Graded Assignment 2 – Data Warehouse

Business Case & Extraction

The first step of this project's implementation was the detection of a fitting dataset. The criteria that were taken under consideration for the selection of the dataset were its size (approximately >100.000 observations), the number of its variables (more than 10) and the diversity of the data included (balance between the number of dimensions and measures). The most important factor considered for the selection of the dataset was its potential for development of an interesting business case. For the selection of the final dataset, research was conducted on a series of websites (for example Kaggle, GitHub and KDNuggets). Also, Google's Dataset Search tool was used in the above research. The final dataset was selected from one of the previously mentioned websites, specifically, it was selected from the website of Kaggle (https://www.kaggle.com/mrferozi/loan-data-for-dummy-bank).

The Irish Bank 'Fion Bank' has requested that a report be composed to evaluate the factors it uses to identify the financial danger of its borrowers. This information is very valuable for the bank since it will be used to adjust its future loan programs. An additional objective of the bank is to identify the relationship between the recoveries of the loans it lends, with the borrower's characteristics. Finally, the bank aims to develop a better understanding in the relationship between the amount of the loans it lends, the loan purpose and the borrower's characteristics.

The extracted dataset included the following variables:

- **ID:** Unique number of each loan
- **Year:** Year the loan was lent
- **Issue_d:** Date the loan was lent
- **Final_d:** Date the loan was fully paid
- **Emp_length_int:** Years of borrower's employment (values from 0 to 10)
- **Home_ownership:** Borrower's property ownership (for example: MORTGAGE, RENT, OWN etc.)
- **Home_ownership_cat:** Property ownership ID
- **Income_category:** Income Category (Low, Medium, High)
- **Income_cat:** Income category ID
- **Annual_inc:** Borrower's annual income
- **Loan_amount:** Loan amount
- **Term:** Loan repayment months
- **Term_cat:** Loan repayment months ID
- **Application_type:** Categorical variable that shows whether it is an individual or a joint loan
- **Application_type_cat:** ID for application type variable
- **Purpose:** Loan purpose (for example: small business, car, credit card etc.)
- **Purpose_cat:** Loan purpose ID
- **Interest_payments:** Category of interest rates
- **Interest_payments_cat:** ID for categories of interest rates
- **Loan_condition:** Categorical variable that shows whether the loan was fully repaid on time

- **Loan_condition_cat:** ID of loan condition
- **Interest_rate:** Interest Rate
- **Grade:** Loan grade attributed from the bank
- **Grade_cat:** ID of loan grade
- **Dti:** A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, divided by the borrower's self-reported monthly income.
- **Total_pymnt:** Borrower's total payments (with fees)
- **Total_rec_prncp:** Borrower's total payments (w/o fees)
- **Recoveries:** The amount recovered when a loan default
- **Installment:** The monthly amount the borrower must pay the bank for the loan
- **Region:** Borrower's region

Transformation

The main tool used to process and clean the data was Microsoft Excel.

Initially, the quality of the dataset's variables was evaluated. It was decided that the following variables must be removed from the dataset.

- **Year:** This variable was removed because its information already exists in another variable (issue_d)
- **Home_ownership_cat, term_cat, Application_type_cat, Purpose_cat, Loan_condition_cat, Grade_cat:** These variables were removed because they do not offer additional information to the current analysis
- **Interest_payments, Interest_payments_cat, Income_cat, Income_category:** These variables were removed because, in their current forms, they did not cover the needs of the current analysis

Upon removing the redundant variables, the remaining variables were examined for missing values and outliers. No missing values were detected, but there were a few outliers in the data of the variable 'dti'. It was identified that the reason behind those values was that some customers had no self-reported income, which caused a malfunction in the division that calculated dti for the borrowers. All outliers detected in the dti variable were replaced with NULL.

Afterwards, it was noticed that the variable 'final_d' had a wrong form that did not match any known date format. This problem was solved by applying excel functions (concat, left, right) on this variable to transform its values to a valid date format. Also, the variables 'Interest_payments' and 'Income_category' of the original dataset were replaced with two new variables named 'Interest_Rate_cat' and 'Annual_Income_Category' respectively. These variables group the interest rates of the loans and the borrower's annual income in more flexible categories. Additionally, the three components of each of the date variables included in the dataset were divided (day, month, year). As a result, the original 'issue_d' and 'final_d' variables were dropped, and six new variables were created. Finally, the processed dataset was saved in a csv format.
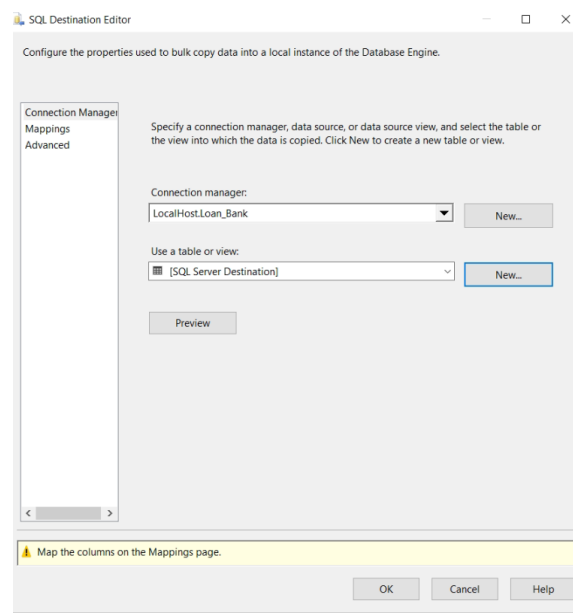
Loading

For the loading process of this project, SQL Server Management Studio (SSMS) and 'Integrations Services' of Microsoft Visual Studio (SSDT) were used.

First, a database was created in SSMS and a new Integration Services Project in SSDT. Afterwards, a new 'Data Flow' task was created in the SSDT project that loads the file through a 'Flat File Connection Manager'. In this step, the data type (int, varchar etc.) for each variable was determined. Inside the data flow task there is also a SQL Destination Editor task, which essentially shows the SSDT the destination it should load the file to.
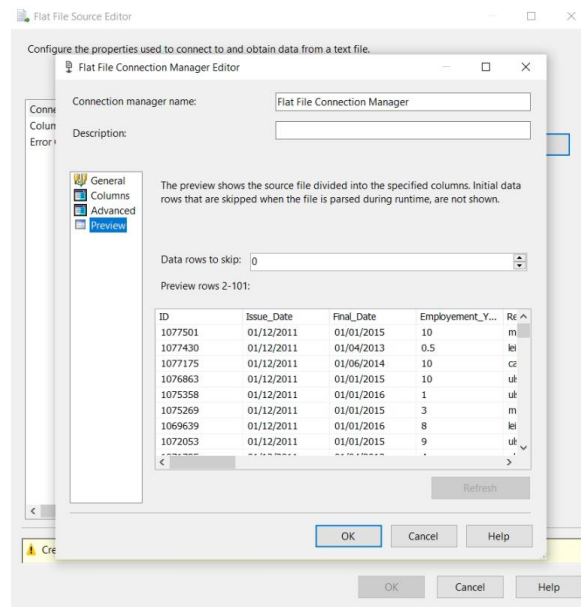
1. Flat File Connection Manager Editor

