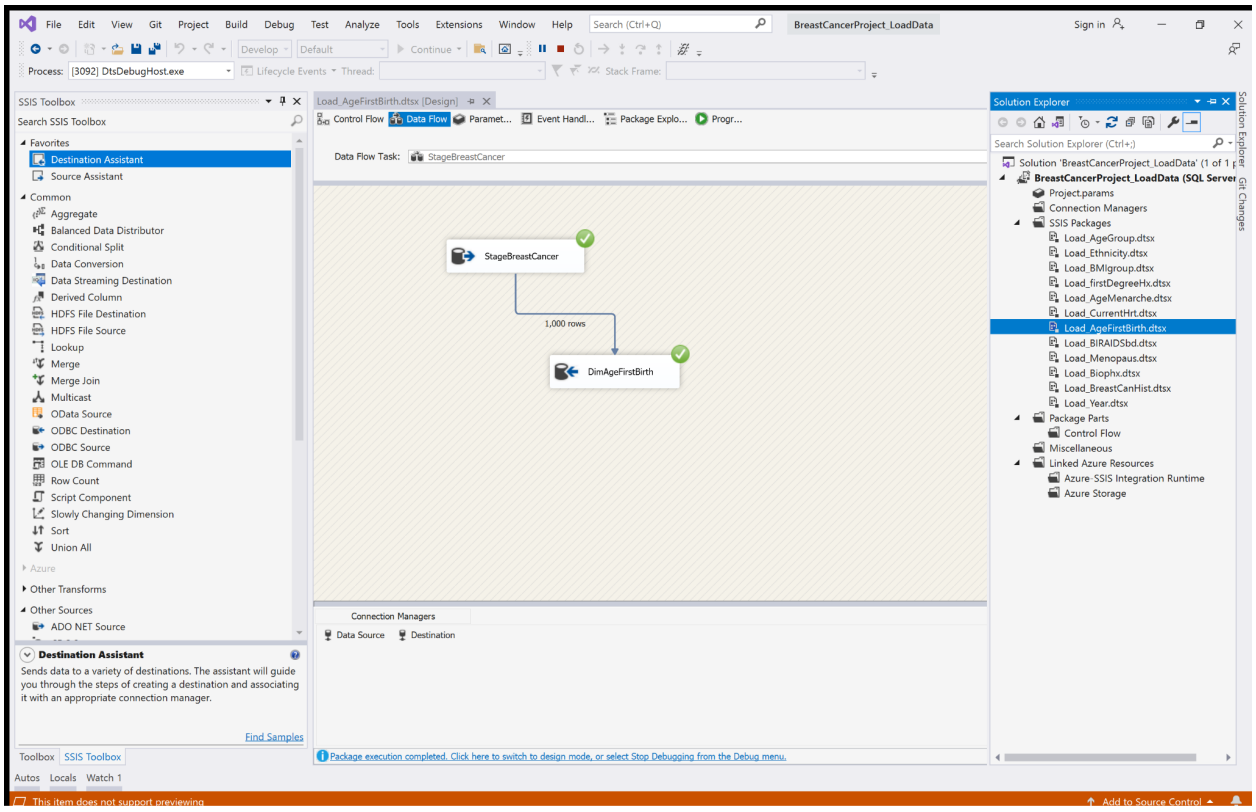
Load DimAgeMenarche:



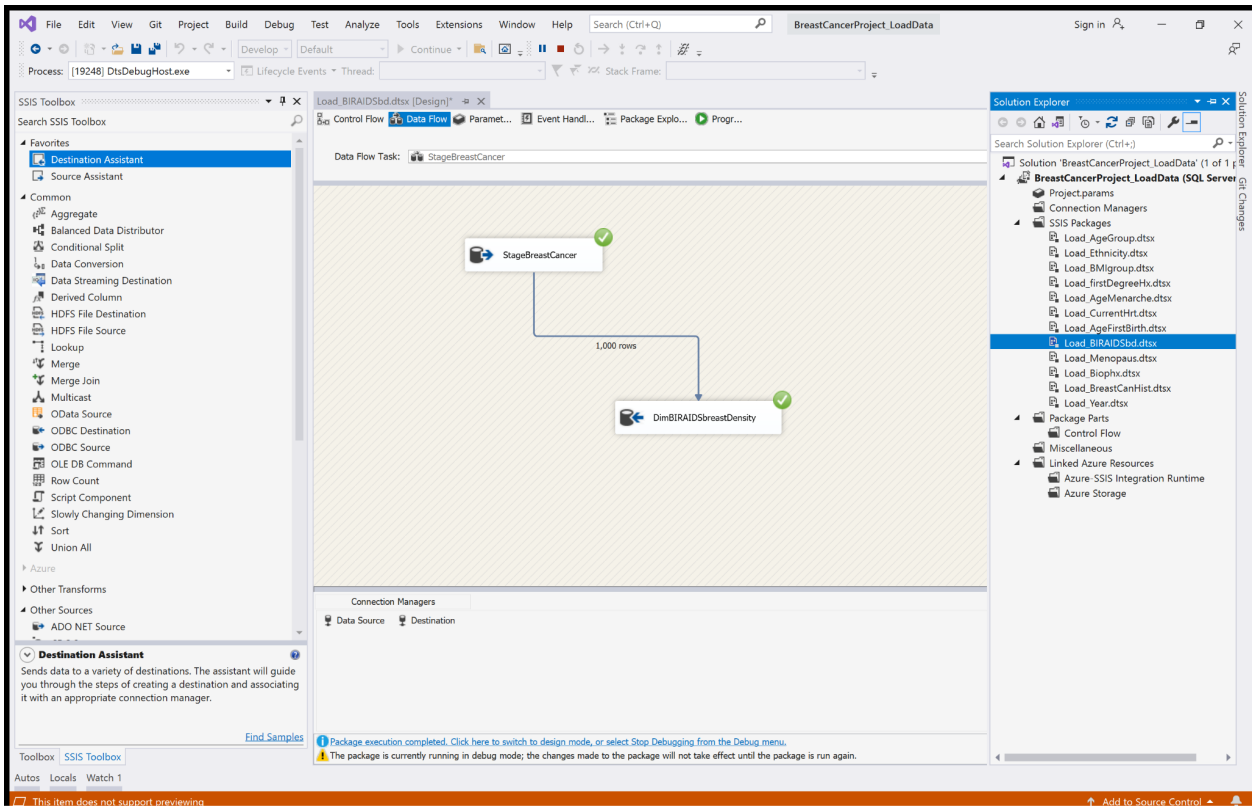
Load DimCurrentHrt:



