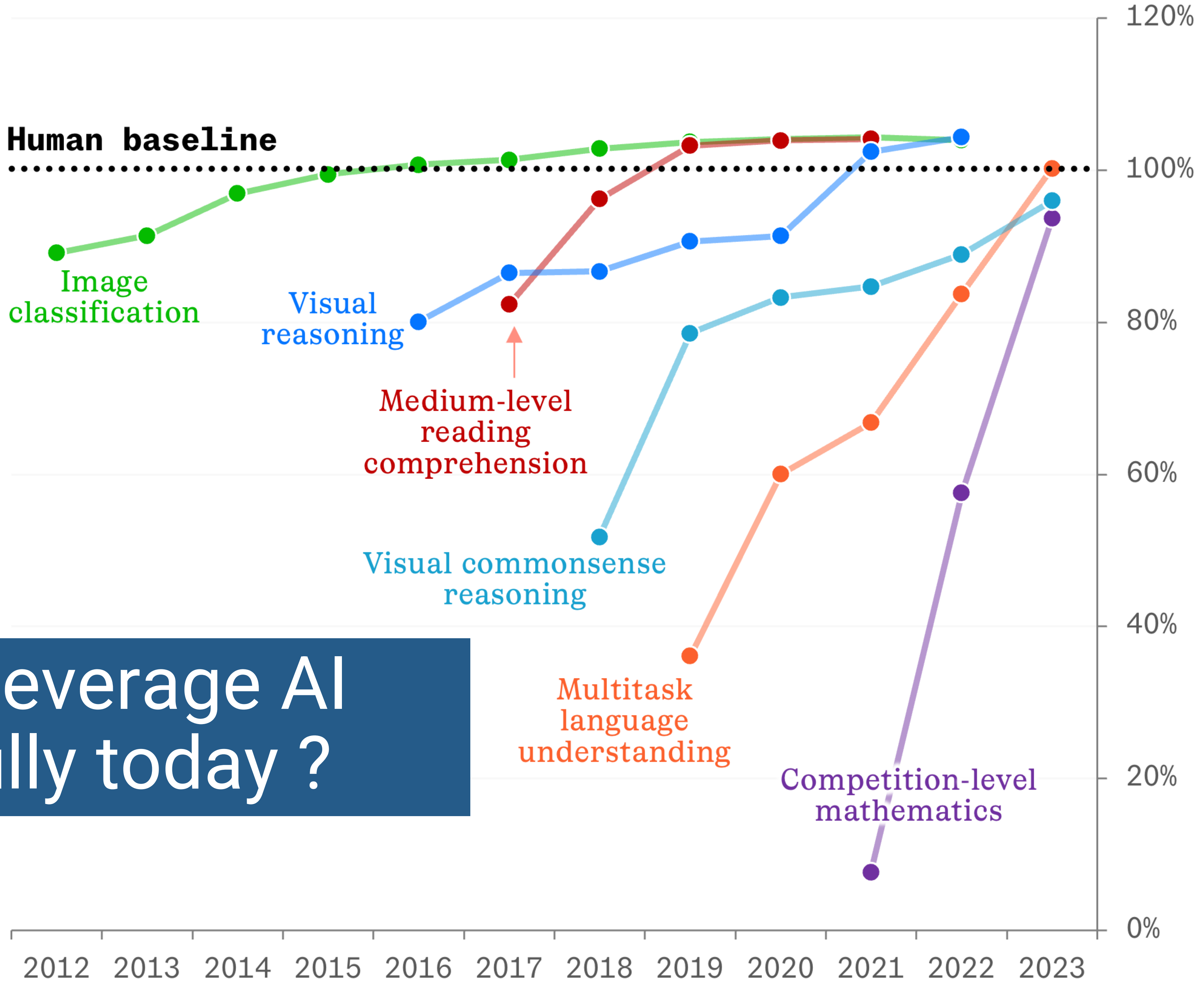
CreaXum

AI & Analytics

Grow, Convince, Inspire

Our Business Intelligence Offering

AI Technical Performance



How can I leverage AI meaningfully today ?

Stanford University AI Index Report 2024

CreaXum

Images

DALL-E	https://openai.com/dall-e-3
Stability	https://stability.ai/stable-image
Adobe Firefly	https://firefly.adobe.com
Midjourney	https://www.midjourney.com

Text

ChatGPT	https://chat.openai.com/
Copilot	https://copilot.microsoft.com/
Gemini	https://gemini.google.com/
Claude	https://www.claude.ai/

Forbes

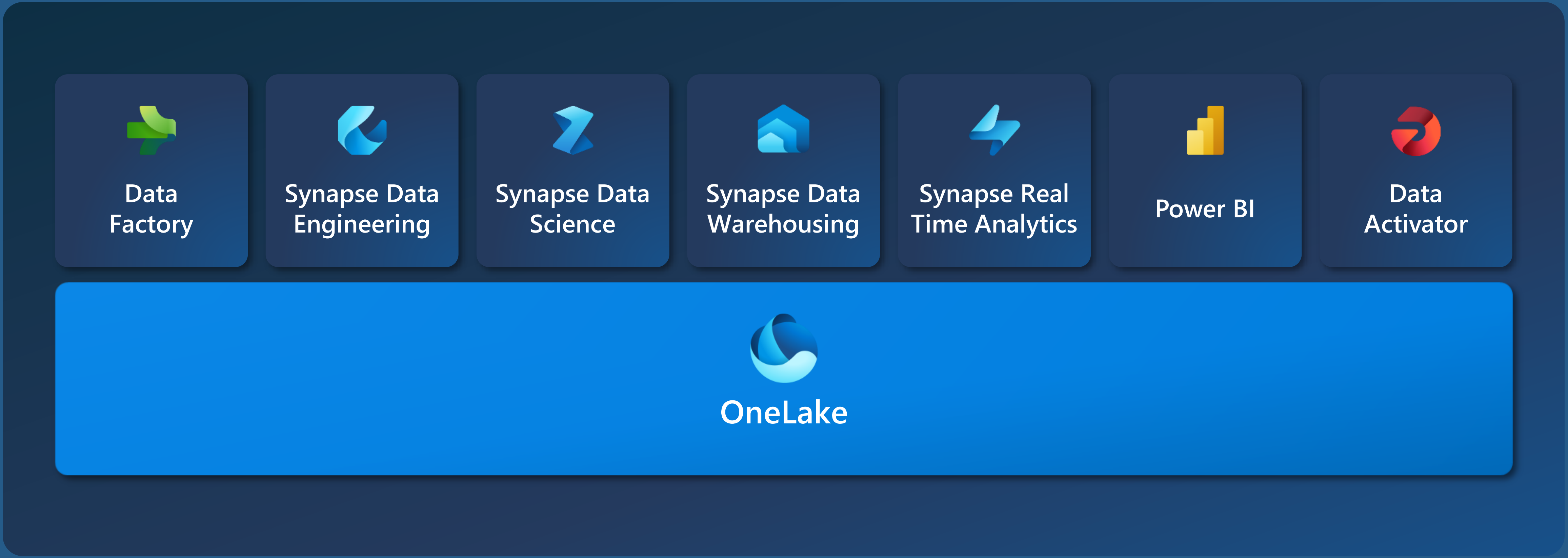
And Analytics ?

"Simplify,
I am the Chief Information Officer and don't want to be the Chief Integration Officer."

