

**UNIVERSITY OF INFORMATION TECHNOLOGY,  
VNUHCM**

**FACULTY OF INFORMATION SYSTEMS**



## **Final project**

**Course:** MSIS4263.P21.CTTT – Decision support and business intelligence application

**Lecturer:** Dr Do Phuc

**TA:** MsC. Nguyen Thi Kim Phung

**Group:**

- |                   |          |        |
|-------------------|----------|--------|
| 1. Lý Tuấn Khoa   | 21522225 | Member |
| 2. Nguyễn Gia Huy | 21522152 | Member |

## **TABLE OF CONTENT**

Chapter 1: SSIS process .....	3
Chapter 2: Analysis and Reporting Process .....	16
Chapter 3: Data mining.....	44

# Chapter 1: SSIS process

## 1.1 Data

**Link data:** <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>

This report utilizes the Telcom Customer Churn dataset to examine customer behavior associated with service cancellation (churn) in the telecommunications industry. Each row in the dataset represents a customer, with 21 features describing attributes such as gender, seniority, and so on.

The primary objective of this project is to determine the key factors that influence whether a customer will leave the service. Understanding these factors can help the company improve service quality and customer retention. The target variable in this dataset is “**churn**”, which indicates whether a customer has left the service.

<b>CustomerId</b>	<b>INT</b>
<b>Gender</b>	<b>NVARCHAR</b>
<b>SenoirCitizen</b>	<b>INT</b>
<b>Dependents</b>	<b>NVARCHAR</b>
<b>Tenure</b>	<b>INT</b>
<b>PhoneService</b>	<b>NVARCHAR</b>
<b>MultipleLines</b>	<b>NVARCHAR</b>
<b>InternetService</b>	<b>NVARCHAR</b>
<b>OnlineSecurity</b>	<b>NVARCHAR</b>
<b>OnlineBackup</b>	<b>NVARCHAR</b>
<b>DeviceProtection</b>	<b>NVARCHAR</b>
<b>TechSupport</b>	<b>NVARCHAR</b>
<b>StreamingTV</b>	<b>NVARCHAR</b>
<b>StreamingMovies</b>	<b>NVARCHAR</b>
<b>Contract</b>	<b>NVARCHAR</b>
<b>PaperlessBilling</b>	<b>NVARCHAR</b>
<b>PaymentMethod</b>	<b>NVARCHAR</b>
<b>MonthlyCharges</b>	<b>FLOAT</b>
<b>TotalCharges</b>	<b>FLOAT</b>
<b>Churn</b>	<b>NVARCHAR</b>
<b>ServiceID</b>	<b>INT</b>
<b>BillingID</b>	<b>INT</b>
<b>ContractID</b>	<b>INT</b>
<b>id</b>	<b>INT</b>

## 1.2 Process of building SSIS

Step 1: Pulling the Excel icon

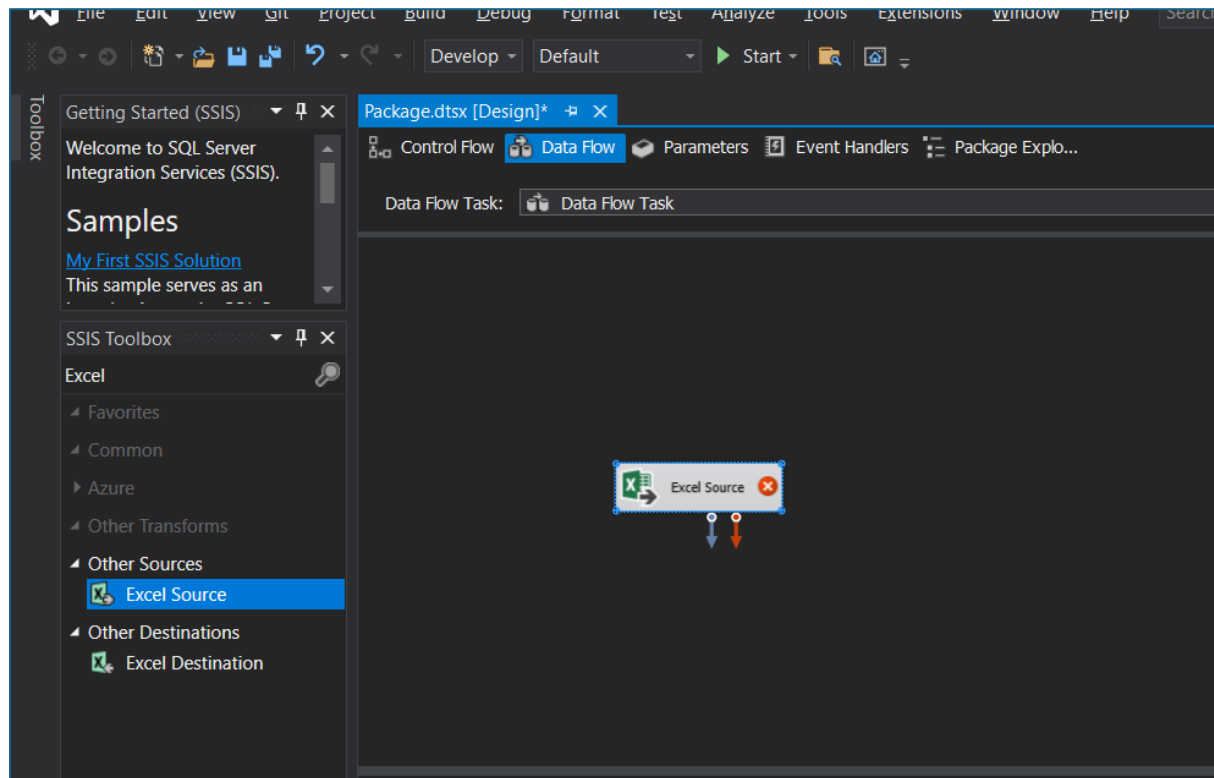


Figure 1. Pulling excel source

After we selected the Excel file then we used multicast to connect it like this picture below:

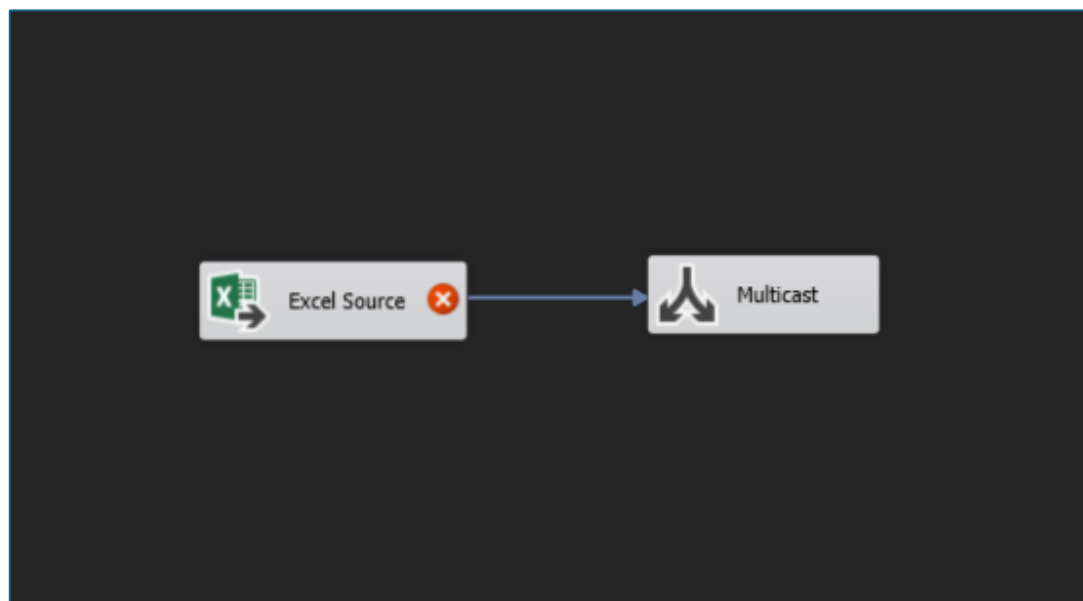


Figure 2. connecting multicast

After this step, we pulled and sorted suitable data contributions for each Dimension of the cube. In this project, we put in 5 dimensions ( 4 dim and 1 fact):

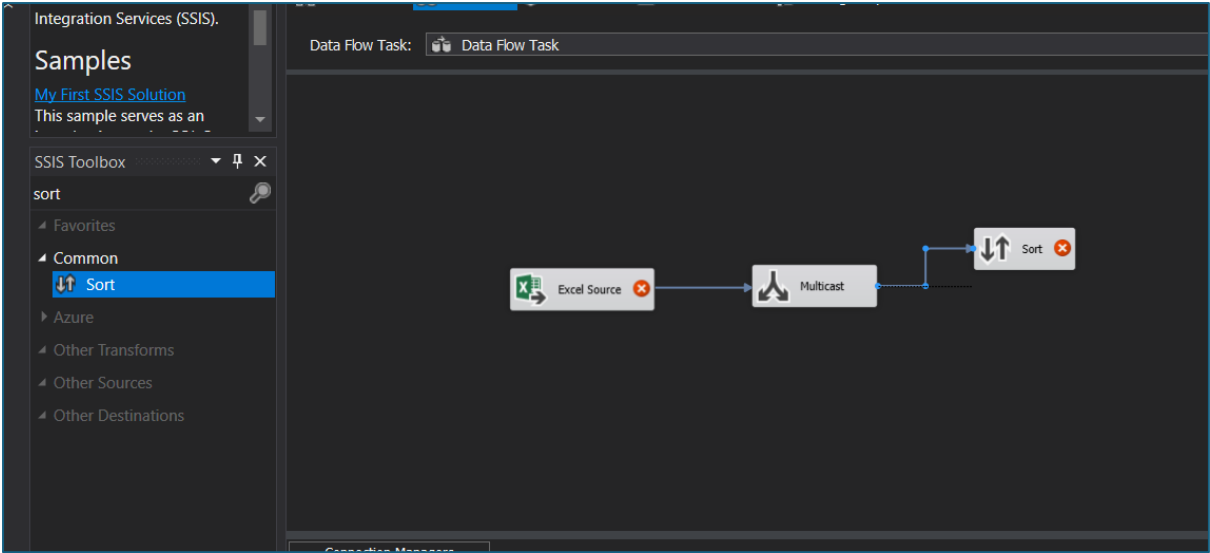


Figure 3. Sorting

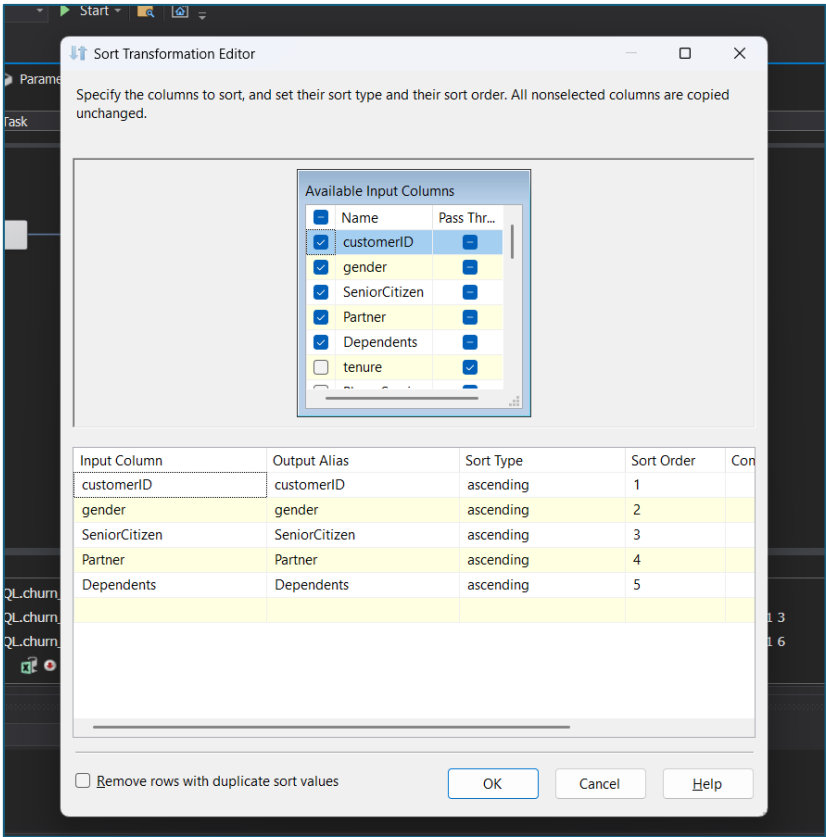


Figure 4. Choosing attribution

In OLE DB, we queried sql commands in it

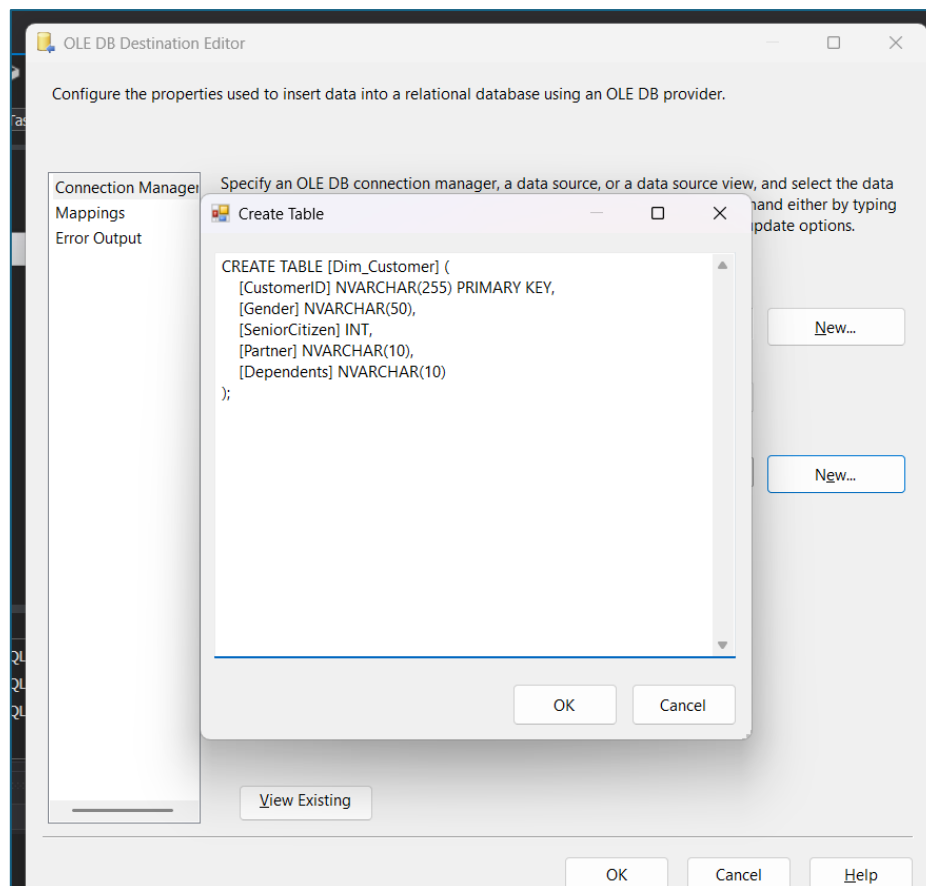


Figure 5. SQL query

To connect and execute in SQL, we moved to “Control flow” and pulled the Execute SQL Task icon

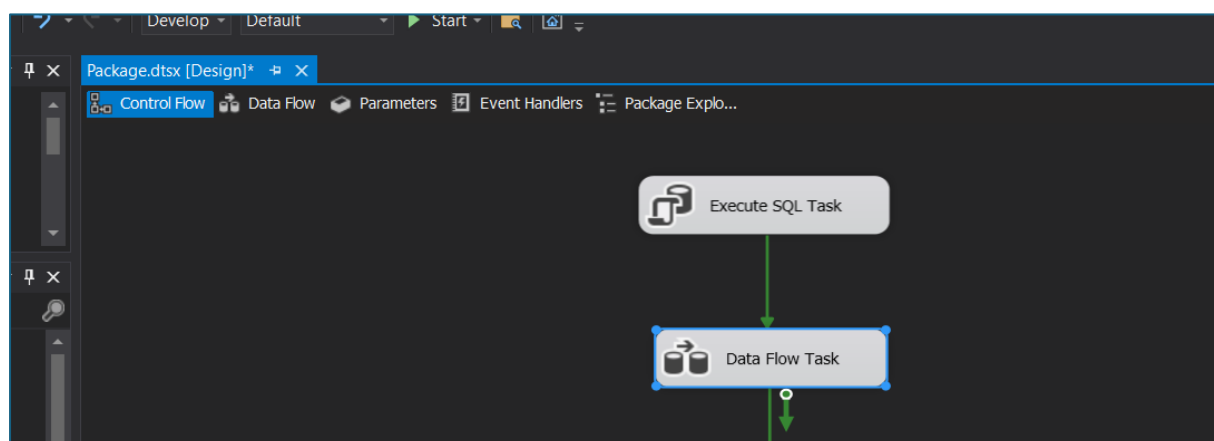


Figure 6. Execute SQL task

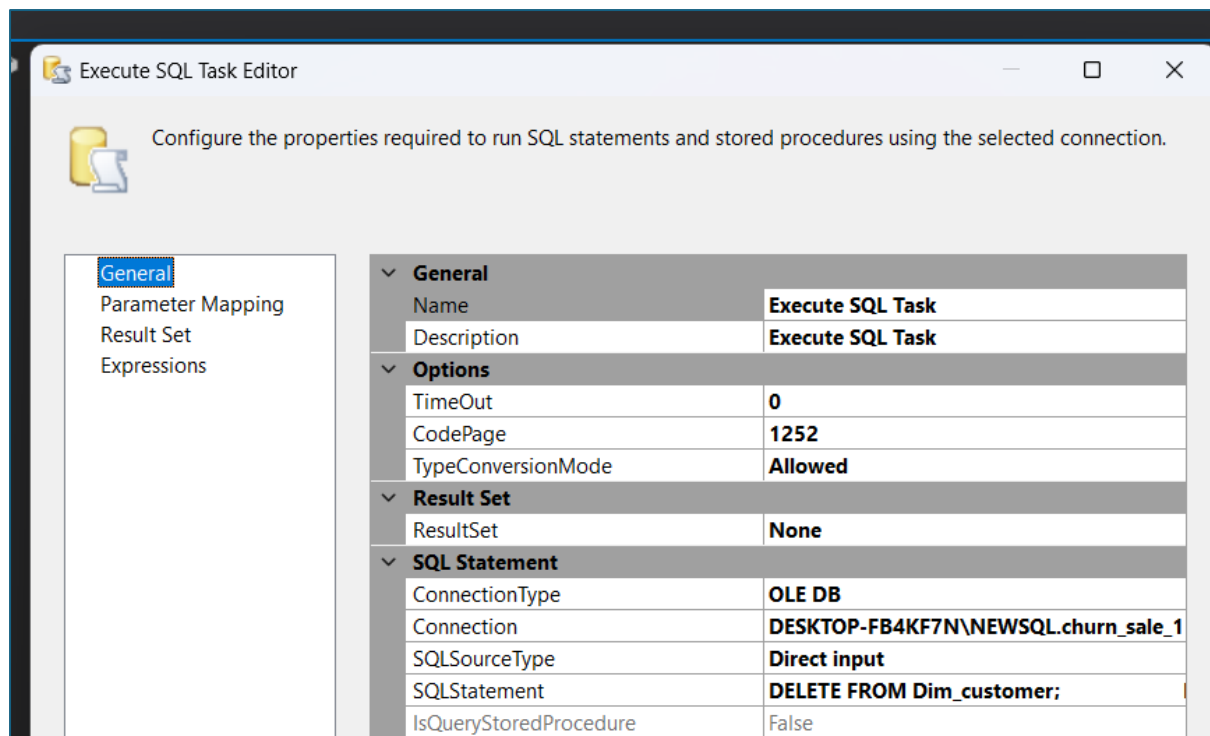


Figure 7. Connecting to SQL management

To execute and connect it, we created the database for it first in SQL management.

