1. Get Data;
2. Transform Data (Open Power Query);
3. Column Profile and Column Quantity;
4. Identification of missing values in the column “Income”;
5. Calculation of the median of this column - **51381, 5**;
6. Replacement Values – “null values” with median;
7. Creating a new custom column for customer age by using “Year\_Birth” and filtering on years < 100 as there are customers which are 120 years old;
8. I will remove these columns - “Z\_CostContact”, ” Z\_Revenue”;
9. Closing Power Query and apply the changes;
10. After that I calculate 25th and 75th percentile to identify outliers in the column “Income”. I calculate IQR and lower and upper bound;
11. Next filter in “Income” column for values greater than up bound – 117418:

A screenshot of a computer

Description automatically generated

The outlier in this case is just ID 9432 with an income 666666 and then I filtered for values less than that value.

1. For verification, I added histogram as a visualization:

A green and blue graph

Description automatically generated

1. Of course in other columns exist outliers but for purposes of this analysis these outliers are necessary;
2. Creating new columns:

* Age\_Group;
* Children – total children in one family;
* Total\_Spend – sum of total amount for all products;
* Total\_Accepted - total accepted of campaigns by customer.

1. Replaced values in columns “Education”, “Marital\_Status”:

* “Graduation” and “2n Cycle” with “Graduated”, “PhD” and “Master” with “Post Graduated”
* “Single”, “Divorced”, “Widow”, “Alone”, “Absurd”, “YOLO” with “No Relationship” and “Married”, “Together” with “In Relationship”;

1. Creation of two more tables – promotion and products by unpivot columns who need;
2. Creating measures for sum of amount of products;
3. For the RFM Analysis I Created fields:

* Frequency – sum of purchases made by different channels;
* Monetary – sum of amount of products;
* R Score – a conditional statement (SWITCH) that checks if the value is less than or equal to the nth percentile and assigns a label. In conclusion, this formula assigns an R Score to each recency value in our dataset based on its percentile rank within the dataset. This can be used to segment customers or events into different categories or tiers based on how recently they occurred, with "5" being the most recent and "1" being the least recent;
* F Score – similar to R Score but is used Frequency field with "1" being the least frequent and "5" being the most frequent;
* M Score – like F Score but is used Monetary field;
* RFM Score – concatenated the three columns R Score, F Score and M Score;
* Segment – In my other file where I use Python these fields are created with regex but in Power BI I use DAX Syntax and again conditional statement SWITCH and extract the first two characters from RFM Score with LEFT(marketing\_campaign[RFM\_Score], 2)
* If the first two characters of RFM\_Score are "11," "12," "21," or "22," the segment is "Hibernating."
* If the first two characters of RFM\_Score are "13," "14," "23," or "24," the segment is "At Risk."
* If the first two characters of RFM\_Score are "15" or "25," the segment is "Cannot Lose Them."
* If the first two characters of RFM\_Score are "31" or "32," the segment is "About To Sleep."
* If the first two characters of RFM\_Score are "33," the segment is "Need Attention."
* If the first two characters of RFM\_Score are "34," "35," "44," or "45," the segment is "Loyal Customers."
* If the first two characters of RFM\_Score are "41," the segment is "Promising."
* If the first two characters of RFM\_Score are "51," the segment is "New Customers."
* If the first two characters of RFM\_Score are "42," "43," "52," or "53," the segment is "Potential Loyalist."
* If the first two characters of RFM\_Score are "54" or "55," the segment is "Champions."

This formula categorizes customers into different segments based on the first two characters of their RFM\_Score. Each segment represents a different type of customer behavior.