Every CIO, Every Enterprise

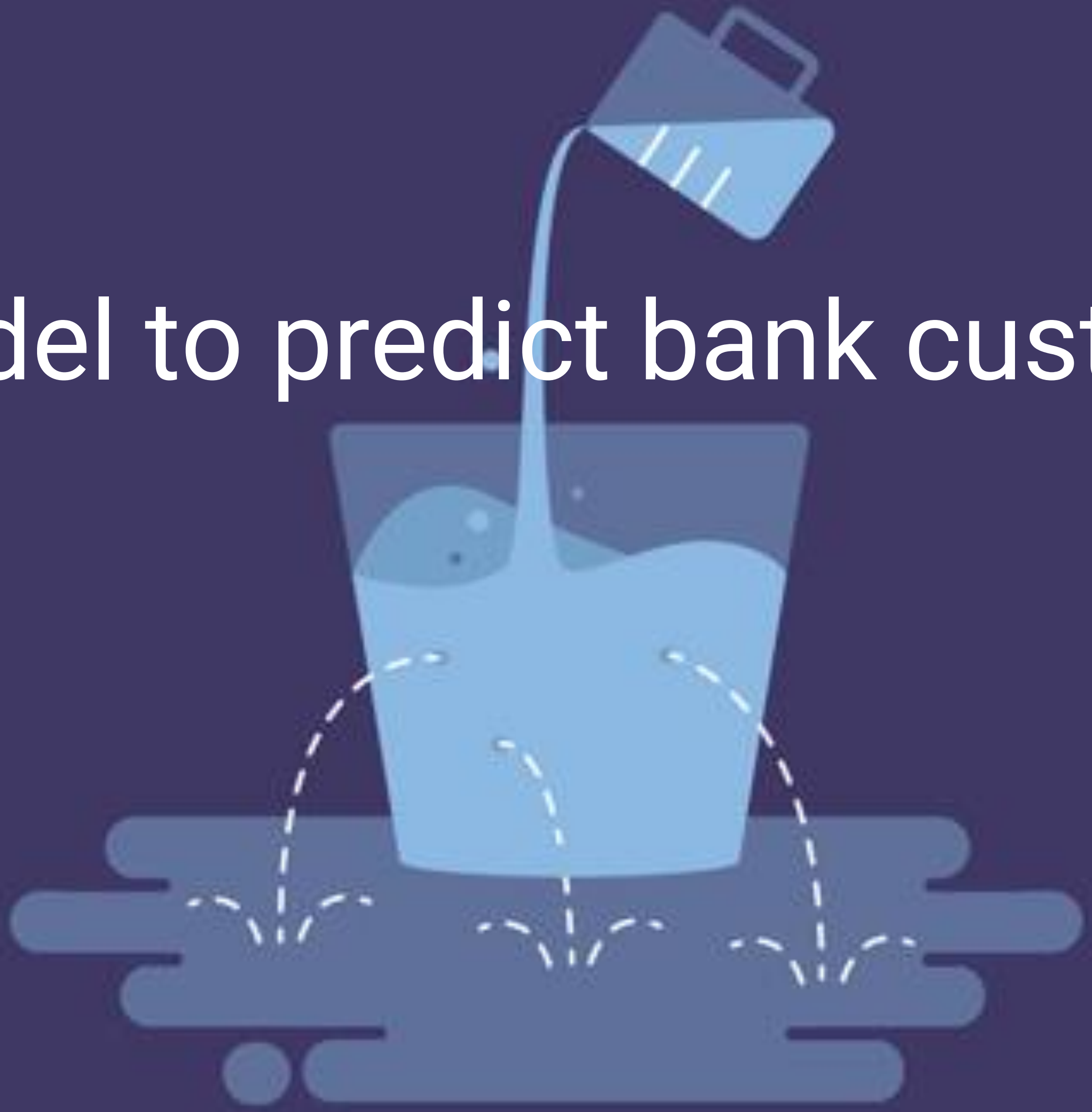
Microsoft Fabric

The unified data platform for the era of AI



Demo time !

Build a model to predict bank customer churn

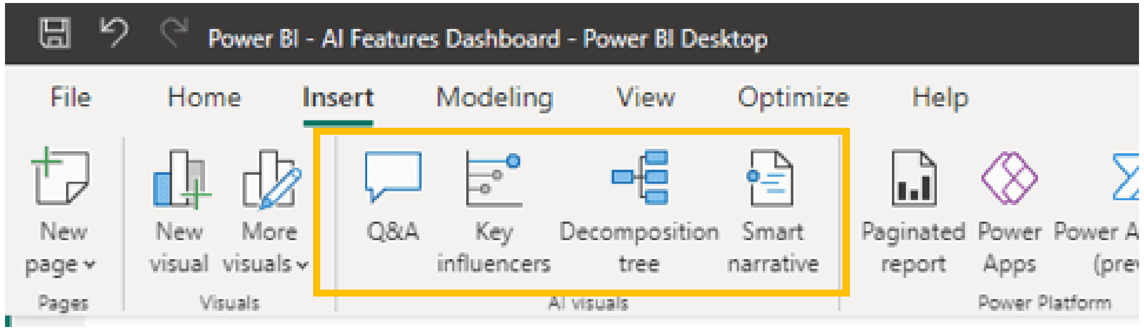


The churn rate, also known as the rate of attrition refers to the rate at which bank customers stop doing business with the bank.

AI visuals

Advanced AI Visuals in Power BI

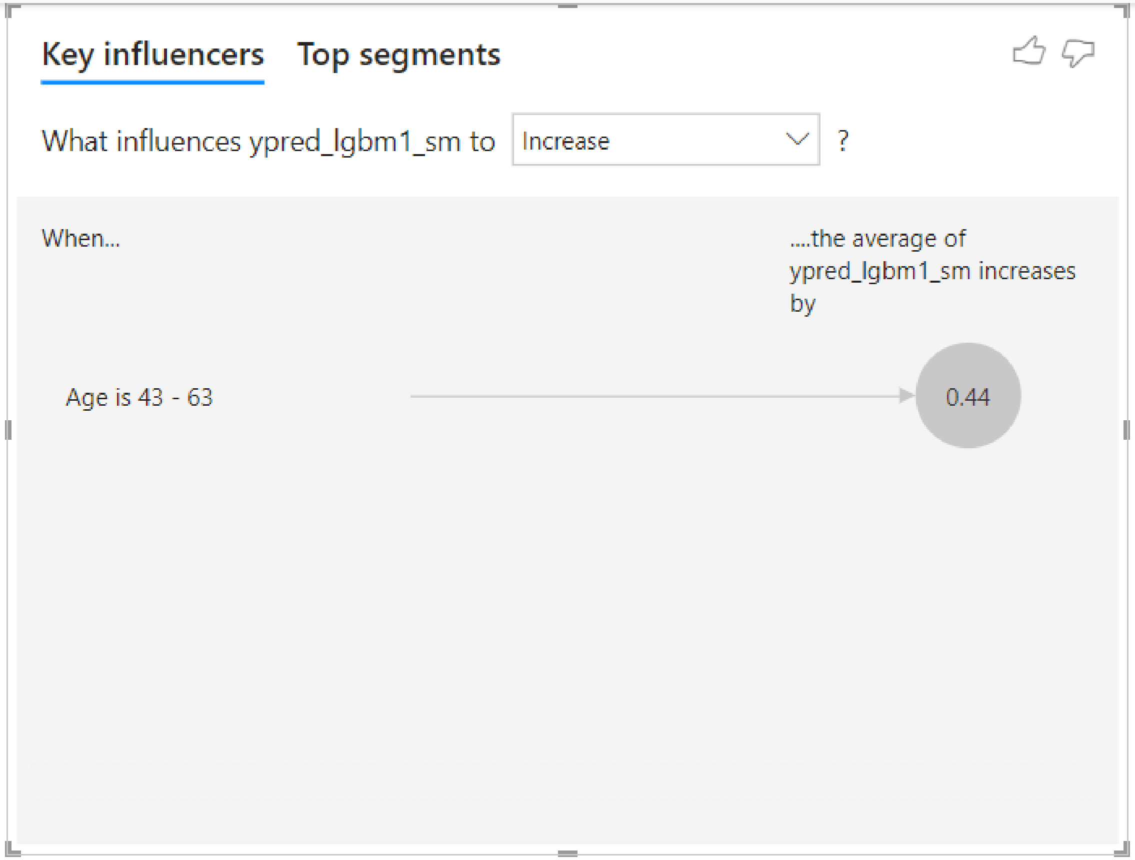
Power BI has 4 different AI powered visuals that are separate from the regular visuals in Power BI. They can be accessed from the Power BI ribbon under **Insert > AI Visuals.**

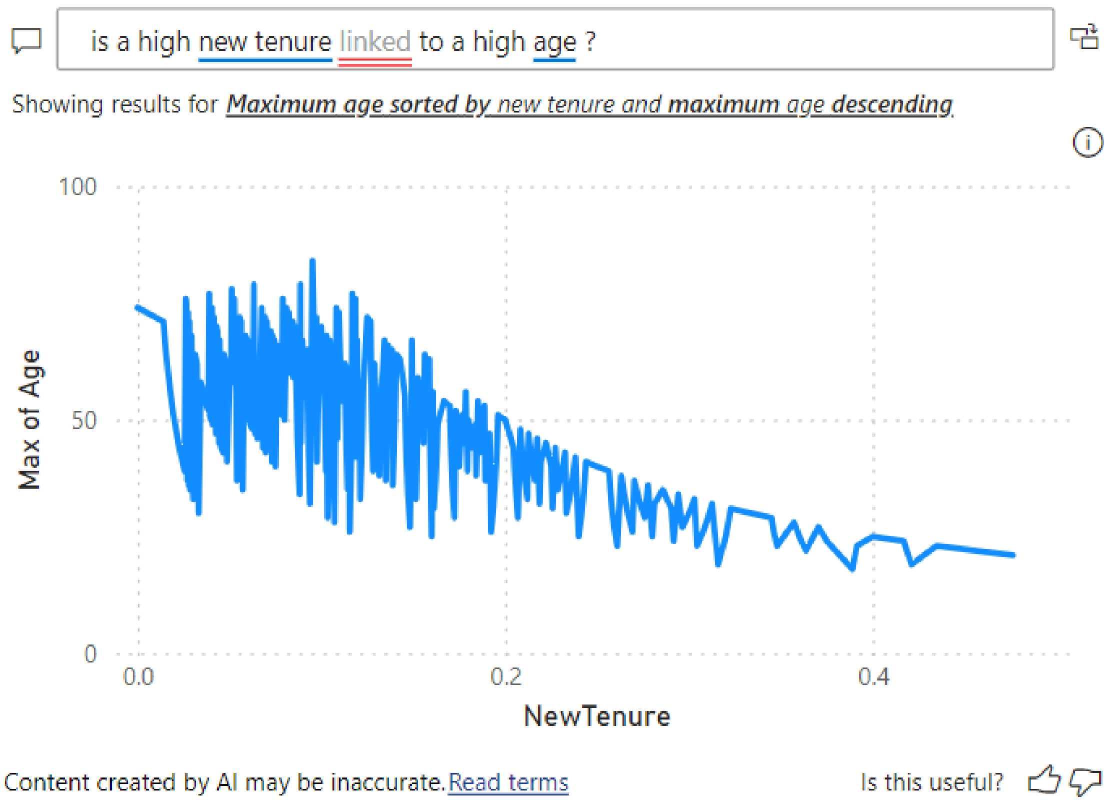


Each one is designed to help users explore their data in different ways.