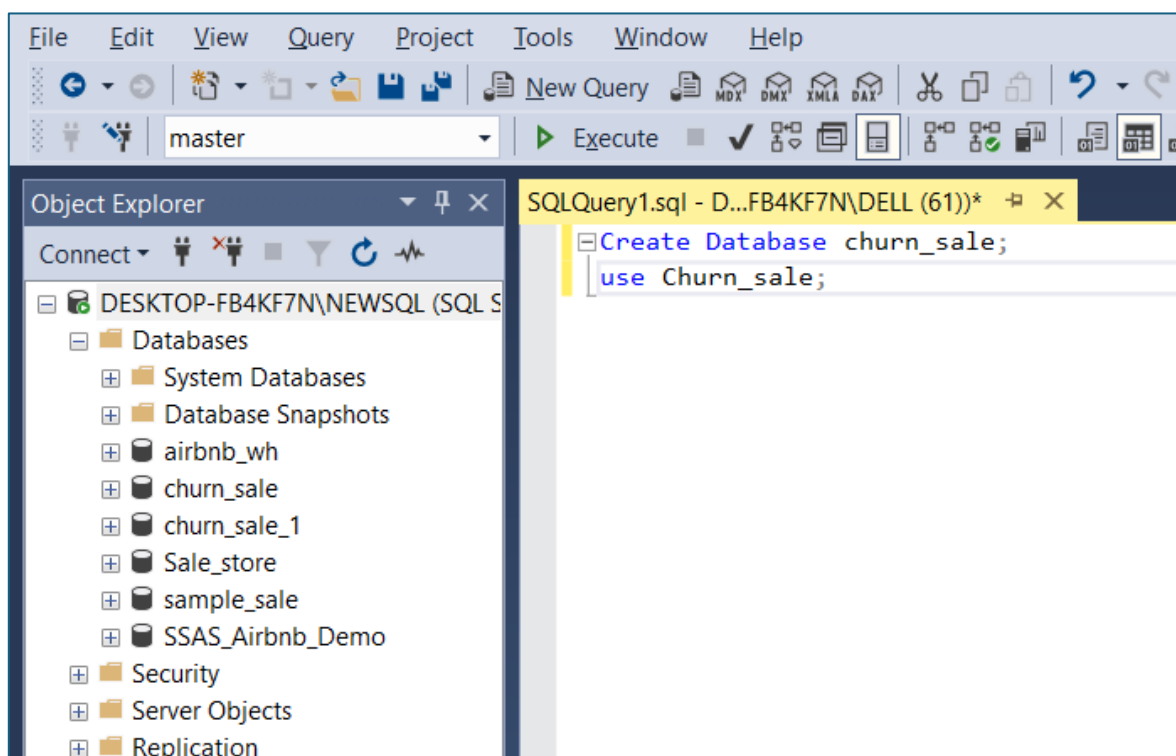


Figure 8. Create Database

And put SQL statements:

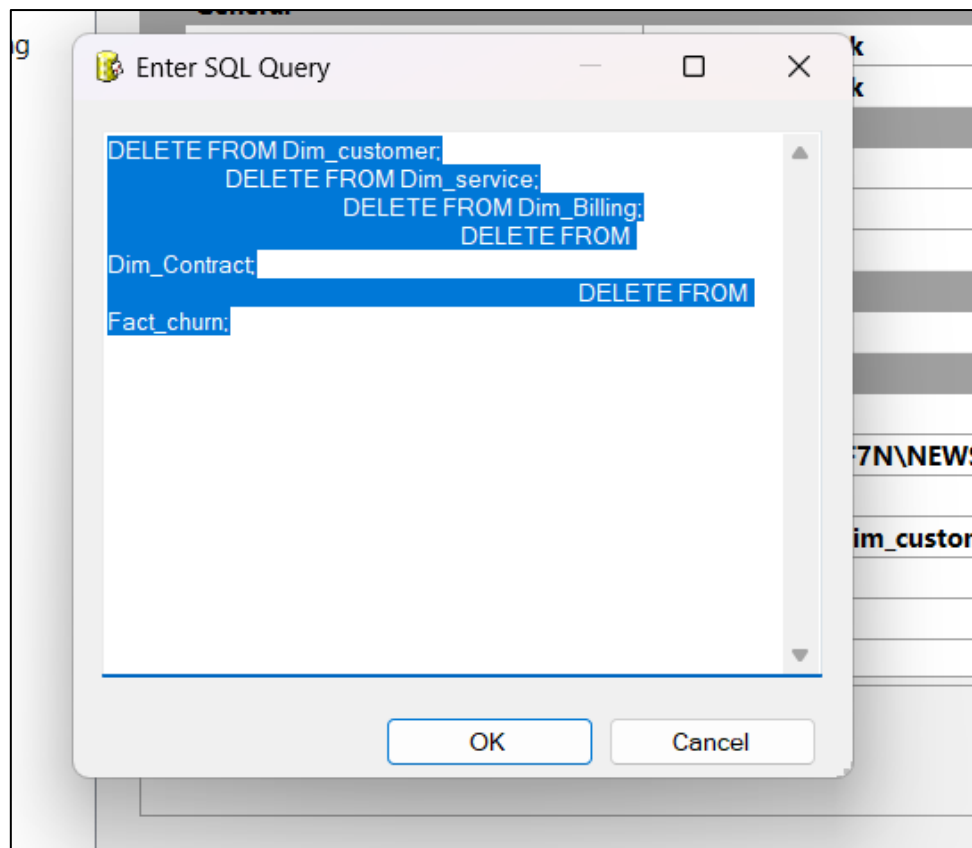


Figure 9. Deleting Query

Finally, do with foreign keys for those Dims

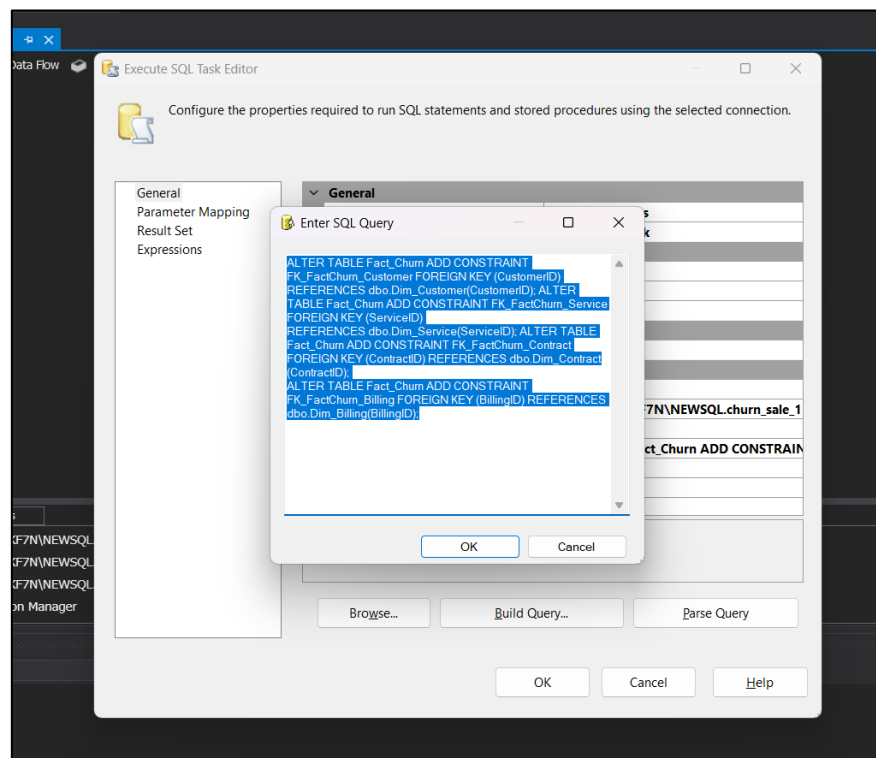


Figure 10. Altering foreign keys



Result:

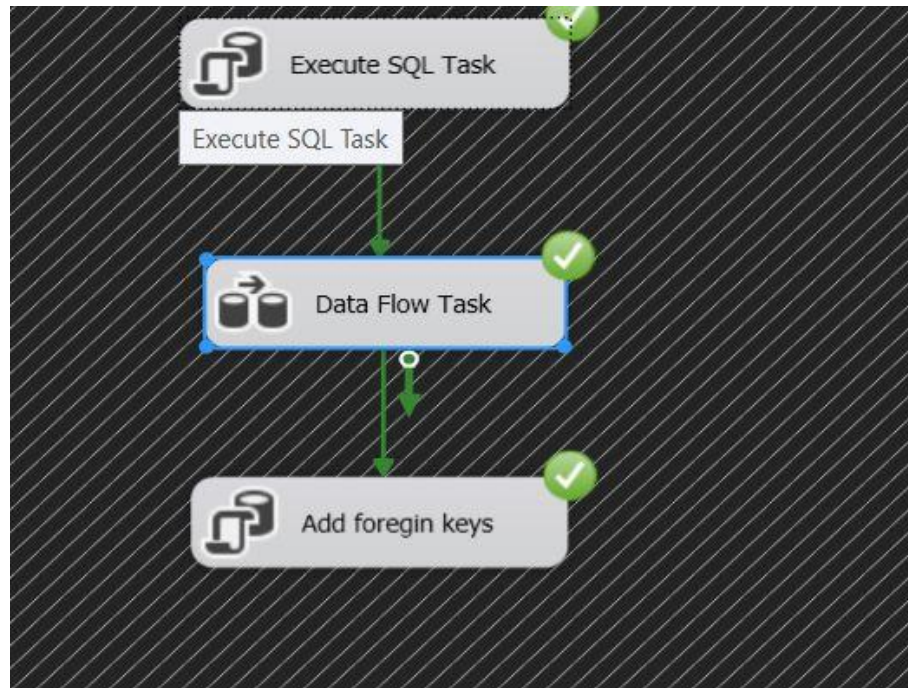


Figure 11. Execute successfully

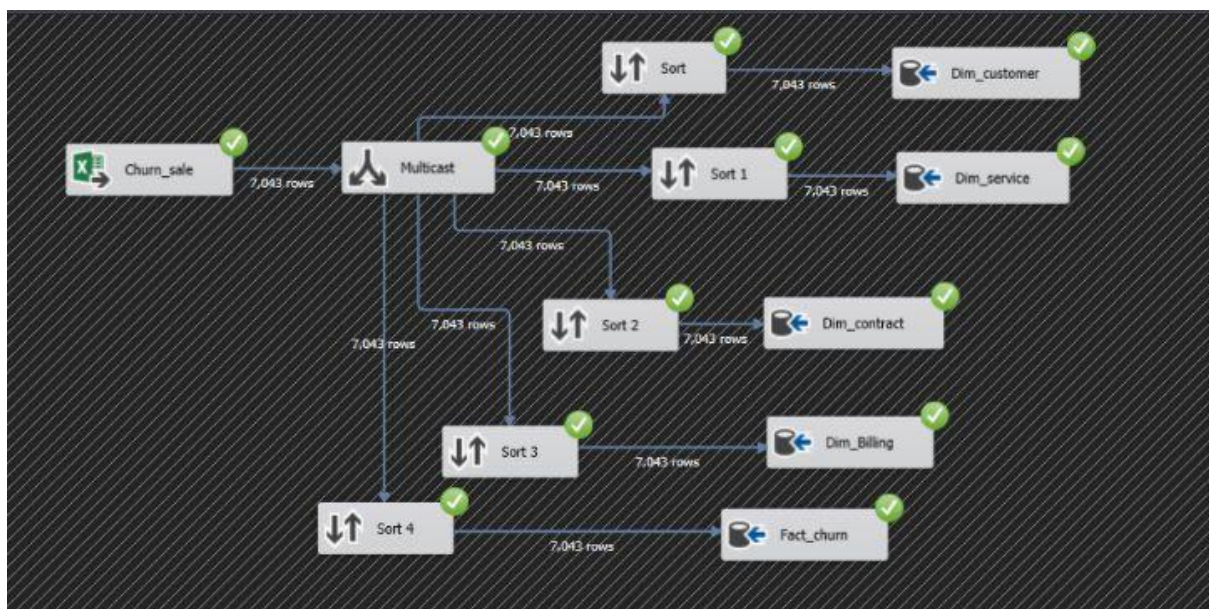


Figure 12. Data flow successfully

The screenshot shows a SQL query execution window. The query entered is `SELECT * FROM Dim_Customer;`. Below the query editor, there is a progress bar at 00% and two tabs: 'Results' and 'Messages'. The 'Results' tab is active, displaying a table with 6 columns: CustomerID, Gender, SeniorCitizen, Partner, and Dependents. The table contains 12 rows of data. The first row, with CustomerID '0002-ORFBO', is highlighted with a dashed border.

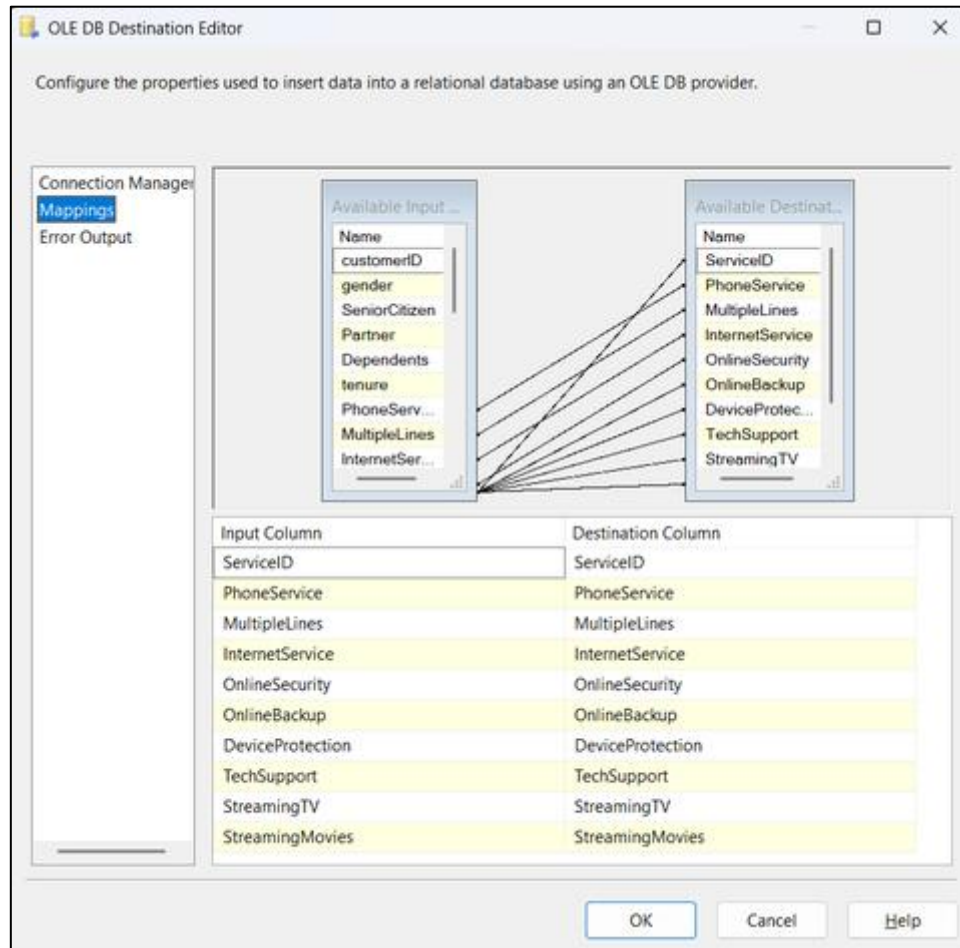
	CustomerID	Gender	SeniorCitizen	Partner	Dependents
1	0002-ORFBO	Female	0	Yes	Yes
2	0003-MKNFE	Male	0	No	No
3	0004-TLHLJ	Male	0	No	No
4	0011-IGKFF	Male	1	Yes	No
5	0013-EXCHZ	Female	1	Yes	No
6	0013-MHZWF	Female	0	No	Yes
7	0013-SMEOE	Female	1	Yes	No
8	0014-BMAQU	Male	0	Yes	No
9	0015-UOCOJ	Female	1	No	No
10	0016-QLJIS	Female	0	Yes	Yes
11	0017-DINOC	Male	0	No	No
12	0017-IUDMW	Female	0	Yes	Yes

Figure 13. Connecting successfully

## Schema overview

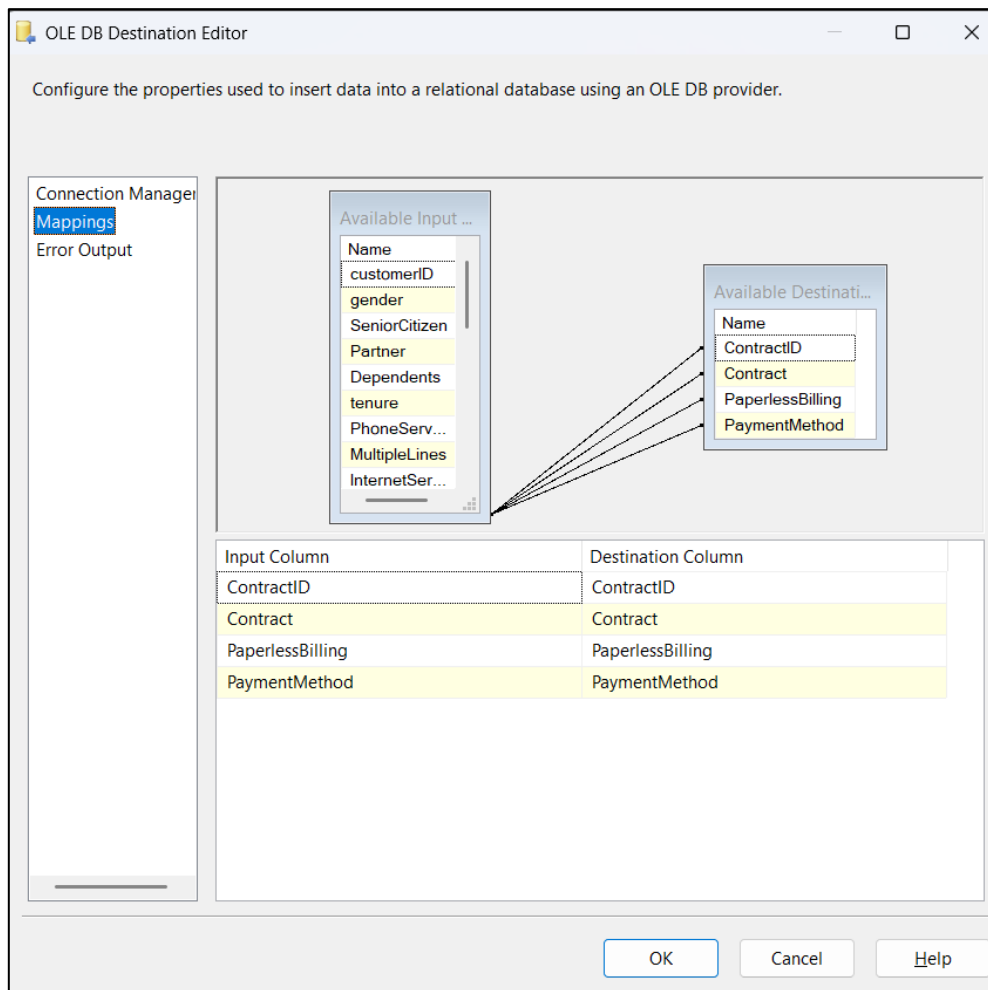
### Dim\_Customer:

Show detail of customer information (ID, phone service,...)



