Data Credit Analistic

This is an analytical project developed as the final paper for the SQL course at EBAC.
Also, the graphics were created in Power BI

I created the database table on AWS S3 Bucket, and for rode the code and download the results used the AWS Athena.

About the Data Base

For this analisy I'm using a small part of the database public on the link:
https://github.com/andre-marcos-perez/ebac-course-utils/tree/main/dataset.

The code in SQL used to created the table was:

```
CREATE EXTERNAL TABLE IF NOT EXISTS default.credito (
  `idade` int,
  `sexo` string,
  `dependentes` int,
  `escolaridade` string,
  `estado_civil` string,
  `salario_anual` string,
  `tipo_cartao` string,
  `qtd_produtos` bigint,
  `iteracoes_12m` int,
  `meses_inativo_12m` int,
  `limite_credito` float,
  `valor_transacoes_12m` float,
  `qtd_transacoes_12m` int
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
WITH SERDEPROPERTIES (
  'serialization.format' = ',',
  'field.delim' = ','
) LOCATION "s3://****/"
TBLPROPERTIES ('has_encrypted_data'='false');
```

The structure of the table is:

- `idade` = age of client
- `sexo` = sex of the client (F or M)

- `dependentes` = number of client dependents
- `escolaridade` = client education level
- `salario_anual` = client's annual income range
- `tipo_cartao` = type of credit card held by the client
- `qtd_produtos` = number of products purchased in the last 12 months
- `iteracoes_12m` = number of interactions/transactions in the last 12 months
- `meses_inativo_12m` = number of months the client was inactive in the last 12 months
- `limite_credito` = client's credit limit
- `valor_transacoes_12m` = value of transactions in the last 12 months
- `qtd_transacoes_12m` = number of transactions in the last 12 months

```
import pandas as pd
from IPython.display import display, Code, Markdown, Image

print("Using the SQL query:" )
display(Code("SELECT * FROM credito LIMIT 10;", language="sql"))
print("This is the found sctruture of the table")

df = pd.read_csv('/kaggle/input/finalproj/query_select.csv', sep = ",")

display(df.head(10))

print("publish interally: https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/query_
print("")
print("The horizontal size of the table is found using this query: ")
display(Code("SELECT COUNT(*) FROM credito;", language="sql"))
count = pd.read_csv('/kaggle/input/finalproj/query_count.csv', sep = ",")

count
```

Using the SQL query:

```
SELECT * FROM credito LIMIT 10;
```

This is the found sctruture of the table

	idade	sexo	dependentes	escolaridade	estado_civil	salario_anual	tipo_cartao
0	45	M	3	ensino medio	casado	$60K - $80K	blue
1	49	F	5	mestrado	solteiro	menos que $40K	blue
2	51	M	3	mestrado	casado	$80K - $120K	blue
3	40	F	4	ensino medio	na	menos que $40K	blue
4	40	M	3	sem educacao formal	casado	$60K - $80K	blue
5	44	M	2	mestrado	casado	$40K - $60K	blue
6	51	M	4	na	casado	$120K +	gold
7	32	M	0	ensino medio	na	$60K - $80K	silver
8	37	M	3	sem educacao formal	solteiro	$60K - $80K	blue
9	48	M	2	mestrado	solteiro	$80K - $120K	blue

publish interally: https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/query_sel

The horizontal size of the table is found using this query:

```
SELECT COUNT(*) FROM credito;
```

	_col0
0	2564

Data Exploration

Describing the variables type ```SQL DESCRIBE credito

idade	int
sexo	string
dependentes	int
escolaridade	string
estado_civil	string
salario_anual	string
tipo_cartao	string
qtd_produtos	bigint
iteracoes_12m	int
meses_inativo_12m	int
limite_credito	float
valor_transacoes_12m	float
qtd_transacoes_12m	int

link

```
unique_values = {col: df[col].unique() for col in df.columns}

# Cria uma Series a partir desse dicionário e aplica pd.Series para expandir a visualizaç
unique_series = pd.Series(unique_values).apply(pd.Series)

# Exibe todas as colunas e linhas no output
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

print("And these are the possible answers for each one of our variables:")
for col in df.columns:
    print(f"{col}: {df[col].unique()[:8]}")
```

```
And these are the possible answers for each one of our variables:
idade: [45 49 51 40 44 32 37 48]
sexo: ['M' 'F']
dependentes: [3 5 4 2 0 1]
escolaridade: ['ensino medio' 'mestrado' 'sem educacao formal' 'na' 'graduacao'
 'doutorado']
estado_civil: ['casado' 'solteiro' 'na' 'divorciado']
salario_anual: ['$60K - $80K' 'menos que $40K' '$80K - $120K' '$40K - $60K' '$120K +'
 'na']
tipo_cartao: ['blue' 'gold' 'silver' 'platinum']
qtd_produtos: [5 6 4 3 2 1]
iteracoes_12m: [3 2 0 1 4 5]
meses_inativo_12m: [1 4 2 3 6 0 5]
limite_credito: [12691.51  8256.96  3418.56  3313.03  4716.22  4010.69 34516.72 29081
valor_transacoes_12m: [1144.9  1291.45 1887.72 1171.56  816.08 1088.07 1330.87 1538.3
qtd_transacoes_12m: [42 33 20 28 24 31 36 32]
```

Is notible analisyng this descrition for possibles results that only 3 variables, "escolaridade", "estado_civil" and "salario_anual", have "na" answer as possible answer.

