ITI FINAL PROJECT

Term Deposit Dataset

Team Members

- Mohamed AlGhaly
- Abdallah Galal
- Radwa Mohamed
- Mohamed Shoeb

Github: https://github.com/al-ghaly/ITI-Project

Step 1 Data Wrangling

Data Validation & Definition:

The aim of this step is to get to know the data and validate each cell.

NOTE: We won't deal with any quality issues here (unless a fatal issue presents itself), because the methodology for dealing with dirty data will depend on what we are aiming to achieve.

We already have a data definition file, but let's dig into the data.

- **We have 31,647 rows & 18 Columns.**
- **The data does not have Nulls but does not necessarily mean no missing values.**
- **IDs are unique.**
- **Age ranges from 18 to 95 with a mean value of 40.**
- **We have 11 different jobs and 206 missing jobs.**
- **We have (Married, Single, and Divorced) clients.**
- **We have Primary, Secondary, and College education with 1314 missing values.**
- **The default column is a perfect Yes | No column.**
- **The balance column ranges from -8k to 102k with a mean value of 1360, and after a deep investigation it appears that negative values are normal here.**
- **Housing & Loan are perfect Yes | No Columns.**
- **Contact column has two values (Cellular and telephone) with 9k missing values.**
- **Day & Month columns are perfect 31 | 12 columns.**
- **The duration column ranges from 0 to 4920 seconds with a mean value of 4 minutes.**
- **Campaign column ranges from 1 to 63.**
- **P days column have 25924 values of -1 which mean something, we can't get.**
- **Previous column ranges from 0 to 275 with 25924 zeros.**
- **P outcome column has 27k missing values.**
- **Subscribed column is a perfect Yes | No column.**
- **After looking at some statistics about the data, all looks good and ready to go.**

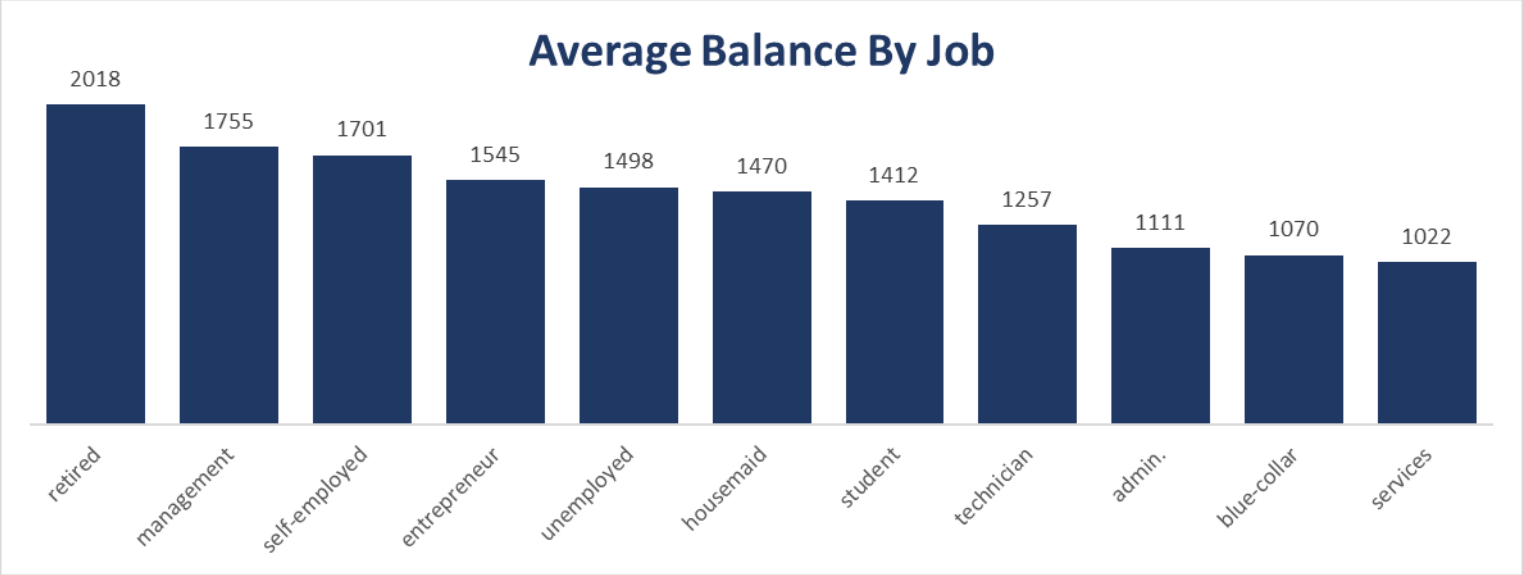
Data cleaning:

This is the typica next step in the data wrangling process, but we will call that off for now, as we are going to analyze the data using various tools and each one needs its own data cleansing techniques.

Step 2 Answering Questions

We will answer some insightful business questions using various tools such as Excel, Python, SQL, and Power BI.

1. How does the average yearly balance vary based on the client's job type?

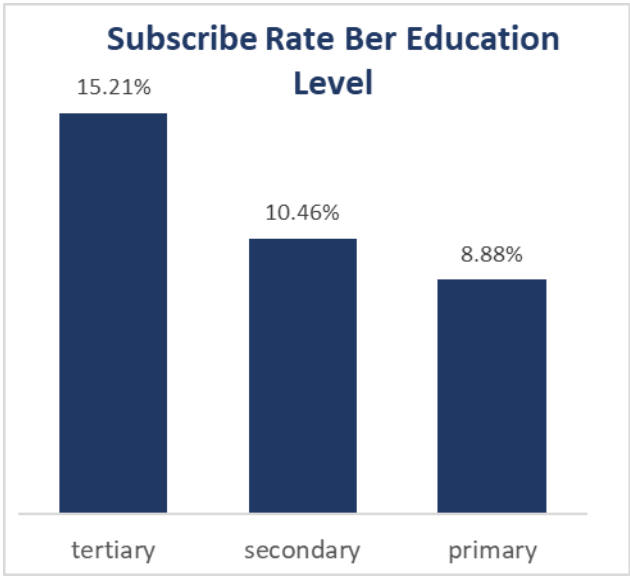


It looks like the average yearly balance **depends mainly** on the job type as:

- Retired clients have the largest balance which makes perfect sense.
- Management-Layer clients are rich.
- Services and Blue-Collar have the lowest average yearly balance.