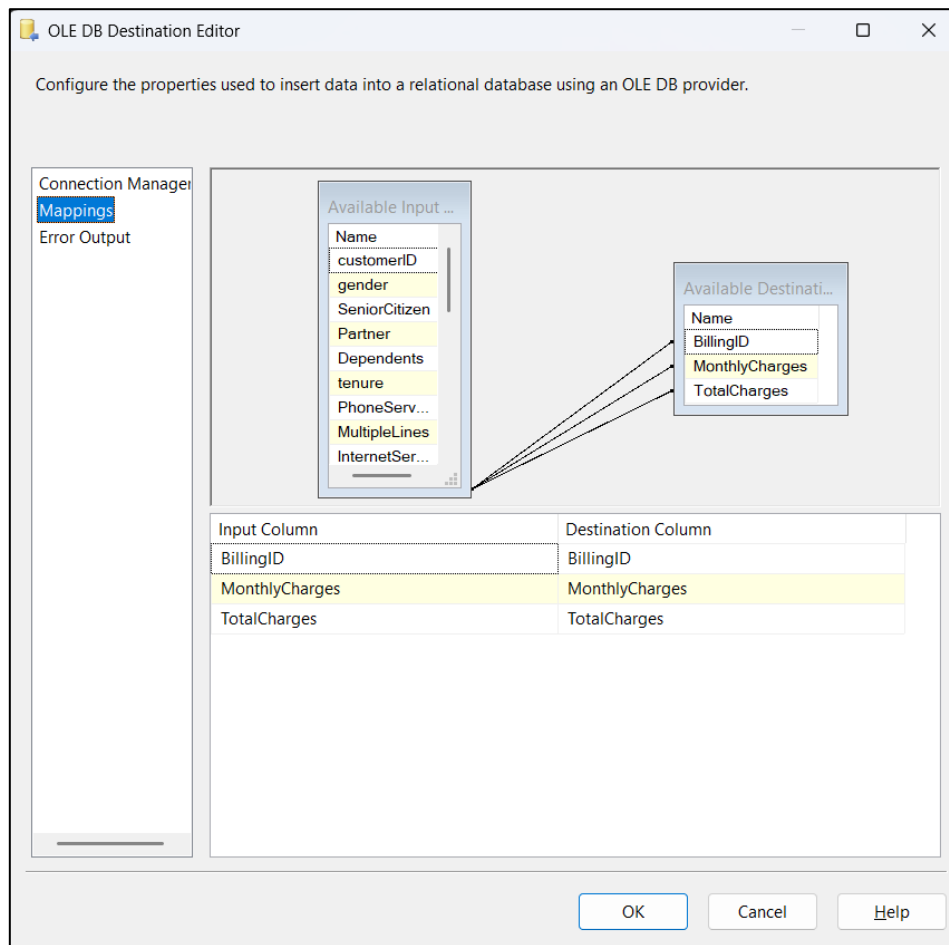
## Dim\_Contact:

This dim store about contact information



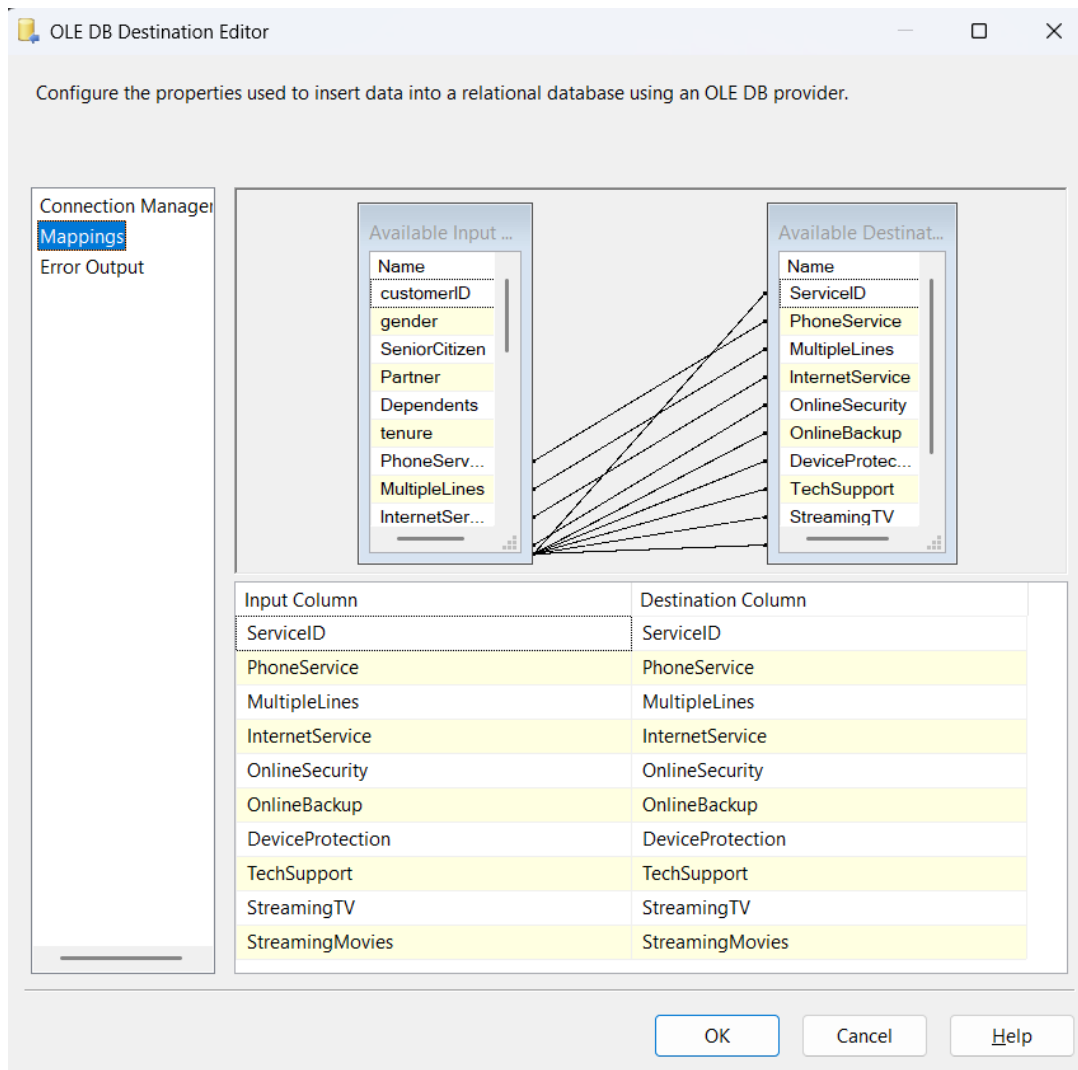
## Dim\_Billing

Show the detail of billing information



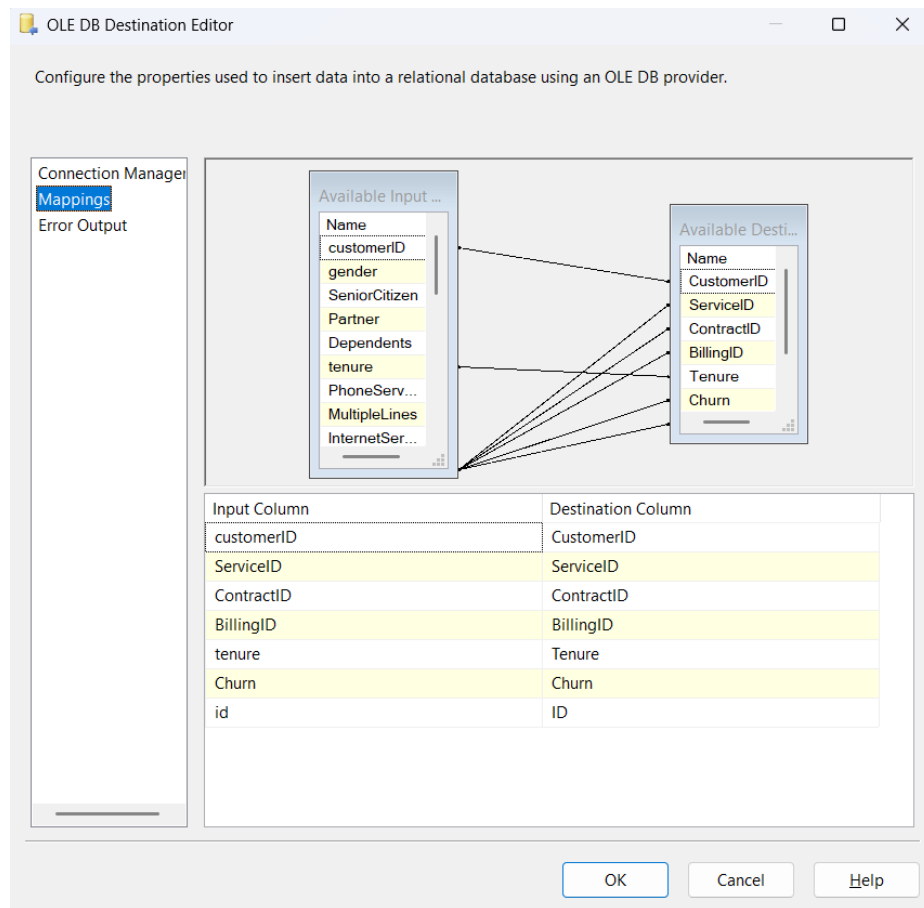
## Dim\_Service

Show the information about service detail (ServiceId, phoneService,...)



## Fact

This fact table is designed to store customer churn data by mapping input columns such as customerID, ServiceID, ContractID, BillingID, tenure, and Churn..



## Chapter 2: Analysis and Reporting Process

### 2.1 Process of building SSAS

First, we create a data source:

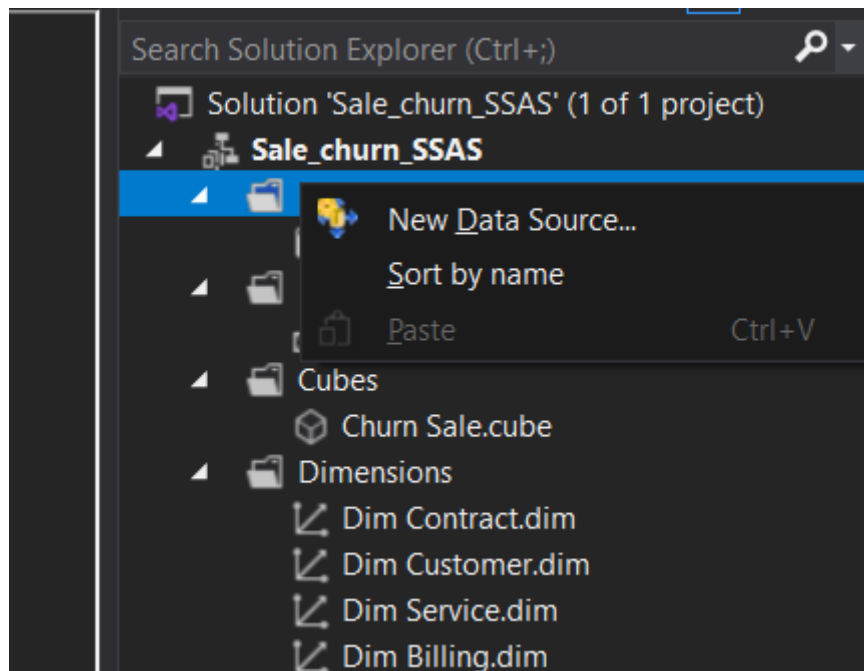


Figure 14. Choosing database

Then with the source view

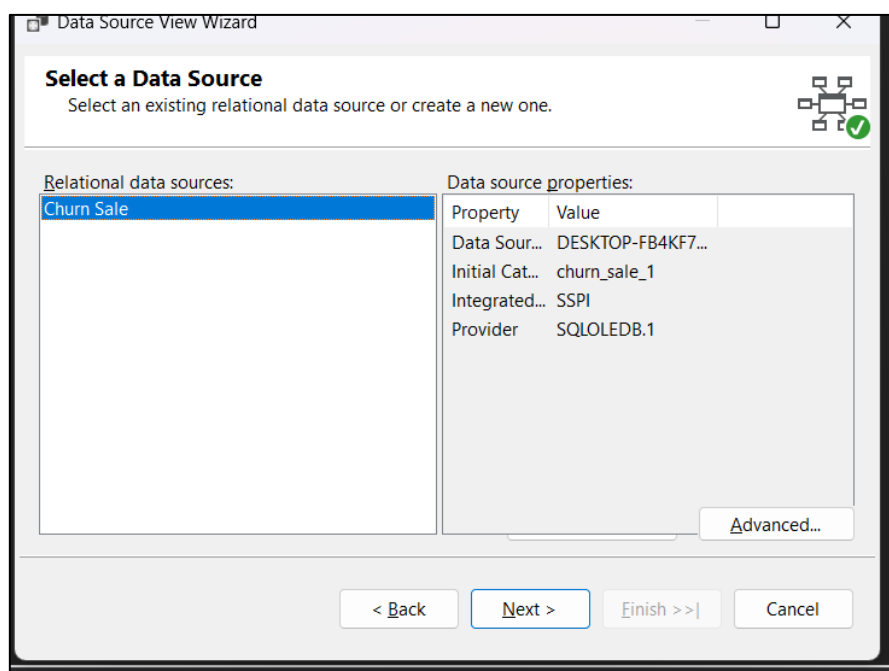


Figure 14. Choosing existing data source



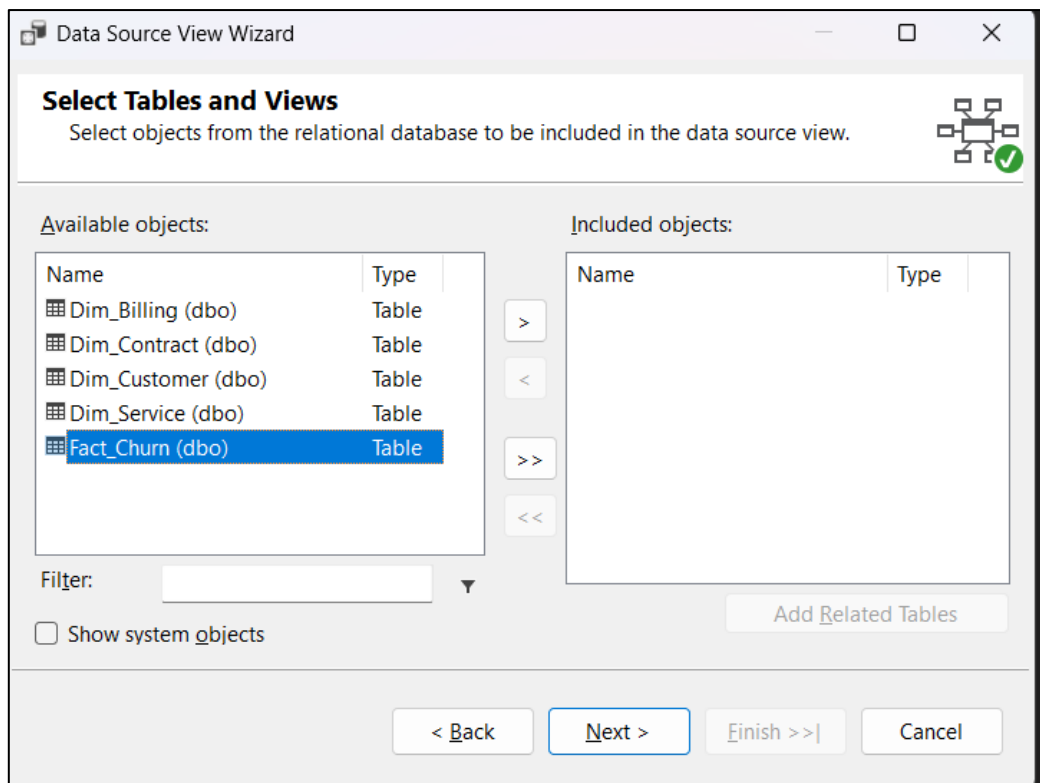


Figure 15. Choosing Fact\_churn table

Select the Fact table the click the “Add Related Tables” and press Finish.

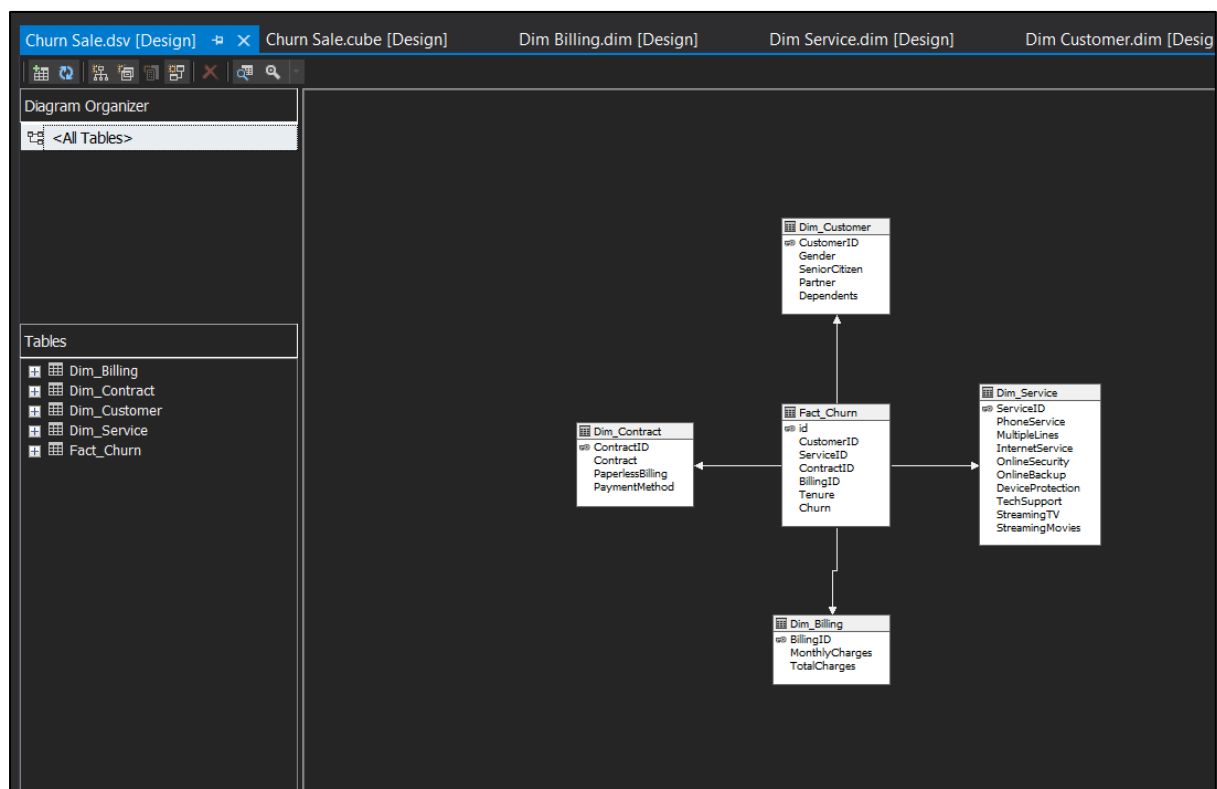


Figure 15. Displaying cube

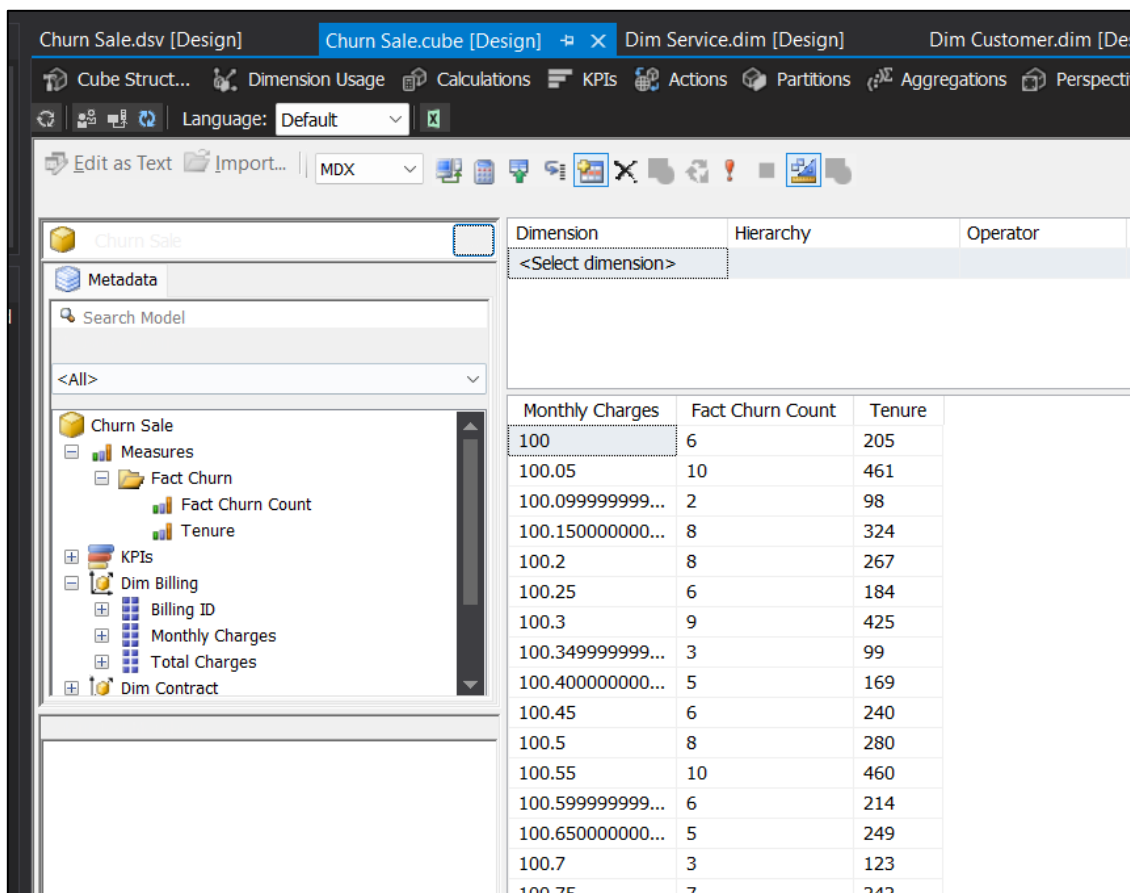
Now we have a cube with 4 dimensions and 1 fact

### Overview:

The data warehouse is designed using a **star schema** structure, with the central **Fact\_Churn** table connected to four-dimension tables: **Dim\_Customer**, **Dim\_Contract**, **Dim\_Service**, and **Dim\_Billing**.

- The **Fact\_Churn** table stores key churn-related metrics, such as customer tenure and churn status.
- Dimension tables provide descriptive attributes for customers, contracts, services, and billing details.
- This structure allows for efficient analysis of customer churn patterns based on demographic, service, and billing factors.

## 2.2 Analysis on SSAS and BI



Dimension	Hierarchy	Operator	Monthly Charges	Fact Churn Count	Tenure
<Select dimension>			100	6	205
			100.05	10	461
			100.099999999...	2	98
			100.150000000...	8	324
			100.2	8	267
			100.25	6	184
			100.3	9	425
			100.349999999...	3	99
			100.400000000...	5	169
			100.45	6	240
			100.5	8	280
			100.55	10	460
			100.599999999...	6	214
			100.650000000...	5	249
			100.7	3	123
			100.75	7	242

Figure 15. Choosing Fact\_churn table

This screenshot shows the sample Churn Sale cube in SQL Server Analysis Services (SSAS). The cube contains key measures from the Fact\_Churn table, including Fact Churn Count and Tenure, and is linked with billing-related dimensions such as Monthly Charges and Total Charges.