2. SQL Destination Editor

A new 'SQL Task' was created to truncate the table 'Loans_Table' before it starts loading the data. Practically, each time this service is executed, first it deletes all the contents of the table, and then executes a full load of the data contained in the csv selected in the Flat File Connection Manager.
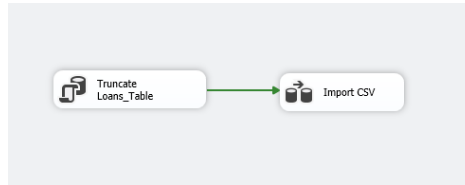
3. Truncate SQL Task

Then, the consistency of the above process was examined through the execution of a variety of SQL Queries in SSMS. An example is provided, in which the total number of rows was selected for the table Loans_Table.
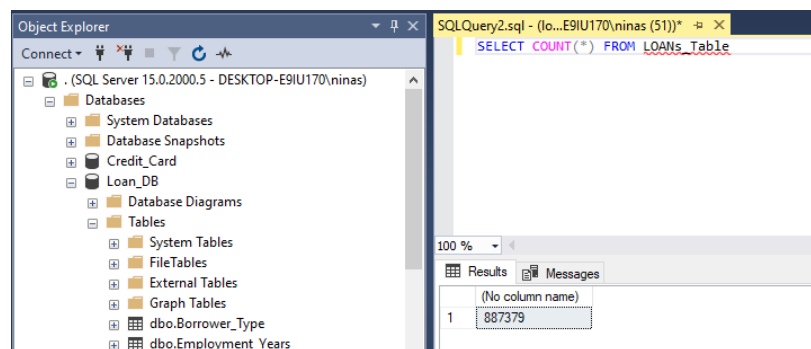
4. SQL Query

Thereafter, all the dimension tables of the dataset were created. Specifically, the following tables were created with their respective variables from the original table:

- **Loan_Purpose**
    - i. Loan_Purpose
- **Region**
    - i. Region
- **Interest_Rate_Category:** The new interest rate categories that were created in the transformation phase
    - i. Interest_Rate_Category
- **Employment_Years**
    - i. Employment_years
- **Issue_Date**
    - i. Issue_day
    - ii. Issue_month
    - iii. Issue_year

- **Final_Date**
    i. Final_day
    ii. Final_month
    iii. Final_year
- **Borrower_Type:** Borrower's income category and application type of the loan
    i. Application_type
    ii. Annual_Income_Category
- **Home_ownership:** Home ownership categories
    i. Home_ownership
- **Loan_class:** Loan characteristics
    i. Loan_Grade
    ii. Loan_Condition
    iii. Terms
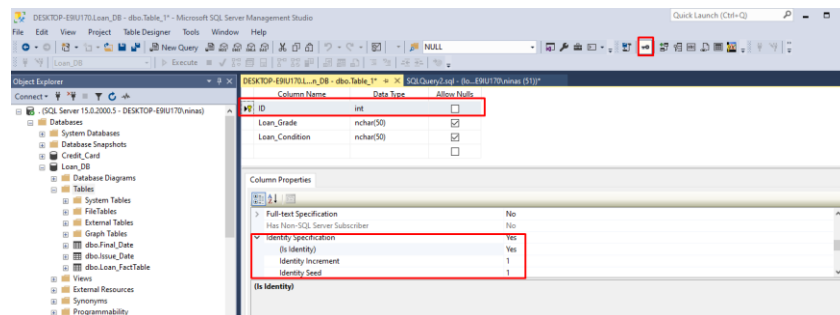
5. Create Tables Example

For each dimension table, primary keys and index constraints were created, to avoid having duplicate values.
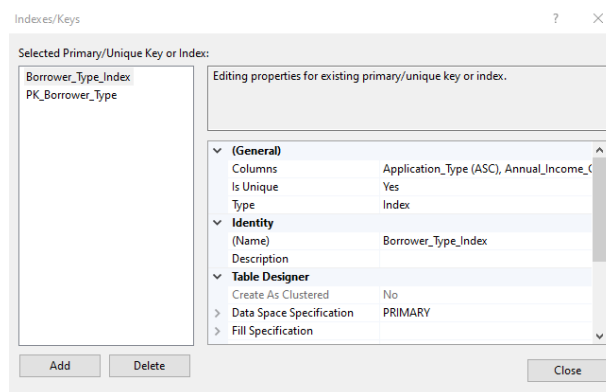
6. Constraints of the table

Traditionally, to pour data into a SSMS table, SQL Queries must be typed. Usually, SQL Queries for these tasks have the following form.

```
INSERT INTO Borrower_Type SELECT DISTINCT Application_Type, Annual_Income_Category FROM Loans_Table
```

100 %

Results | Messages

	Application_Type	Annual_Income_Category
1	JOINT	20k-35k
2	JOINT	50k-75k
3	INDIVIDUAL	50k-75k
4	JOINT	100k-250k
5	INDIVIDUAL	Less than 20k
6	INDIVIDUAL	100k-250k
7	INDIVIDUAL	Over 250k
8	JOINT	Less than 20k
9	JOINT	75k-100k
10	INDIVIDUAL	20k-35k

7. Insert statement through an SQL Query

A SQL Task was created for each one of the dimension tables, which updates each dimension table only with the data that belong to it from the initial table created (Loans_Table).
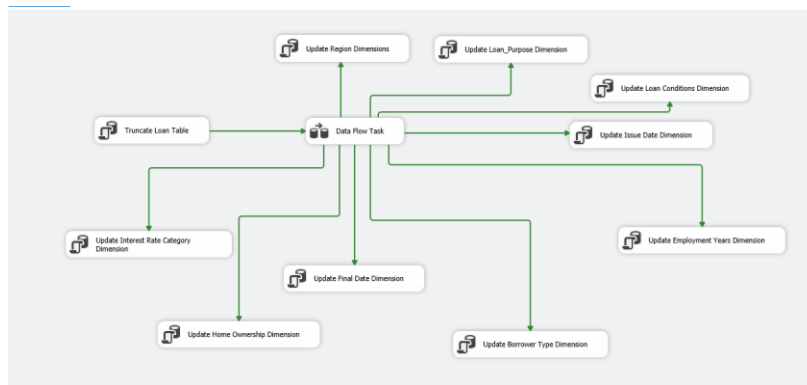
8. Dimension Tables

In the next step, a Fact Table was created in the SSMS. The fact table connects with all the dimension tables that were created in the previous step using foreign keys. This table also contains all the measures of the initial dataset, which will be used to conduct calculations and grouping of the data. The fact table will be used as the canvas on which the analysis will be conducted.