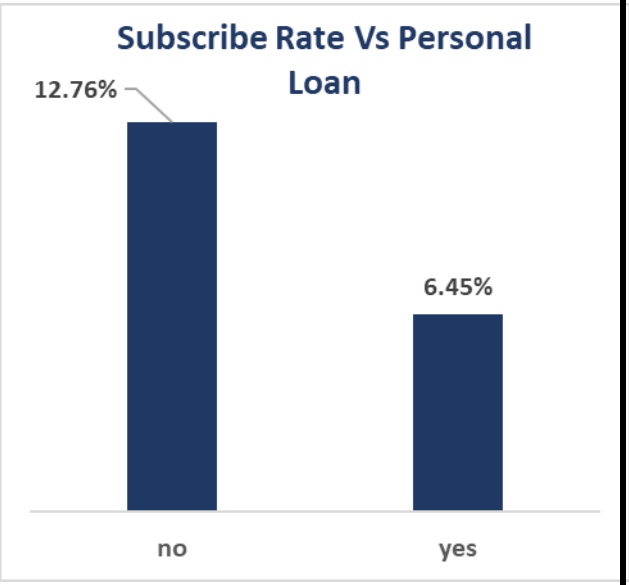
2. Is there a relationship between the client's education level and their decision to subscribe?

Clearly **higher-educated** clients are **more likely** to subscribe.
This insight can be very helpful as we may need to target highly educated clients in our campaign.

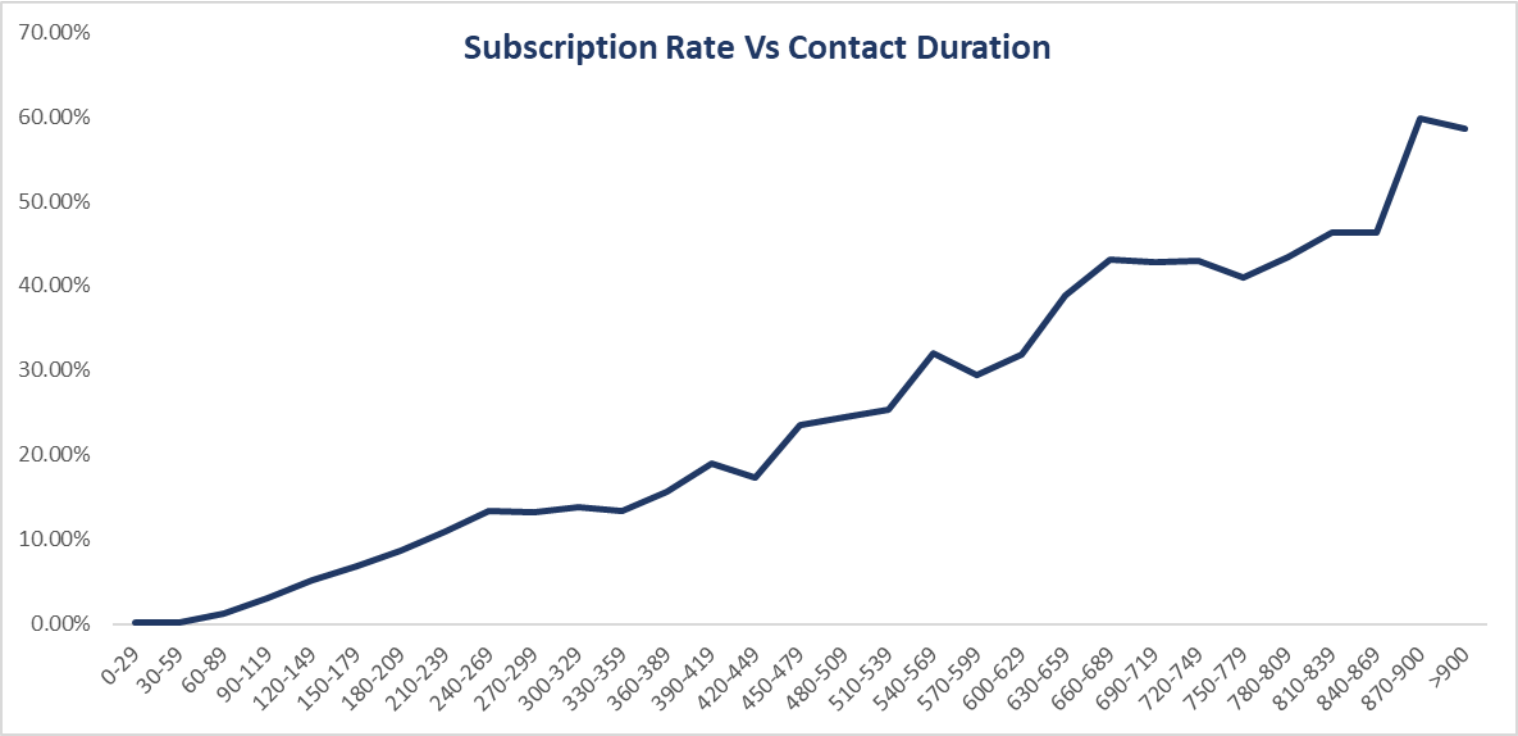


3. Do clients with a personal loan tend to subscribe more or less frequently compared to those without a loan?

Clearly **clients with a personal loan** are less likely to subscribe.
This insight can be very helpful as we may need to target clients without a personal loan in our campaign.



4. Are there any notable differences in the contact duration for subscribed and non-subscribed clients?

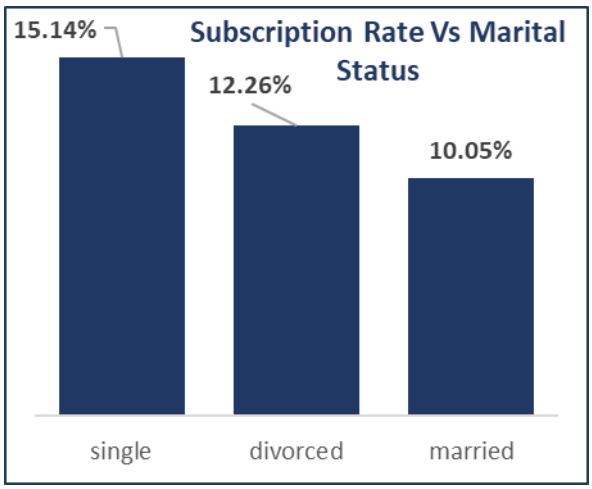


A blind man can see that as the duration of the contact increases the subscription **rate increases**.

