Customer Profile Analysis

About:

Analysis of the database to identify the customer profile.

Proposal:

Using information extracted from a **.csv** format database, a profile analysis will be conducted, taking into consideration the factors that have the most impact on customer ratings. An analysis will be performed to identify the key factors influencing customer ratings. The final solution will consist of a set of graphs from which insights based on these analyses will be extracted.

Repository Structure:

- **data:** Here you will find the **.csv** file containing the data used for the analysis.
- **img:** This is where the screenshots obtained during the analysis can be found.
- **notebook:** This directory contains the Jupyter notebook where the analysis was conducted and the results obtained.
- **readme_translated:** This repository contains the **PDF** with the report translated into English.

Language Used:

Libraries Used:

Methodology:

Initially, we used the **Pandas** library to import and read the database. We encountered an encoding issue because the database contained special characters and accents, so we resolved the issue using the **encoding='latin'** property and the **sep=';'** property to organize and better visualize the DataFrame.

	ClienteID	Origem	Idade	Salário Anual (R$)	Nota (1-100)	Profissão	Experiência Trabalho	Tamanho Família	Unnamed: 8
0	1	Normal	19	15000	39	Saúde	1	4	NaN
1	2	Normal	21	35000	81	Engenheiro	3	3	NaN
2	3	Promoção	20	86000	2	Engenheiro	1	1	NaN
3	4	Promoção	23	59000	73	Advogado	0	2	.
4	5	Promoção	31	38000	48	Entretenimento	2	6	NaN
5	6	Promoção	22	58000	84	Artista	0	2	NaN
6	7	Promoção	35	31000	2	Saúde	1	3	NaN
7	8	Promoção	23	84000	90	Saúde	1	3	NaN
8	9	Normal	64	97000	3	Engenheiro	0	3	NaN
9	10	Promoção	30	98000	80	Artista	1	4	NaN

The next step was data preprocessing and cleaning. We used the **.info** method to check the data types we were working with and identified the presence of **35 rows** with null values and the **"Annual Salary (R$)"** column that was not in numeric format. We transformed the data, removed the **"Unnamed: 8"** column, and finally **eliminated empty rows** that did not contain relevant values.

```
<class 'pandas.core.frame.DataFrame'>
Index: 1965 entries, 0 to 1999
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   ClienteID             1965 non-null   int64
 1   Origem                1965 non-null   object
 2   Idade                 1965 non-null   int64
 3   Salário Anual (R$)    1965 non-null   float64
 4   Nota (1-100)          1965 non-null   int64
 5   Profissão             1965 non-null   object
 6   Experiência Trabalho  1965 non-null   int64
 7   Tamanho Família       1965 non-null   int64
dtypes: float64(1), int64(5), object(2)
memory usage: 138.2+ KB
```

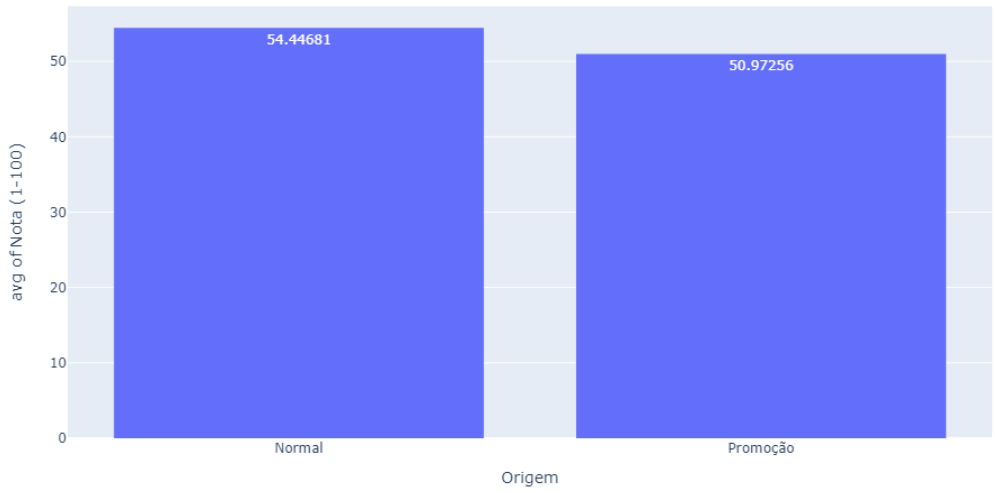
Next, we conducted an initial analysis and created the graphs. We used the **.describe** method to obtain a summary of the information and understand how the database functions. The **Plotly** library with a **for** loop was used to create the graphs.