- **Q&A –** Allows you to use plain language to ask questions about data.
- **Key Influencers** – Helps identify factors that drive a metric of interest.
- **Decomposition Tree** – Explore data across multiple dimensions and easily drill into details.
- **Smart Narrative** – Summarizes data and places insights into plain language.

CreaXum

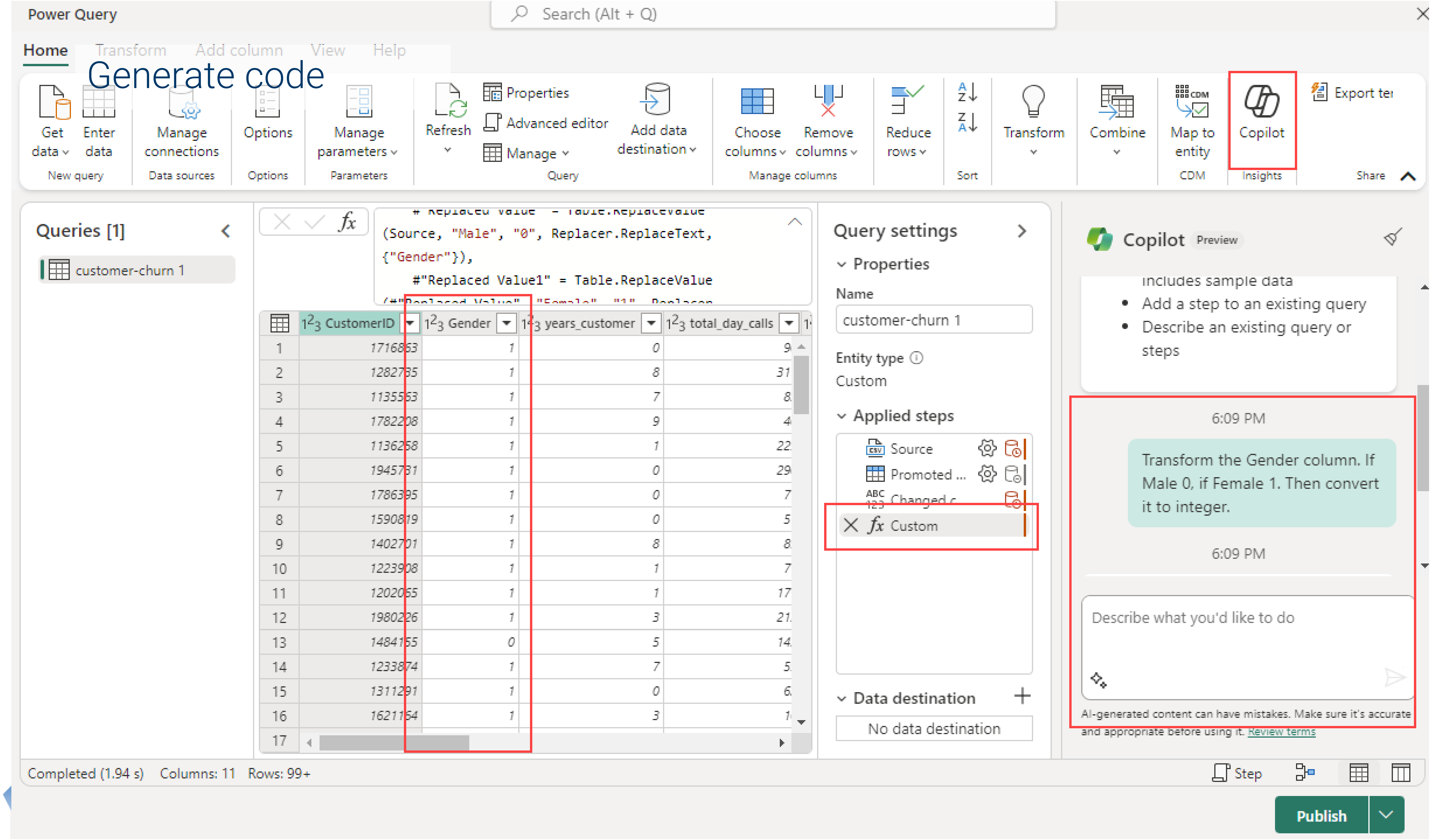




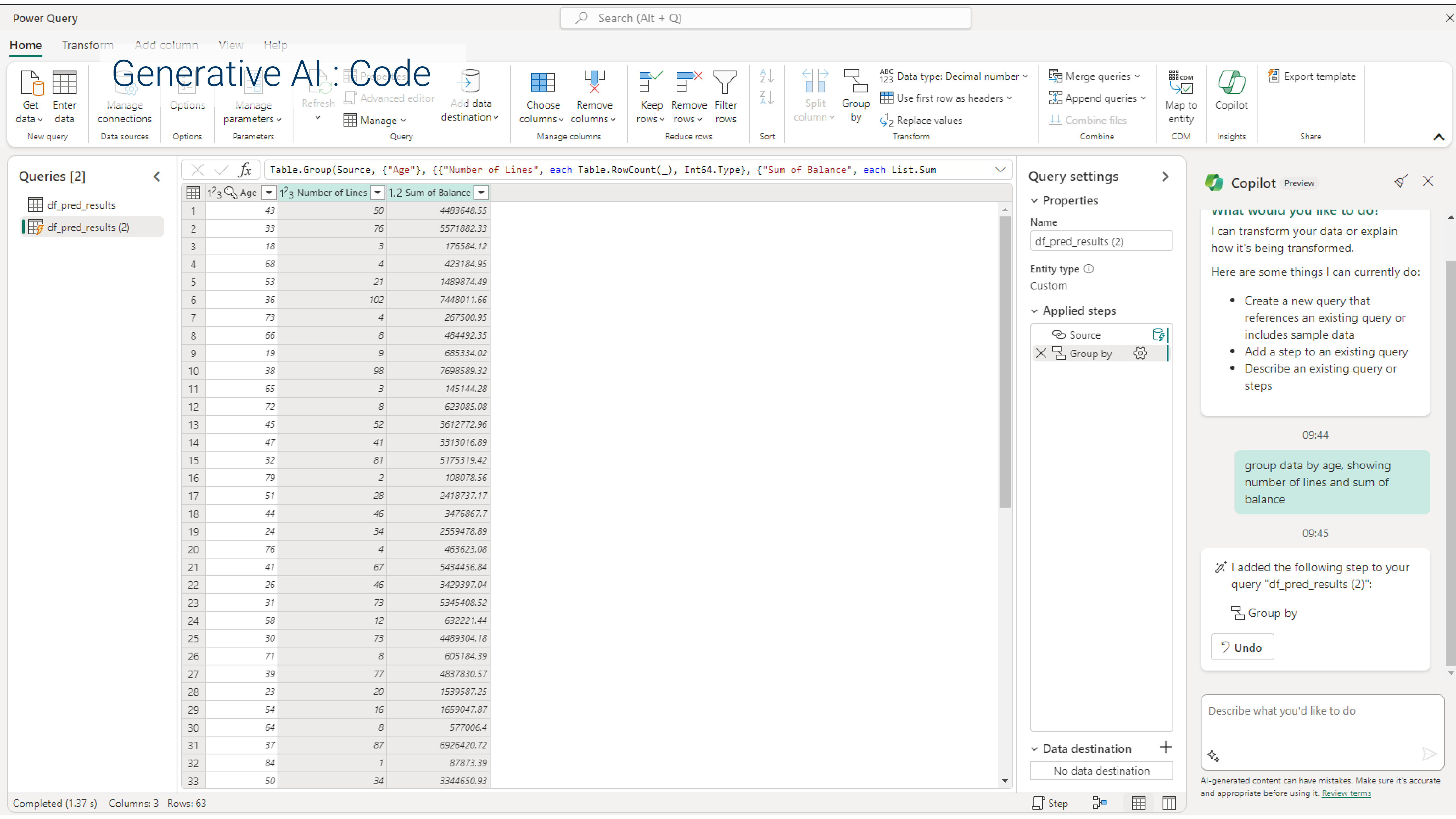
The key influencer visual presents data that focuses on age as a significant metric. The key takeaway from the available data is that when the age is between 43 and 63, there is a noticeable increase in the average ypred_lgbm1_sm by approximately 0.44 units, which is higher compared to other age values. Furthermore, this particular age group comprises around 23.10% of the total data set. **The correlation between age and ypred_lgbm1_sm can be significant for decision-making processes in business settings.** [1]