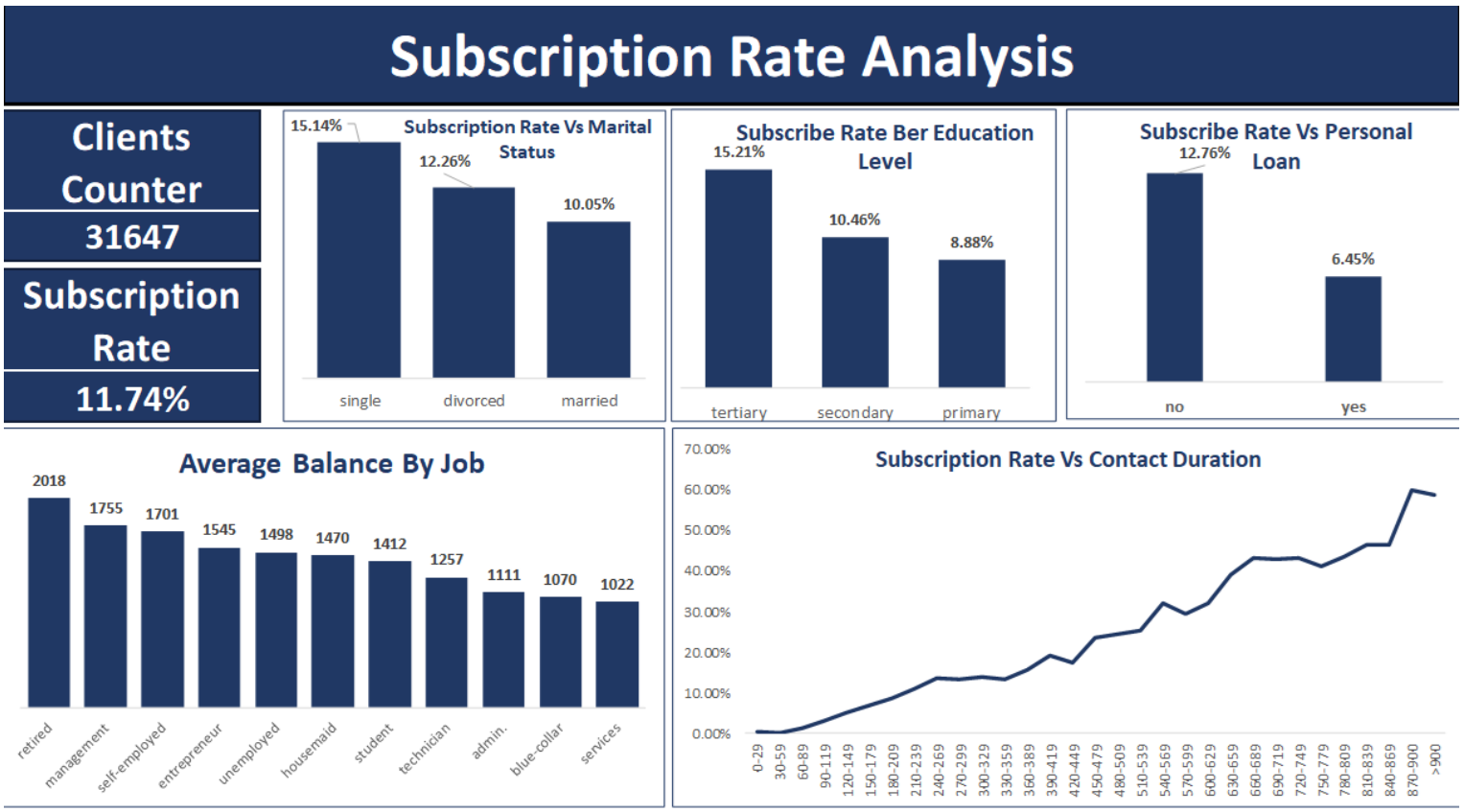
5. Is there a relationship between the client's Marital Status and their decision to subscribe?

Clearly **single clients** are more likely to subscribe.

This insight can be very helpful as we may need to target single in our campaign.



Dashboard to Communicate the Insights



Step 3 Data Warehousing

Step one will be defining business process:

1- Marketing Campaign (Marketing Team)

- In this process we are going to model the company's marketing campaigns, including the telephonic marketing campaign.
- This process is mainly concerned with the marketing team of the bank.

2- Transactions (Finance Team)

- This process is mainly concerned with analyzing term deposits and loans.
- It will also involve Analyzing the behavior of clients.
- The main purpose of this process is to manage deposits and loans to maximize our profits.

3- Customer Services: (Customer Support Team)

- This process involves three main relative branches:
  - Customer Inquiries.
  - Customer Complaints.
  - Customer Feedback.
- Mainly concerned with customer satisfaction.

Bus Matrix

Facts	Customer	Employee	Date	Branch	Services
Transactions	√	×	√	√	√
Customer Services	√	√	√	√	√
Marketing Campaign	√	×	√	×	√

Step two

will be defining which questions we want our model to tackle:

Questions (Will be answered along with a specific time period)

1- Which customers respond to our campaigns the most (Age/Gender/Education/and so on).
2- What factors help us increase the conversion/subscription rate.
3- What deposit types attract the most clients.
4- How often each client rolls over/cancels their deposit, and what factors affect its decision.
5- What are the most profitable deposit types.
6- What are the trends of the revenue/expenses of the deposits.
7- What are the most common deposit types.
8- What is the effect of changing the interest rate on the profit.
9- What is the cancellation rate of our term deposits.
10- Which customers are the most profitable (Age/Gender/Education/and so on).
11- What is the impact of investment amount, the duration of their investment, the interest rate they are offered on the revenue.
12- What are the factors that affect the decision of the clients on whether to roll over or not.
13- What is the channel of communication that is most common.
14- How many complaints are we getting.
15- How to improve customer satisfaction.
16- How to improve our business.
17- What loans are most rewarding in terms of profit.
18- What are the possible ways to reduce expenses.
19- What are the best ways to maximize revenue.
20- Which loan category is most rewarding.
21- What is the annual net profit for the bank.
22- What are the factors that make us decline a specific loan.
23- What is our total revenue/expenses.
24- What is the net amount of money we have.
25- Way, way more, but just keep up with this.