The displayed result allows users to analyze the relationship between Monthly Charges, customer tenure, and the churn count, supporting data-driven insights for customer retention strategies and so on.

## 2.3 Analysis (MDX)

### 1. Roll-up: Customer Churn Analysis by Internet Service Type

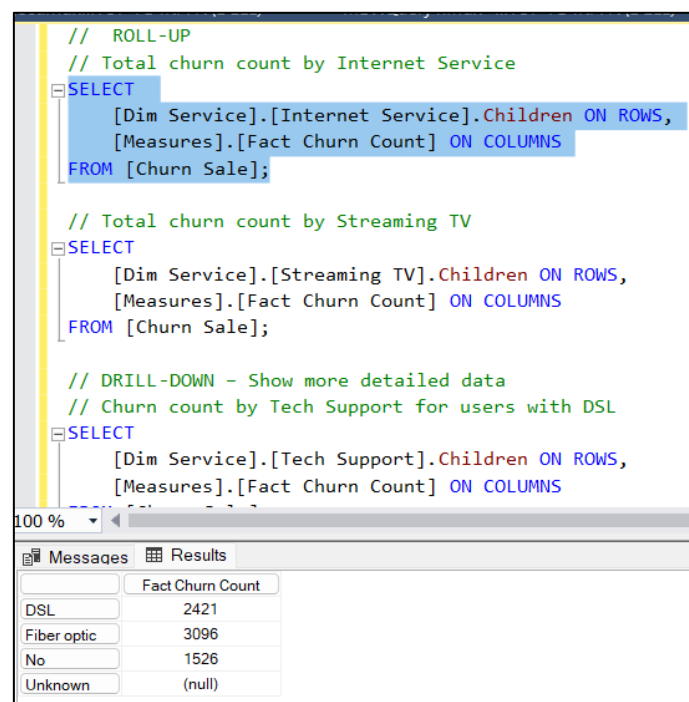


Figure 15. Choosing Fact\_churn table

Rows: [Dim Service].[Internet Service].Children → Types of Internet Services (DSL, Fiber optic, No internet service, Unknown).

Columns: [Measures].[Fact Churn Count] → The number of customers who churned.

Data Source: [Churn Sale] cube.

Customers with Fiber Optic service have the highest churn count (3,096), followed by DSL (2,421). Those with no internet service churn less, possibly due to lower expectations or service interactions.

## 2. Slice: Customer Churn Distribution by Streaming TV Subscription Status

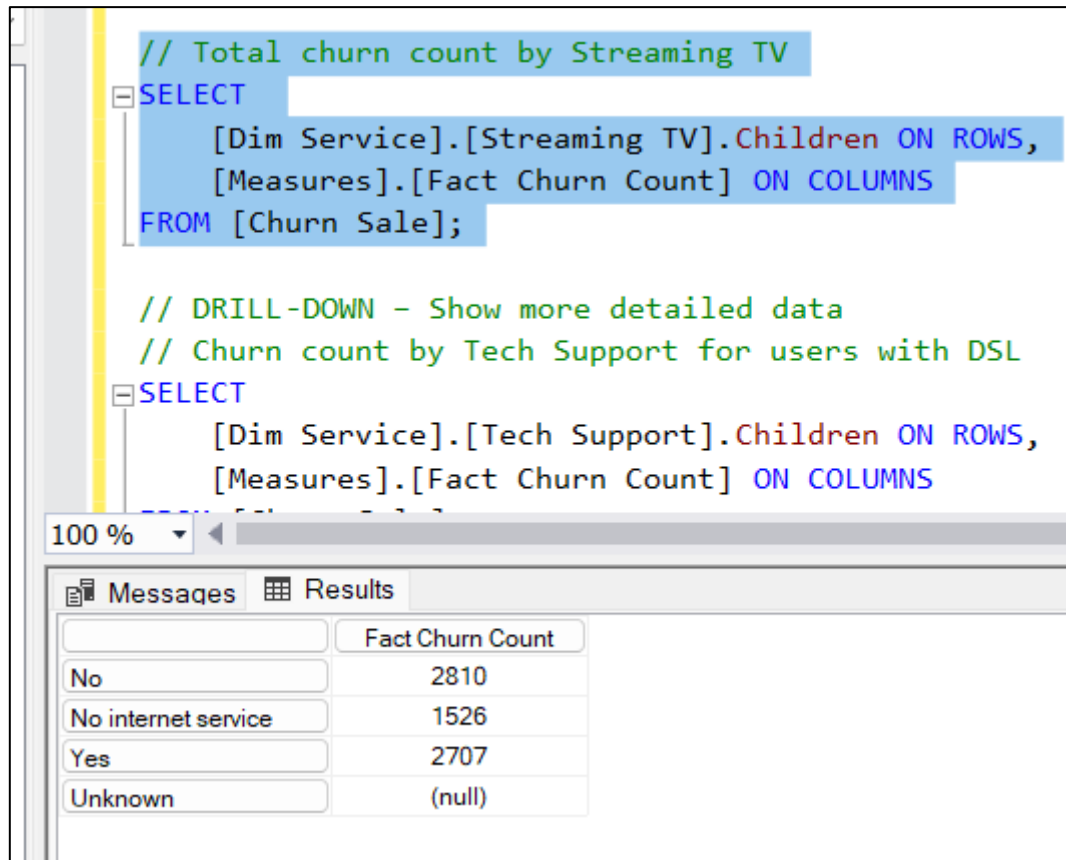


Figure 16. Querying Total churn\_sale

Rows: [Dim Service].[Streaming TV].Children → Whether customers use streaming TV or not.

Columns: [Measures].[Fact Churn Count] → Number of churned customers.

Data Source: [Churn Sale] cube.

Customers who do not use streaming TV slightly churn more than those who do. This suggests streaming TV might contribute to retention, but the difference is not very large.

### 3. Drill-Down: Churn Distribution by Tech Support (DSL Users Only)

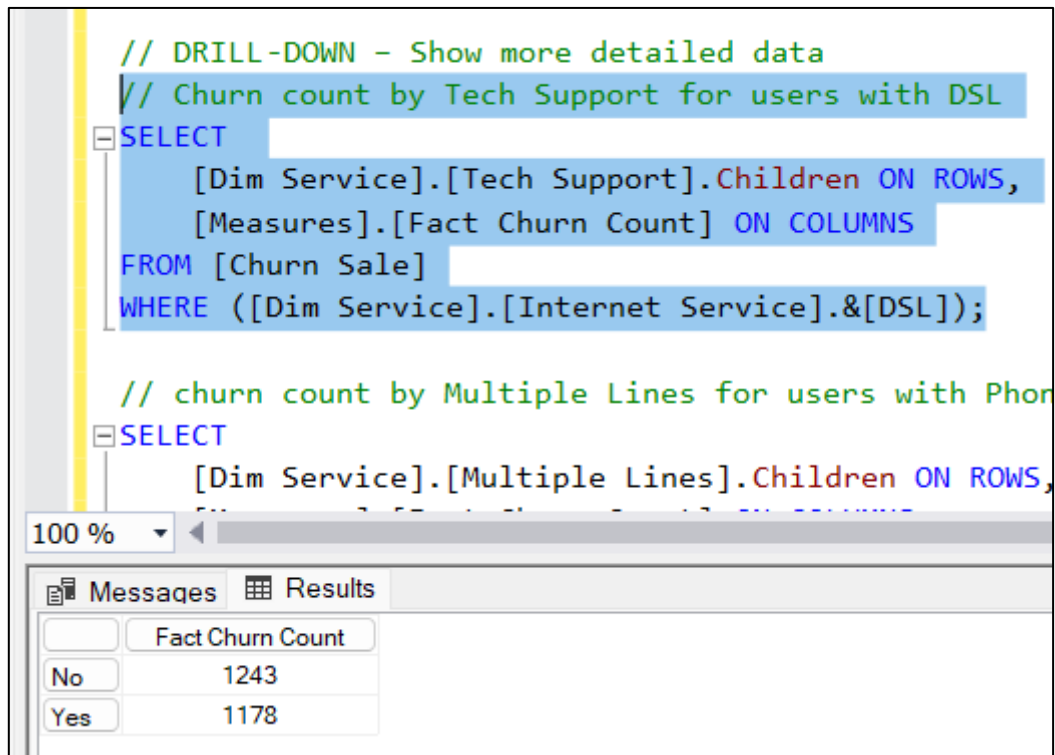


Figure 17. Querying Total churn\_sale

Rows: "Yes" or "No" for Tech Support.

Columns: Number of customers churned (Fact Churn Count).

Filter (WHERE): Only includes users with DSL Internet service.

Analyze the churn behavior of customers using DSL Internet service, segmented by whether they have Tech Support or not.

#### 4. Churn by Multiple Line Usage Among Customers with Phone Service

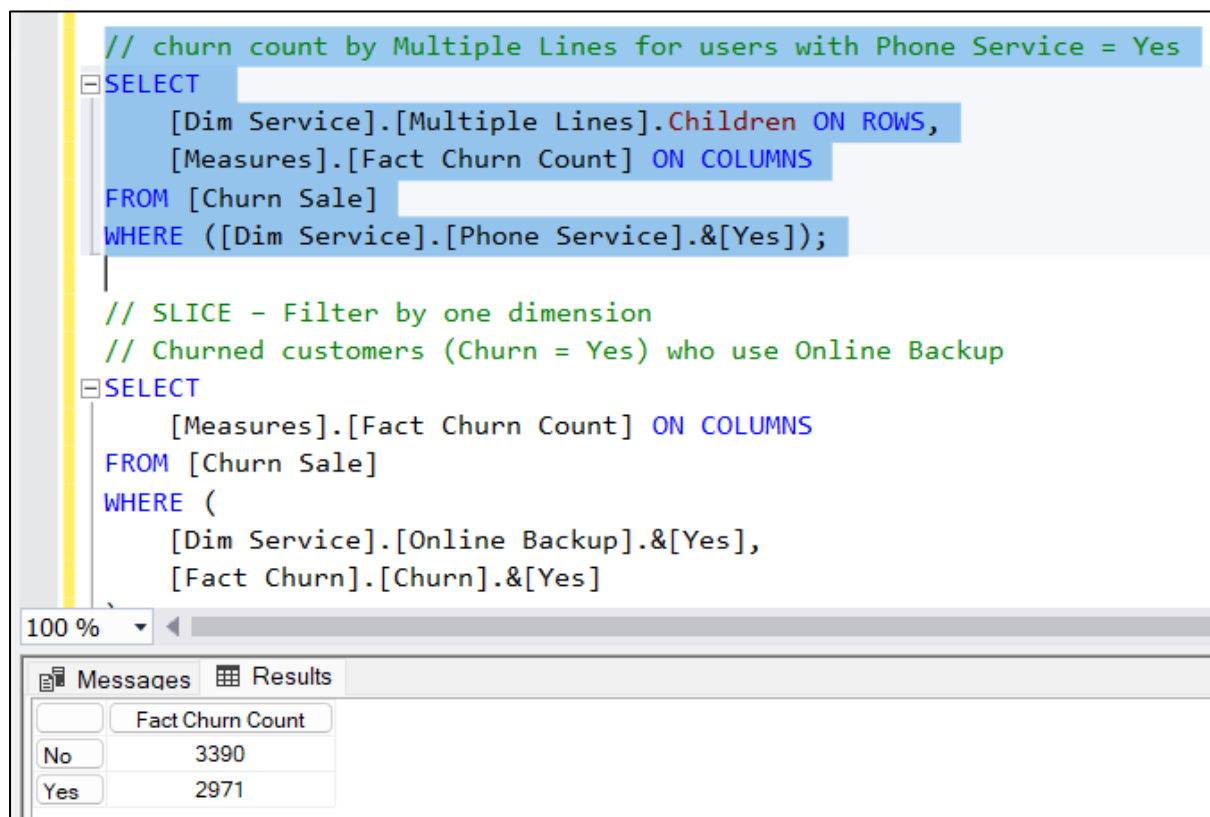


Figure 18. Querying Total churn\_sale

Rows: "Yes" or "No" for Multiple Lines (for users with Phone Service = Yes).

Columns: Number of customers churned (Fact Churn Count).

Filter (WHERE): Only includes users with Phone Service = Yes.

Analysis:

Among users with Phone Service, those with Multiple Lines ("Yes") have a specific churn count, while those without ("No") have a different count (exact numbers depend on additional data).

This segmentation suggests that the presence of Multiple Lines may influence churn behavior, potentially indicating higher or lower retention depending on the count.

## 5. Dice: Non-Churned Customers Who Use Streaming Movies But Not Online Security

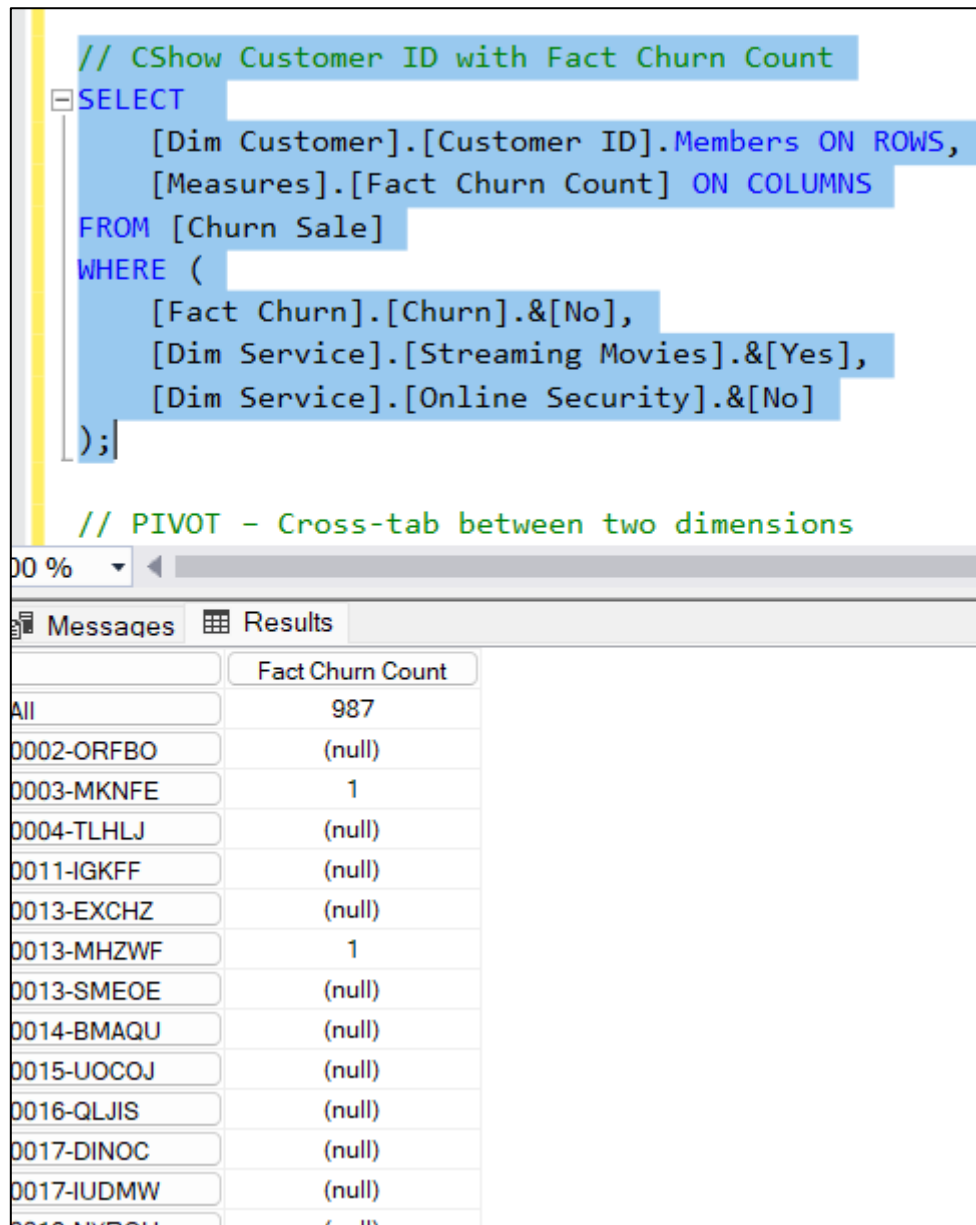


Figure 19. Querying Total churn\_sale

Rows: Customer ID.

Columns: Number of customers churned (Fact Churn Count).

Filter (WHERE): Only includes users with Streaming Movies = Yes and Online Security = No

Analysis:

Specific customers (e.g., 0002-ORFBO with 987 churns) show high churn counts, while others (e.g., 0003-MKNFE with 1) have minimal or no churn.

This suggests variability in churn behavior among customers with Streaming Movies but without Online Security, possibly indicating individual factors driving churn.

## 6. Non-Churned Customers with Streaming Movies Enabled and Online Security Disabled

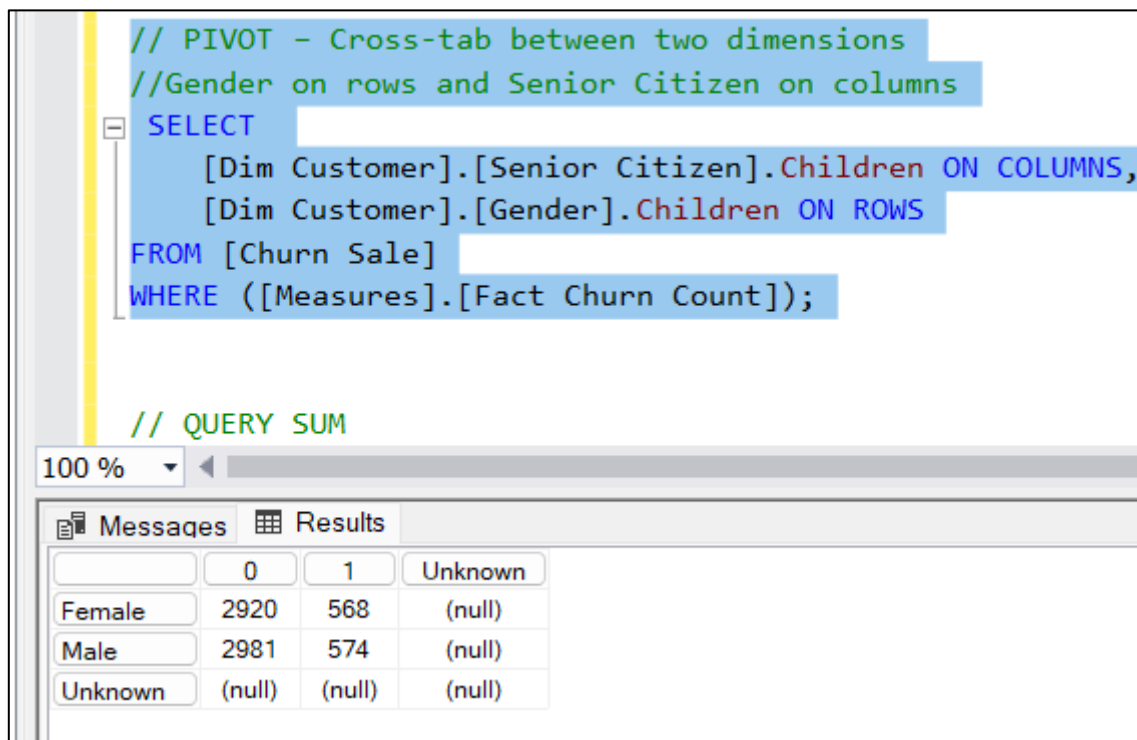


Figure 20. Querying Total churn\_sale

Rows: Gender

Columns: Senior Citizen (0 or 1).

Filter (WHERE): No specific filter beyond Fact Churn Count

Analysis:

Females show 2920 non-senior citizens (0) and 568 senior citizens (1), while males show 2981 and 574 respectively.

Churn counts are similar across genders and senior status, suggesting that gender and senior citizenship may not strongly influence churn behavior.