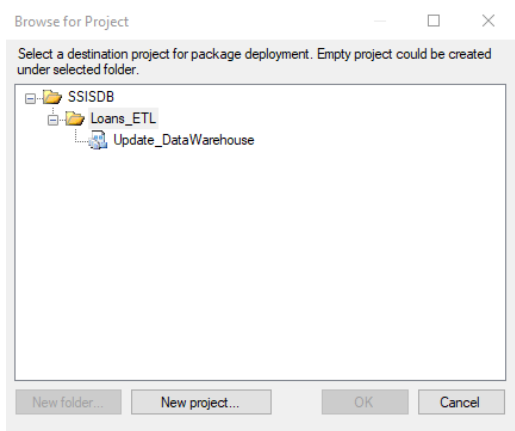| Column Name	Data Type	Allow Nulls
ID	int	☐
Issue_Date	int	☐
Final_Date	int	☐
Borrower_Type	int	☐
Employment_Years	int	☐
Home_Ownership	int	☐
Loan_Class	int	☐
Loan_Purpose	int	☐
Interest_Rate_Category	int	☐
Region	int	☐
Annual_Income	real	☑
Loan_Amount	real	☑
Interest_Rate	real	☑
dti	real	☑
Total_Payment	real	☑
Total_rec_prncp	real	☑
Recoveries	real	☑
Installment	real	☑
		☐

9. Fact Table
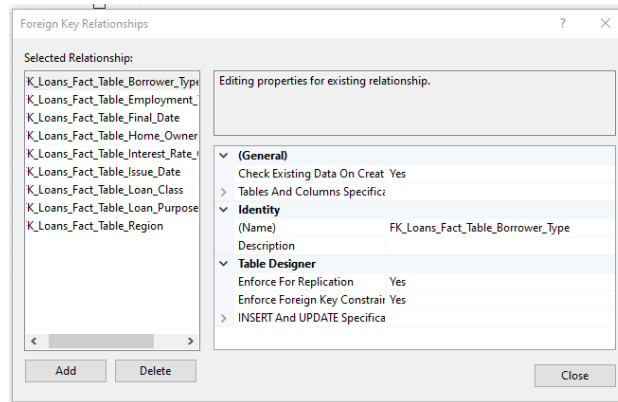
10. Foreign Keys

Every time the SSDT series of tasks are executed, all the dimension tables included in the SQL tasks are updated with all the new values contained in the csv file that is read from the Flat File Connection Manager. In order to update the fact table automatically every time the service is executed, a new SQL Task is created. This task updates the fact table with the foreign keys of the new data in the dimension tables it is connected to.
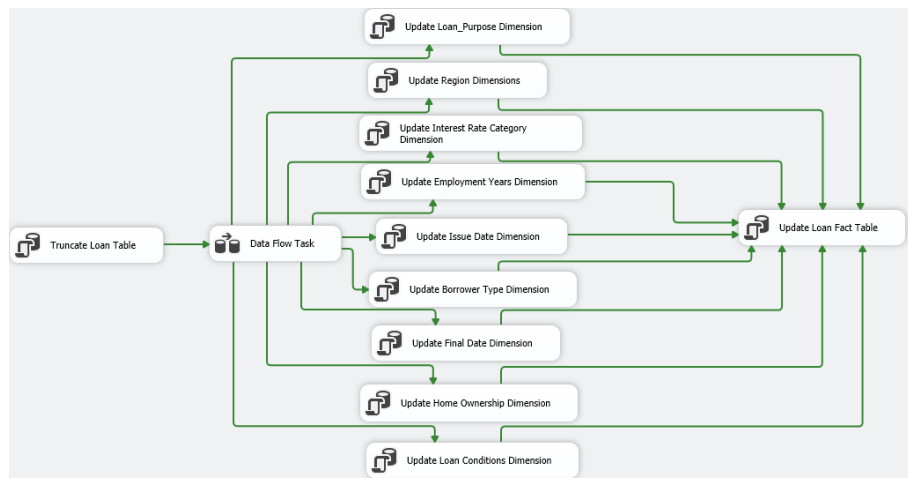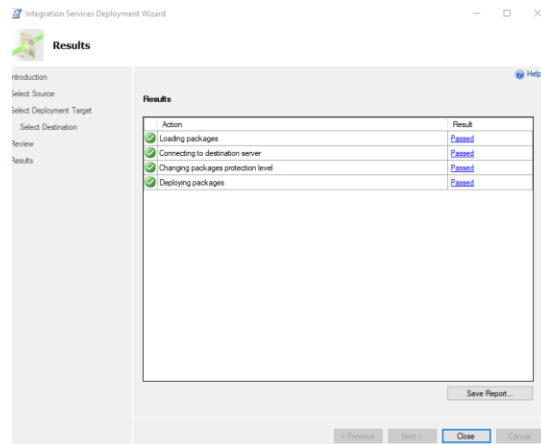

11. Process flow of SSDT

Subsequently, a shortcut of the above process was created inside the SSMS. First, to activate that shortcut, a new project must be created inside the Integration Services Catalog of SSMS. Then, that project must be connected to the Integration Services Project of SSDT. As a result, an administrator user can execute the SSDT process flow through the SSMS. Finally, a daily schedule task was created for the above process, which executes the process automatically every day.
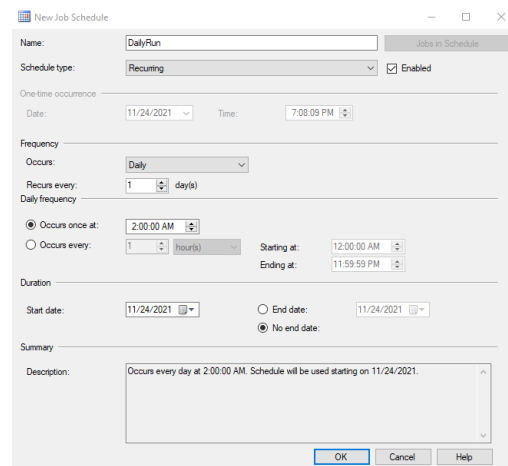
12. Integration Services Catalog

13. SSDT Project Deployment

14. Service results in the SSMS view

15. Service Schedule

Multidimensional Model

In the following steps, the multidimensional model (cube) that would host the fact table was created. The 'Analysis Services Multidimensional and Data Mining' project off SSDT was used to create the cube. First, the new project connected with the Loan_DB database in SSMS that includes the fact table and all the dimensions. Then, the fact table and its corresponding dimension tables were selected as data sources that would fill the cube. Also, in this step of the process the star-schema was created. During the previous step it was also confirmed that all the dimensions and measures were identified correctly in the project.
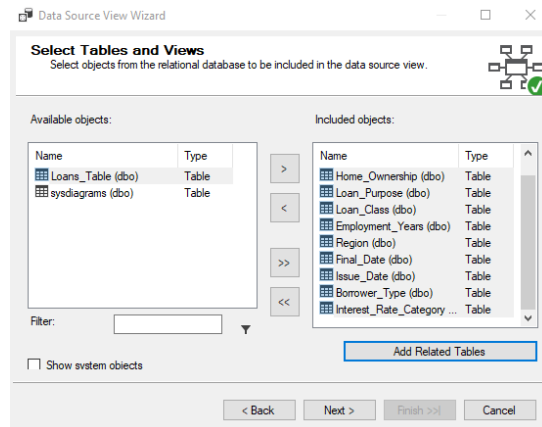
16. Fact Table & Dimension Tables
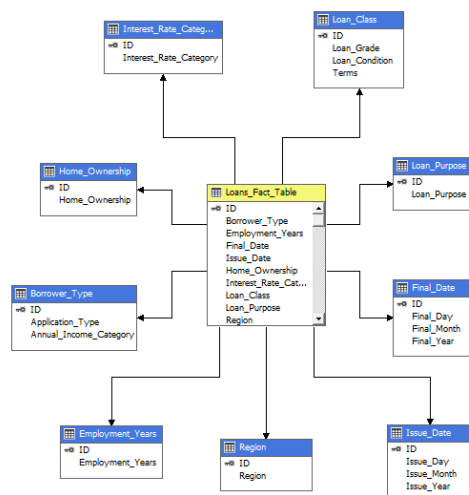
17. Star Schema