STEP Three

Step three will be defining granularity for each business process:

Marketing Campaign

- The most detailed grain is each marketing action (Phone call).
- I preferred just for the simplicity of the modeling to separate all marketing related processes into a single business process, in which will analyze the marketing performance, by analyzing each individual action taken in response to a marketing campaign.

Transactions

- The most detailed grain is each interaction or transaction made by or to any client, on a specific date.
- Transaction or interaction here refers to:
    - Opening a new deposit account.
    - Renewal of a deposit.
    - Cancellation of a deposit.
    - Loan request.
    - Approval or refusal of a loan.
    - Renewal or Cancellation of a loan.

Customer Services

- The most detailed grain is the combination of an individual customer care action (Inquiry, Feedback, Complaint), on a specific deposit or loan for a specific customer, on a specific branch, on a specific employee, at a given date.

STEP Four

Step 4 will be the capstone for the project which involves determining both facts and dimensions.

Marketing Campaign

- In this fact table we want to analyze the behavior of our clients who invest on term deposits or loans, and assess the marketing team performance, to maximize performance of our marketing campaigns.

Measurements/Attributes

- Subscribed

Dimensions

- Customer
- Services
- Date

Transactions

- In this fact table we want to analyze any interaction or transaction made by or to any client.
    - The term deposits we offer to our clients.
    - The loans we offer to our clients.
    - The process of assessing clients asking for loans, we want to be 100 % sure any loan given to any client would be PAID BACK.
    - The revenue analysis of the bank.
    - The profit analysis.
    - Monitoring our expenses.

Measurements/Attributes

- Renewed
- Cancelled
- Expenses
- Approved

Dimensions

- Customer
- Service

- Date
- Branch

Customer Services

- This fact table helps us keep track of customer satisfaction and provides ways to improve the company's performance by responding to clients' needs.

Measurements/Attributes

- Channel (channel of interaction)
- Action Type (Inquiry, Feedback, or Complaint)
- Severity

Dimensions

- Customer
- Service
- Branch
- Employee
- Date

DIMENSIONS:

1- Customer

- A dimension holding data about all the potential and current clients of our bank.

2- Service

- Holding data about all term deposits and loans offered by the bank.

3- Date

- Typical calendar dimension for any DWH.

4- Branch

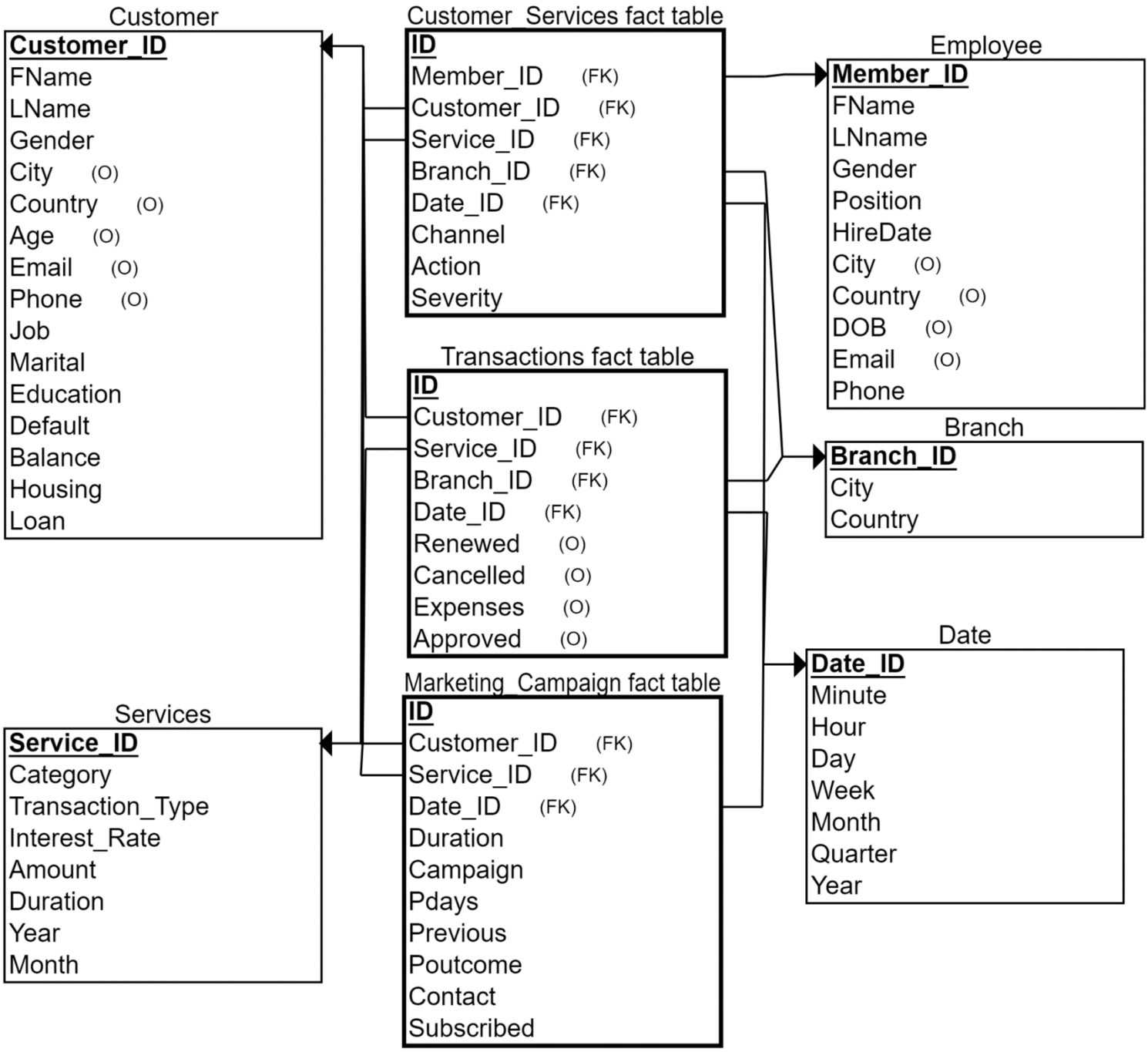
- A dimension holding data about all the branches of the bank.