## 7. Sum: Total Tenure by Internet Service Type

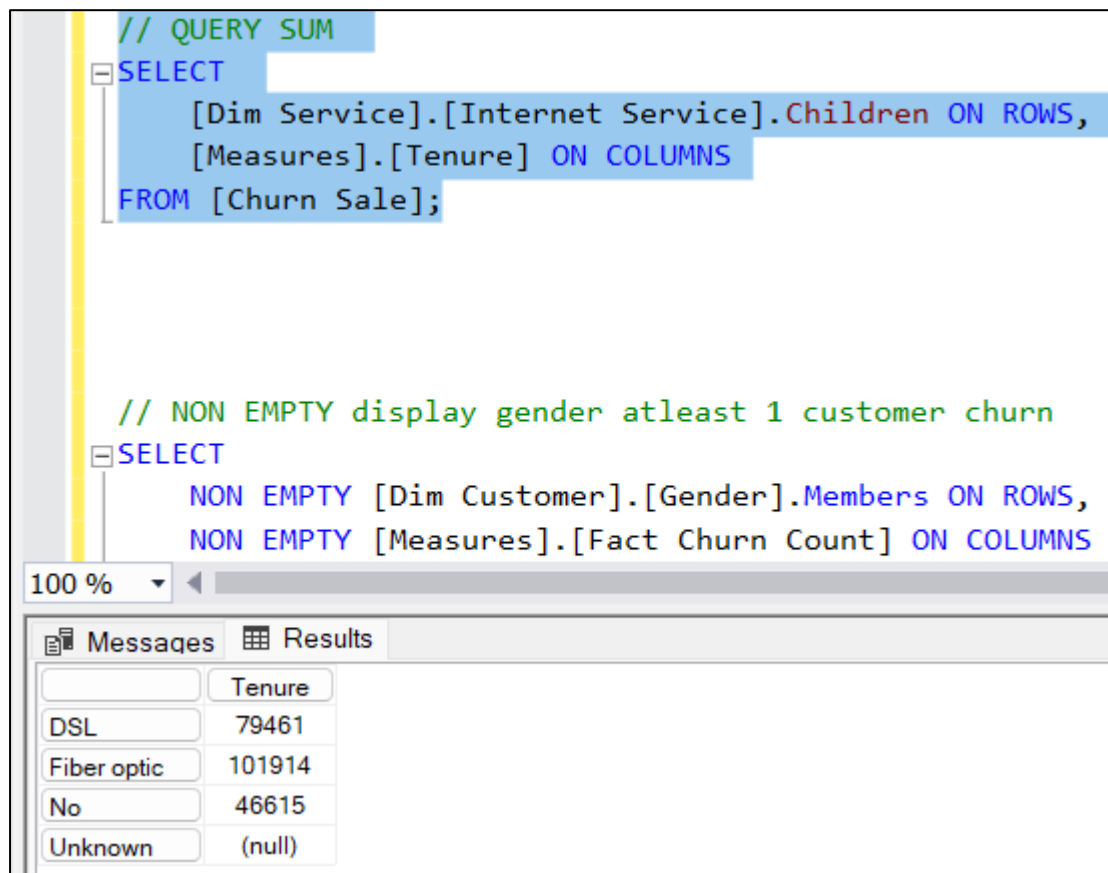


Figure 21. Querying Total churn\_sale

Rows: Internet Service type.

Columns: Tenure.

Filter (WHERE): No specific filter.

Analysis:

DSL users have a tenure of 79461, Fiber optic users 101914, and No internet service users 46615.

Higher tenure values (especially for Fiber optic) suggest longer customer retention, while lower values (No internet) may indicate shorter engagement, though churn data is needed for confirmation.

## 8. Top 5 Internet Service Types with the Highest Customer Churn

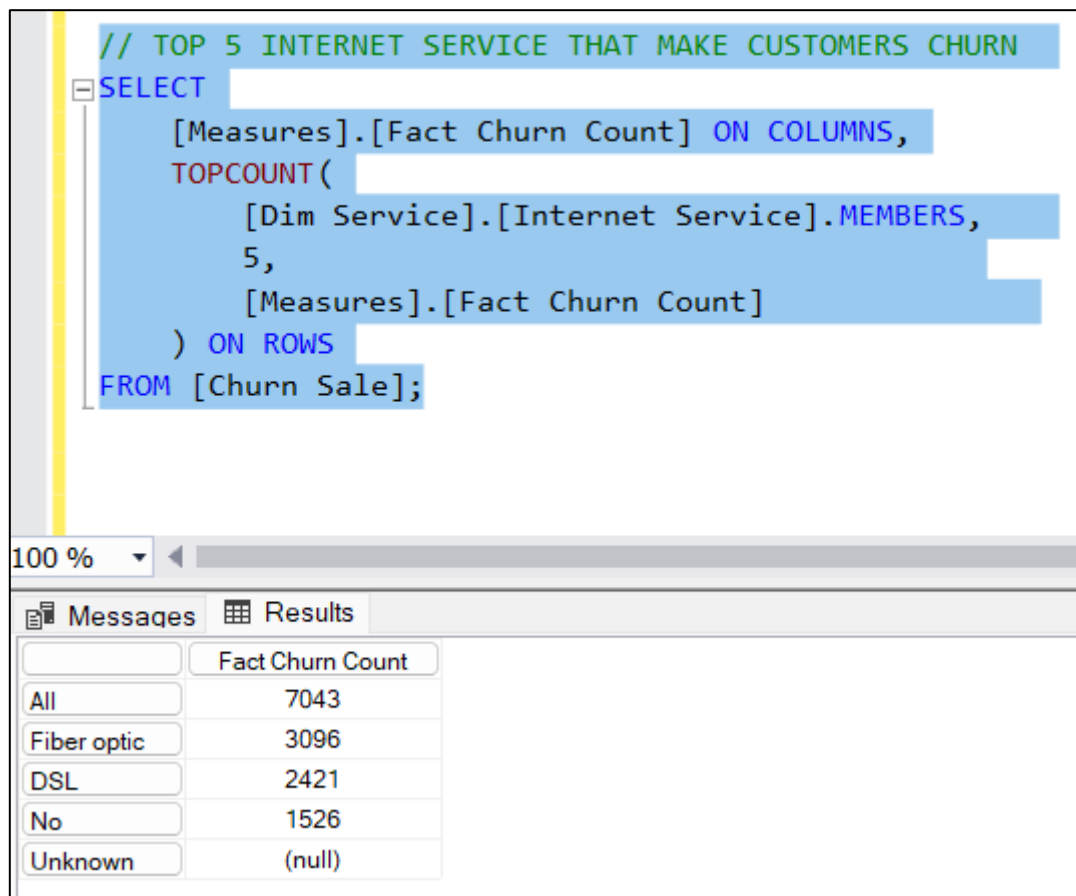


Figure 22. Querying Total churn\_sale

Rows: Internet Service type.

Columns: Number of customers churned (Fact Churn Count)

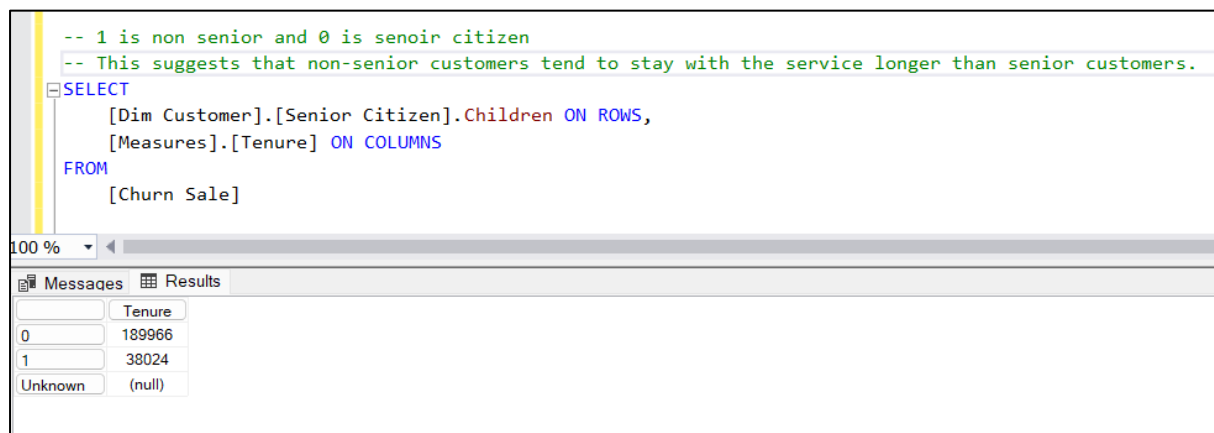
Filter (WHERE): No specific filter.

Analysis:

Fiber optic has the highest churn count (3096), followed by DSL (2421) and No internet (1526).

This indicates that Fiber optic users are more likely to churn, possibly due to service issues or other factors, compared to DSL or No internet users.

## 9. Children: Comparison of Total Tenure Between Senior and Non-Senior Customers ( 1 is non senior and 0 is senior)



```
-- 1 is non senior and 0 is senoir citizen
-- This suggests that non-senior customers tend to stay with the service longer than senior customers.
SELECT
    [Dim Customer].[Senior Citizen].Children ON ROWS,
    [Measures].[Tenure] ON COLUMNS
FROM
    [Churn Sale]
```

	Tenure
0	189966
1	38024
Unknown	(null)

Rows: Senior Citizen (0 or 1).

Columns: Tenure.

Filter (WHERE): No specific filter.

Analysis:

Non-senior citizens (0) have a tenure of 18996, while senior citizens (1) have 38024.

This suggests that non-senior customers tend to stay longer, potentially indicating better retention among younger users.

## 10. Total Revenue by Internet Service Type

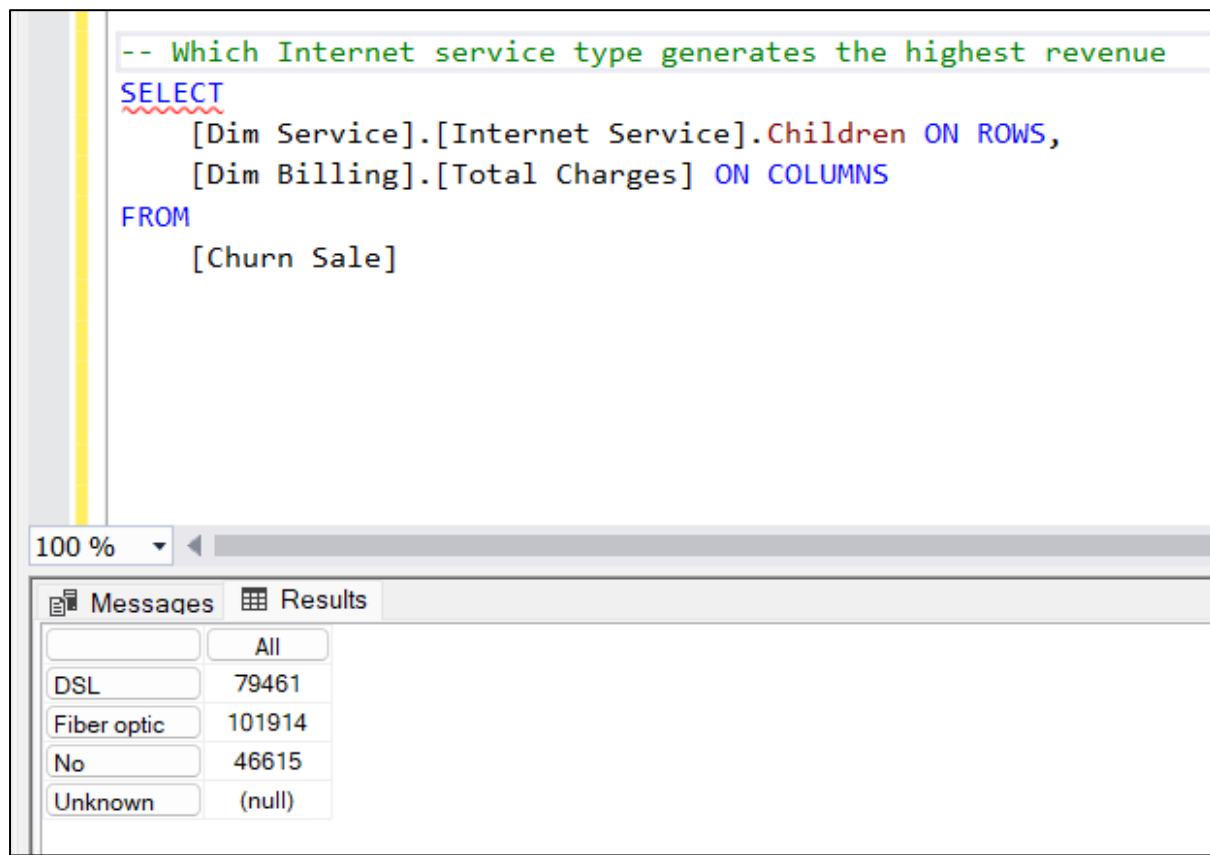


Figure 23. Querying Total churn\_sale

Rows: Internet Service type.

Columns: Total Charges.

Filter (WHERE): No specific filter.

Analysis:

Fiber optic generates the highest revenue (101914), followed by DSL (79461) and No internet (46615).

This highlights Fiber optic as the most profitable service type, though high revenue may correlate with higher churn