CreaXum

Generate code



Generative AI : Code



Generative AI : Report

Build your first report

- Add and prepare your data
- Generate a premade report
- Customize to suit your needs

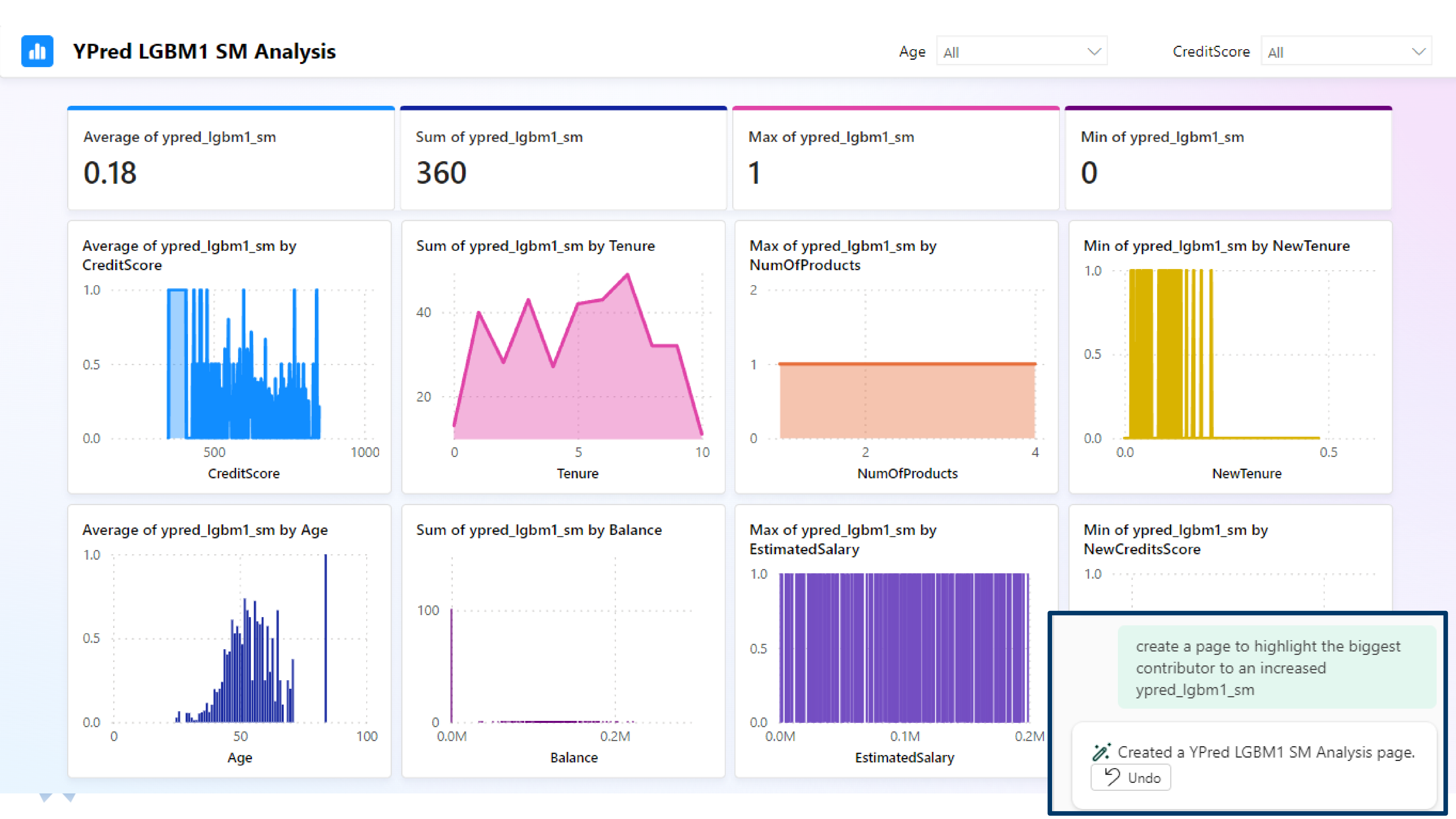
Add data to start building a report

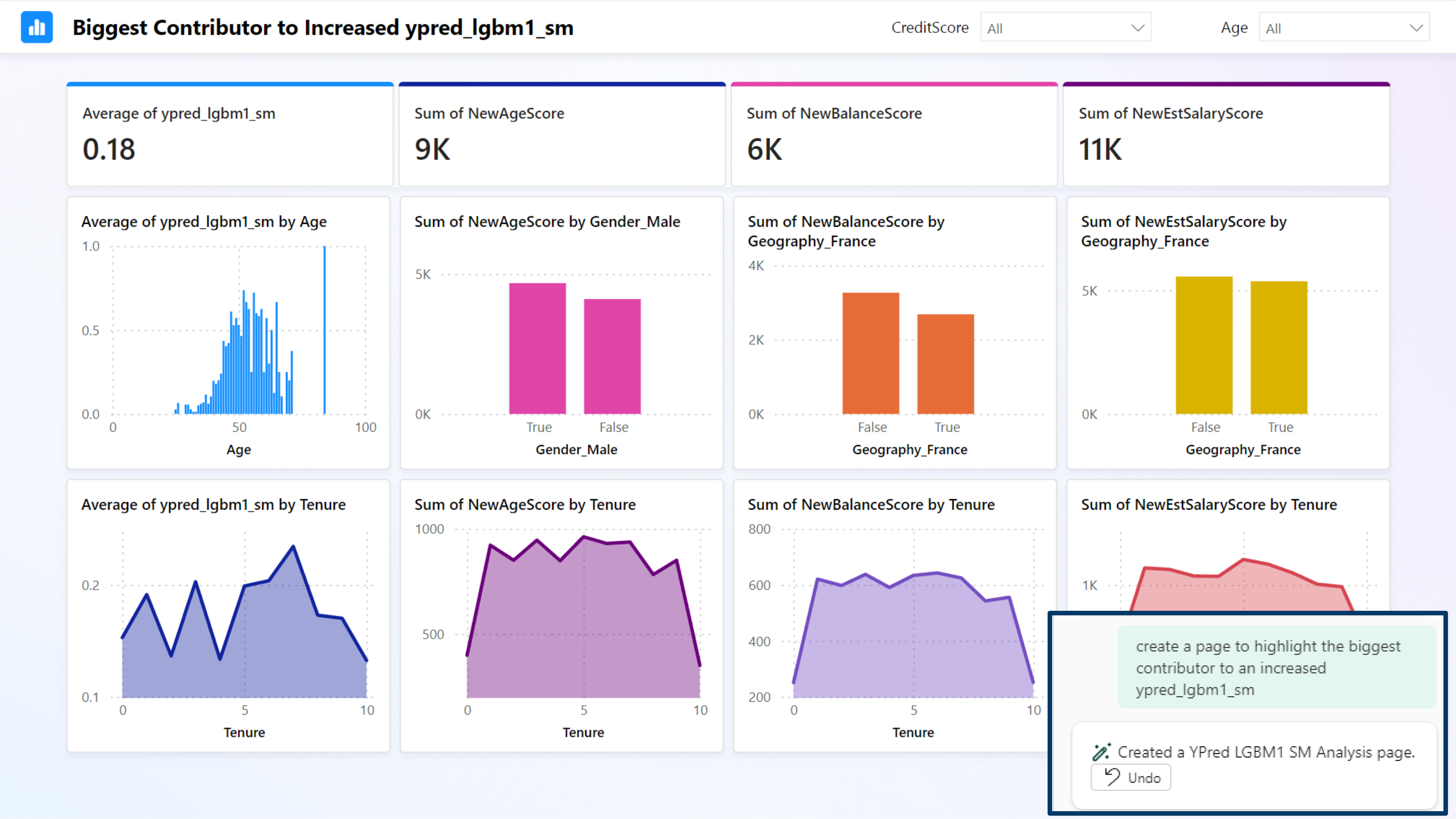
Excel (Preview)

CSV (Preview)

Paste or manually enter data

Pick a published semantic model


YPred LGBM1 SM Analysis
Age
All
CreditScore
All
Average of ypred_lgbm1_sm
0.18
Sum of ypred_lgbm1_sm
360
Max of ypred_lgbm1_sm
1
Min of ypred_lgbm1_sm
0
Average of ypred_lgbm1_sm by CreditScore
1.0
0.5
0.0
500
1000
CreditScore
Sum of ypred_lgbm1_sm by Tenure
40
20
0
5
10
Tenure
Max of ypred_lgbm1_sm by NumOfProducts
2
1
0
2
4
NumOfProducts
Min of ypred_lgbm1_sm by NewTenure
1.0
0.5
0.0
0.0
0.5
NewTenure
Average of ypred_lgbm1_sm by Age
1.0
0.5
0.0
0
50
100
Age
Sum of ypred_lgbm1_sm by Balance
100
0
0.0M
0.2M
Balance
Max of ypred_lgbm1_sm by EstimatedSalary
1.0
0.5
0.0
0.0M
0.1M
0.2M
EstimatedSalary
Min of ypred_lgbm1_sm by NewCreditsScore
1.0
create a page to highlight the biggest contributor to an increased ypred_lgbm1_sm
Created a YPred LGBM1 SM Analysis page.
Undo