The previous steps were completed using a Microsoft account. Though, to process the Multidimensional model successfully a local user had to be used. To resolve that issue, a local user was created with whom all the following steps were executed. The local user was assigned administrator authorizations on the SSMS database and the SSDT.
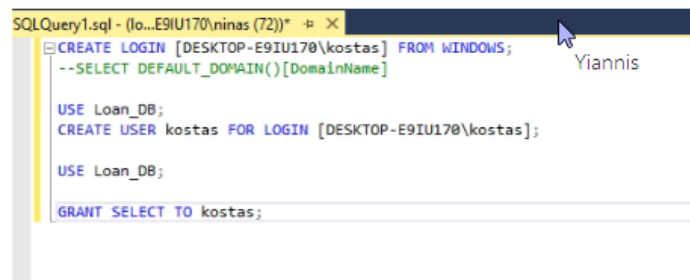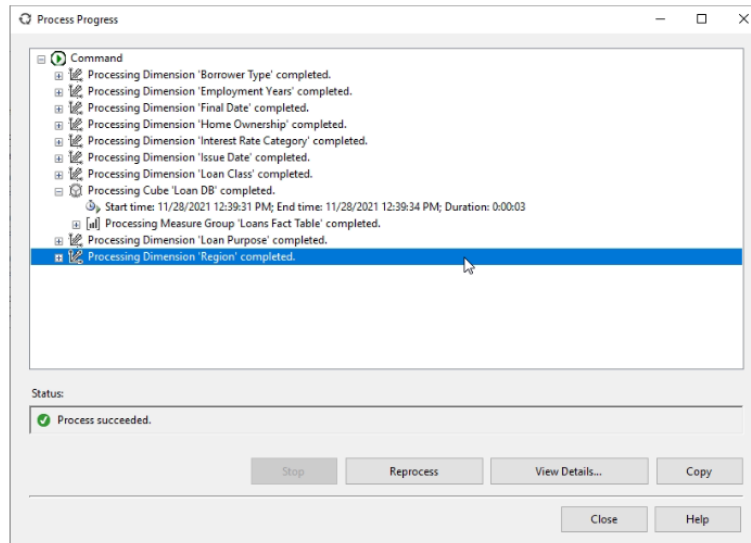

18. Grant authorizations to local user

19. Cube Deployment

Through examination of the cube, it was identified that all the dimensions of the fact table included only their primary keys. To deal with this issue, all the dimension tables inside the cube were processed through SSDT. Essentially, for every variable of each dimension, a drag and drop process was executed, where all the variables except ID were moved from 'Data Source View' table to 'Attributes' table.
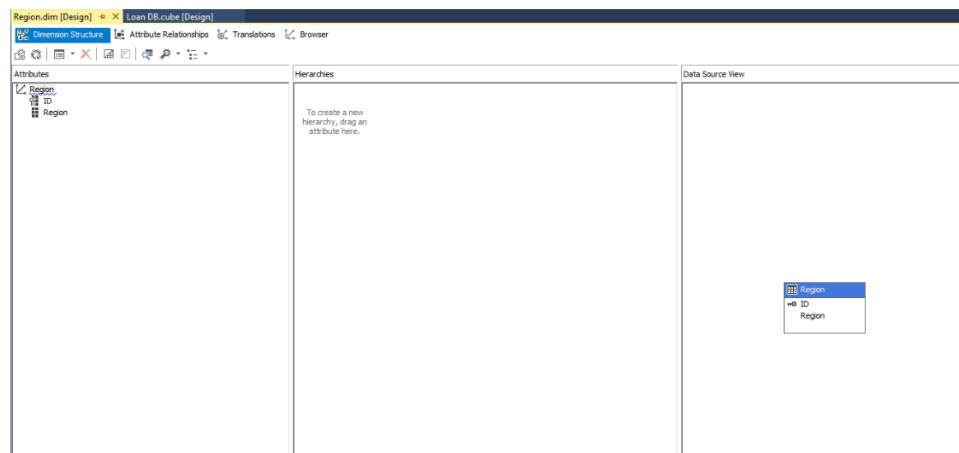
20. Revealing the variables of dimensions

To broaden the flexibility of the analysis, some calculated measures were created, that originate from existing measures in the fact table of SSDT. These measures calculate values such as the mean or the difference between two or more existing measures. Indicatively, some of the measures created were the mean for loan amounts, the mean for interest rates and the balance of bank's borrowers (loan amount – total amount paid). These calculations were conducted in the 'Calculations' tab of SSDT.
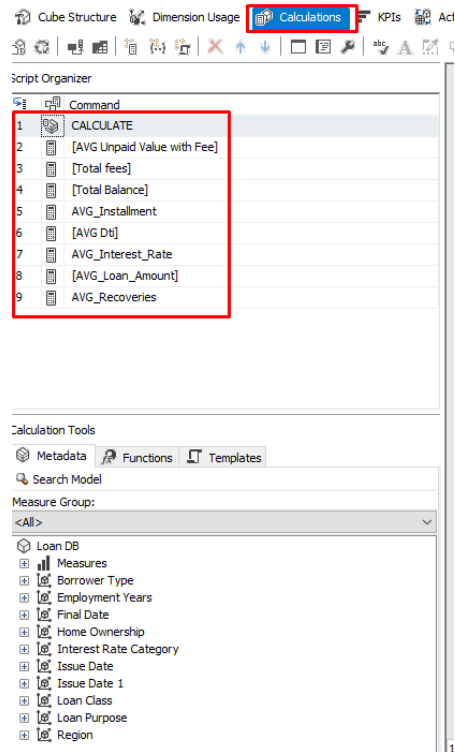
21. Calculation of new measures

To examine the consistency of the data, a lot of SQL Queries were executed in SSMS, from whom the results were compared with the results of the same query in SSDT.