## 2.4 Pivot table and Excel

### 1 Roll-up: Customer Churn Analysis by Internet Service Type

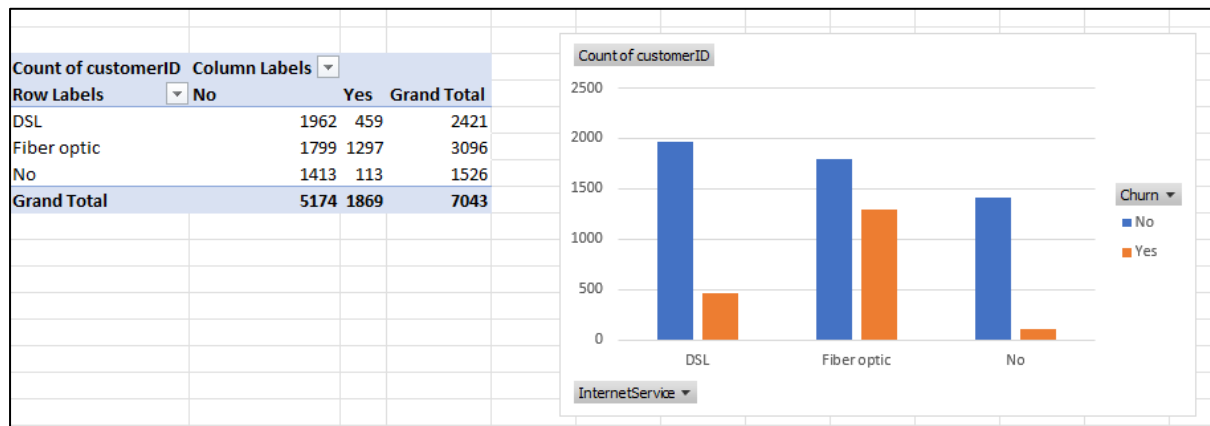


Figure 23. Querying Total churn\_sale

### 2 Slice: Customer Churn Distribution by Streaming TV Subscription Status

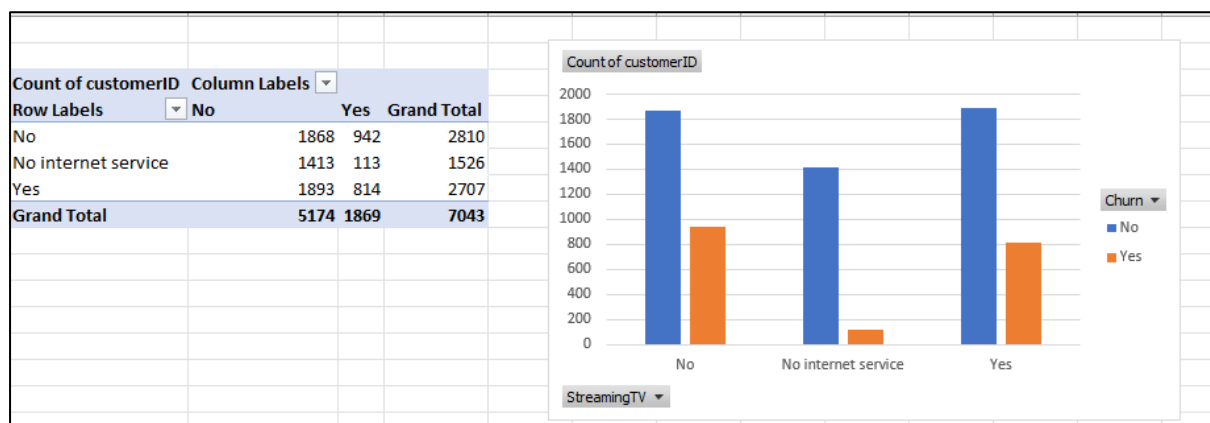


Figure 24. Querying Total churn\_sale

### 3 Slice: Customer Churn Distribution by Streaming TV Subscription Status

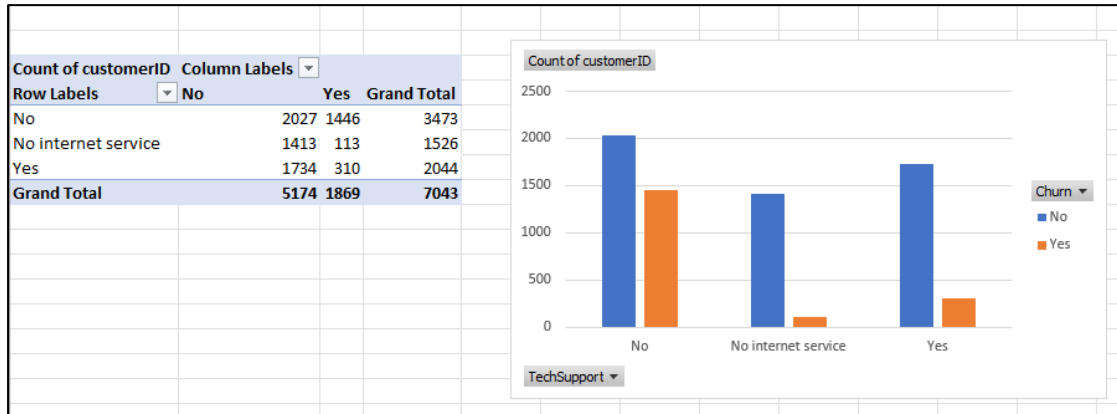


Figure 25. Querying Total churn\_sale

### 4 Churn by Multiple Line Usage Among Customers with Phone Service

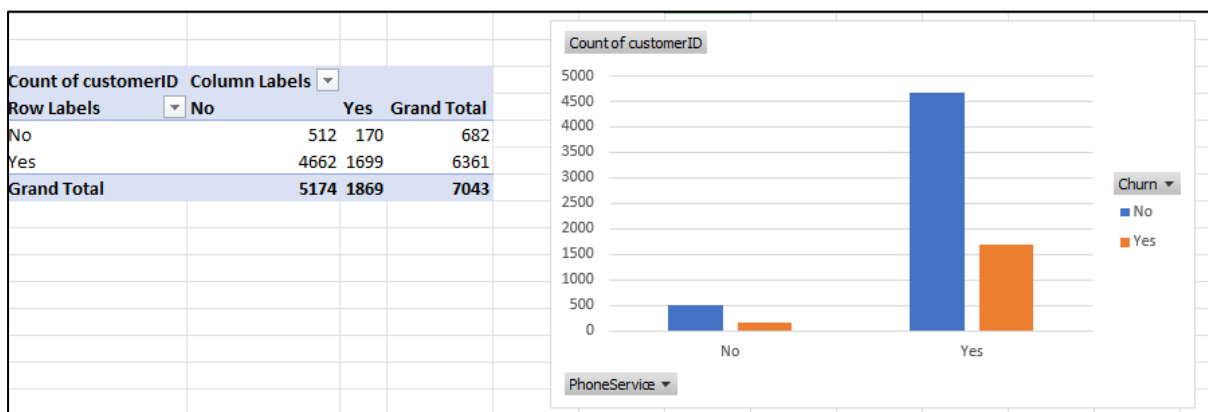


Figure 26. Querying Total churn\_sale

## 5 Customer Churn Distribution by Contract Type

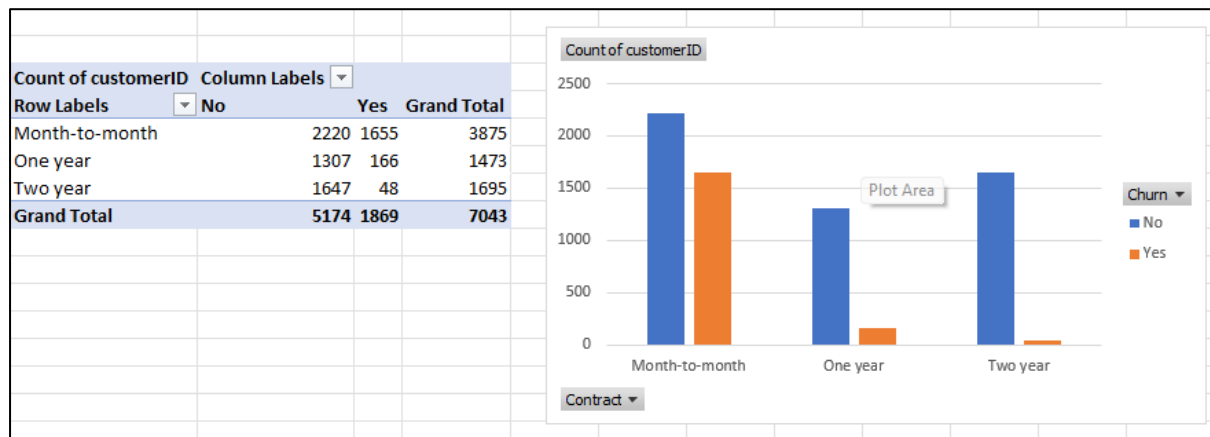


Figure 27. Churn Distribution

## 6 Non-Churned Customers with Streaming Movies Enabled and Online Security Disabled

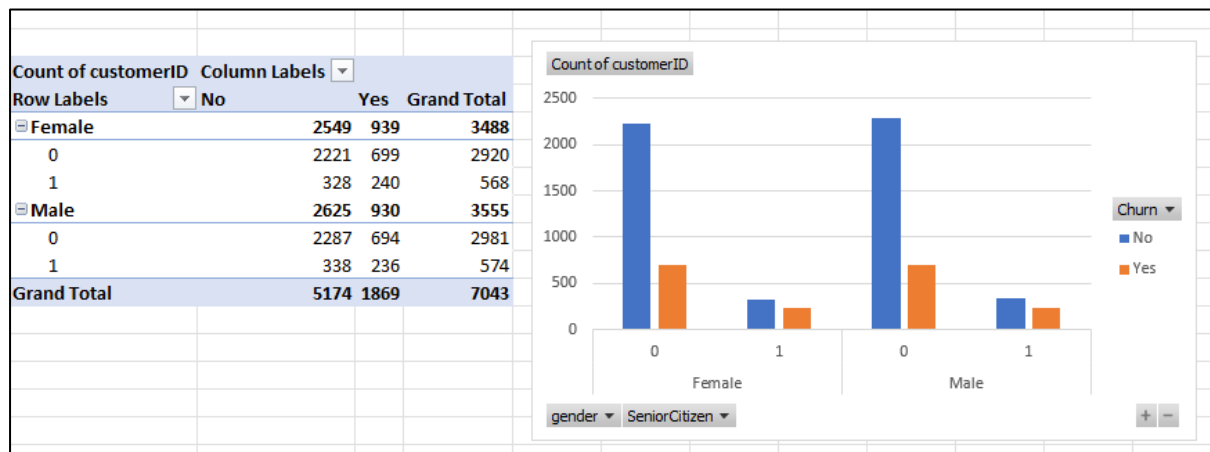


Figure 28. Non\_churned Customer Streaming movies

## 7 Total Customer Tenure by Internet Service Type

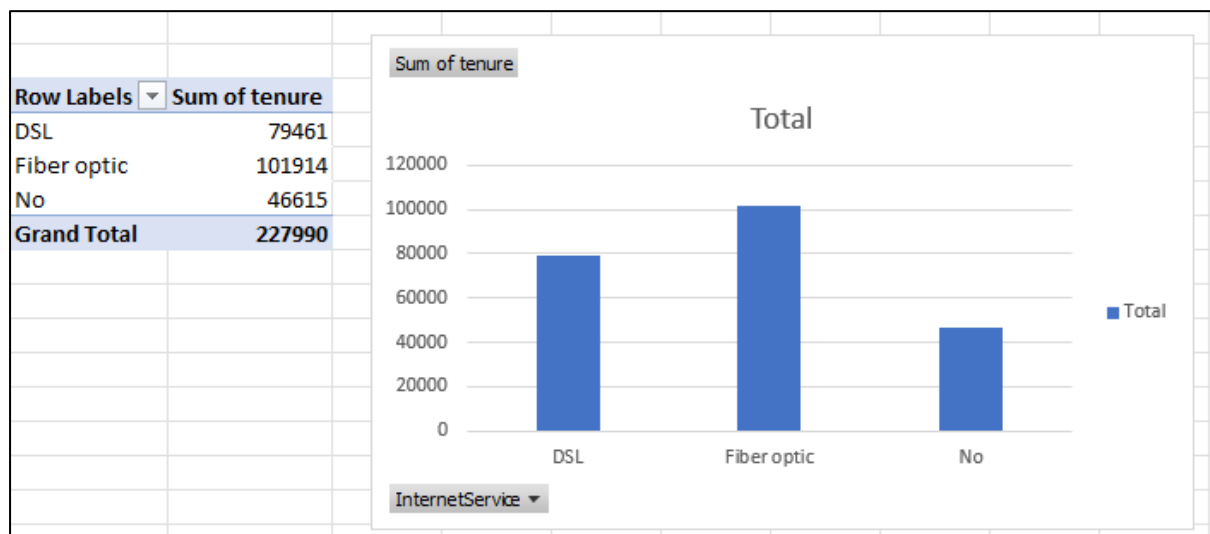


Figure 29. Total Customer Tenure by Internet Service Type

## 8 Churn Analysis Across Internet Service Categories

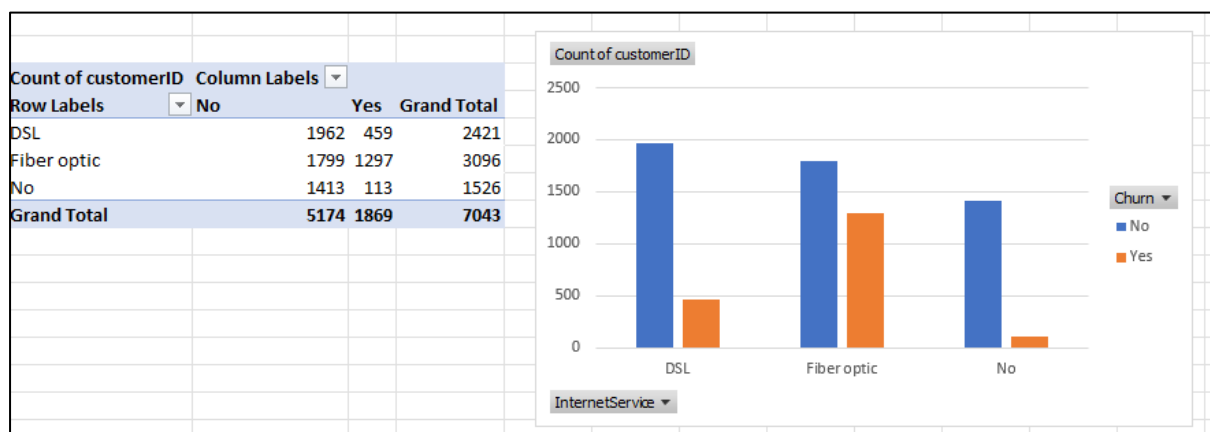


Figure 29. Customer Churn Distribution by Internet Service Type



## 9 Customer Churn by Senior Citizen Status

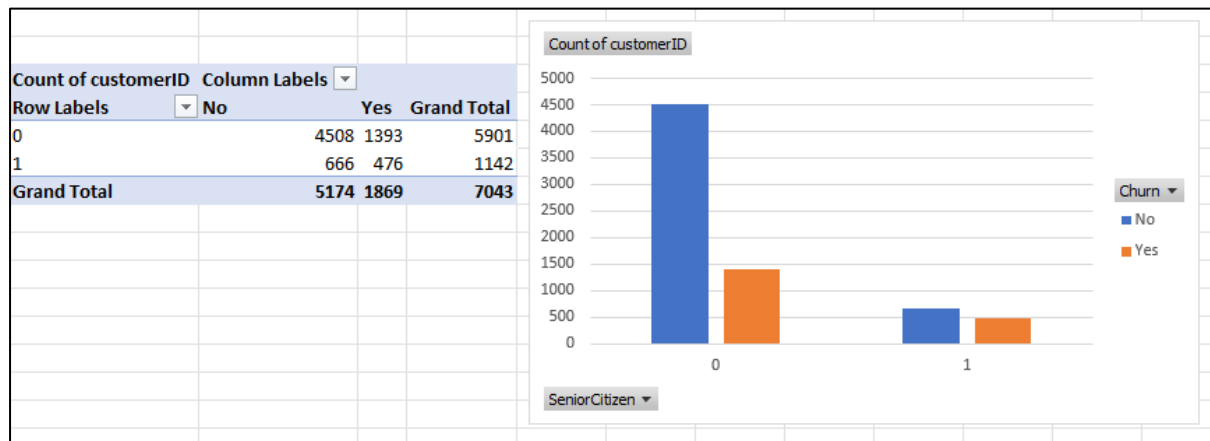


Figure 30. Churn Comparison Between Senior and Non-Senior Customers

The chart shows that **non-seniors (0)** have a higher customer base and churn volume, but **seniors (1)** have a **higher churn rate** proportionally (476 out of 1,142  $\approx$  41.7%).