5- Employee

- A dimension holding data about all the employees in the bank.

STEP Five

Step 5 will be the schema modeling, and we will use a star schema.



STEP Six

Step 6 is a discussion about step 5.

Why did we choose star schema modeling?

A star schema is perfect for our design for some reasons:

1- Simplified query performance: Star schema modeling allows us for simplified and optimized query performance. Since the fact table is at the center of the schema and connected to the dimension tables through foreign keys, queries can be executed efficiently and quickly, without the need for complex joins or subqueries.

2- Improved data analysis: Star schema modeling provides a simplified and intuitive way to analyze data from multiple dimensions. Analysts can easily drill down into data by navigating through the dimension tables, allowing for more complex analysis and insights.

3- Easier maintenance: Star schema modeling is easier to maintain than other modeling approaches. Since each dimension table is connected directly to the fact table, changes to one dimension table will not affect other tables in the schema. This makes it easier to modify and update the data warehouse over time.

4- Scalability: Star schema modeling is highly scalable and can handle large amounts of data. By separating the data into smaller, more manageable tables, the schema can accommodate large amounts of data without impacting query performance or data analysis.

Overall, star schema modeling is a popular and effective approach for designing data warehouses. It provides a simple, intuitive, and scalable way to organize and analyze data, making it easier for analysts and business users to get the insights they need from their data.

It mainly consists of 3 components:

- Dimensions: integral companions to a fact table. containing the textual context associated with a business process measurement event.
- Facts: stores the performance measurements resulting from an organizations' business process events.
- Measurements: The actual measurements stored in the fact tables.

STEP 7

The Physical Model

Marketing Campaign Fact		
Column	**Data Type**	**Index**
ID (PK)	Number	Clustered
Customer ID (FK)	Number	B_Tree
Service ID (FK)	Number	B_Tree
Date ID (FK)	Number	B_Tree
Duration	Number	
Campaign	Number	
Pdays	Number	
Previous	Number	
Poutcome	Varchar	Bit Map
Contact	Varchar	
Subscribed	Number	Bit Map

Services Dimension		
Column	**Data Type**	**Index**
Service ID (PK)	Number	Clustered
Category	Varchar	Non Clustered
Transaction_Type	Varchar	Bit Map
Intereste rate	Number	
Amout	Number	
Duration	Number	
Year	Number	Non Clustered
Month	Number	

Branch Dimension		
Column	**Data Type**	**Index**
Branch_ID (PK)	Number	Clustered
City	Varchar	
Country	Varchar	Bit Map

Transactions Fact		
Column	**Data Type**	**Index**
ID (PK)	Number	Culstered
Customer ID (FK)	Number	B_Tree
Service ID (FK)	Number	B_Tree
Branch ID (FK)	Number	B_Tree
Date ID (FK)	Number	B_Tree
Renewed	Number	Bit Map
Cancelled	Number	Bit Map
Expenses	Number	
Approved	Number	Bit Map

Employee Dimension		
Column	**Data Type**	**Index**
Member ID (PK)	Number	Clustered
FName	Varchar	
LName	Varchar	
DOB	Date	
Gender	Varchar	Bit Map
City	Varchar	
Country	Varchar	
Email	Varchar	
Position	Varchar	
Hire date	Date	
Phone	Varchar	

Customer Services Fact		
Column	**Data Type**	**Index**
ID (PK)	Number	Culstered
Customer ID (FK)	Number	B_Tree
Service ID (FK)	Number	B_Tree
Date ID (FK)	Number	B_Tree
Member ID (FK)	Number	B_Tree
Branch ID (FK)	Number	B_Tree
Channel	Varchar	Bit Map
Action	Varchar	Bit Map
Serverity	Number	

Date Dimension		
Column	**Data Type**	**Index**
Date_ID (PK)	Number	Clustered
Minute	Number	
Hour	Number	
Day	Number	
Week	Number	
Month	Number	
Quarter	Number	Bit Map
Year	Number	Non Clustered

Customer Dimension		
Column	**Data Type**	**Index**
Customer_ID (PK)	Number	Clustered
FName	Varchar	
LName	Varchar	
Gender	Varchar	Bit Map
City	Varchar	
Country	Varchar	
Age	Number	
Email	Varchar	
Phone	Varchar	
Job	Varchar	
Marital	Varchar	Bit Map
Education	Varchar	Bit Map
Default	Number	
Balance	Number	
Housing	Number	
Loan	Number	

STEP 8

In step 8 let's talk for a little bit about each type of index we used and

NOTE:

Taking about each index and the algorithm used to implement it would be a lot of fun, but it is way out of the scope of this project and will take a long time, so we will just give a hint on each one, Enjoy.

INDEX TYPE	WHEN	WHY
CLUSTERED	Primary Key Columns	physically order the data in a table based on the indexed column.
B_TREE	Foreign Keys Columns	highly efficient for range-based queries.
UNIQUE	unique columns	This index enforces uniqueness for the indexed column or columns.
HASH	Columns used in equality filtering	Uses hash-function to retrieve data in constant time so it is used for equality filtering and join conditions.
NON-CLUSTERD	Categorial Data (With wider value range)	stored separately from the data and contains a copy of the indexed column, along with a pointer to the corresponding data.
BITMAP	Categorial Data (With 2-5 values)	It uses a bitmap to represent the data, with each bit representing a possible value, so it is used on columns with a small number of distinct values.

For me, I prefer using a clustered index on the PK columns, non-clustered index on categorial data with wider range of values, b-tree for categorial data with 2-5 values, unique index for any unique column, and the last one is my personal favorite which can do magic, hash index can use a hash function to retrieve a piece of data from a table containing a zillion row in just O(1) time so I prefer using such index on any column used a lot in where statement with an equal sign maybe for filtering or joining tables, If I am about to implement a database engine, Hash Index will be forced on both primary and foreign keys.

Let's Start with INDEXES.

WHAT is an Index:

In a database, an index is a data structure that provides quick access to rows in a table based on the values in one or more columns.

WHY do we need indexes:

Indexes can greatly improve the performance of database queries, especially for large tables with many rows. They allow the database to quickly narrow down the rows that match a query condition, which can greatly reduce the amount of time and resources required to execute the query.