22. Query results in SSDT

23. Query results in SSMS

After the verification of the consistency of the results, a new task was created in the initial Integration Services Project. The new task processes the multidimensional model automatically with all the new data from the csv file in the data source every time the Integration Services project is executed. These steps ensure that the cube's data are always up to date with all the data included in its dimensions, without requiring a user to execute the process manually every time.
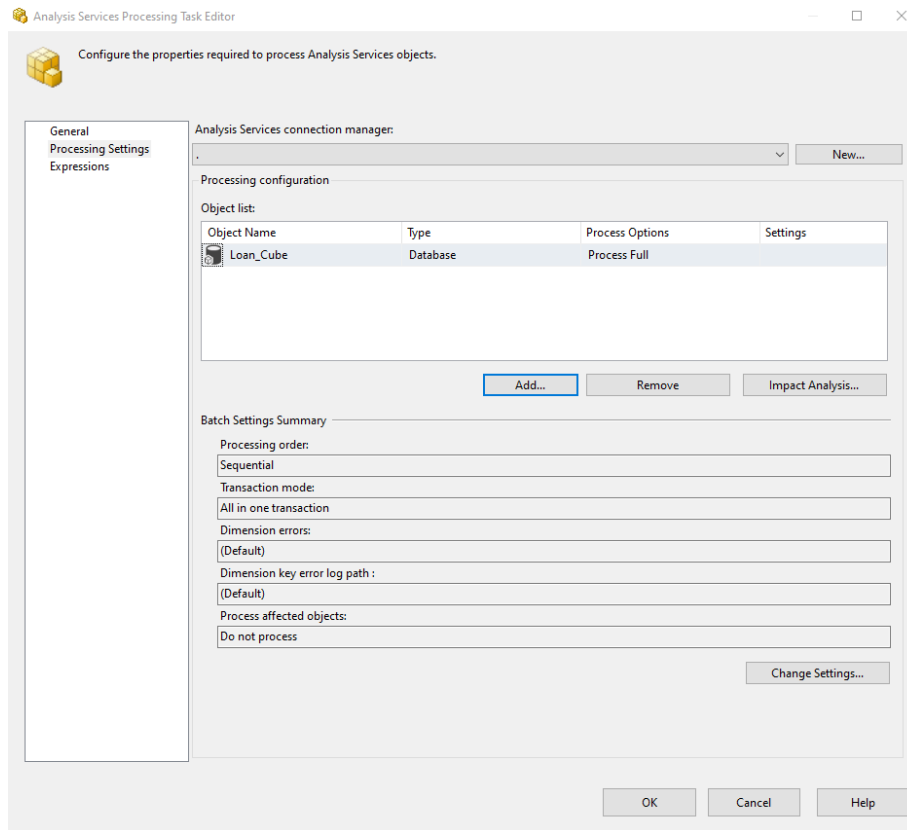
24. Analysis Services Processing Task
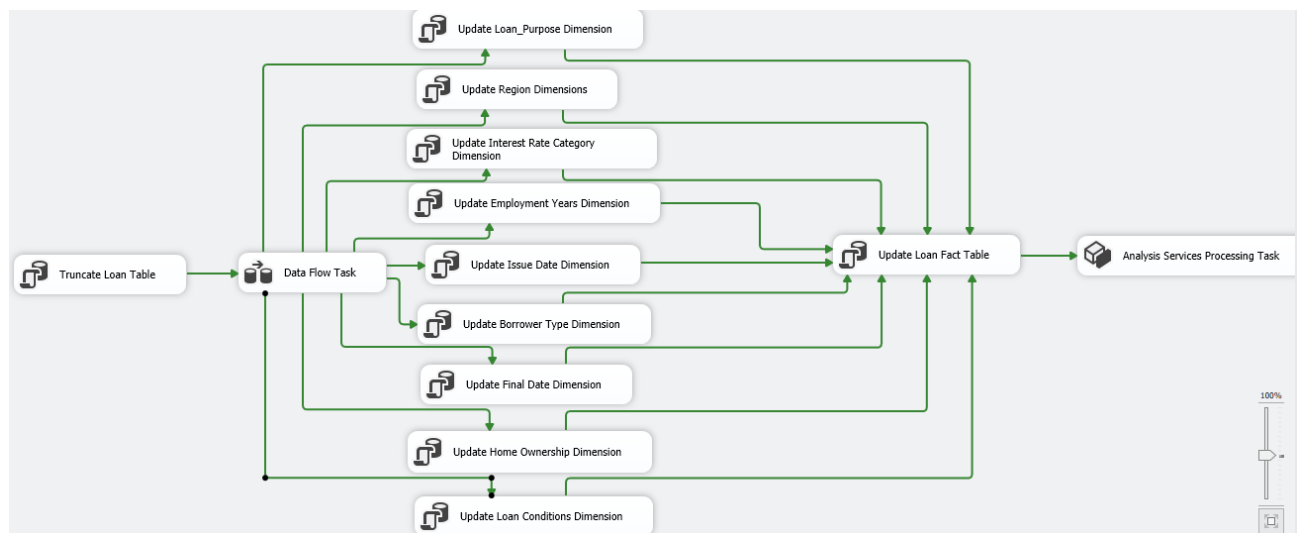
25. Integration Services Final Flow

Reporting

The multidimensional model that was created in the above steps, can be used to prepare a plethora of reports in no time. The creation of such reports is easier through the cube because writing such sql queries can prove to be very complex and time consuming. In the 'Browser' tab of SSDT, by choosing the Excel icon, the multidimensional model is exported in Excel format.
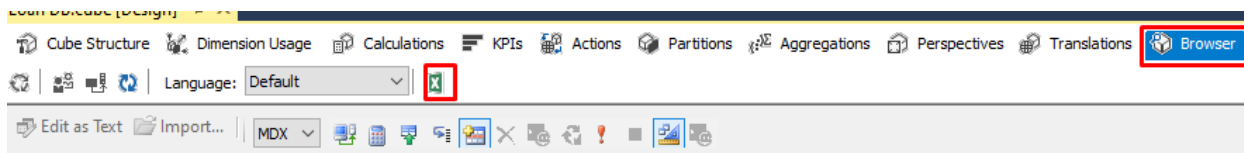
26. Export to Microsoft Excel from SSDT
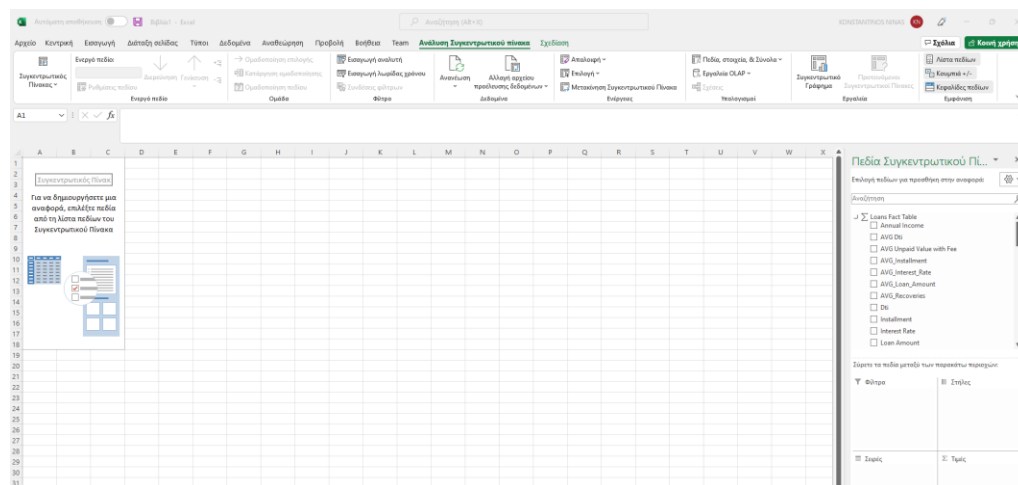
27. Multidimensional Model data through excel view