Biggest Contributor to Increased ypred_lgbm1_sm
CreditScore All
Age All
Average of ypred_lgbm1_sm
0.18
Sum of NewAgeScore
9K
Sum of NewBalanceScore
6K
Sum of NewEstSalaryScore
11K
Average of ypred_lgbm1_sm by Age
Age
Sum of NewAgeScore by Gender_Male
True
False
Gender_Male
Sum of NewBalanceScore by Geography_France
False
True
Geography_France
Sum of NewEstSalaryScore by Geography_France
False
True
Geography_France
Average of ypred_lgbm1_sm by Tenure
Tenure
Sum of NewAgeScore by Tenure
Tenure
Sum of NewBalanceScore by Tenure
Tenure
Sum of NewEstSalaryScore by Tenure
create a page to highlight the biggest contributor to an increased ypred_lgbm1_sm
Created a YPred LGBM1 SM Analysis page.
Undo

Create, evaluate, and score a churn prediction model

Introduction

In this notebook, you'll see a Microsoft Fabric data science workflow with an end-to-end example. The scenario is to build a model to predict whether bank customers would churn or not. The churn rate, also known as the rate of attrition refers to the rate at which bank customers stop doing business with the bank.

The main steps in this notebook are:

1. Install custom libraries
2. Load the data
3. Understand and process the data through exploratory data analysis and demonstrate the use of Fabric Data Wrangler feature
4. Train machine learning models using `Scikit-Learn` and `LightGBM`, and track experiments using MLflow and Fabric Autologging feature
5. Evaluate and save the final machine learning model
6. Demonstrate the model performance via visualizations in Power BI

Step 1: Install custom libraries

When developing a machine learning model or doing ad-hoc data analysis, you may need to quickly install a custom library (e.g., `imblearn` in this notebook) for the Apache Spark session. To do this, you have two choices.

1. You can use the in-line installation capabilities (e.g., `%pip`, `%conda`, etc.) to quickly get started with new libraries. Note that this installation option would install the custom libraries only in the current notebook and not in the workspace.

```
# Use pip to install libraries
%pip install <library name>

# Use conda to install libraries
%conda install <library name>
```

2. Alternatively, you can follow the instructions here to learn how to create an environment which allows you to install libraries from public sources or upload custom libraries built by you or your organization.

For this notebook, you'll install the `imblearn` using `%pip install`. Note that the PySpark kernel will be restarted after `%pip install`, thus you'll need to install the library before you run any other cells.

```
# Use pip to install imblearn for SMOTE
%pip install imblearn
```

✓ 31 sec -Command executed in 3 sec 237 ms by Michel Aebischer on 2:08:17 PM, 3/06/24 PySpark (Python)

Step 2: Load the data

Dataset

The dataset contains churn status of 10,000 customers along with 14 attributes that include credit score, geographical location (Germany, France, Spain), gender (male, female), age, tenure (years of being bank's customer), account balance, estimated salary, number of products that a customer has purchased through the bank, credit card status (whether a customer has a credit card or not), and active member status (whether an active bank's customer or not).

The dataset also includes columns such as row number, customer ID, and customer surname that should have no impact on customer's decision to leave the bank. The event that defines the customer's churn is the closing of the customer's bank account, therefore, the column `exit` in the dataset refers to customer's abandonment. Since you don't have much context about these attributes, you'll proceed without having background information about the dataset. Your aim is to understand how these attributes contribute to the `exit` status.

Out of the 10,000 customers, only 2037 customers (around 20%) have left the bank. Therefore, given the class imbalance ratio, it is recommended to generate synthetic data.

- churn.csv

"CustomerID"	"Surname"	"CreditScore"	"Geography"	"Gender"	"Age"	"Tenure"	"Balance"	"NumOfProducts"	"HasCrCard"	"IsActiveMember"	"Es
15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101
15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112

Introduction to SMOTE

The problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary. Synthetic Minority Oversampling Technique (SMOTE) is the most widely used approach to synthesize new samples for the minority class. Learn more about SMOTE here and here.

You will be able to access SMOTE using the `imblearn` library that you installed in Step 1.

Step 3: Exploratory Data Analysis

Display raw data

Explore the raw data with `display`, do some basic statistics and show chart views. You first need to import required libraries for data visualization such as `seaborn` which is a Python data visualization library to provide a high-level interface for building visuals on dataframes and arrays. Learn more about seaborn.

```
import seaborn as sns
sns.set_theme(style="whitegrid", palette="tab10", rc = {'figure.figsize':(9,6)})
import matplotlib.pyplot as plt
import matplotlib.ticker as mticker
from matplotlib import rc, rcParams
import numpy as np
import pandas as pd
import itertools
```

✓ 12 sec -Command executed in 12 sec 144 ms by Michel Aebischer on 2:08:41 PM, 3/06/24 PySpark (Python)

Log

```
display(df, summary=True)
```

✓ 3 sec -Command executed in 3 sec 564 ms by Michel Aebischer on 2:08:45 PM, 3/06/24 PySpark (Python)

The five-number summary

Show the five-number summary (the minimum score, first quartile, median, third quartile, the maximum score) for the numerical attributes, using box plots.

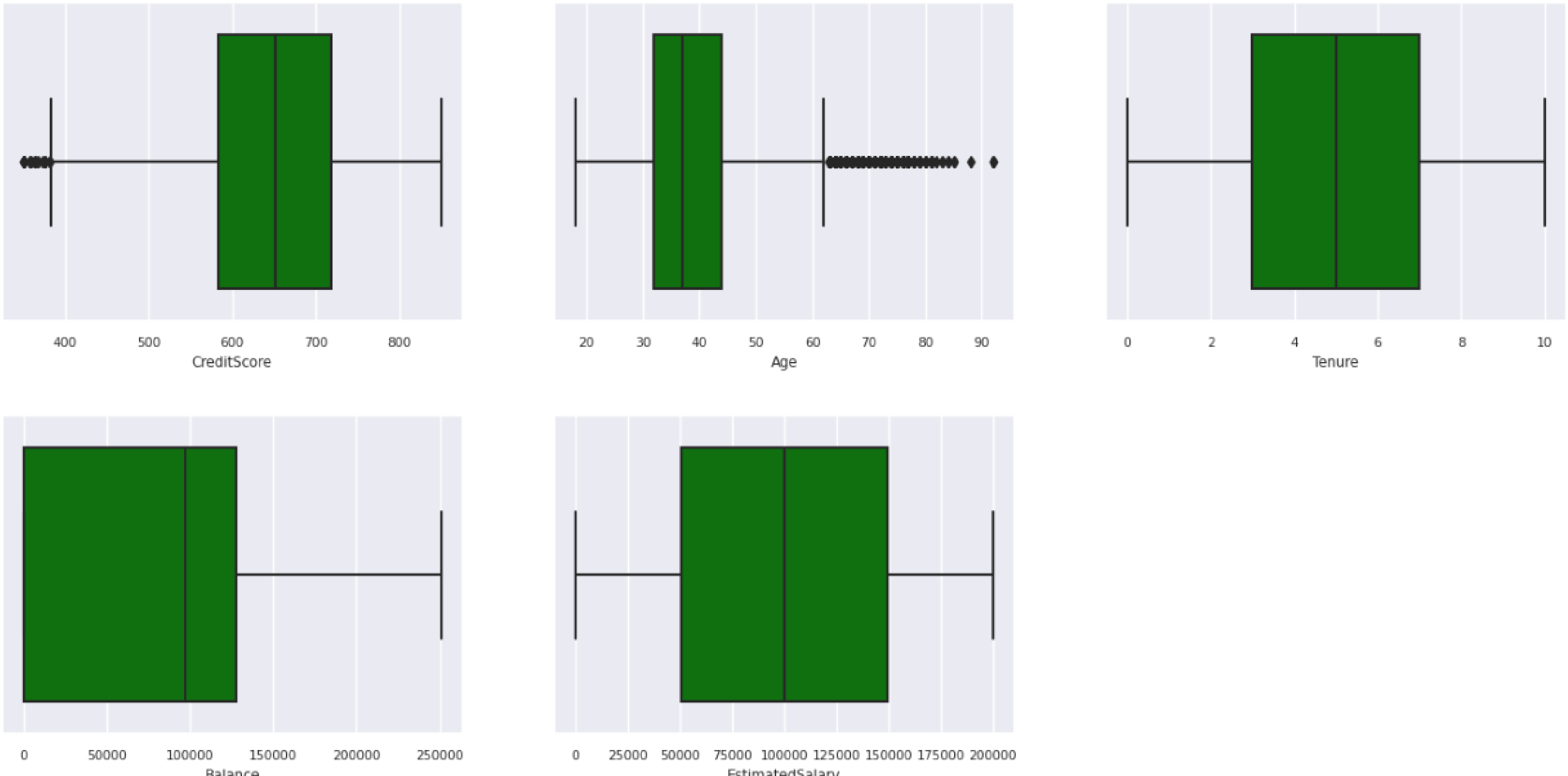
```
df_num_cols = df_clean[numeric_variables]
sns.set(font_scale = 0.7)
fig, axes = plt.subplots(nrows = 2, ncols = 3, gridspec_kw =  dict(hspace=0.3), figsize = (17,8))
fig.tight_layout()
for ax,col in zip(axes.flatten(), df_num_cols.columns):
    sns.boxplot(x = df_num_cols[col], color='green', ax = ax)
# fig.suptitle('visualize and compare the distribution and central tendency of numerical attributes', color = 'k', fontsize = 12)
fig.delaxes(axes[1,2])