Let's know a little something about each index and when to use it:

1. **B-tree index:** A B-tree index is a balanced tree structure that allows for quick access to data based on a single column or a combination of columns. It is the most common type of index used in databases and is well-suited for range queries, such as queries that use greater than or less than operators.
2. **Hash index:** A hash index is a data structure that uses a hash function to map keys to values. It is typically used for equality queries, such as queries that use the equals operator. Hash indexes are very fast for lookups but are not well-suited for range queries or sorting.
3. **Bitmap index:** A bitmap index is a data structure that uses a bitmap to represent the presence or absence of values in a column. Bitmap indexes are well-suited for columns with a low number of distinct values and are particularly useful for data warehousing and decision support systems.
4. **Clustered index:** A clustered index is an index that determines the physical order of the data in a table. This means that the rows in the table are physically sorted based on the indexed columns. Each table can have only one clustered index, and it is typically created on the primary key column(s) of the table. Because the data is physically sorted, clustered indexes are very efficient for range queries and queries that retrieve a large number of rows.
5. **Non-clustered index:** A non-clustered index is an index that does not determine the physical order of the data in a table. Instead, it creates a separate data structure that contains the indexed columns and a pointer to the corresponding rows in the table. Non-clustered indexes are typically created on columns that are frequently searched, but not used for sorting or grouping. Each table can have multiple non-clustered indexes.
6. **Unique index:** A unique index is an index that enforces a constraint that ensures that the values in the indexed column(s) are unique. This means that no two rows in the table can have the same values in the indexed column(s). Unique indexes can be clustered or non-clustered, and they are useful for enforcing data integrity and optimizing queries that search for unique values.

Is it all good?

However, indexes also have some drawbacks. They require additional storage space and can slow down insert, update, and delete operations, as the database must update the index as well as the main table. Additionally, creating too many indexes or using them improperly can actually decrease query performance, as the database must spend more time maintaining and updating the indexes.

As a result, it's important to carefully consider the use of indexes in a database and to create them only for columns that are frequently queried and have a high selectivity (i.e., a large number of unique values). It's also important to regularly monitor and optimize the use of indexes to ensure that they are providing a net benefit to the database's performance.

Finally

IN DWH we use large volumes of data, so many rows, so many tables, and so many heavy joins.

In addition to this we do not manipulate our data so often so indexes will be a great help for DWH.

DWH.

WE HAVE ADDED THE INDEXES TO THE SWH and A query that took 1 minute takes no time now.

STEP 9

Let's get some insights from the DWH.

QUESTION 1: What is the cancellation rate of our term deposits?

It looks like we have a cancellation rate of 12.6 % over our term deposits.

```
WITH deposit_cancelled AS
(
SELECT DISTINCT cancelled, COUNT(*) OVER(PARTITION BY cancelled) / COUNT(*) OVER() cancellation_rate
FROM transactions_fact t
JOIN services_dim s
ON t.service_id = s.service_id
WHERE transaction_type = 'Deposit'
)
SELECT cancellation_rate
FROM deposit_cancelled
WHERE cancelled = 1;
```

Data Grid | Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer | CodeXpert | Explain Plan | Script Output | Cancel

CANCELLATION_RATE
0.125852233198831

QUESTION 2: What is the channel of communication that is most common?

```
SELECT channel_, COUNT(*) most_common_channel
FROM customer_services_fact
GROUP BY channel_
ORDER BY most_common_channel DESC;
```

Looks like the phone is the most common way.

Data Grid | Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer | Cancel

CHANNEL_	MOST_COMMON_CHANNEL
Phone	1706
In Person	1656
Online	1638

QUESTION 3: What is the clients subscription rate per job?

Looks like our marketing campaign is working best for students.

```
SELECT job_, Avg(subscribed) rate
FROM marketing_campaign_fact m
JOIN customer_dim c
ON m.customer_id = c.customer_id
GROUP BY job_
ORDER BY rate DESC;
```

Data Grid | Data Grid | Auto Trace | DBMS Output (disabled) | Query Vie | Cancel

JOB_	RATE
student	0.286614173228346
retired	0.229987293519695
unemployed	0.142541436464088
management	0.139026961891851
unknown	0.12621359223301
self-employed	0.1246660730187
admin.	0.124483613329661
technician	0.111927642736009
housemaid	0.0903890160183066
services	0.0874956941095419
entrepreneur	0.0843253968253968
blue-collar	0.071470330312774

QUESTION 4: What is the total profit?

Over the past 5 years we have made a net income of 120 M, which is good for an international bank (over 3 countries and 53 cities), having 300 k transactions.

```
SELECT SUM((INTEREST_RATE*AMOUNT)*
  (CASE
    WHEN TRANSACTION_TYPE = 'Deposit' AND cancelled = 0 THEN -1
    WHEN TRANSACTION_TYPE = 'Loan' AND approved = 1 AND cancelled = 0 THEN 1
    ELSE 0
  END)
   - NVL(EXPENSES,0)) AS profit
FROM transactions_fact t
JOIN services_dim s
ON t.service_id = s.service_id;
```

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer | CodeXpert | Explain Plan | Script Ou

Cancel

PROFIT
120632074

QUESTION 5: What are the most common services we offer?

```
/* Most common services */
  SELECT transaction_type, category, COUNT (*) AS frequency
    FROM services_dim JOIN Transactions_Fact USING (Service_ID)
GROUP BY transaction_type, category
ORDER BY frequency DESC, transaction_type;
```

Housing loans are on fire.

ata Grid

Messages | Data Grid | Trace | DBMS Output (disabled) | Query Viewer | Explain Plan

TRANSACTION_TYPE	CATEGORY	FREQUENCY
Loan	Housing Loan	47958
Loan	Gold Loan	35949
Deposit	Bullet deposit	35467
Deposit	Fixed Deposit	33488
Loan	Personal Loan	30149
Deposit	Callable Deposit	26883
Deposit	Flexi deposit	26707
Loan	Vehicle Loan	24151
Deposit	Recurring Deposit	21195
Loan	Home Loan	18054