After the extraction of the cube in excel format, a variety of reports were created, such as the calculation of the loan amounts granted, and the total payments collected by the borrower's income category. From this report it can be deducted that the greatest percentage of the bank's borrowers has an income between 50-75K€. As expected, these customers also cover the biggest percentage of earnings for the bank. Additionally, it is observed that there is a relationship between the average loan amount and the income category of the borrowers. Specifically, the average loan amount tends to rise for borrowers with greater income.

| Ετικέτες γραμμής	Loan Amount	Total Payment	Total Rec Prncp	Total Balance	Total fees	Loans Fact Table Count	AVG_Loan_Amount
100k-250k	3,127,929,856.00 €	1,546,750,208.00 €	1,189,538,688.00 €	- 1,581,179,648.00 €	357,211,520.00 €	149,793	20,881.68 €
20k-35k	775,394,176.00 €	409,868,352.00 €	315,304,256.00 €	- 365,525,824.00 €	94,564,096.00 €	92,551	8,378.02 €
35k-50k	2,062,785,408.00 €	1,094,318,592.00 €	830,503,936.00 €	- 968,466,816.00 €	263,814,656.00 €	185,480	11,121.34 €
50k-75k	3,930,502,656.00 €	2,031,537,792.00 €	1,533,937,280.00 €	- 1,898,964,864.00 €	497,600,512.00 €	274,755	14,305.48 €
75k-100k	2,898,245,120.00 €	1,480,109,056.00 €	1,127,680,256.00 €	- 1,418,136,064.00 €	352,428,800.00 €	163,726	17,701.80 €
Less than 20k	96,682,800.00 €	48,341,024.00 €	36,905,132.00 €	- 48,341,776.00 €	11,435,892.00 €	13,107	7,376.42 €
Over 250k	202,010,944.00 €	96,619,600.00 €	75,399,616.00 €	- 105,391,344.00 €	21,219,984.00 €	7,967	25,355.96 €
Γενικό Άθροισμα	13,093,440,512.00 €	6,707,547,648.00 €	5,109,267,456.00 €	- 6,385,892,864.00 €	1,598,280,192.00 €	887,379	14,755.18 €

28. Loan and payment amounts per income category

Also, a report depicting the top 5 loan purposes for granting a loan in terms of average recoveries for the bank in case the borrower bankrupts. From the report, it can be assumed that the loans that don't get repaid on time are accompanied with higher interest rates, higher loan amounts and the borrowers owe higher monthly payments to the bank. In each of these five cases, it can be observed that the absolute number of borrowers that don't repay their loan on time (or at all) is at most equal to 1/5 of the number of all borrowers with this specific loan purpose.

Ετικέτες γραμμής	AVG_Recoveries	AVG_Installment	AVG_Loan_Amount	AVG_Interest_Rate	Loans Fact Table Count
⊟ educational					
Bad Loan	508.70 €	233.14 €	7,337.78 €	13.33	88
Good Loan	- €	209.37 €	6,426.49 €	11.80	335
⊟ house					
Bad Loan	815.95 €	461.71 €	14,785.19 €	17.82	417
Good Loan	- €	458.03 €	14,799.74 €	15.78	3290
⊟ renewable_energy					
Bad Loan	766.38 €	352.53 €	10,957.29 €	16.06	72
Good Loan	- €	312.55 €	9,803.88 €	15.30	503
⊟ small_business					
Bad Loan	889.85 €	511.30 €	16,374.83 €	17.28	1781
Good Loan	- €	481.23 €	15,222.69 €	16.02	8596
⊟ wedding					
Bad Loan	867.00 €	367.28 €	11,539.53 €	15.91	289
Good Loan	- €	325.76 €	10,318.25 €	13.90	2058
Γενικό Άθροισμα	130.56 €	447.33 €	14,227.04 €	15.78	17,429

29. Rank – 5 best loan purposes in terms of average recoveries

Last but not least, a report depicting the average loan amount in terms of the borrowers' home ownership and the loan grade assigned from the bank. It is observed that the mean loan amount tends to rise while the loan grade drops. In other words, the loans that get assigned with lower grades tend to be accompanied with higher loan amounts, regardless of the home ownership of the borrower and the repayment status of the loan.

AVG_Loan_Amount Ετικέτες στήλης								
Ετικέτες γραμμής	A	B	C	D	E	F	G	Γενικό Άθροισμα
⊟ Good Loan								
MORTGAGE	15,035.30 €	15,116.68 €	16,444.07 €	17,696.35 €	20,187.37 €	21,286.85 €	22,437.89 €	16,341.46 €
NONE	8,085.00 €	8,840.00 €	13,827.78 €	17,275.00 €	35,000.00 €	24,550.00 €	25,000.00 €	13,638.69 €
OTHER	6,142.71 €	10,764.10 €	10,460.81 €	12,065.22 €	12,364.58 €	18,900.00 €	13,337.50 €	10,611.11 €
OWN	13,940.59 €	13,309.48 €	13,830.93 €	14,777.52 €	17,081.53 €	18,004.07 €	19,815.27 €	14,232.33 €
RENT	12,437.51 €	11,930.35 €	12,455.12 €	13,356.76 €	15,784.42 €	16,888.80 €	18,453.46 €	12,844.89 €
⊟ Bad Loan								
MORTGAGE	13,735.42 €	14,255.99 €	15,904.49 €	17,420.61 €	20,069.42 €	21,726.70 €	22,273.24 €	16,844.44 €
NONE		10,143.75 €	9,000.00 €	18,287.50 €		15,000.00 €		12,643.75 €
OTHER	13,100.00 €	11,479.17 €	10,542.86 €	11,814.29 €	13,540.00 €	22,500.00 €	2,100.00 €	11,564.47 €
OWN	12,498.74 €	12,402.51 €	12,901.84 €	14,150.76 €	16,830.90 €	17,374.26 €	20,988.77 €	14,100.14 €
RENT	11,064.31 €	11,453.35 €	12,054.37 €	12,916.42 €	15,719.63 €	17,573.32 €	18,882.51 €	13,136.32 €
Γενικό Άθροισμα	14,038.86 €	13,643.00 €	14,466.23 €	15,459.45 €	17,998.15 €	19,139.16 €	20,538.56 €	14,755.24 €

30. Average loan amount in terms of home ownership and loan grade

Data Visualization

Aiming to conduct a more detailed and identify possible patterns in the data, a connection was created between the multidimensional model and a data visualization tool. The tool used to perform the above steps was Microsoft Power BI.