```

✓ 1 sec -Command executed in 1 sec 520 ms by Michel Aebischer on 2:08:48 PM, 3/06/24 PySpark (Python)

Log

```
/tmp/ipykernel_7006/2095287195.py:4: UserWarning: This figure includes Axes that are not compatible with tight_layout, so results might be incorrect.
  fig.tight_layout()
```

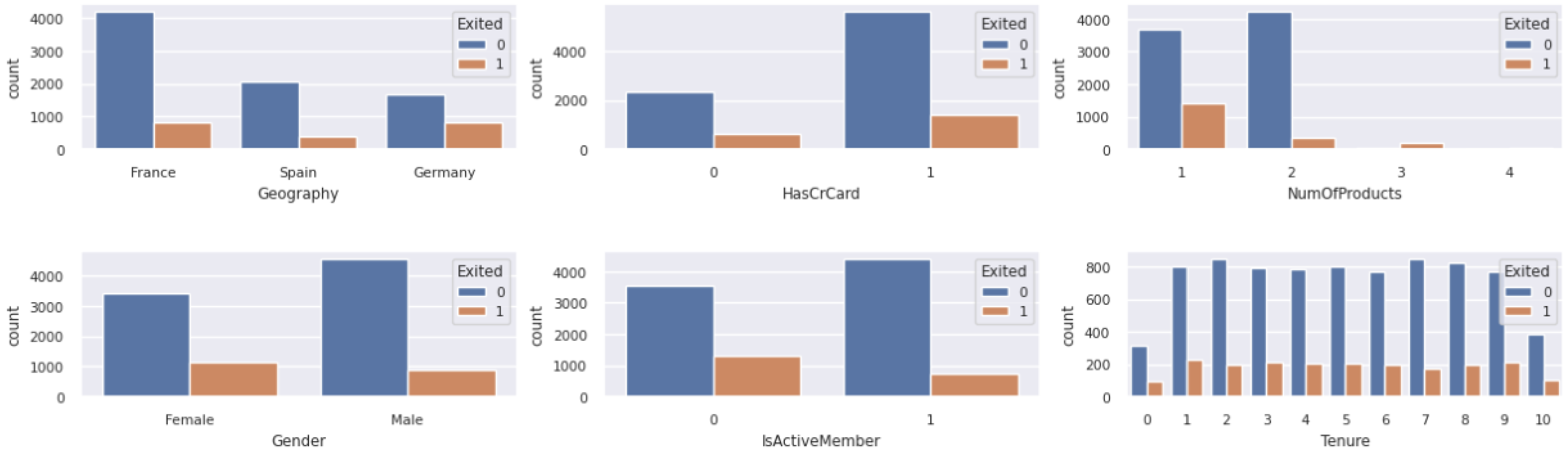


Distribution of exited and non-exited customers

Show the distribution of exited versus non-exited customers across the categorical attributes.

```python
attr_list = ['Geography', 'Gender', 'HasCrCard', 'IsActiveMember', 'NumOfProducts', 'Tenure']
fig, axarr = plt.subplots(2, 3, figsize=(15, 4))
for ind, item in enumerate (attr_list):
    sns.countplot(x = item, hue = 'Exited', data = df_clean, ax = axarr[ind%2][ind//2])
fig.subplots_adjust(hspace=0.7)
```

✓ 1 sec -Command executed in 1 sec 543 ms by Michel Aebischer on 2:08:50 PM, 3/06/24 PySpark (Python)

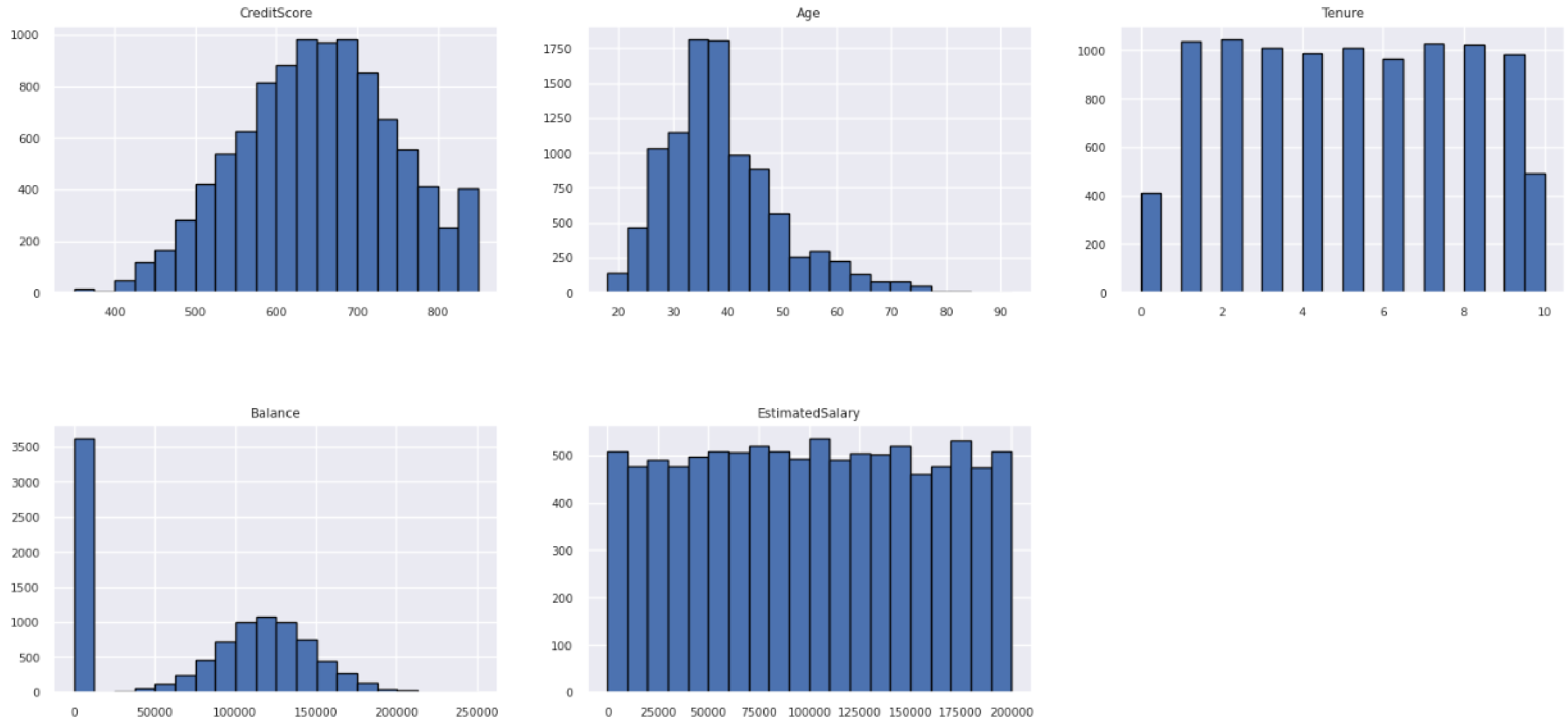


Distribution of numerical attributes

Show the the frequency distribution of numerical attributes using histogram.

```
columns = df_num_cols.columns[: len(df_num_cols.columns)]
fig = plt.figure()
fig.set_size_inches(18, 8)
length = len(columns)
for i,j in itertools.zip_longest(columns, range(length)):
    plt.subplot((length // 2), 3, j+1)
    plt.subplots_adjust(wspace = 0.2, hspace = 0.5)
    df_num_cols[i].hist(bins = 20, edgecolor = 'black')
    plt.title(i)
# fig = fig.suptitle('distribution of numerical attributes', color = 'r' ,fontsize = 14)
plt.show()
```

14] ✓ 1 sec -Command executed in 1 sec 500 ms by Michel Aebischer on 2:08:52 PM, 3/06/24 PySpark (Python)