QUESTION 6: What is the amount of profit we get from each customer?

```
SELECT fname||' '||lname client_name, SUM((INTEREST_RATE*AMOUNT)*
  (CASE
    WHEN TRANSACTION_TYPE = 'Deposit' AND cancelled = 0 THEN -1
    WHEN TRANSACTION_TYPE = 'Loan' AND approved = 1 AND cancelled = 0 THEN 1
    ELSE 0
  END)
   - NVL(EXPENSES,0)) AS profit
FROM customer_dim c
JOIN transactions_fact t
ON c.customer_id = t.customer_id
JOIN services_dim s
ON t.service_id = s.service_id
GROUP BY fname||' '||lname
ORDER BY profit DESC;
```

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer | CodeXpert | Explain Plan

Cancel

CLIENT_NAME	PROFIT
Cleveland Foux	1303978
Willie Windeatt	1302470
Miner Pacitti	1234963
Frederica O' Meara	1207083
Flory Gerrelts	1187001
Hamlen Forrestill	1181946
Raye Reasce	1160811
Dacy Crumby	1151695
Lauritz Willshire	1150987

QUESTION 7: What are the cancellation rates of each term deposit?

```sql
WITH category_cancelled AS
(
SELECT DISTINCT cancelled, category, COUNT(*) OVER(PARTITION BY cancelled, category) / COUNT(*) OVER(PARTITION BY category) category_cancellation_rate
FROM transactions_fact t
JOIN services_dim s
ON t.service_id = s.service_id
WHERE transaction_type = 'Deposit'
)
SELECT category, category_cancellation_rate
FROM category_cancelled
WHERE cancelled = 1
ORDER BY category;
```

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer | CodeXpert | Explain Plan | Script Output

Cancel

CATEGORY	CATEGORY_CANCELLATION_RATE
Bullet deposit	0.124199960526687
Callable Deposit	0.123684112636239
Fixed Deposit	0.126433349259436
Flexi deposit	0.12697045718351
Recurring Deposit	0.129039867893371

QUESTION 8: Which customers respond to our campaigns the most (Age/Gender/Education/and so on).

```sql
--1-  Which customers respond to our campaigns the most (Age/Gender/Education/and so on).
---gender
 SELECT gender, AVG (subscribed) subscribtion_rate
   FROM    Customer_Dim c
       JOIN
          Marketing_Campaign_Fact m
       ON c.customer_id = m.customer_id
GROUP BY gender
ORDER BY subscribtion_rate DESC;
```

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer | CodeXpert | Explain Plan | Script O

Cancel

GENDER	SUBSCRIBTION_RATE
Female	0.11875483100274
Male	0.116277443714166

Gender

No difference between female male in subscription rate.

Education

Customers who have tertiary certificates subscribe the

```sql
---education

 SELECT c.education, AVG (SUBSCRIBED) subscribtion_rate
   FROM    Customer_Dim c
       JOIN
          Marketing_Campaign_Fact m
       ON c.customer_id = m.customer_id
GROUP BY c.education
ORDER BY subscribtion_rate DESC;

---job
```

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer

Cancel

EDUCATION	SUBSCRIBTION_RATE
tertiary	0.152134179120525
secondary	0.104598126232742
primary	0.088810316139767 1

```sql
---job

 SELECT c.job_, AVG (SUBSCRIBED) subscribtion_rate
   FROM    Customer_Dim c
       JOIN
          Marketing_Campaign_Fact m
       ON c.customer_id = m.customer_id
GROUP BY c.job_
ORDER BY subscribtion_rate DESC;
```

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer

Cancel

JOB_	SUBSCRIBTION_RATE
student	0.286614173228346
retired	0.229987293519695
unemployed	0.142541436464088
management	0.139026961891851
unknown	0.12621359223301
self-employed	0.1246660730187
admin.	0.124483613329661
technician	0.111927642736009
housemaid	0.0903890160183066

Job

Students have higher subscription rate.

----age

```
----age
-- avg age of subscribed customers

 SELECT SUBSCRIBED, AVG (age) Age_average
   FROM    Customer_Dim c
       JOIN
         Marketing_Campaign_Fact m
       ON c.customer_id = m.customer_id
GROUP BY SUBSCRIBED;
```

Data Grid

Data Grid | Auto Trace | DBMS Output (disabled) | Query

Cancel

SUBSCRIBED	AGE_AVERAGE
1	41.6721399730821
0	40.8621652584849

```
---Marital

 SELECT c.Marital, AVG (SUBSCRIBED) subscribtion_rate
   FROM    Customer_Dim c
       JOIN
         Marketing_Campaign_Fact m
       ON c.customer_id = m.customer_id
GROUP BY c.Marital
ORDER BY subscribtion_rate DESC;

--country
```

Data Grid

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viev

Cancel

MARITAL	SUBSCRIBTION_RATE
single	0.151423447657476
divorced	0.122589531680441
married	0.100497512437811

Marital

Single clients have higher subscription rate.

--country

```
--country

 SELECT c.country, AVG (SUBSCRIBED) subscribtion_rate
   FROM    Customer_Dim c
       JOIN
         Marketing_Campaign_Fact m
       ON c.customer_id = m.customer_id
GROUP BY c.country
ORDER BY subscribtion_rate DESC;
```

Data Grid

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer

Cancel

COUNTRY	SUBSCRIBTION_RATE
Saint Pierre and Miquelon	1
Falkland Islands	1
Andorra	1
Norfolk Island	0.5
Guadeloupe	0.5
Algeria	0.5
Trinidad and Tobago	0.4
Vanuatu	0.4
Suriname	0.333333333333333

---Housing loan

```
---Hosing loan

 SELECT c.Housing, AVG (SUBSCRIBED) subscribtion_rate
   FROM    Customer_Dim c
       JOIN
         Marketing_Campaign_Fact m
       ON c.customer_id = m.customer_id
GROUP BY c.Housing
ORDER BY subscribtion_rate DESC;
```

Data Grid

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer

Cancel

HOUSING	SUBSCRIBTION_RATE
0	0.168171798336059
1	0.0767743403093722

-- personal loan

```
-- pesonal loan

 SELECT c.loan, AVG (SUBSCRIBED) subscribtion_rate
   FROM    Customer_Dim c
       JOIN
         Marketing_Campaign_Fact m
       ON c.customer_id = m.customer_id
GROUP BY c.loan
ORDER BY subscribtion_rate DESC;
```

Data Grid

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer

Cancel

LOAN	SUBSCRIBTION_RATE
0	0.127621058983255
1	0.0645098421360359

Customers with no personal loan have a higher subscription rate.