## 10 Total Revenue by Internet Service Type

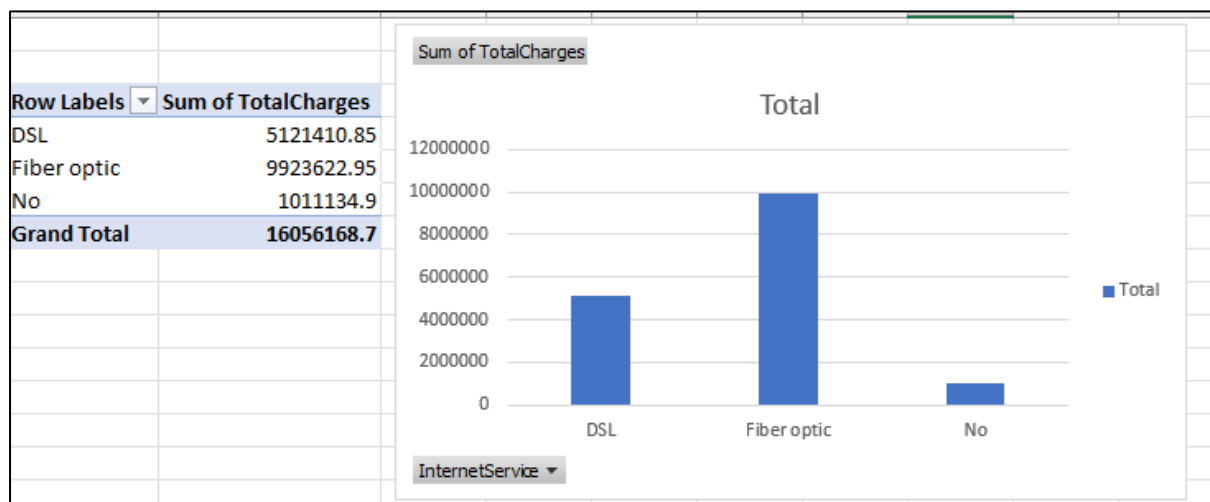


Figure 31. Total Charges Collected by Internet Service Type

## Power BI

### 1 Customer Churn Breakdown by Internet Service Type

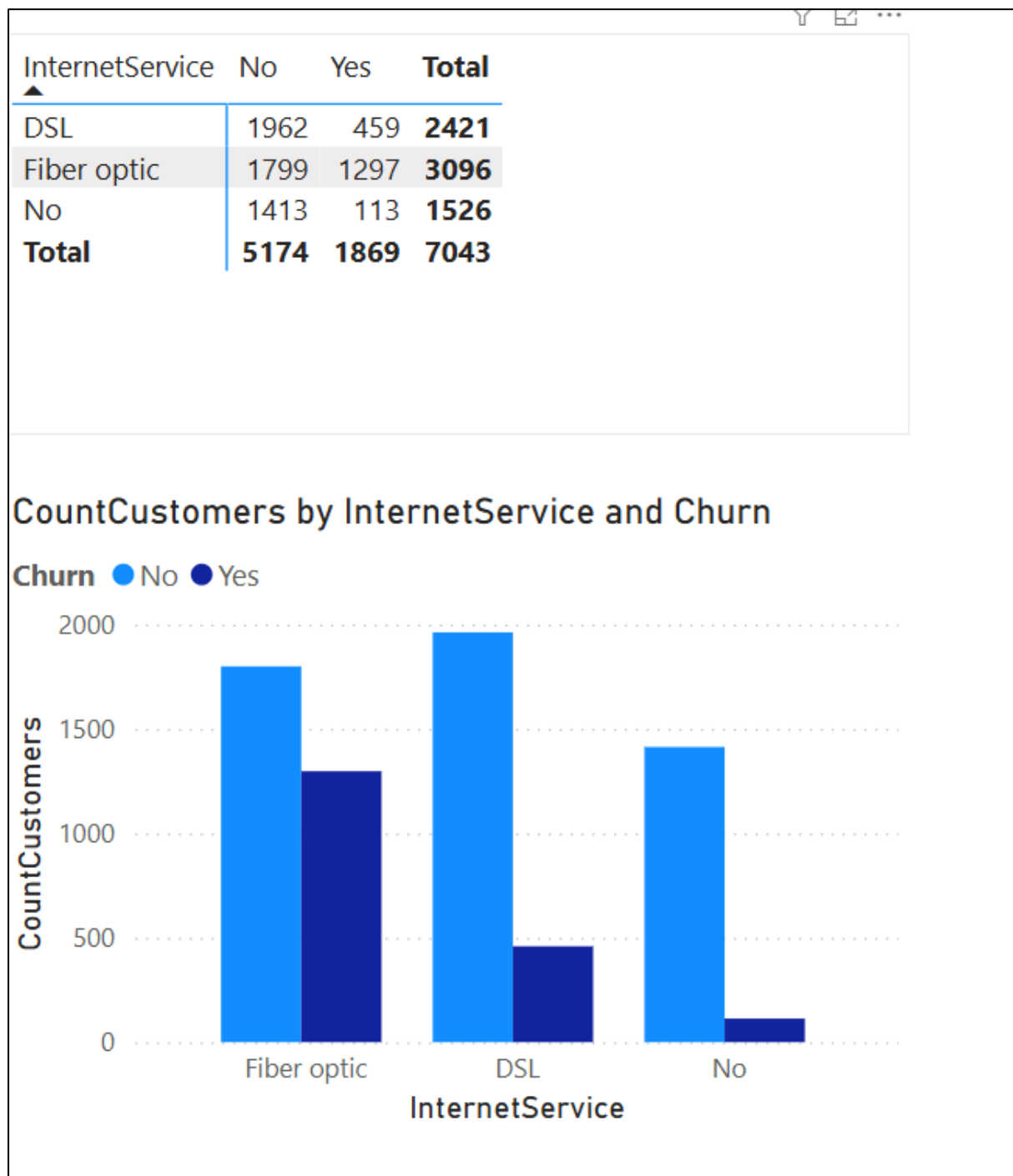


Figure 32. Number of Churned and Retained Customers by Internet Service

## 2 Total Revenue by Internet Service Type

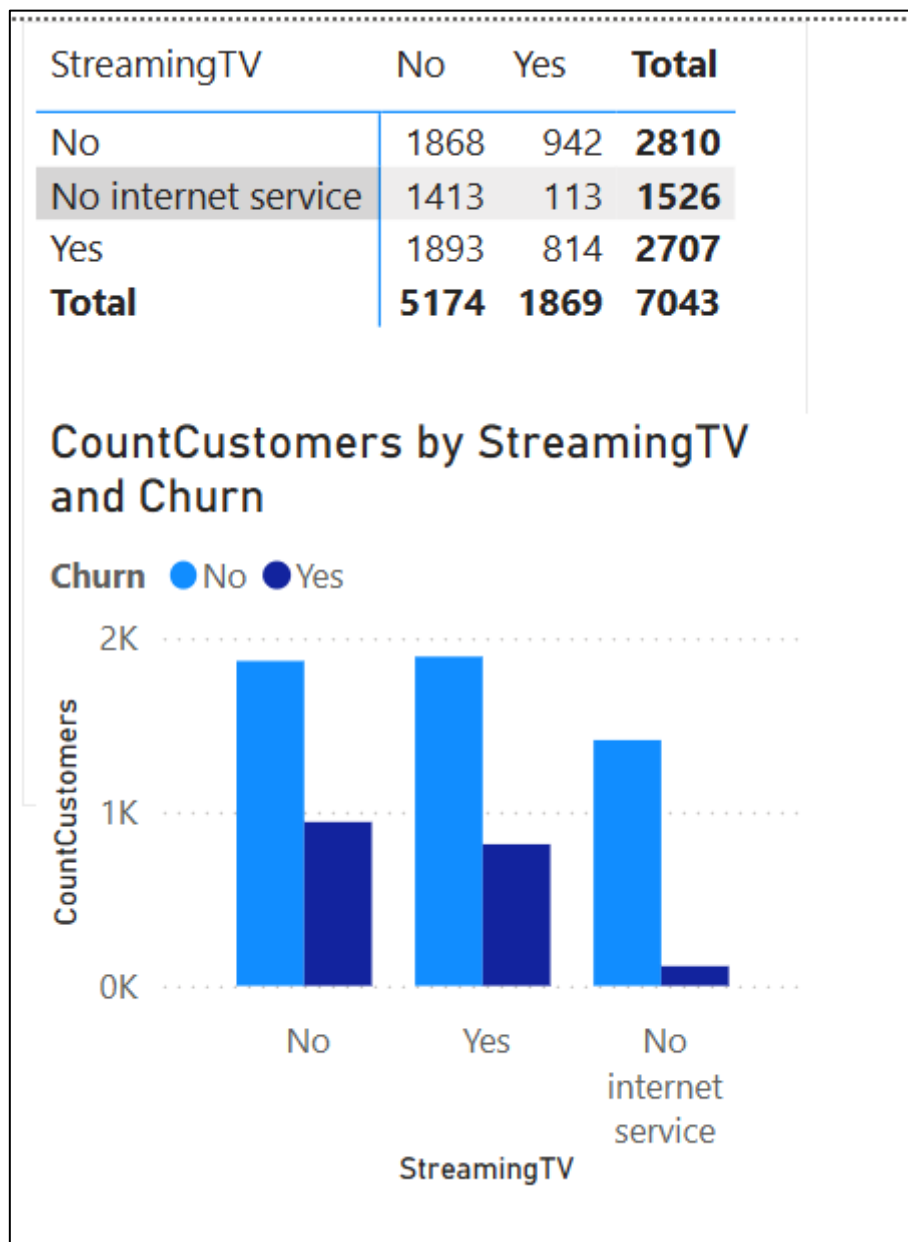


Figure 33. Total Charges Collected by Internet Service Type

### 3 Customer Churn by Tech Support Availability

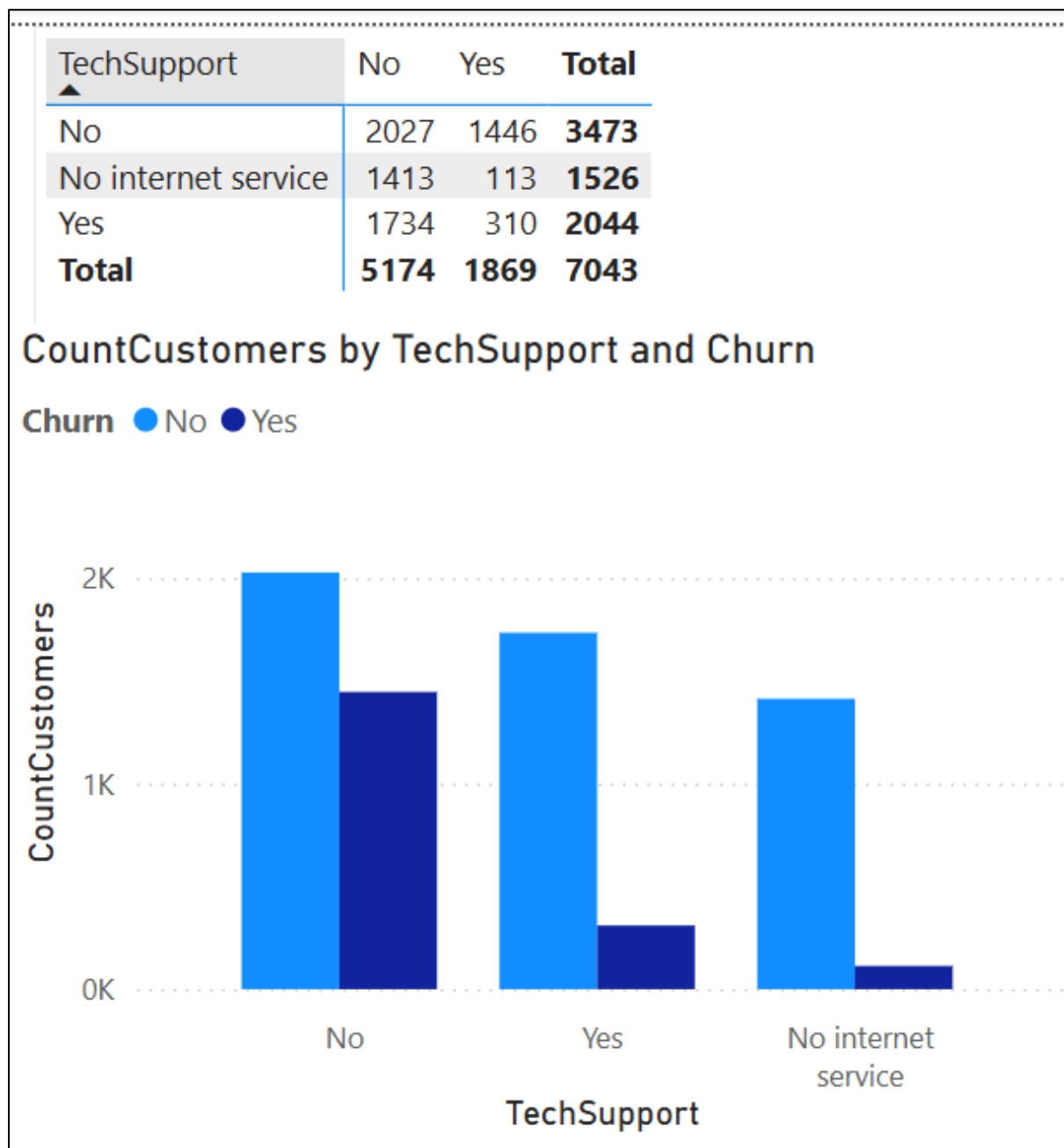


Figure 34. Churn Distribution Based on Access to Tech Support Services

#### 4 Customer Churn by Phone Service Subscription

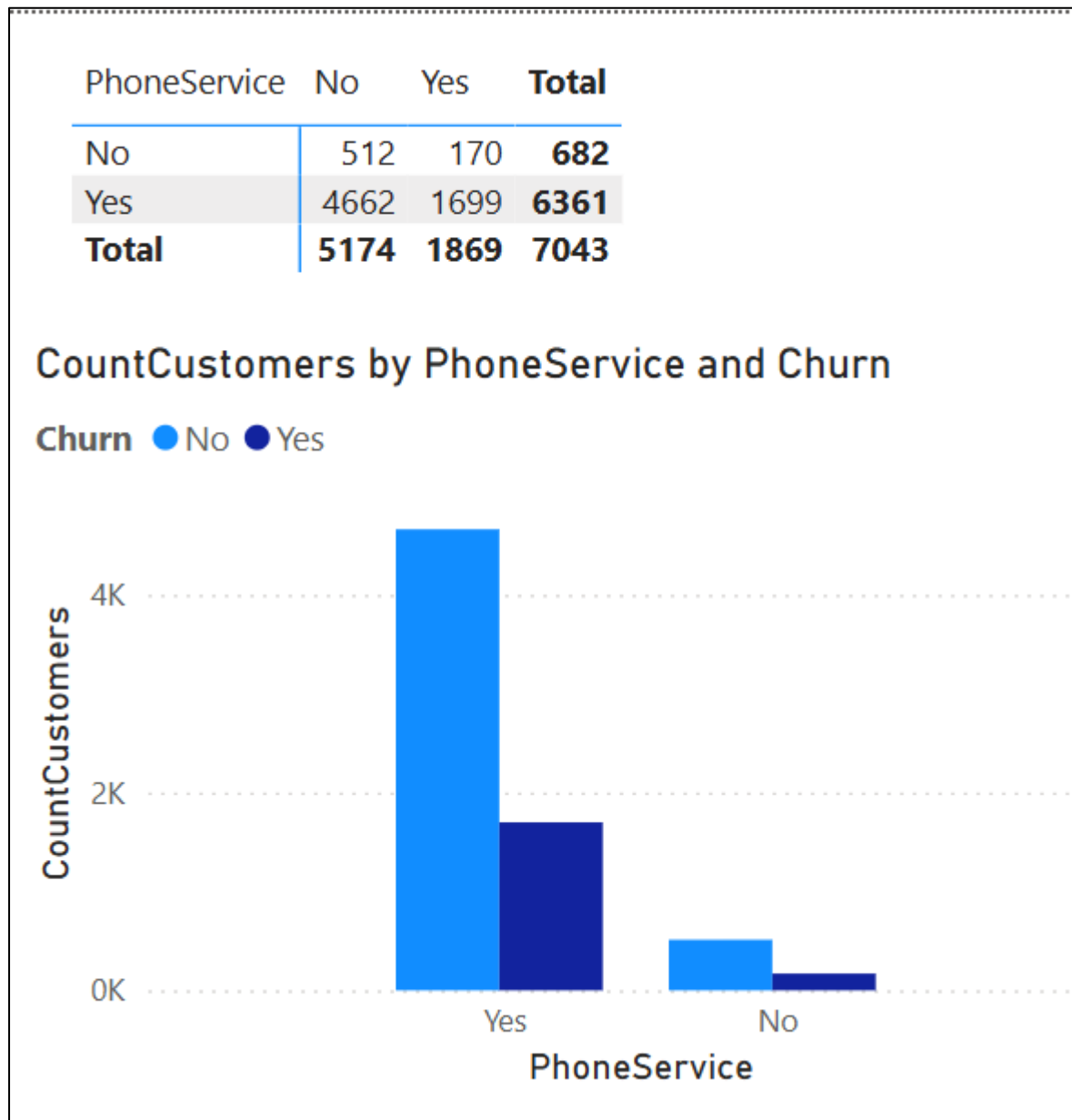


Figure 35. Comparison of Churned and Retained Customers by Phone

## 5 Customer Churn by Contract Type

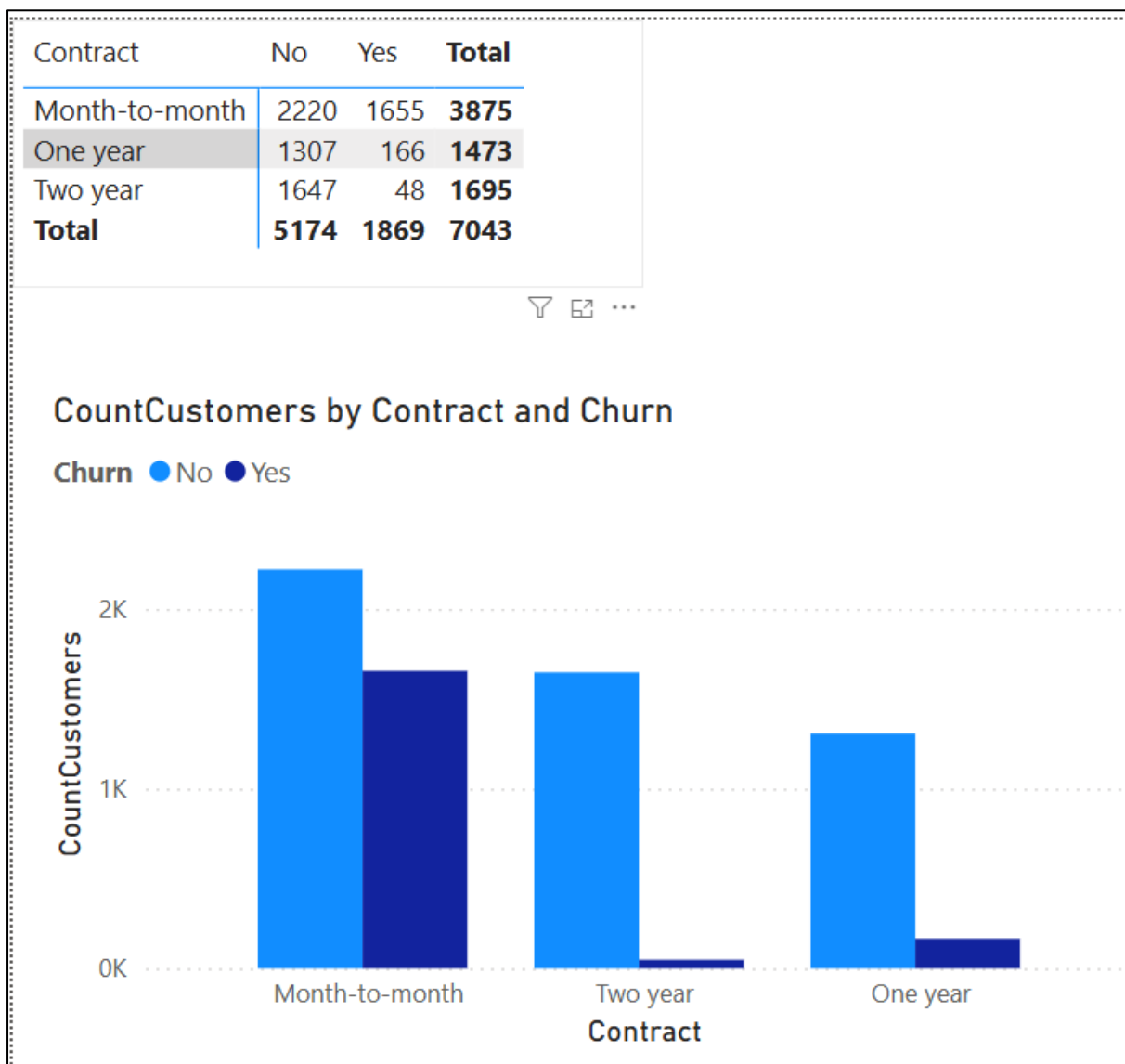


Figure 37. Churn Distribution Across Contract Durations

## 6 Customer Churn by Gender and Senior Citizen Status

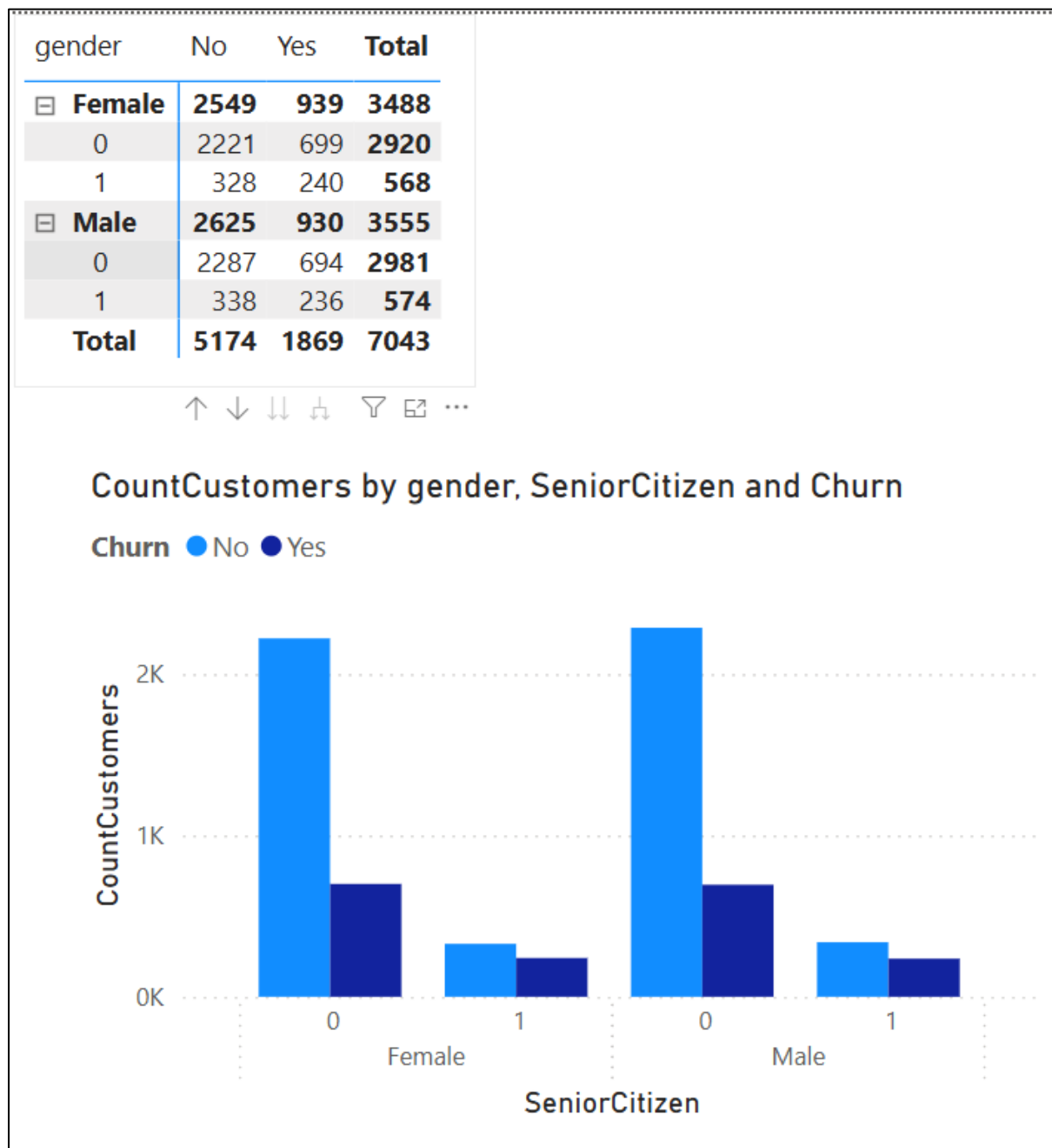


Figure 38. Churn Distribution by Gender Combined with Senior Citizen Status

## 7 Total Customer Tenure by Internet Service Type

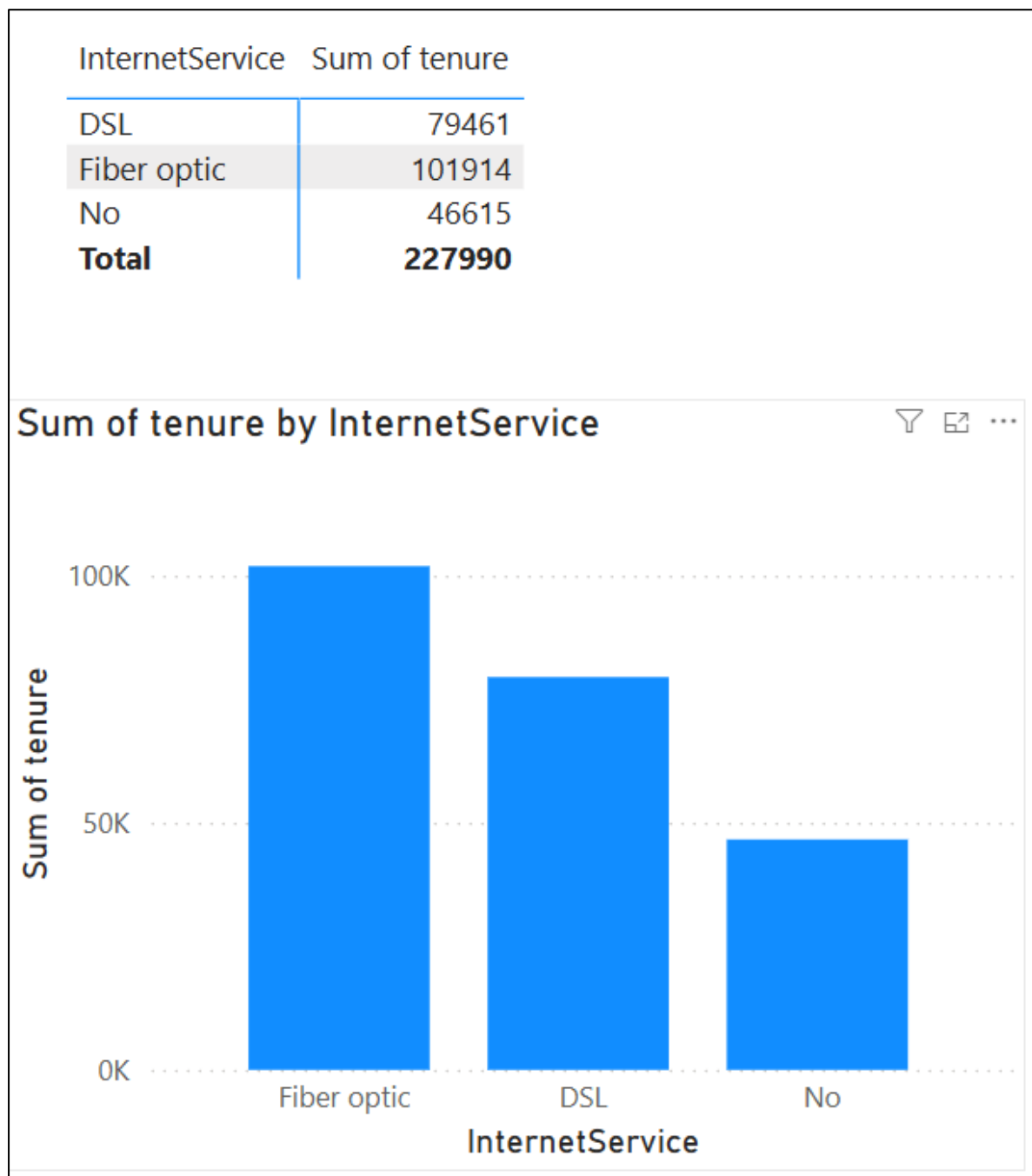


Figure 39. Aggregated Tenure of Customers Based on Internet Service



## 8 Customer Churn Distribution by Internet Service Type

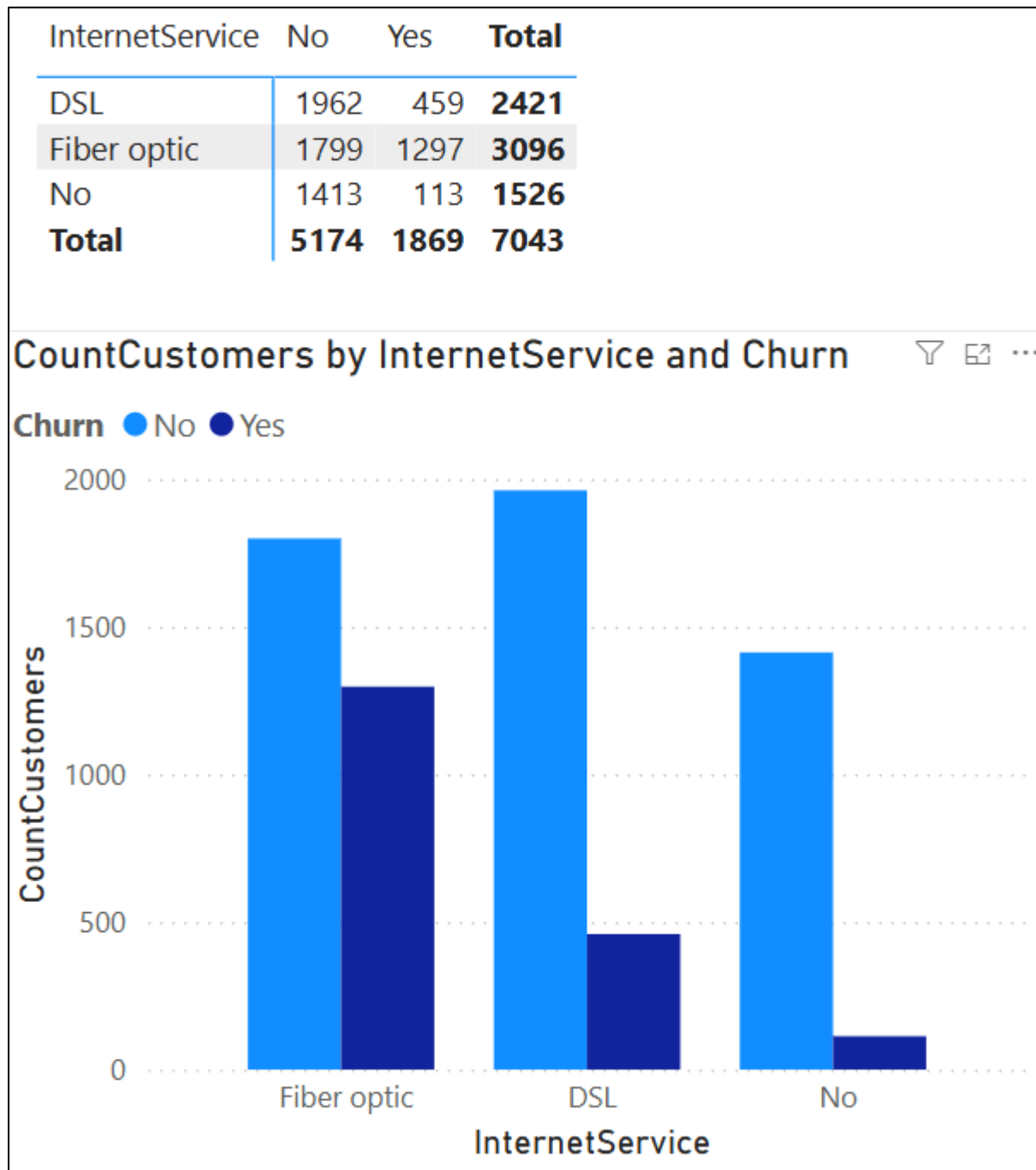


Figure 40. Count of Churned and Retained Customers per Internet Service Category

## 9 Customer Churn by Senior Citizen Status

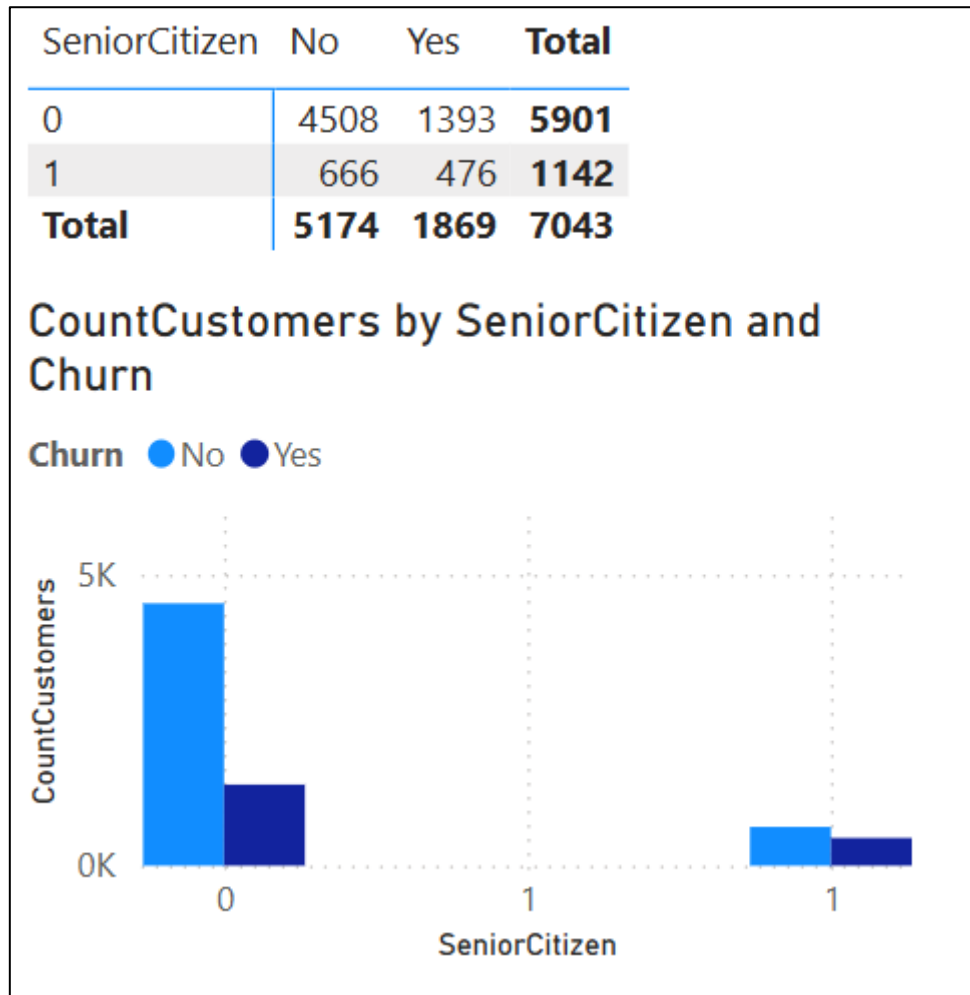


Figure 41. Count of Churned and Retained Customers Based on Senior Citizen Classification