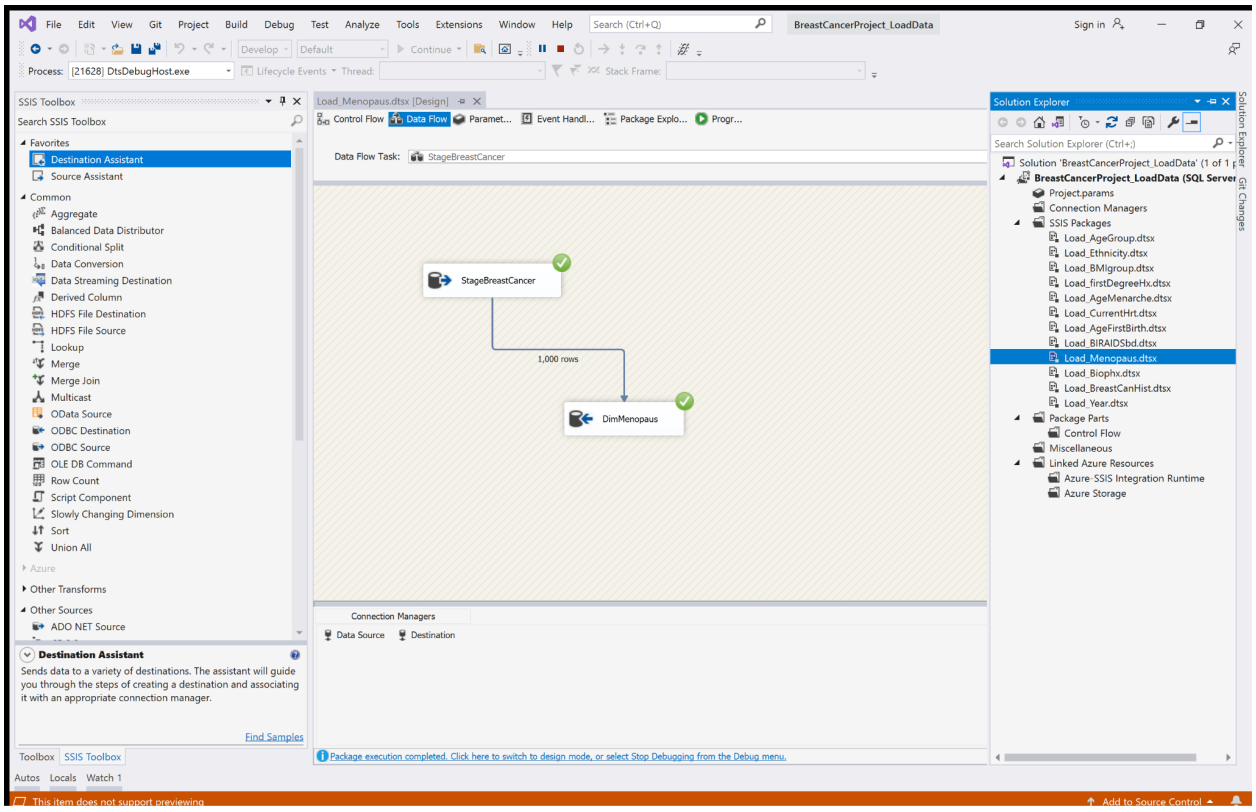
Load DimAgeFirstBirth:



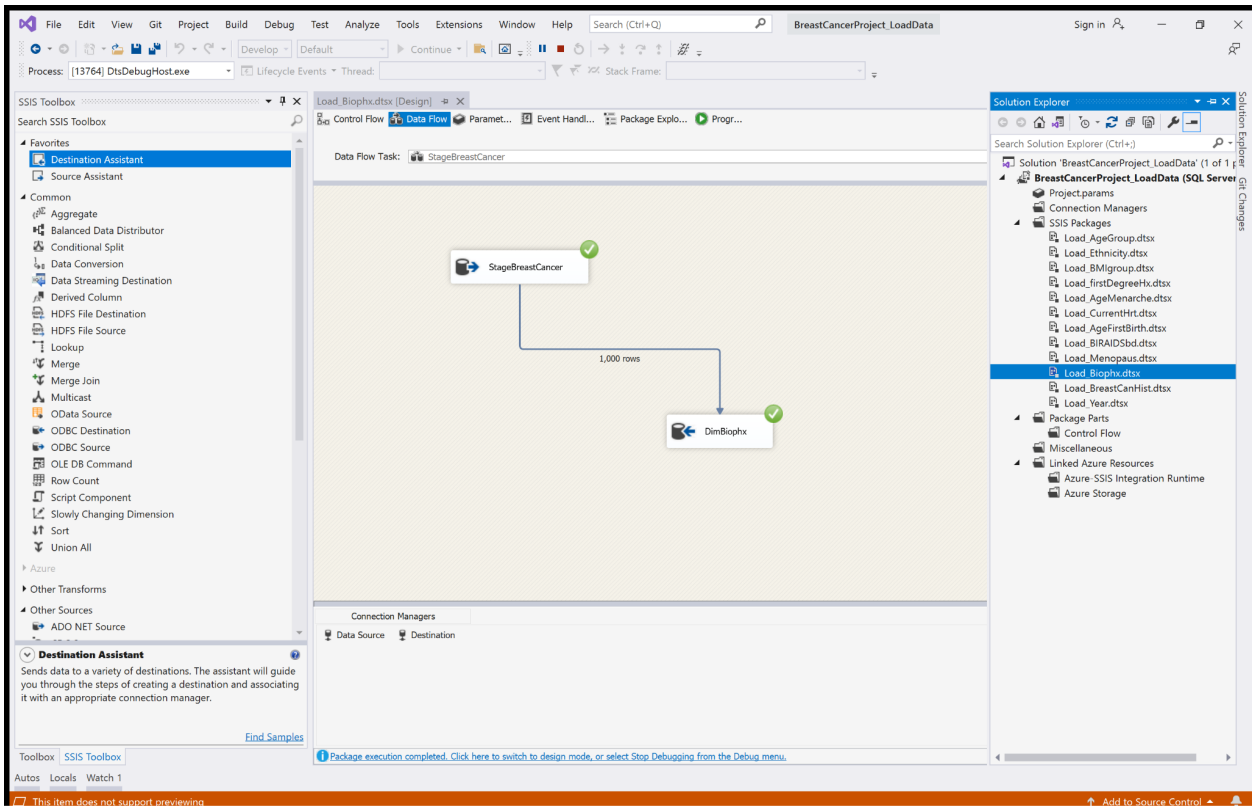
Load BIRAIDSbd:



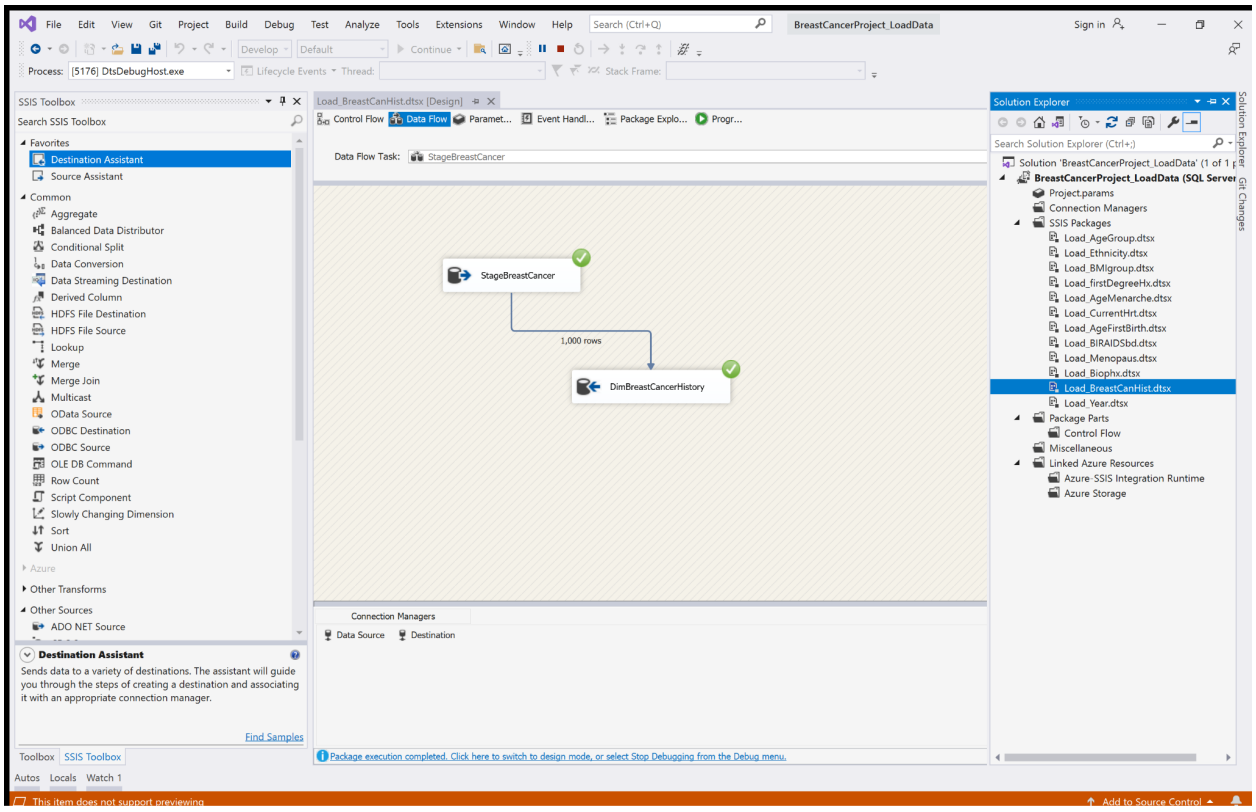
Load Menopaus:



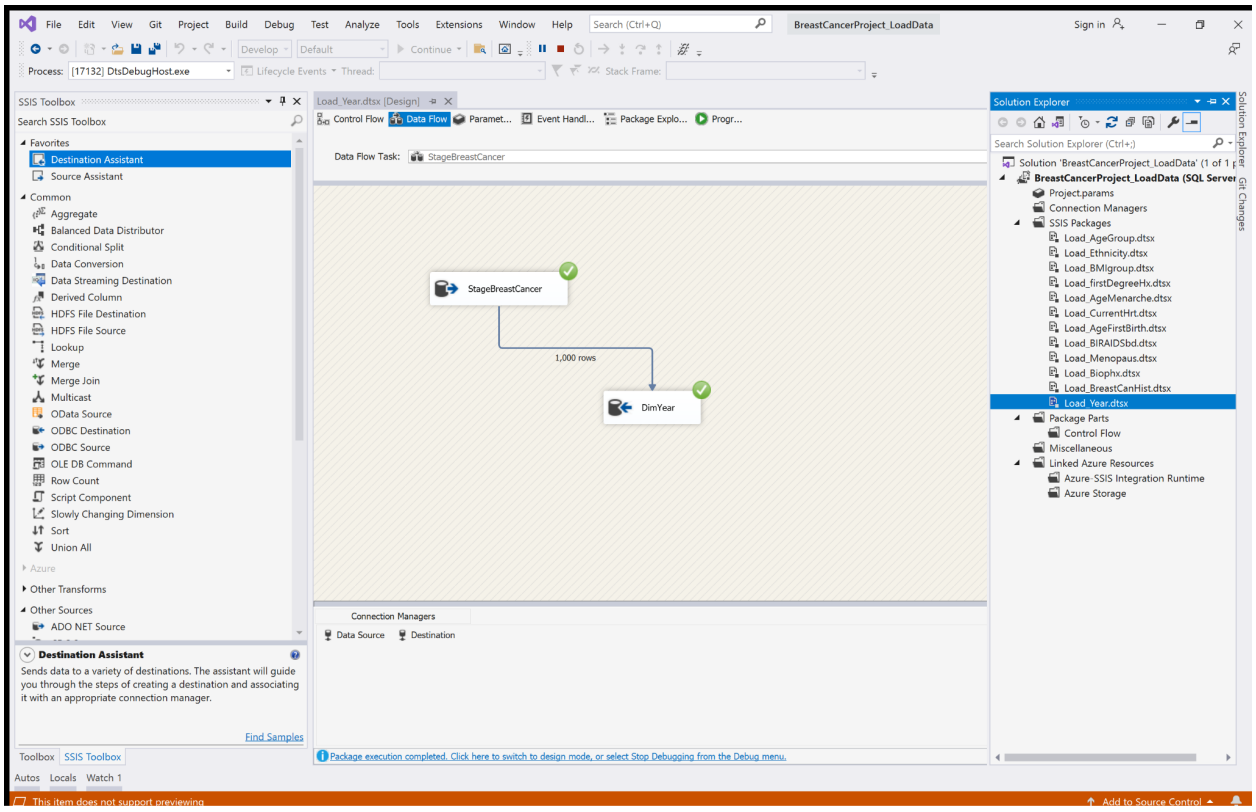
Load DimBiophx:



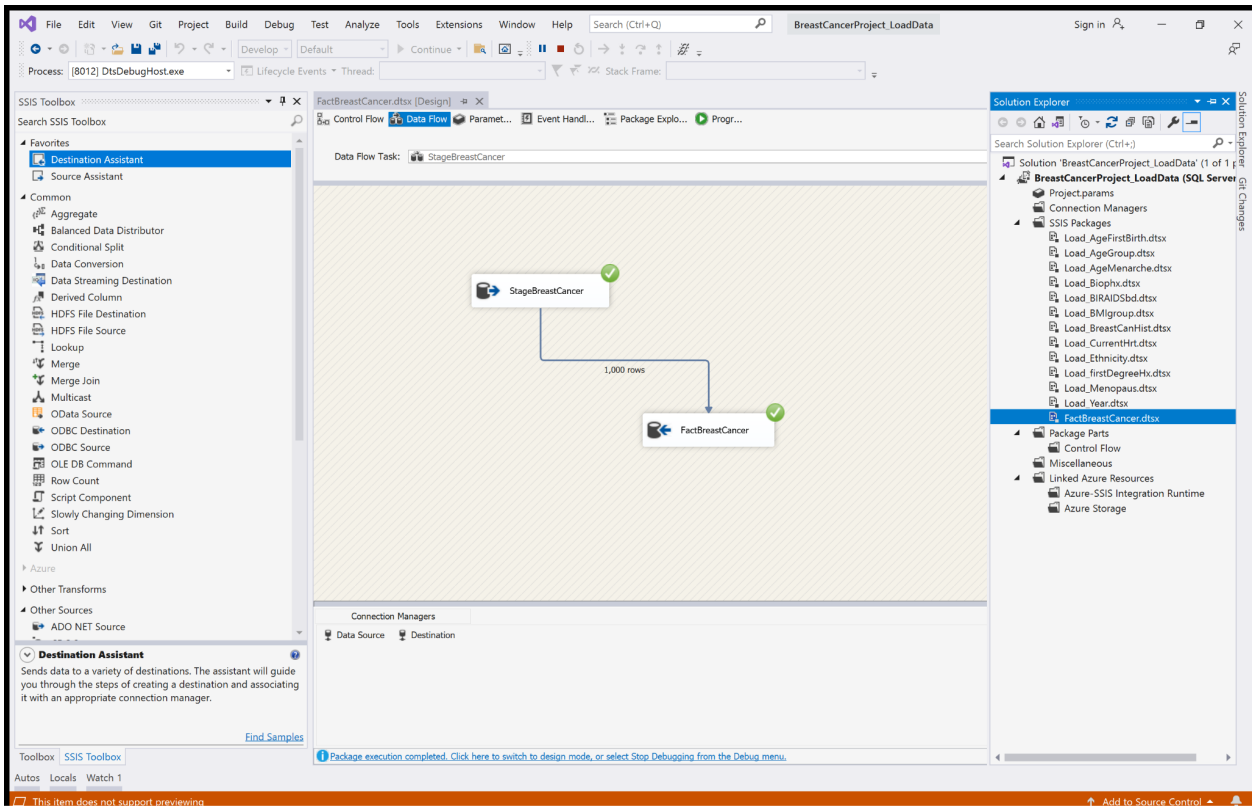
Load DimBreastCancerHistory:



Load DimYear:

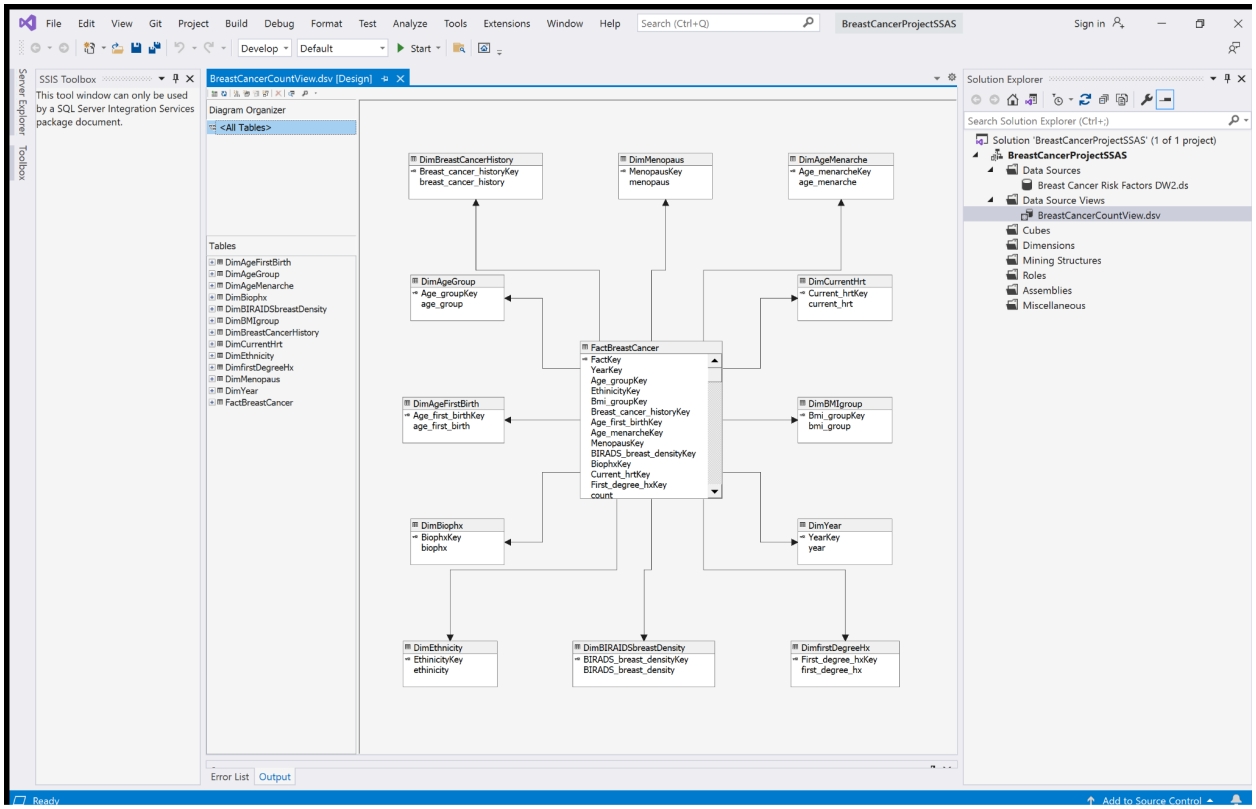


Load FactBreastCancer:

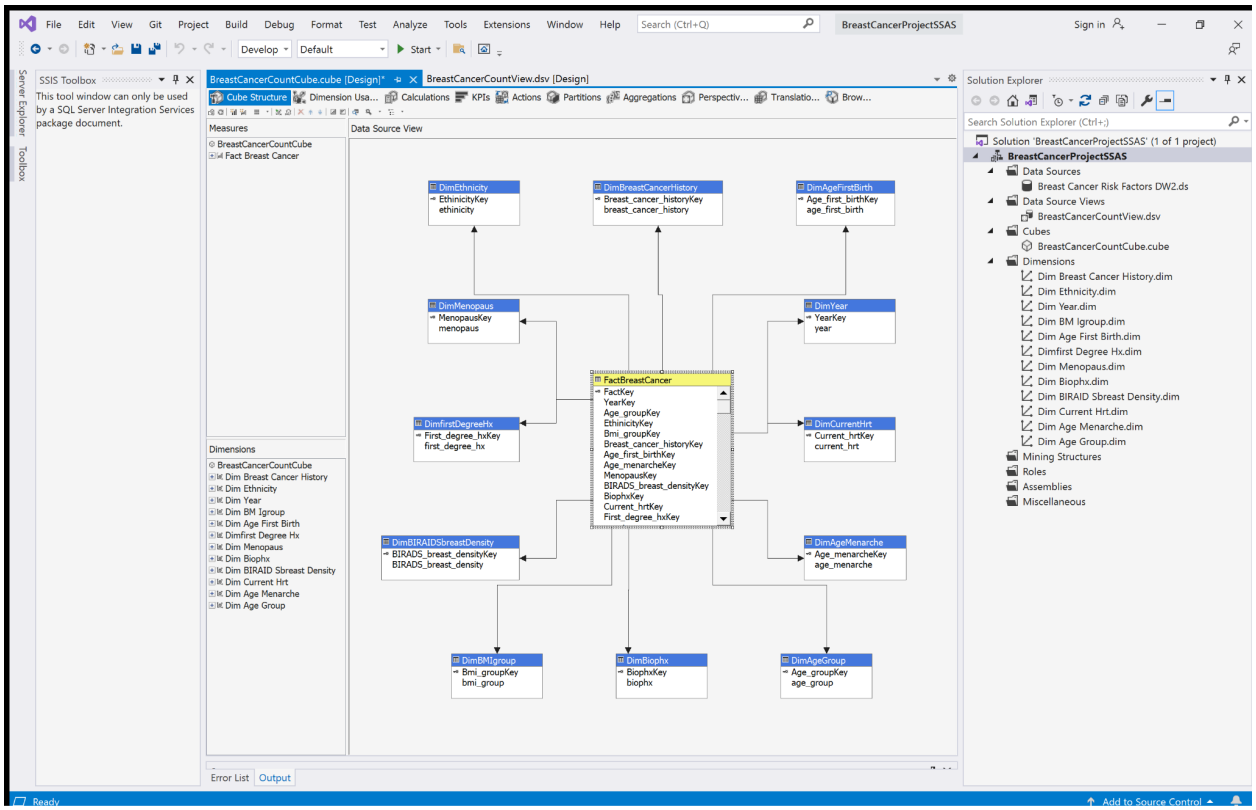


SSAS Design

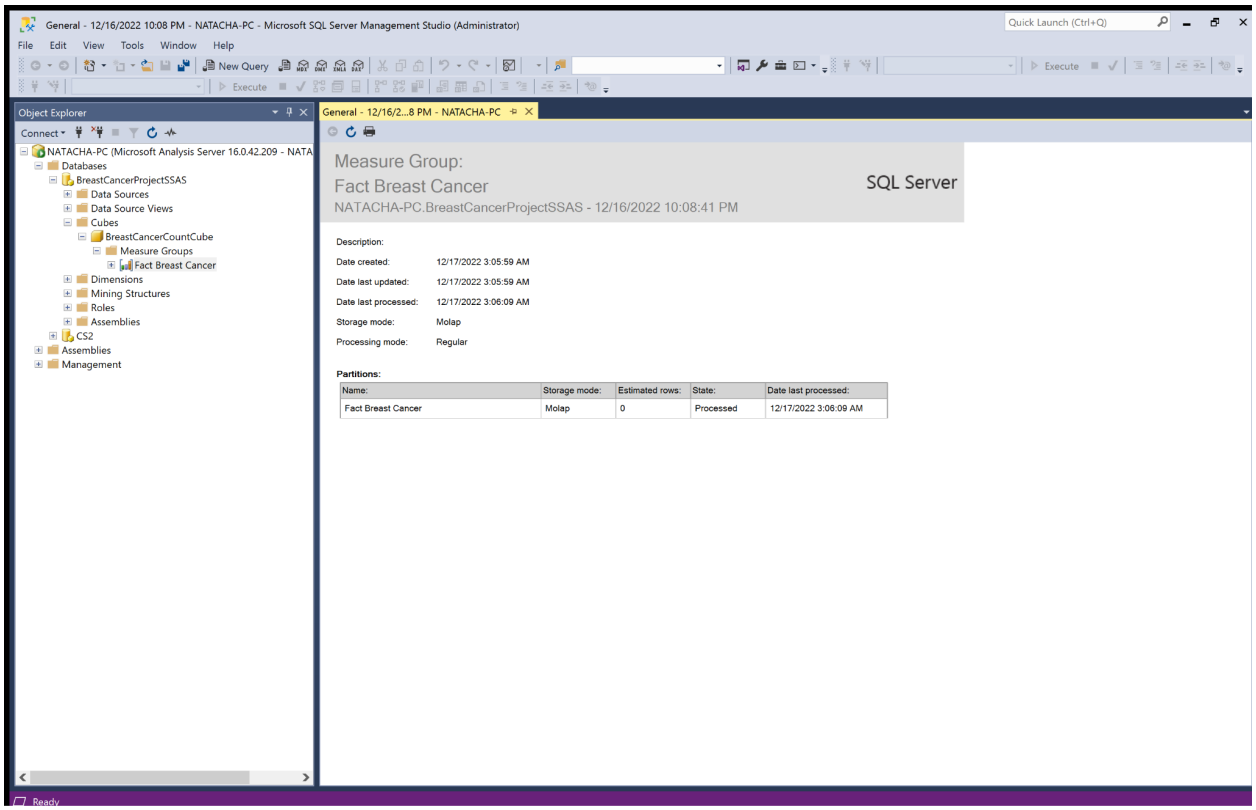
Create View:



Create Cube:

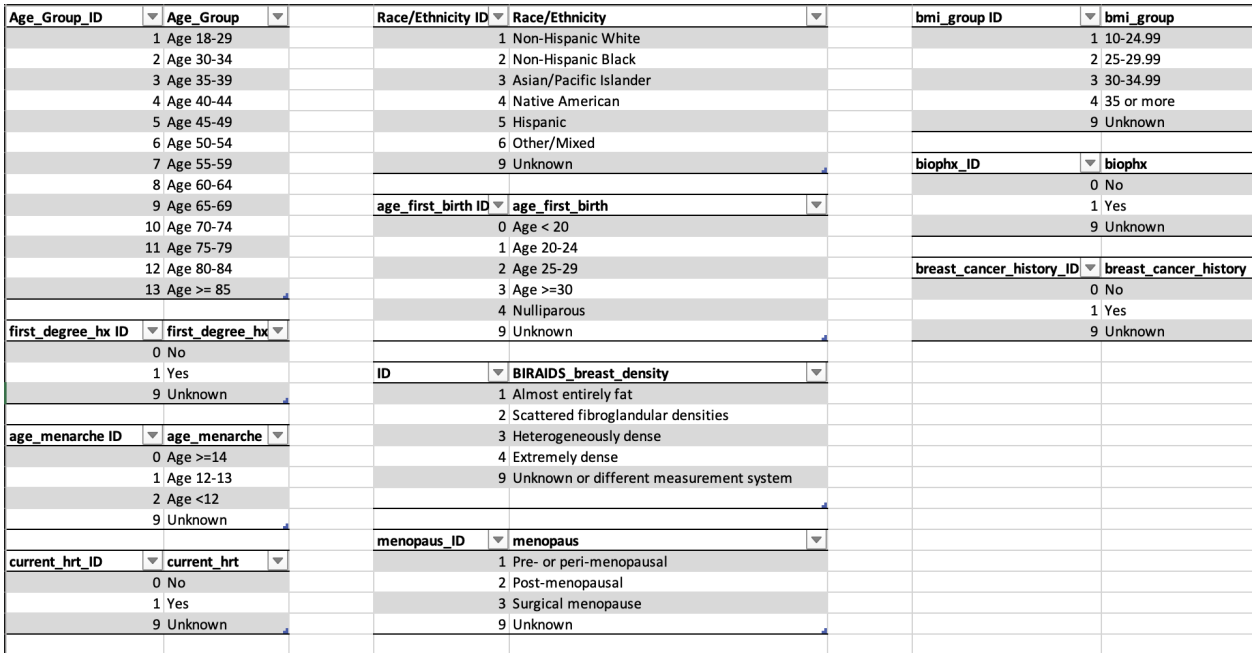

File Edit View Git Project Build Debug Format Test Analyze Tools Extensions Window Help
BreastCancerProjectSSAS
SSIS Toolbox
This tool window can only be used by a SQL Server Integration Services package document.
BreastCancerCountCube.cube [Design]*
BreastCancerCountView.dsv [Design]
Cube Structure
Measures
BreastCancerCountCube
Fact Breast Cancer
Data Source View
Dimensions
Dim Breast Cancer History
Dim Ethnicity
Dim Year
Dim BM Igroup
Dim Age First Birth
Dimfirst Degree Hx
Dim Menopaus
Dim Biophx
Dim BIRAID Sbreast Density
Dim Current Hrt
Dim Age Menarche
Dim Age Group
DimEthnicity
EthinicityKey
ethinicity
DimBreastCancerHistory
Breast_cancer_historyKey
breast_cancer_history
DimAgeFirstBirth
Age_first_birthKey
age_first_birth
DimMenopaus
MenopausKey
menopaus
DimYear
YearKey
year
FactBreastCancer
FactKey
YearKey
Age_groupKey
EthinicityKey
Bmi_groupKey
Breast_cancer_historyKey
Age_first_birthKey
Age_menarcheKey
MenopausKey
BIRADS_breast_densityKey
BiophxKey
Current_hrtKey
First_degree_hxKey
DimfirstDegreeHx
First_degree_hxKey
first_degree_hx
DimCurrentHrt
Current_hrtKey
current_hrt
DimBIRAIDSbreastDensity
BIRADS_breast_densityKey
BIRADS_breast_density
DimAgeMenarche
Age_menarcheKey
age_menarche
DimBMIgroup
Bmi_groupKey
bmi_group
DimBiophx
BiophxKey
biophx
DimAgeGroup
Age_groupKey
age_group
Solution Explorer
Solution 'BreastCancerProjectSSAS' (1 of 1 project)
BreastCancerProjectSSAS
Data Sources
Breast Cancer Risk Factors DW2.ds
Data Source Views
BreastCancerCountView.dsv
Cubes
BreastCancerCountCube.cube
Dimensions
Dim Breast Cancer History.dim
Dim Ethnicity.dim
Dim Year.dim
Dim BM Igroup.dim
Dim Age First Birth.dim
Dimfirst Degree Hx.dim
Dim Menopaus.dim
Dim Biophx.dim
Dim BIRAID Sbreast Density.dim
Dim Current Hrt.dim
Dim Age Menarche.dim
Dim Age Group.dim
Mining Structures
Roles
Assemblies
Miscellaneous
Error List
Output
Ready
Add to Source Control

Report from SQL Management Studio:



Report

The diagram below shows the variables used in the dataset:



Age_Group_ID	Age_Group
1	Age 18-29
2	Age 30-34
3	Age 35-39
4	Age 40-44
5	Age 45-49
6	Age 50-54
7	Age 55-59
8	Age 60-64
9	Age 65-69
10	Age 70-74
11	Age 75-79
12	Age 80-84
13	Age >= 85

first_degree_hx ID	first_degree_hx
0	No
1	Yes
9	Unknown

age_menarche ID	age_menarche
0	Age >=14
1	Age 12-13
2	Age <12
9	Unknown

current_hrt_ID	current_hrt
0	No
1	Yes
9	Unknown

Race/Ethnicity ID	Race/Ethnicity
1	Non-Hispanic White
2	Non-Hispanic Black
3	Asian/Pacific Islander
4	Native American
5	Hispanic
6	Other/Mixed
9	Unknown

age_first_birth ID	age_first_birth
0	Age < 20
1	Age 20-24
2	Age 25-29
3	Age >=30
4	Nulliparous
9	Unknown

ID	BIRAIDS_breast_density
1	Almost entirely fat
2	Scattered fibroglandular densities
3	Heterogeneously dense
4	Extremely dense
9	Unknown or different measurement system

menopaus_ID	menopaus
1	Pre- or peri-menopausal
2	Post-menopausal
3	Surgical menopause
9	Unknown