The first summarized interpretations excluded from the analysis of the data are related to some descriptive insights on the activities of the bank in a 9-year term (2007-2015). It is observed that the bank has granted more than 880.000 loans to its borrowers, with a total amount greater than 13 billion €. In fact, it is worth mentioning that the number of loans granted rose greatly in the last 3 years documented, almost reaching 90% of total loans granted. Up to this day, the bank has collected almost 6.7 billion €, of which, 1.6 billion € are fees collected from the bank's borrowers. The greatest percentage of the bank's borrowers have medium to high incomes, in Ireland. The income of the bank's most frequent customers is between 50-75K€. On the other hand, the borrowers that get granted the least number of loans belong either to the greatest, or the lowest income category. Finally, it is observed that the greatest amount of loans granted can be found in borrowers that have medium and large incomes.
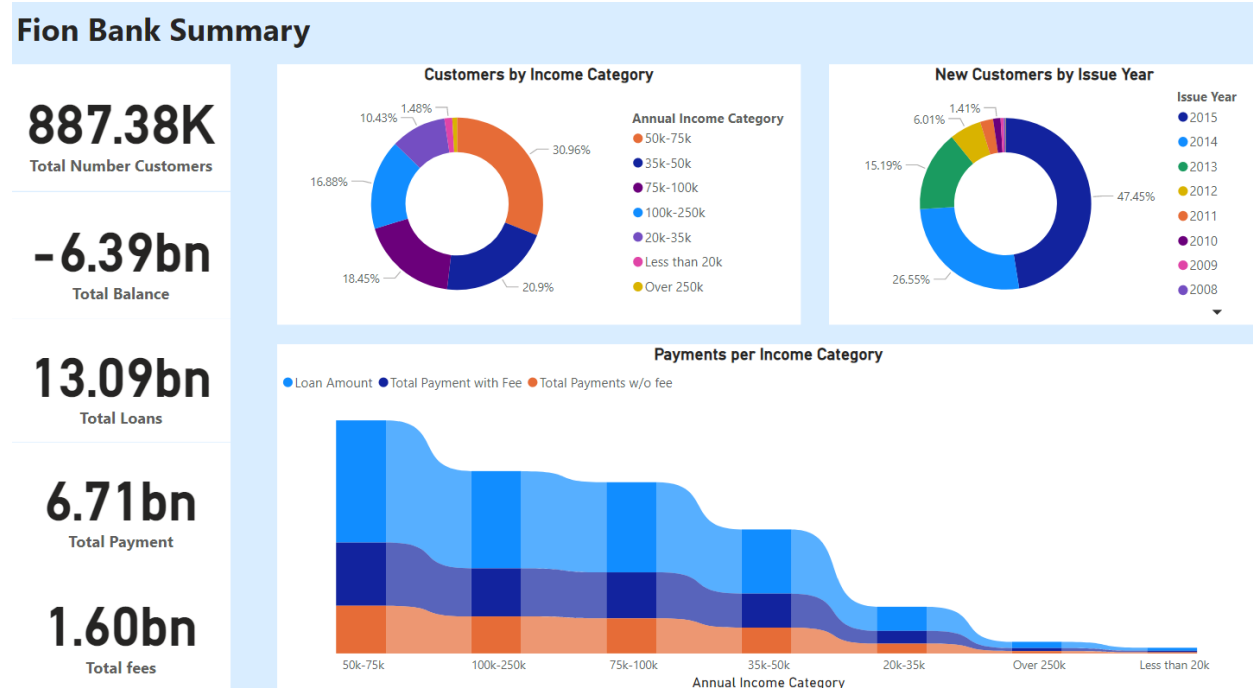

31. Fion Bank Summary

It is of great importance for Fion Bank to understand the characteristics of its borrowers, so as to adjust in a better manner the conditions of the loans it grants to them. First, it is identified that almost half of the bank's borrowers have a mortgage in Fion Bank. At the same time, a high percentage of its borrowers either rent or own their place of residence. Also, it is observed that only 2% of the total borrowers do not own or possess any place of residence, or it has not been recorded from the bank. It is noticed that borrowers with greater incomes usually receive loans of higher average value. In fact, the rise of the average loan amount by the borrowers' income category is scaling. In the third graph, two clear patterns can be observed related to the dti index and the borrowers' average income. The first pattern projects an upward trend in the borrowers' average income and a downward trend in the average dti, while the loan

grade is improving (D-A). On the other hand, the exact reverse behavior is observed for loans with grades from E to G. Additionally, it is noticed that the borrowers that owe higher average monthly payments, have received either business or mortgage loans. The high value of these installments is a result of the high, in average, loan amounts and the high interest rates that accompany those loans. It is worth mentioning that the borrowers with incomes that are less than 20K € appear to have greater inclination in receiving loans to renovate their house in comparison to borrowers with slightly higher incomes. Concluding, the borrowers with incomes between 100-250K € appear to have greater inclination in receiving educational loans in comparison to the rest of the borrowers, regardless of their income.
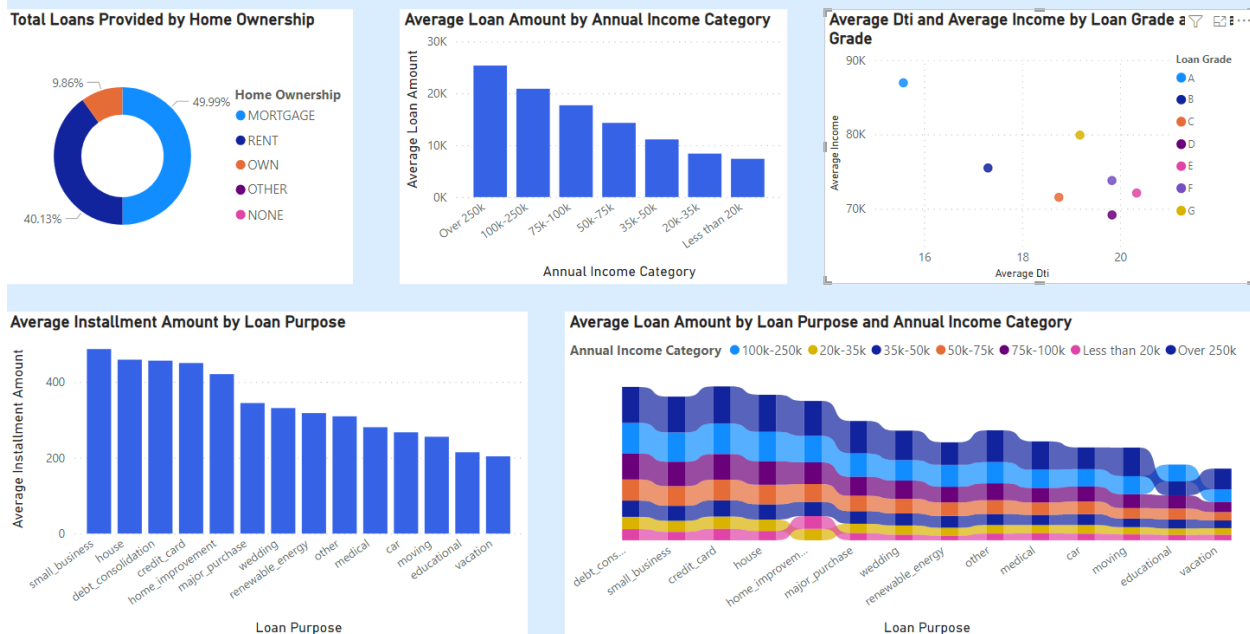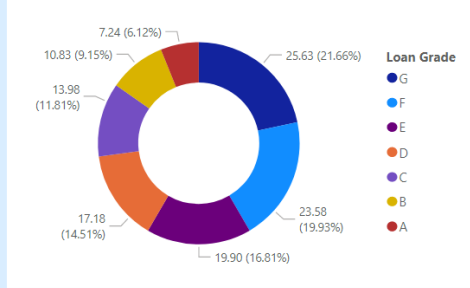