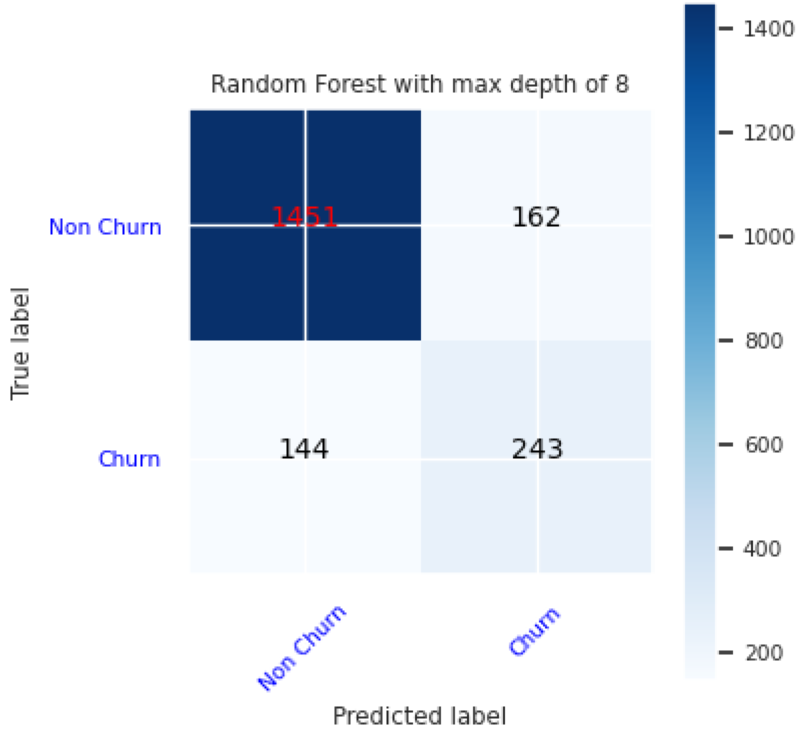


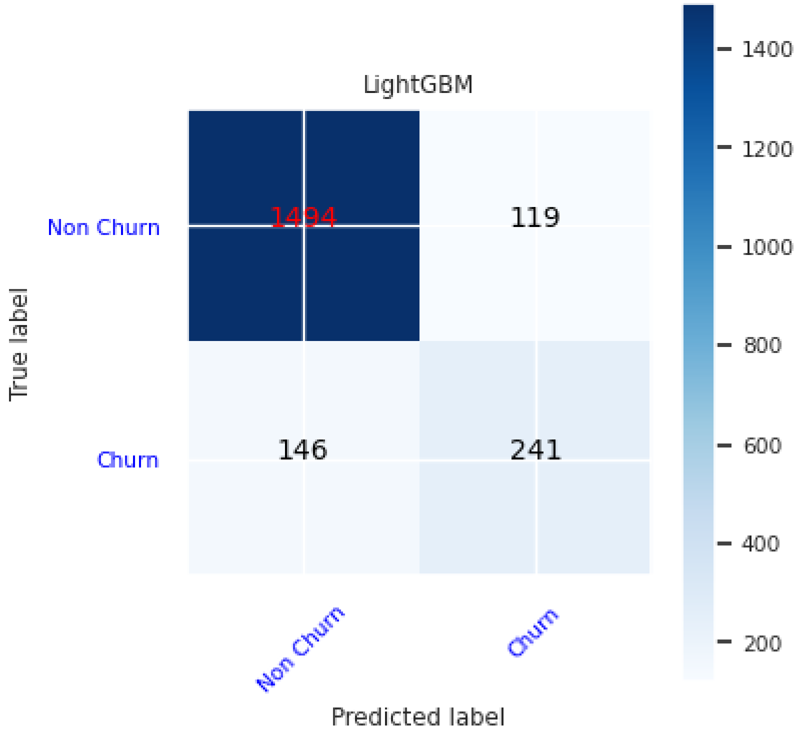
Summary of observations from the exploratory data analysis

- Most of the customers are from France comparing to Spain and Germany, while Spain has the lower churn rate comparing to France and Germany.
- Most of the customers have credit cards.
- There are customers whose age and credit score are above 60 and below 400, respectively, but they can't be considered as outliers.
- Very few customers have more than two of the bank's products.
- Customers who aren't active have a higher churn rate.
- Gender and tenure years don't seem to have an impact on customer's decision to close the bank account.