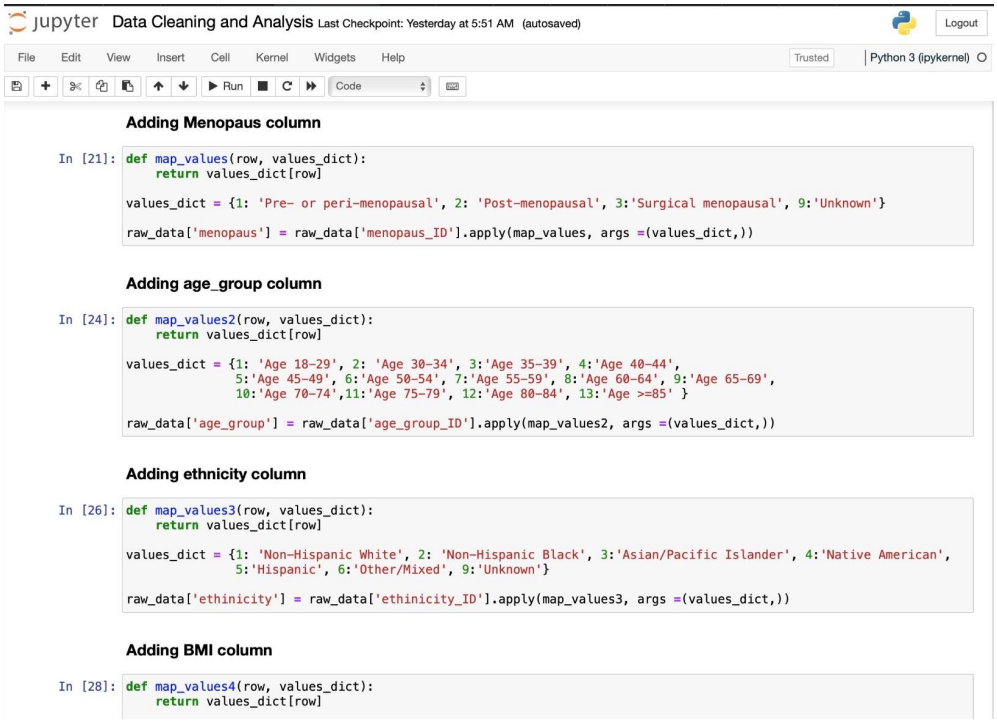
bmi_group ID	bmi_group
1	10-24.99
2	25-29.99
3	30-34.99
4	35 or more
9	Unknown

biophx_ID	biophx
0	No
1	Yes
9	Unknown

breast_cancer_history_ID	breast_cancer_history
0	No
1	Yes
9	Unknown

To be able to analyze the dataset all the numerical variables were matched to their categorical variables.

Jupyter Notebook and Pandas Library were utilized for the data cleaning process. The pictures below show how the new columns were added and saved into a new CSV file.



In [47]:
```
column_names = list(raw_data.columns.values)
print(column_names)
```

```
['year', 'age_group_ID', 'ethinicity_ID', 'first_degree_hx_ID', 'age_menarche_ID', 'age_first_birth_ID', 'BIRADS_breast_density_ID', 'current_hrt_ID', 'menopaus_ID', 'bmi_group_ID', 'biophx_ID', 'breast_cancer_history_ID', 'count', 'menopaus', 'age_group', 'ethinicity', 'bmi_group', 'breast_cancer_history', 'age_first_birth', 'age_menarche', 'BIRADS_breast_density', 'biophx', 'current_hrt', 'first_degree_hx']
```

In [48]:
```
new_raw_data = raw_data[['year', 'age_group', 'ethinicity', 'bmi_group', 'breast_cancer_history',
                         'age_first_birth', 'age_menarche','menopaus','BIRADS_breast_density',
                         'biophx', 'current_hrt', 'first_degree_hx','count']]
```

In [49]:
```
new_raw_data.head()
```

Out[49]:

	year	age_group	ethinicity	bmi_group	breast_cancer_history	age_first_birth	age_menarche	menopaus	BIRADS_breast_density	biophx	current_hrt	first
0	2005	Age 18-29	Non-Hispanic White	Unknown	No	Age < 20	Age >=14	Pre- or peri-menopausal	Scattered fibroglandular densities	No	No	
1	2005	Age 18-29	Non-Hispanic White	10-24.99	No	Age < 20	Age >=14	Pre- or peri-menopausal	Extremely dense	No	No	
2	2005	Age 18-29	Non-Hispanic White	25-29.99	No	Age < 20	Age >=14	Pre- or peri-menopausal	Extremely dense	No	No	
3	2005	Age 18-29	Non-Hispanic White	Unknown	No	Age < 20	Age >=14	Pre- or peri-menopausal	Extremely dense	No	No	
4	2005	Age 18-29	Non-Hispanic White	35 or more	No	Age 20-24	Age >=14	Pre- or peri-menopausal	Extremely dense	No	No	

Save new CSV files without the old columns

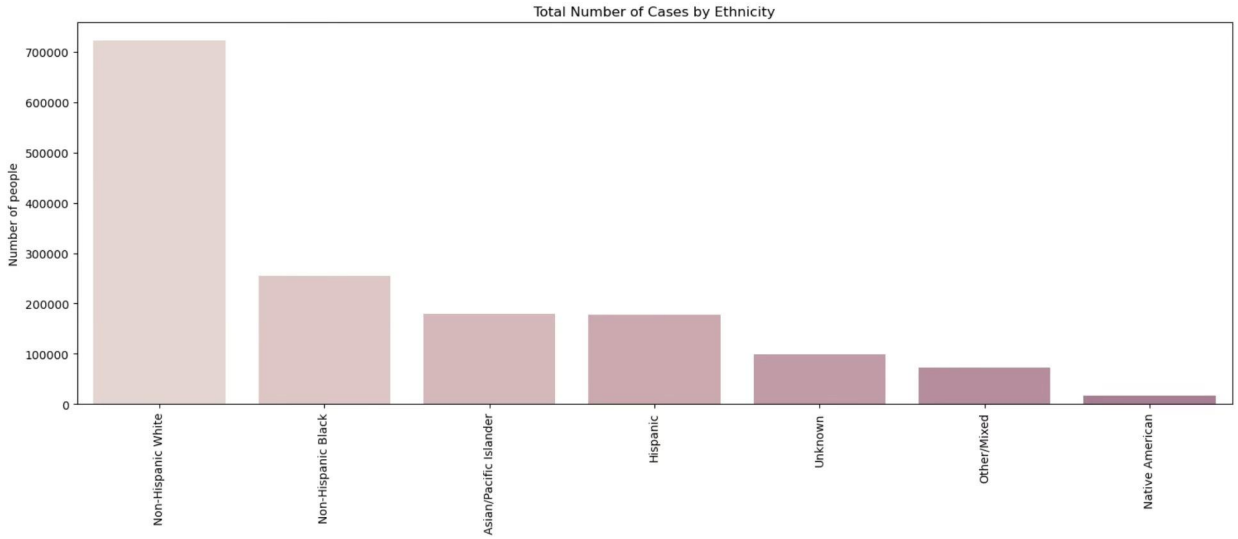
In [50]:
```
new_raw_data.to_csv("new_raw_data.csv", index=False)
```

This is the **Number of Breast Cancer cases by Ethnicity** available in the data set:

Non-Hispanic White	721859
Non-Hispanic Black	255476
Asian/Pacific Islander	179926
Hispanic	177069
Unknown	98329