32. Borrower Profile

It is a major issue for Fion Bank to correctly evaluate the factors that affect the financial danger of its borrowers. Initially, it is perceived that the interest rates that accompany the loans are inversely proportional to the borrowers' income and the loan grade assigned from the bank. In other words, the interest rate tends to rise for borrowers with lower income and for loans with a low loan grade, because the bank considers that there is a greater chance of non-repayment of the loan for these individuals. As expected, it is observed that the borrowers that own or rent a place of residence receive lower average interest rates. The most favorable interest rates are granted to borrowers that have mortgages in the bank. It is worth mentioning that usually, the bank grants higher average interest rates to loans that were not repaid on time, or at all (bad loans). Though, the bank shouldn't have granted these loans in the first place, which leads to the conclusion that the bank must improve its borrowers' evaluation model. Finally, it is perceived that mortgage and small-business loans receive the highest average interest rates. This choice is based on the fact that the bank considers these two types of loans as the most high-risk in terms of not being repaid from the borrowers. Thus, for loans that are deduced as low risk in terms of possible not repayment, the bank grants slightly lower interest rates. According to Fion Bank, in the category of low-risk loans belong loans that are related to credit cards, to car purchases and to educational purposes.
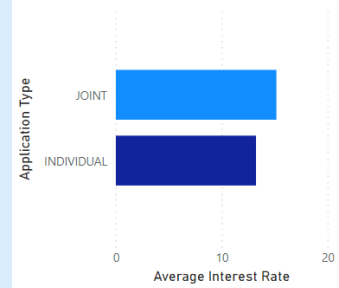
Customer Risk

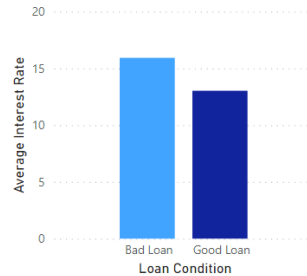**Average Interest Rate by Annual Income Category**
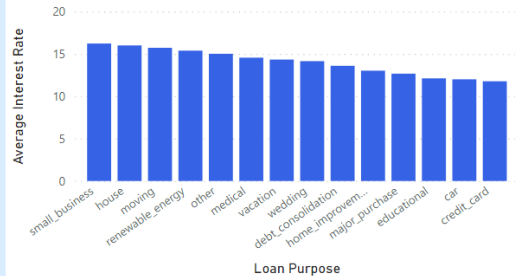
Average Interest Rate by Loan Grade

Loan Grade
- G
- F
- E
- D
- C
- B
- A

7.24 (6.12%)
10.83 (9.15%)
13.98 (11.81%)
25.63 (21.66%)
23.58 (19.93%)
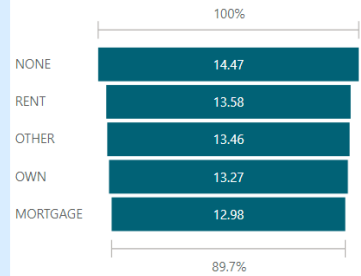19.90 (16.81%)
17.18 (14.51%)

Average Interest Rate by Application Type

Average Interest Rate by Loan Condition

Average Interest Rate by Loan Purpose

Average Interest Rate by Home Ownership

|	
NONE	14.47
RENT	13.58
OTHER	13.46
OWN	13.27
MORTGAGE	12.98

33. Customer's Financial Risk

It is essential for the bank to understand the connection between the recovery amount, in case a loan defaults, with the characteristics of the borrower. According to the analysis of the data, the first thing that is observed, is that for the borrowers that belong to categories of middle to high income (>50K €), there is a higher average recovery amount compared to the borrowers with lower incomes. In addition, it is observed that loans granted to small businesses that default, have the highest average recoveries, while also having the second highest average loan amount. On the other hand, consumer loans granted for vacations, where there is both a low average loan amount and low recovery amount. It is worth mentioning that educational loans have a low average loan amount but high average recoveries. It is also observed that loans with higher credit ratings and lower interest rates have lower average recovery amounts, highlighting the fact that borrowers with better solvency ratings from the bank go bankrupt less frequently. In fact, this occurs in scale according to the loan grade. Finally, as expected, borrowers with no real estates are found to have the highest average recovery amount due to the higher probability of bankruptcy on average, while at the same time they have a high average amount of monthly payments. In contrast, borrowers who own assets appear to have a much lower average recovery amount while they also have relatively large average monthly payment amounts.
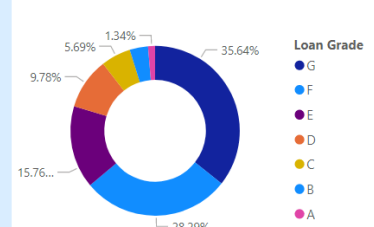
Recoveries

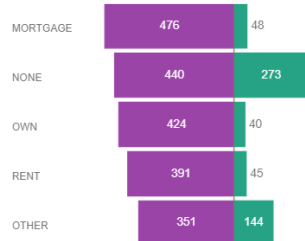### Average Recoveries by Annual Income Category
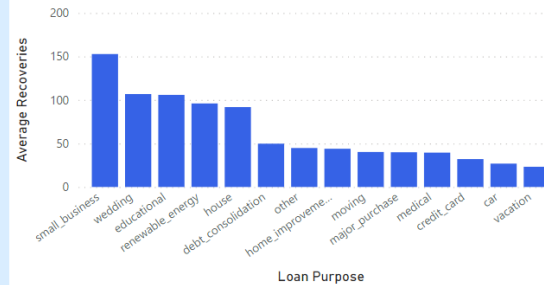

Average Recoveries by Interest Rate Category
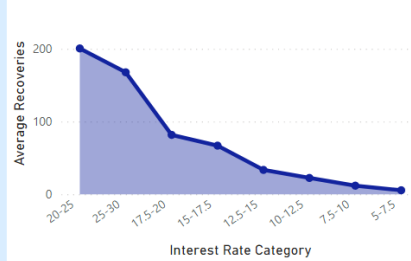
Average Recoveries by Loan Grade

Loan Grade
● G
● F
● E
● D
● C
● B
● A

Average Installments and Average Recoveries by Home Ownership

Average Recoveries by Loan Purpose

Loan Purpose	Average Loan Amount
vacation	6,289.54
educational	6,616.08
moving	7,891.57
car	8,882.43
medical	9,019.07
other	9,889.36
renewable_energy	9,948.30
wedding	10,468.63
major_purchase	11,600.00
home_improvement	14,320.94
house	14,798.10
credit_card	15,338.88
small_business	15,420.43
debt_consolidation	15,428.41
Total	**14,755.22**

34. Recoveries