## 10 Total Revenue by Internet Service Type

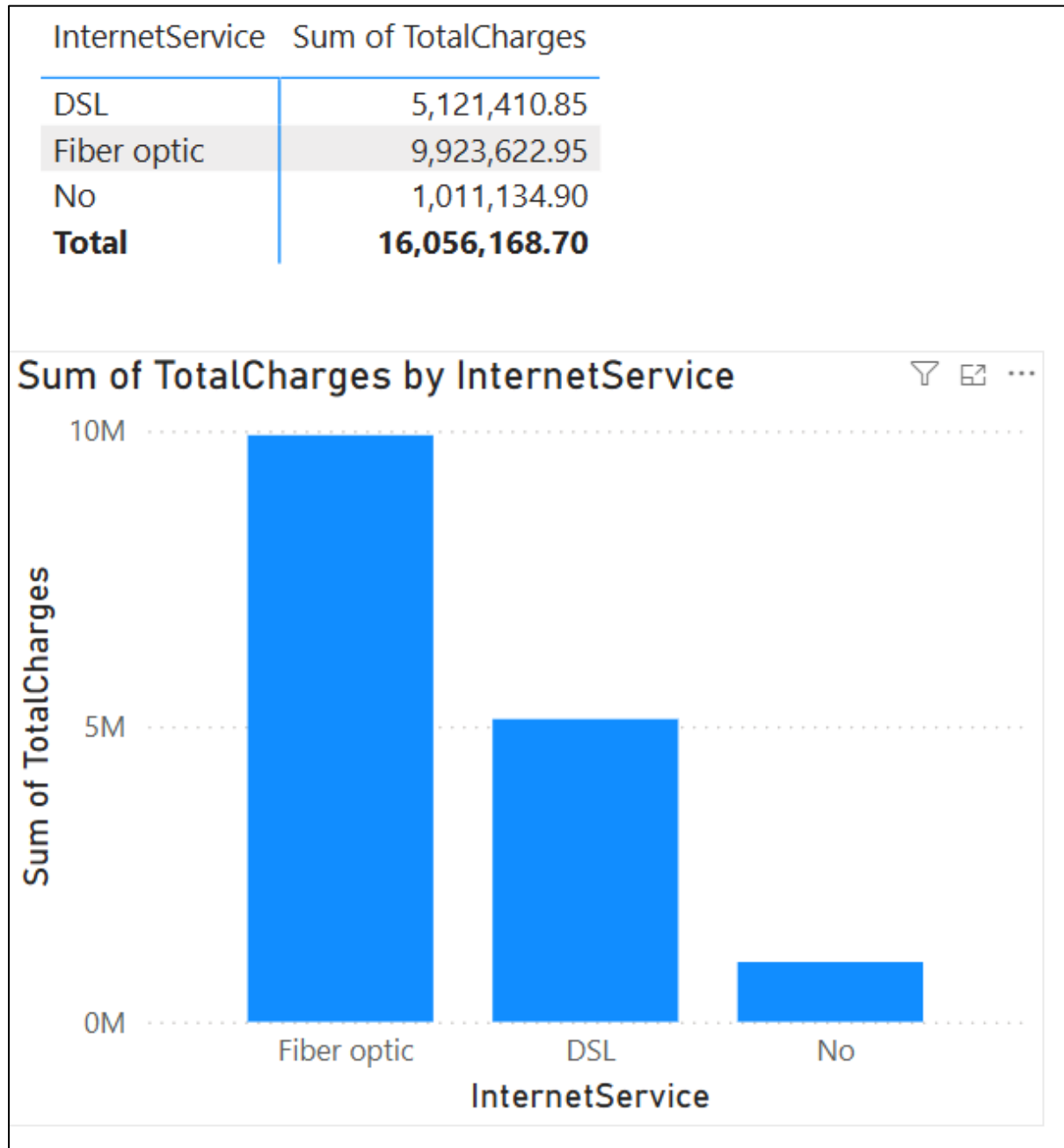


Figure 42. Sum of Total Charges Across Internet Service Categories

## **Chapter 3: Data mining**

### **3.1 Algorithms and Deep learning**

#### **Logistic Regression**

Description: A linear classification algorithm that predicts the probability of a binary target variable (e.g., Churn = Yes/No) using the sigmoid function. Suitable for problems with a linear relationship between features and the outcome.

#### **Gradient Boosting**

Description: An ensemble-based machine learning algorithm that builds decision trees sequentially, where each tree corrects the errors of the previous one by optimizing the loss function using gradient descent.

#### **LightGBM**

Description: An optimized variant of Gradient Boosting that uses a histogram-based tree structure and a leaf-wise approach (focusing on developing the leaf node with the largest loss). Developed by Microsoft, it is highly efficient for large datasets.

### **3.2 Results**

#### **Initial influencing factor**

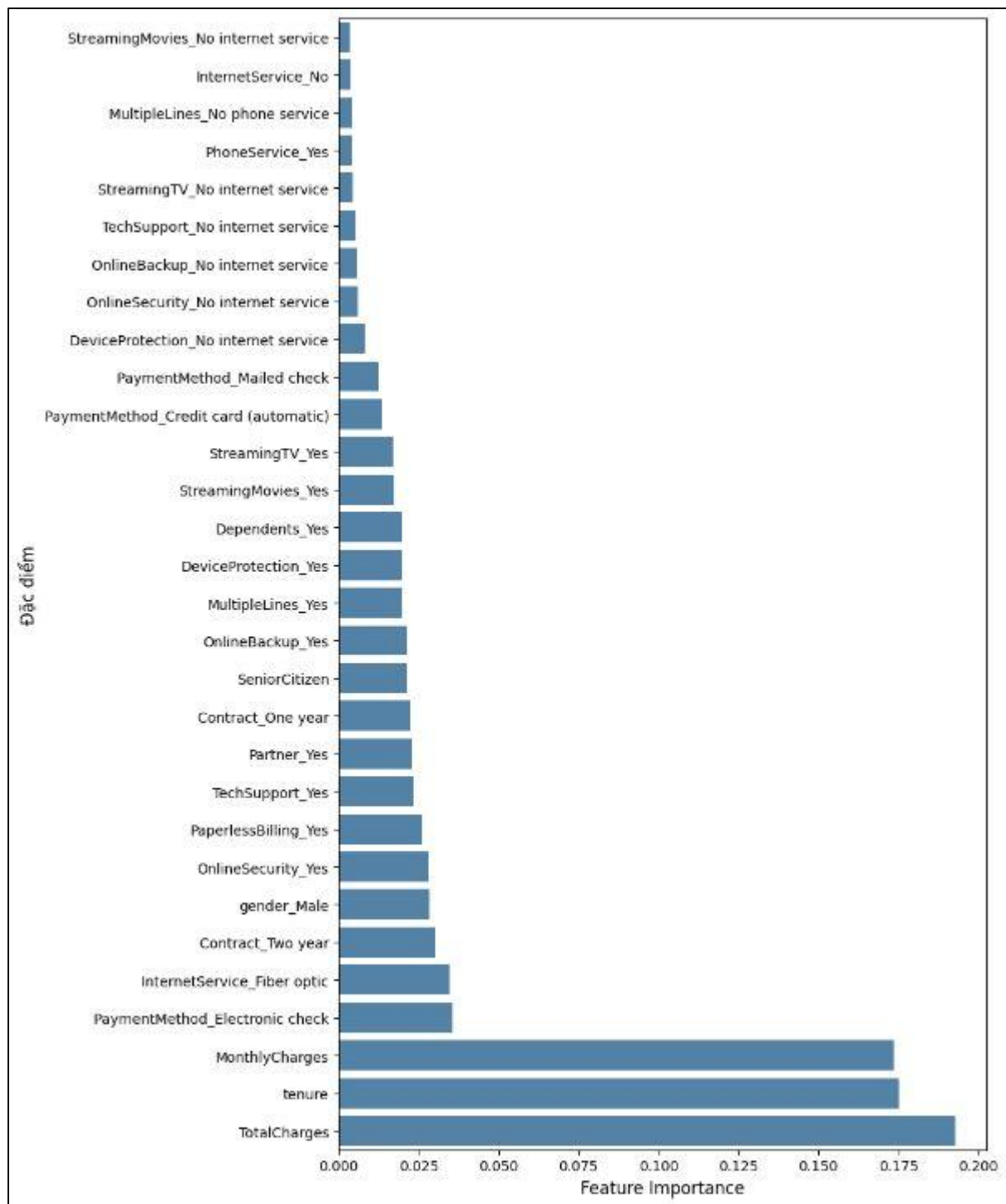


Figure 43. Ranked Feature Importance for Churn Prediction Model

## Model influencing factors

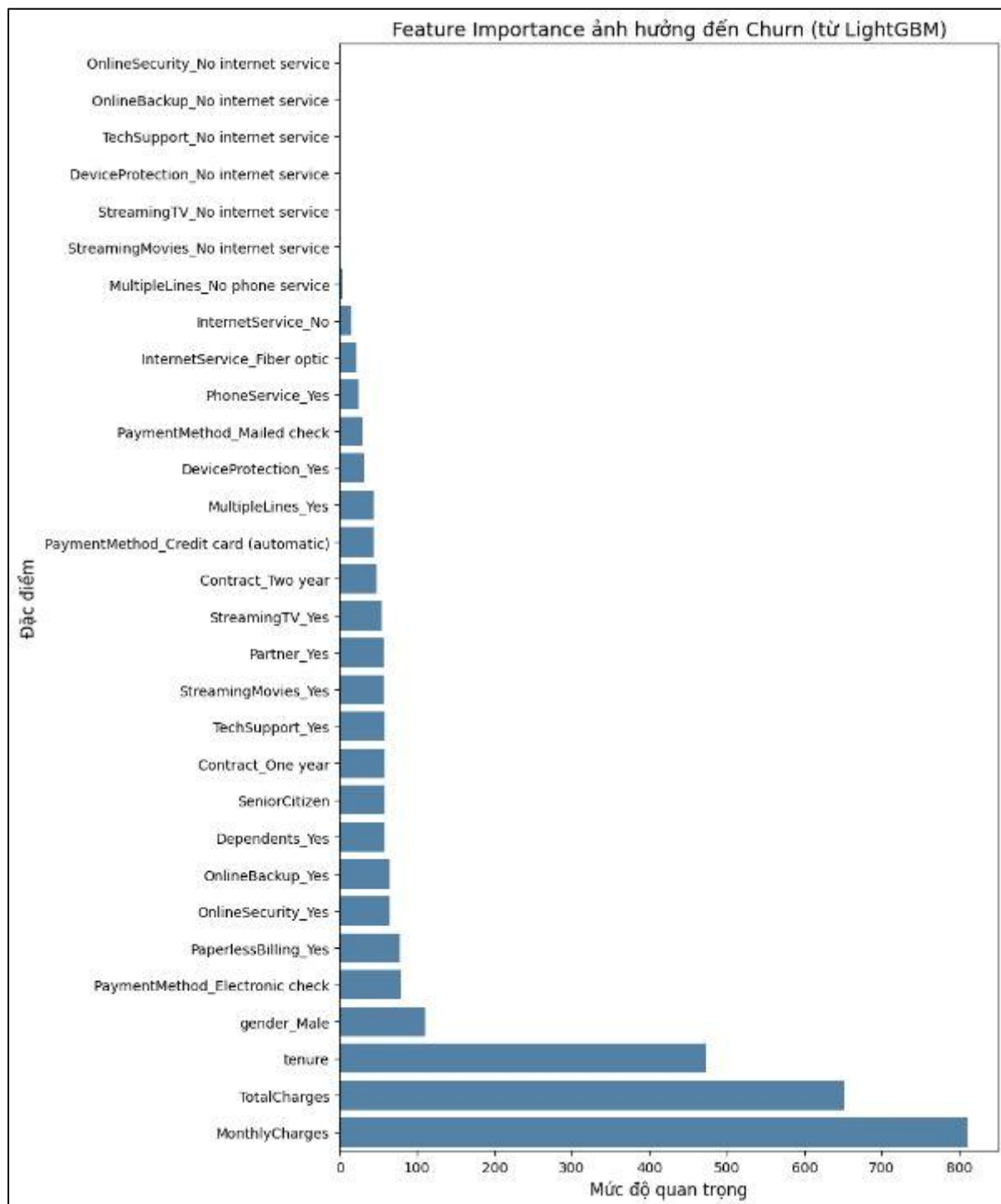


Figure 44. Feature Importance for Customer Churn Prediction (LightGBM Model)

The three most important features are:

- **MonthlyCharges**
- **TotalCharges**

- **Tenure**

These billing-related variables play a dominant role in determining customer churn. Notably, MonthlyCharges is the most influential, suggesting that how much a customer is currently being billed is a strong signal of churn behavior.

Moderate Influencers:

Features with moderate influence include:

- **Gender\_Male**
- **PaymentMethod\_Electronic check**
- **PaperlessBilling\_Yes**
- **OnlineSecurity\_Yes**
- **SeniorCitizen**
- **Contract\_One year / Two year**

These reflect customer demographics, subscription terms, and service-related behaviors.

Low Impact Features:

Several service-related binary flags (mostly "No internet service" combinations) show very low importance, such as:

- **StreamingMovies\_No internet service**
- **StreamingTV\_No internet service**
- **OnlineBackup\_No internet service**
- **OnlineSecurity\_No internet service**

This suggests that the absence of internet service renders related streaming and protection options irrelevant for churn prediction.

## Prediction results

```

Logistic Regression Accuracy: 0.7875
Gradient Boosting Accuracy: 0.7896
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Info] Number of positive: 1495, number of negative: 4130
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.001435 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 637
[LightGBM] [Info] Number of data points in the train set: 5625, number of used features: 30
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.265778 -> initscore=-1.016151
[LightGBM] [Info] Start training from score -1.016151
LightGBM Accuracy: 0.7910

Best Model: LightGBM với accuracy 0.7910

So sánh Churn gốc và Churn dự đoán:
Churn predicted_Churn
0 0 1
1 0 0
2 1 0
3 0 0
4 1 1
5 1 1
6 0 0
7 0 0
8 1 1
9 0 0

Tỷ lệ dự đoán đúng trên toàn bộ dữ liệu: 0.8639

```

Figure 45. Model Accuracy

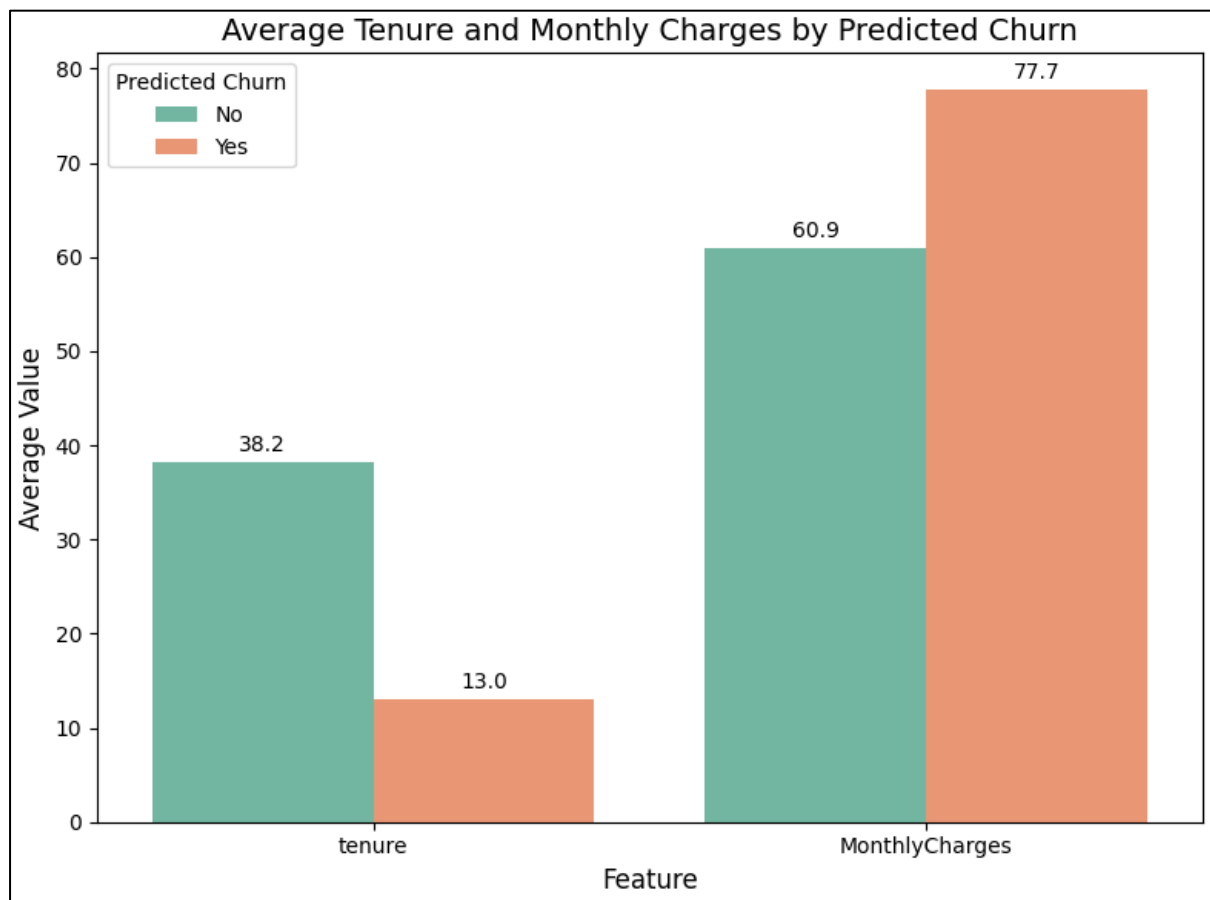
LightGBM outperformed other models and achieved a prediction accuracy of 86.39% on the full dataset, making it a strong candidate for customer churn prediction based on the given features and data.

## Model Accuracy Comparison

Model	Accuracy
Logistic Regression	78.75%
Gradient Boosting	78.96%
LightGBM	79.10%



## Prediction graph to make law



### Tenure Rule:

Customers with an average duration of less than 13 months (churn mean) are significantly more likely to churn than those with a duration of more than 38 months (non-churn mean).

Rule: The shorter the duration of service, the more likely the customer is to churn.

### Monthly Charges Rule:

Customers with an average monthly cost above \$77.7 (churn mean) are more likely to churn than those with a cost below \$60.9 (non-churn mean).

Rule: The higher the monthly cost, the more likely the customer is to churn.