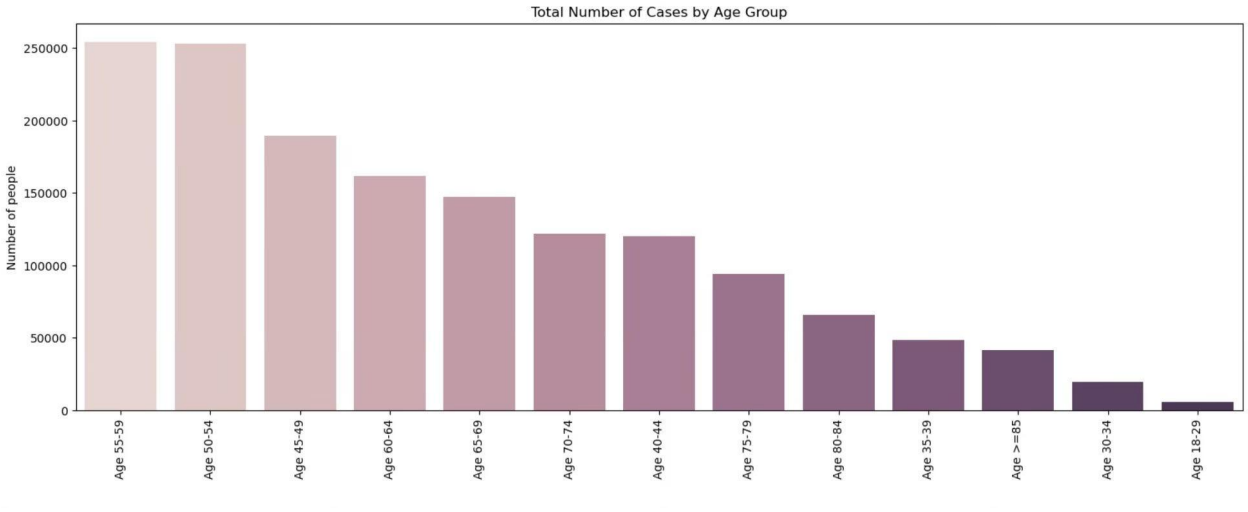
Data visualization libraries, **Matplotlib** and **Seaborn**, were utilized to plot the charts.

Bar Chart showcasing the **Number of Breast Cancer Cases by Ethnicity**:

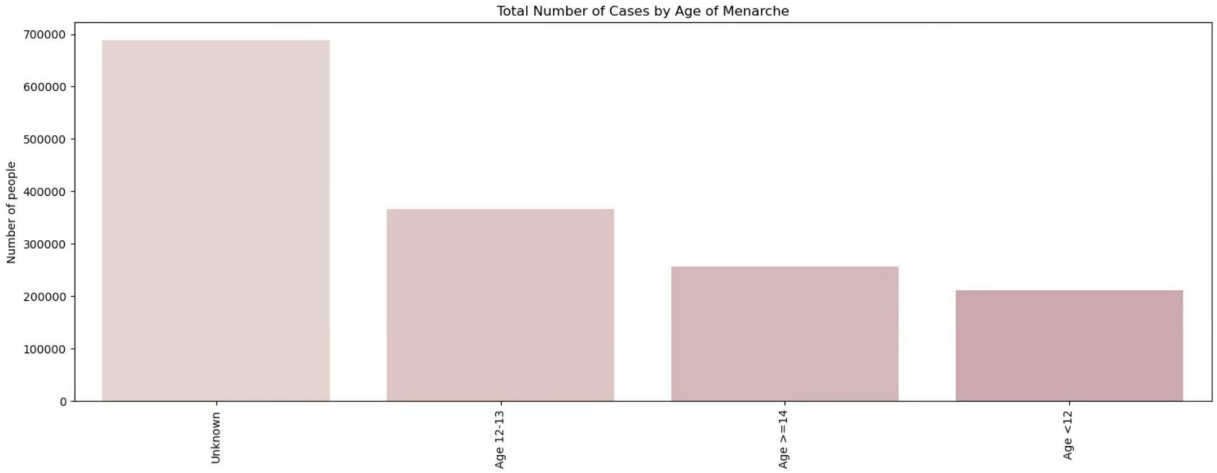


Based on the table and the chart it can be observed that Non-Hispanic White is the ethnicity that has the highest number of breast cancer patients.

This Bar Chart shows the **Number of Breast Cancer cases by Age Group**:



This Bar Chart shows the **Number of Breast Cancer cases by Age of Menarche**:



This chart illustrates that the frequency of women who had breast cancer and had their first menstruation between the age of 12-13 were higher in the mammograms taken during 2005-2017.

This table showcases **the Differences in Breast Cancer History instances between the different Ethnicities**:

ethinicity	breast_cancer_history	
Asian/Pacific Islander	**No**	119796
	Unknown	33778
	Yes	26352
Hispanic	**No**	128408
	Unknown	32666
	Yes	15995
Native American	**No**	12968
	Unknown	2264
	Yes	1273
Non-Hispanic Black	**No**	146469
	Unknown	78581
	Yes	30426
Non-Hispanic White	**No**	441436
	Unknown	149357
	Yes	131066
Other/Mixed	**No**	55875
	Unknown	10626
	Yes	6675
Unknown	**No**	64710
	Unknown	23630
	Yes	9989

Non-Hispanic White has the **highest rate** of **Breast Cancer History (previous diagnosis)** compared to the other ethnicities.

Conclusion

This dataset and topic includes a lot of intriguing features, and with additional time, more insights can be discovered. The following are some interesting questions to consider:

- What are the risk factors that were the most common?
- What are the risk factors that are most common by ethnicity?
- Which state has the population with the higher breast cancer rates?

In the future, I would like to find more interesting datasets that are related to breast cancer diagnosis to have a more detailed analysis.

References

"Data collection and sharing was supported by the National Cancer Institute (P01CA154292; U54CA163303), the Patient-Centered Outcomes Research Institute (PCS-1504-30370), and the Agency for Health Research and Quality (R01 HS018366-01A1). We thank the participating women, mammography facilities, and radiologists for the data they have provided for this study.
You can learn more about the BCSC at: http://www.bcsc-research.org/."

SQL Server Management Studio 2022 [Computer software]. (2022). Retrieved from https://learn.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver16

Visual Studio 2019 [Computer Software]. (2019). Retrieved from https://visualstudio.microsoft.com/downloads/

Jupyter Notebook [Computer Software]. Retrieved from https://jupyter.org/install

Matplotlib [Data Visualization library]. Retrieved from https://matplotlib.org/stable/index.html

Pandas [Data Analysis library]. Retrieved from https://pandas.pydata.org/

Numpy Library [Python Library]. Retrieved from https://numpy.org/install/

Seaborn Library [Data Visualization library]. Retrieved from https://seaborn.pydata.org/installing.html