QUESTION 9: What deposit types attract the most clients?

```
97  --------
98  --3-   What deposit types attract the most clients. Transaction fat join with service dim
99  ---- NOTE THE SAME CATEGORRY HAVE DIFFERENT SERVICE ID
100 ---BECAUSE  EACH CATEGORY CAN CHANGE EVERY YEAR WITH RESPECT INTEREST RATE
101
102   SELECT CATEGORY, COUNT (*) AS count_
103     FROM Services_Dim s JOIN Transactions_fact f ON S.SERVICE_ID = F.SERVICE_ID
104    WHERE transaction_type = 'Deposit'
105 GROUP BY CATEGORY
106 ORDER BY count_ DESC;
107
```

Data Grid

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer | CodeXpert | Explain Plan

Cancel

CATEGORY	COUNT_
Bullet deposit	35467
Fixed Deposit	33488
Callable Deposit	26883
Flexi deposit	26707
Recurring Deposit	21195

Bullet deposit attracts the most clients.

QUESTION 9: What deposit types attract the most clients?

```
118
119 /*  net amount of money we have */
120
121 SELECT total - LEAD (total) OVER (ORDER BY total DESC) AS net
122   FROM (  SELECT SUM (amount) total, transaction_type
123             FROM services_dim S JOIN Transactions_Fact T USING (service_id)
124            WHERE approved = 1 OR transaction_type = 'Deposit'
125         GROUP BY transaction_type);
126
```

Data Grid

Data Grid | Auto Trace | DBMS Output (disabled) | Query Viewer | CodeXpert

Cancel

NET
1892056000

We have about 1892 M $.

Step 4 Machine Learning

We have fitted 6 different classification models on our data and achieved overall accuracy of 90.2 %.
We have used:

- Logistic Regression Model.
- Decision Tree Model.
- Random Forest Model.
- Support Vector Machines Model (SVM)
- Ada Boost Model.
- Gradient Boosting Model

We have got an average accuracy (F1 Score) of 90 on test data and 90.2 on validation data.

Why did we use (F1 Score) to measure the model performance?

- The target label (Subscribed) is biased having subscription rate of 10 %
- F1 score is a useful metric to use when you want to balance the model's precision and recall in binary classification problems. It is particularly useful when the classes in the dataset are imbalanced, meaning one class has significantly more samples than the other. In such cases, accuracy can be misleading as a metric because a model that always predicts the majority class can have a high accuracy but may not be useful in practice.
- F1 score is a weighted harmonic mean of precision and recall. It considers both false positives and false negatives and provides a single score that summarizes the model's performance. The F1 score ranges from 0 to 1, where 1 is the best possible score.
- So overall we chose F1 score because the data is imbalanced/binary, and we want to balance between recall and precision.

Recall is more important because we are more considered with predicting who would subscribe to our services, to target the most appropriate clients, and calling a client who won't subscribe will cause no harm.

After applying hyper tunning parameters, we have discovered that the default parameters are the most efficient way to go.

Data Cleaning:

We have cleaned the data in a DYNAMIC way.

- The transform function is the function responsible for data cleaning and it has hyper parameters for you to try and get the most efficient model.
- It has the following parameters:
  - Data: The data frame you want to clean.
  - Inplace: Whether you want to clean the data frame in place or as a new data frame.
  - Trim_long: Whether you want to remove calls longer than 1000 seconds or just clip those to 1000-second call indicating long calls.
  - trim_poutcome: The p outcome column has 80% missing values, but it is an important feature, so you have the choice whether to drop the column or remove missing values and work with the 20% available.
  - Trim_age: Whether you want to remove ages higher than 70 years or just clip those to 70 years indicating old people.
  - Log_transformation: The balance column is right skewed, so we applied log transformation to handle outliers, but even this is optional you may do it or pass.

```
Signature:
transform(
    data: pandas.core.frame.DataFrame,
    inplace=False,
    trim_long_calls=False,
    trim_poutcome=False,
    trim_age=False,
    log_transformation=True,
) -> pandas.core.frame.DataFrame | None
```

```
Docstring:
This function cleans the data frame INPLACE or as a new dataframe.
:param data: The data frame to clean
:param inplace: Whether to transform the dataframe in place or as a new dataframe
:param  drop_long_calls: whether or not you want to drop calls longer than 1000 seconds
:param trim: Whether tor not you want to drop umknown previous campaign outcome
:return: The transformed data frame as a new dataframe or None if clean happens in plac
e
File:      c:\users\al-ghaly\appdata\local\temp\ipykernel_16532\3743148402.py
Type:      function
```

Step 5 Power Bi Dashboard

We are communicating with 3 teams:

- Financial Team
- Marketing Team
- Customer Support Team

In This Dashboard, we are trying mainly to improve our business by increasing revenue, decreasing costs, and improving customer satisfaction by analyzing each business process carefully looking for opportunities to seize.

The dashboard consists mainly of 4 pages:

Marketing Campaigns Dashboard:

- It is mainly considered with the marketing team aiming to improve its performance to help them target clients who are more likely to use our services.
- This takes you into more detail about Marketing Campaigns results.

Customer Dashboard:

- This is mainly considered with the customer support team to help us achieve better customer satisfaction.
- Discuss complaints/Inquiries /Feedback actions and you can get a lot of insights from it.

Renewal & Cancellation Dashboards:

- Those dashboards are mainly considered with the financial team to help us improve our profit.
- Discuss what factors affect a client's decision to renew or cancel our services loan/deposit.

STEP 6 Desktop Application

Let's put it all together.

We have achieved a lot, worked on so much and gained so many insights and we always succeeded to find the proper way to communicate each insight and to log each step throughout the project as this file (The project documentations) covers in boring details every little step/challenge/output we went throughout the project, **BUT** what if a non-technical-non-patient student wants to see the project and have a look on what we did, the project docs will be a lot for him and that is why we have decided we need a single point to collect all our work together and that will be a **DESKTOP APPLICATION**.

- We will not bother you with any technical details here as the GUI is designed to be user-friendly and needs no guidance on how to navigate it.
- Feel free to go through the application to see the glory we have achieved throughout our work.

MAIN WINDOW