	ClienteID	Idade	Salário Anual (R$)	Nota (1-100)	Experiência Trabalho	Tamanho Família
count	1965.000000	1965.000000	1965.000000	1965.000000	1965.000000	1965.000000
mean	1000.309924	48.894656	110616.009669	52.385242	3.675318	3.757252
std	578.443714	28.414889	45833.860195	28.593269	3.909676	1.968335
min	1.000000	0.000000	0.000000	1.000000	0.000000	1.000000
25%	498.000000	25.000000	74350.000000	29.000000	0.000000	2.000000
50%	1000.000000	48.000000	109759.000000	52.000000	1.000000	4.000000
75%	1502.000000	73.000000	149095.000000	77.000000	7.000000	5.000000
max	2000.000000	99.000000	189974.000000	100.000000	17.000000	9.000000

We will use the observed average rating of **52** as the basis for the analysis.

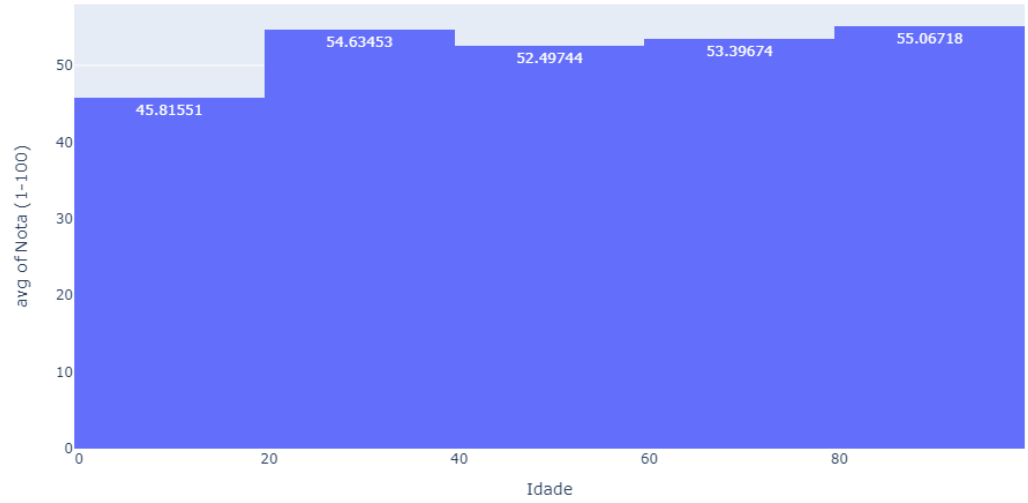
Exploratory Analysis:

- In the first graph, we compared the customer's origin (organic or through promotion) with their rating:



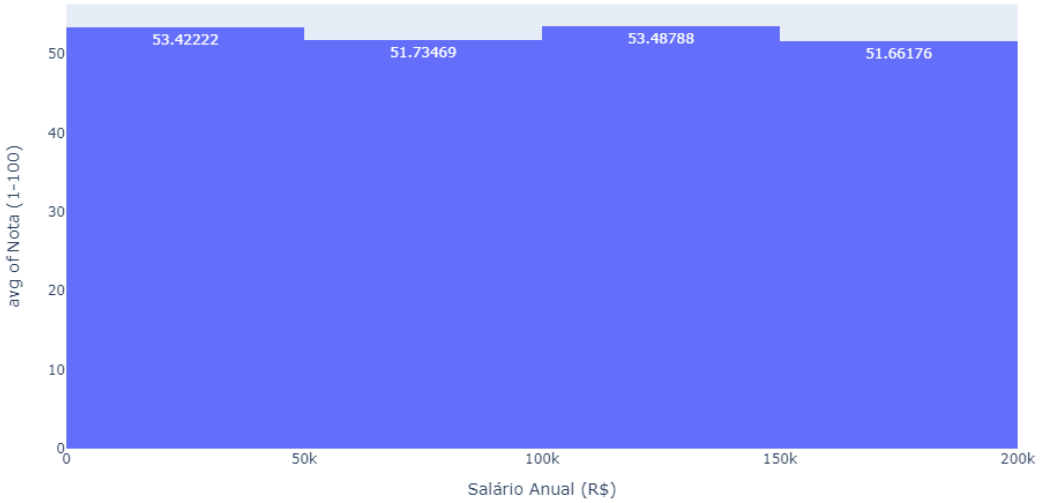
We can observe that there is little difference in ratings between customers from organic sources and those from promotions.

- In the second graph, we compared the age of customers with their ratings:



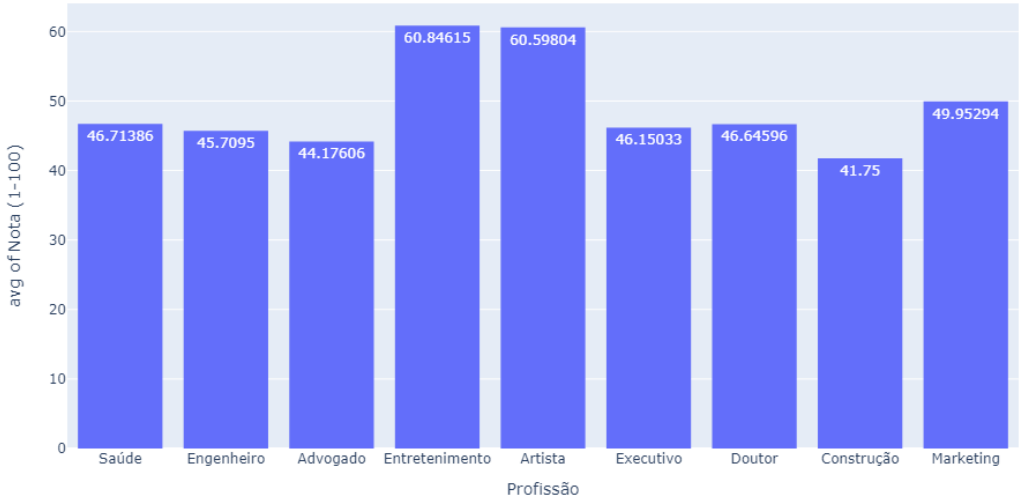
We observed a drop in ratings among people below the age of 20.

- In the third graph, we compared the annual salary of customers with their ratings:



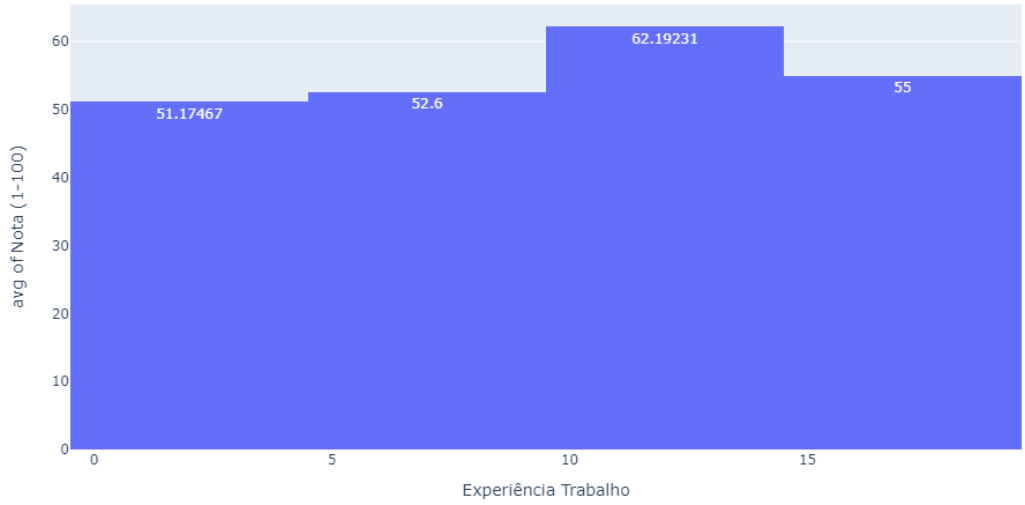
From the analysis of this graph, we can conclude that the salary range is not a significant factor because, although there is a lower rating in the salary range between 50,000 to 100,000 and above 150,000, the difference is not significant compared to other salary ranges.

- In the fourth graph, we compared the profession of customers with their ratings:



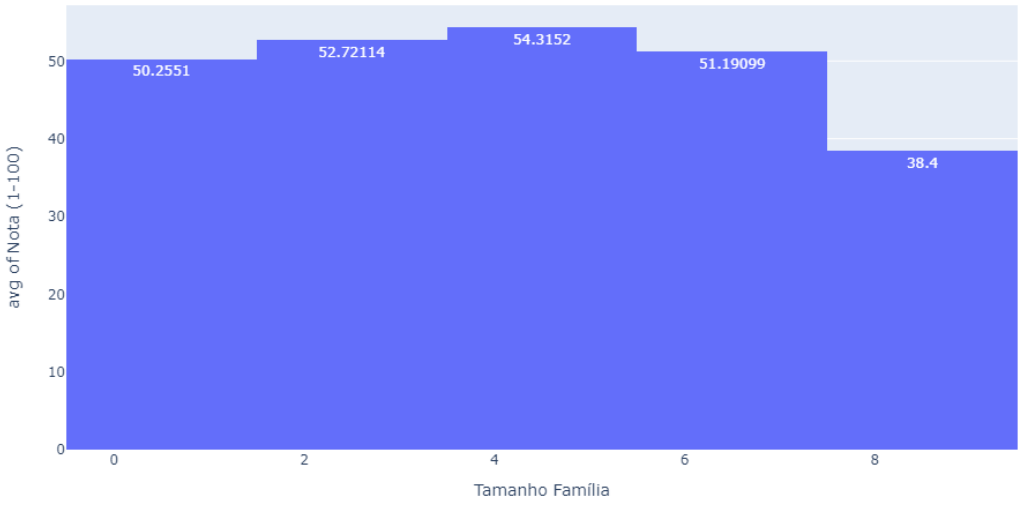
There is a clear difference in ratings between professions. Entertainment and art professionals have higher ratings, indicating that people in these professions tend to have higher ratings than others.

- In the fifth graph, we compared the work experience of customers with their ratings:



We observed relevant information, showing that customers with work experience between 10 to 15 years have higher ratings.

- Finally, in the last graph, it was observed that customers with a family size larger than 7 tend to have very low ratings, providing information about which customers are more likely to receive low ratings.



Conclusion:

After conducting the analyses, we conclude that the customer profile consists of individuals above the age of 20, working in the entertainment or art industry, with work experience between 10 to 15 years, and a family size of less than 7 members. The customer's origin and salary range do not appear to be determining factors for the analysis. Additionally, we found that professionals in the construction industry and families with more than 7 members have the lowest average ratings.