CreaXum

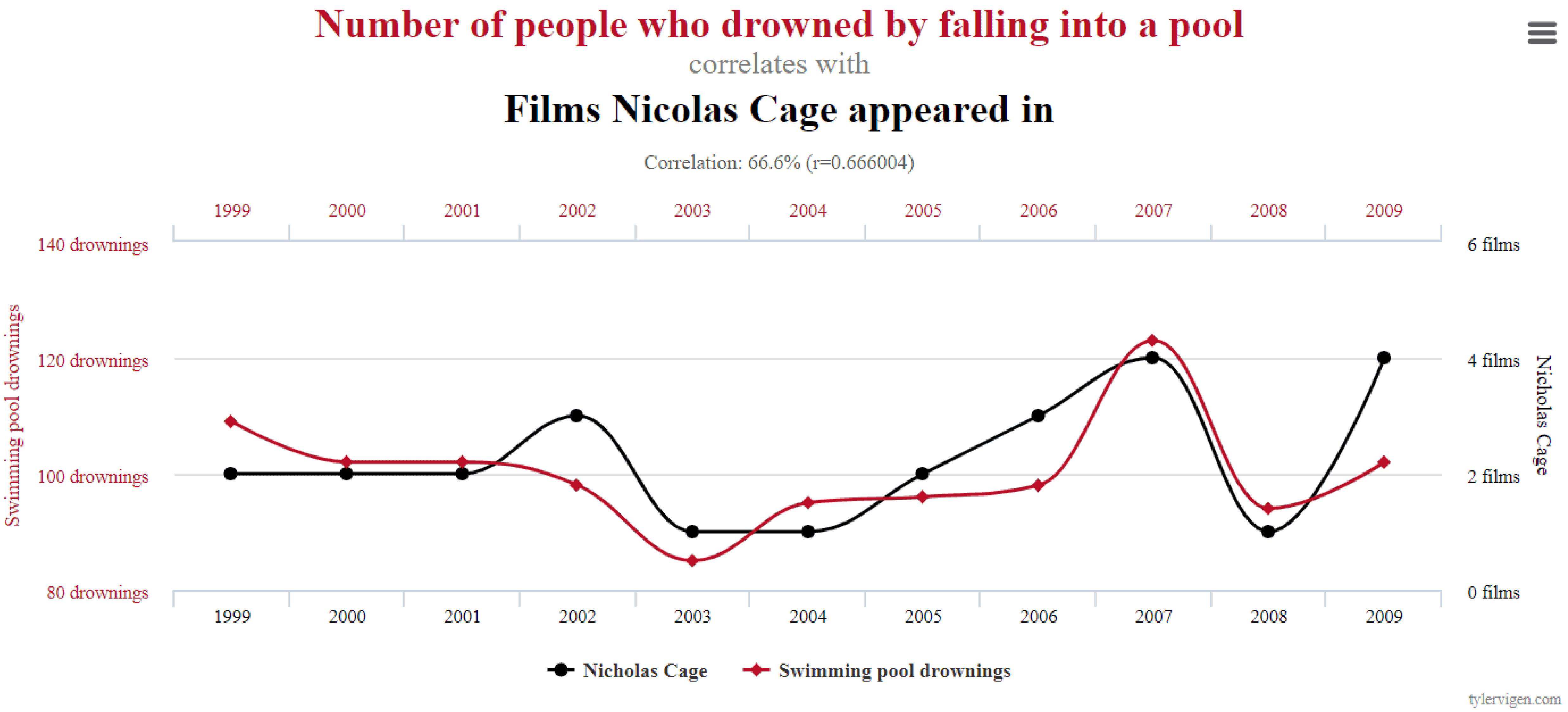
And in the end, the predicted results





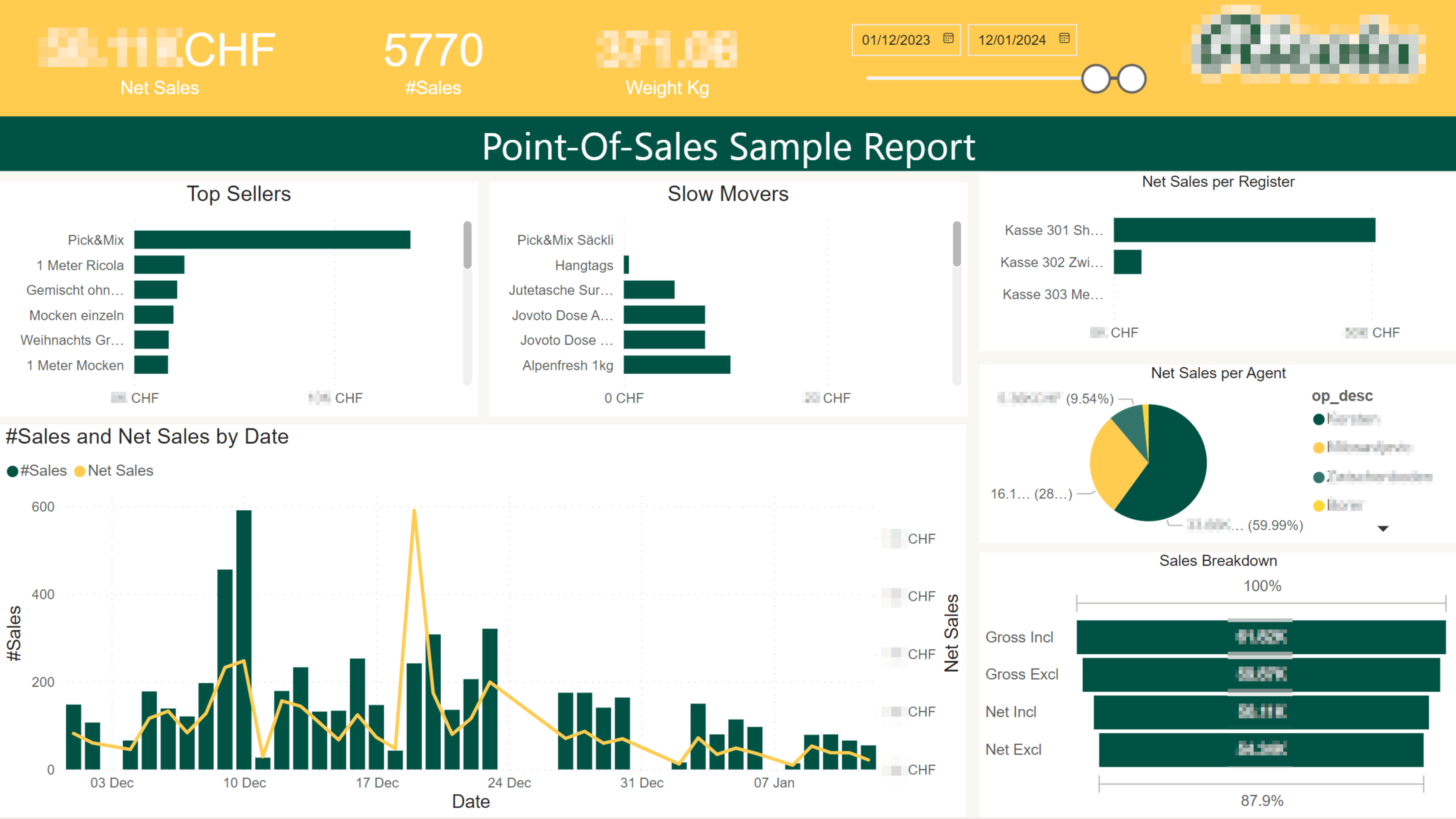
CreaXum

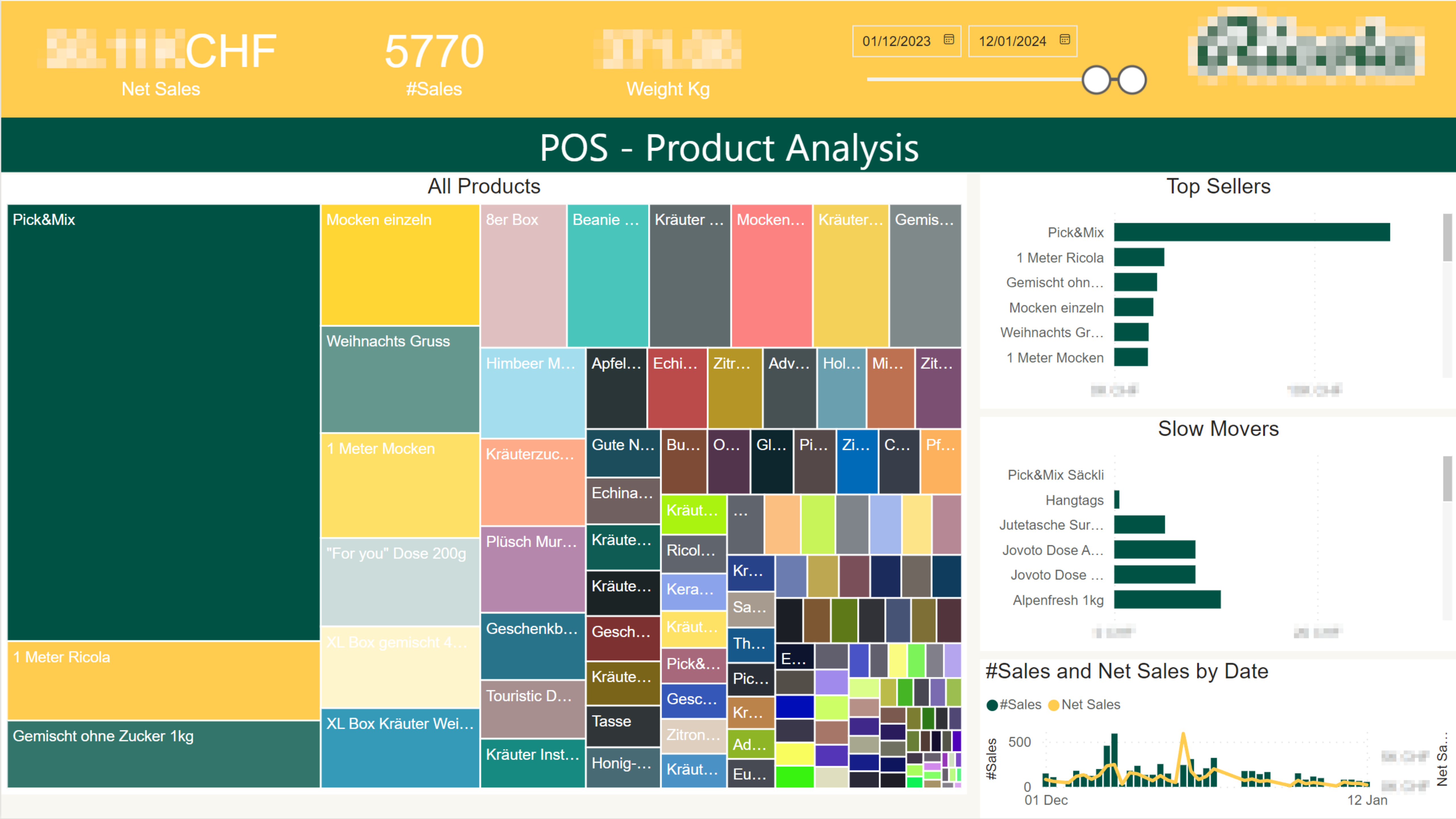
Correlation does not imply causation !



CreaXum

Tylervigen


CHF
Net Sales
5770
#Sales
Weight Kg
01/12/2023
12/01/2024
Point-Of-Sales Sample Report
Top Sellers
Pick&Mix
1 Meter Ricola
Gemischt ohn…
Mocken einzeln
Weihnachts Gr…
1 Meter Mocken
CHF
CHF
Slow Movers
Pick&Mix Säckli
Hangtags
Jutetasche Sur…
Jovoto Dose A…
Jovoto Dose …
Alpenfresh 1kg
0 CHF
CHF
Net Sales per Register
Kasse 301 Sh…
Kasse 302 Zwi…
Kasse 303 Me…
CHF
CHF
Net Sales per Agent
(9.54%)
16.1… (28…)
… (59.99%)
op_desc
#Sales and Net Sales by Date
#Sales
Net Sales
600
400
200
0
#Sales
CHF
CHF
CHF
CHF
CHF
Net Sales
03 Dec
10 Dec
17 Dec
24 Dec
31 Dec
07 Jan
Date
Sales Breakdown
100%
Gross Incl
Gross Excl
Net Incl
Net Excl
87.9%


CHF
Net Sales
5770
#Sales
Weight Kg
01/12/2023
12/01/2024
POS - Product Analysis
All Products
Pick&Mix
Mocken einzeln
8er Box
Beanie …
Kräuter …
Mocken…
Kräuter…
Gemis…
Weihnachts Gruss
Himbeer M…
Apfel…
Echi…
Zitr…
Adv…
Hol…
Mi…
Zit…
1 Meter Mocken
Kräuterzuc…
Gute N…
Bu…
O…
Gl…
Pi…
Zi…
C…
Pf…
Echina…
Kräut…
"For you" Dose 200g
Plüsch Mur…
Kräute…
Ricol…
Kr…
Kräute…
Kera…
Sa…
1 Meter Ricola
XL Box gemischt 4…
Geschenkb…
Gesch…
Kräut…
Th…
E…
Kräute…
Pick&…
Pic…
XL Box Kräuter Wei…
Touristic D…
Gesc…
Kr…
Tasse
Gemischt ohne Zucker 1kg
Kräuter Inst…
Zitron…
Ad…
Honig-…
Kräut…
Eu…
Top Sellers
Pick&Mix
1 Meter Ricola
Gemischt ohn…
Mocken einzeln
Weihnachts Gr…
1 Meter Mocken
Slow Movers
Pick&Mix Säckli
Hangtags
Jutetasche Sur…
Jovoto Dose A…
Jovoto Dose …
Alpenfresh 1kg
#Sales and Net Sales by Date
#Sales
Net Sales
500
0
01 Dec
12 Jan
#Sales
Net Sa…

Michel Aebischer
Founding Partner

CreaXum
Rue du Ronzier 3
CH - 1260 Nyon

+41 79 616 98 24

michel.aebischer@creaxum.com

They trust CreaXum