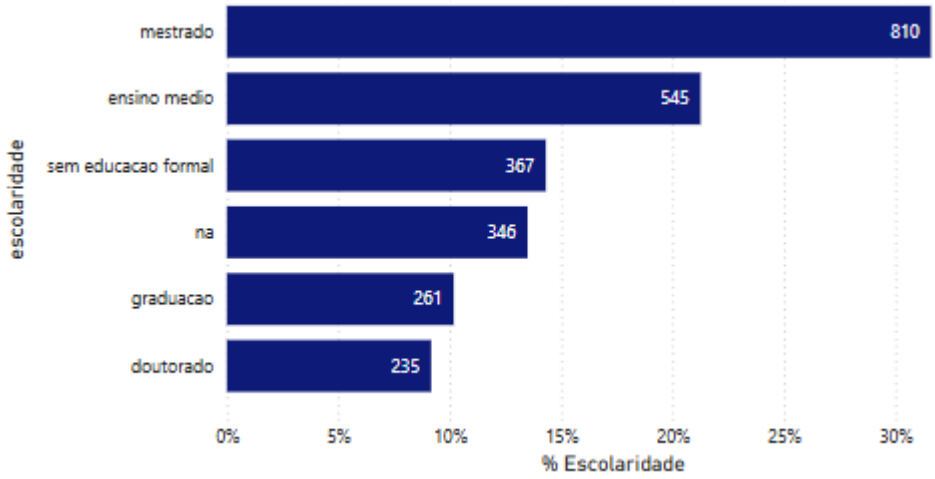
Now analisying the possibles results of each variables and results for it:

```
esc = pd.read_csv("/kaggle/input/finalproj/query_escolaridade.csv", sep = ",")
print("For selected all possibles responses for variable 'escolaridade' was used the foll
display(Code("SELECT escolaridade, COUNT(*) FROM credito GROUP BY escolaridade", language
display(esc)
display(Markdown("[link](https://raw.githubusercontent.com/MahFr115/EBAC_SQL_FinalProj/ma
```

```
For selected all possibles responses for variable 'escolaridade' was used the follow
SELECT escolaridade, COUNT(*) FROM credito GROUP BY escolaridade
```

	escolaridade	_col1
0	na	346
1	ensino medio	545
2	sem educacao formal	367
3	mestrado	810
4	graduacao	261
5	doutorado	235

link



link

We can notice that most of the clientes has a formal uni certification, beeing the mos exprecive the master one.
Also is important to notice that who didn't answer this question ("na") is a significant group, this way it is not efective just excluded they from our analisys, so is important to lear who treat and consider this data.

To understand the behavel of the data considering variables"salario anual" was used:

```
SELECT salario_anual, COUNT() FROM credito GROUP BY salario_anual
```

and the results is:

```
sal_anual = pd.read_csv("/kaggle/input/finalproj/query_salarioanual.csv", sep = ",")
display(sal_anual)
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/query_s
```

	salario_anual	_col1
0	$60K - $80K	451
1	$40K - $60K	467
2	$120K +	222
3	menos que $40K	701
4	na	235
5	$80K - $120K	488

link

Considering this variable we can notice that the response "na" is less present, this way is possible desconsider these values in futures analises considereing this variable.

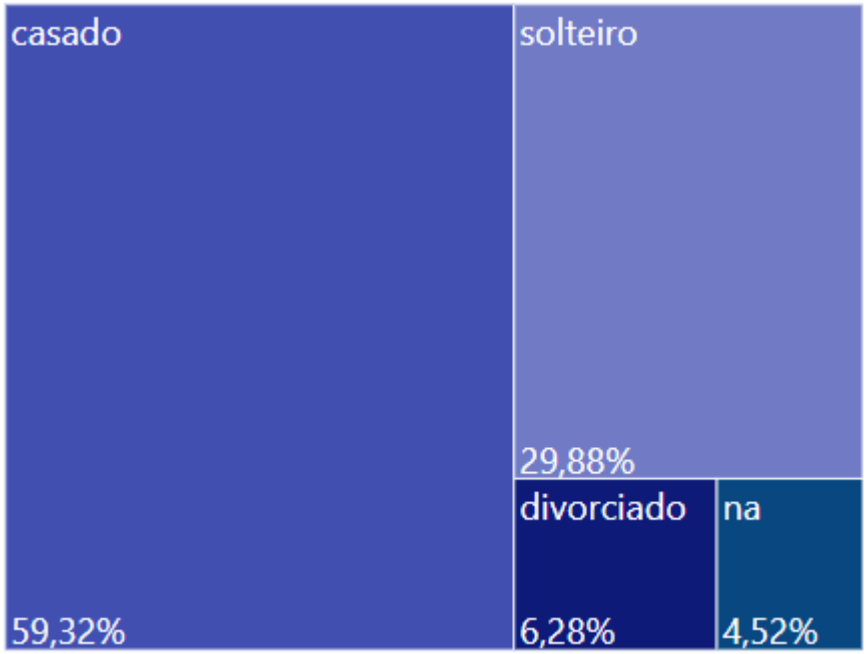
As already mentioned the last variable that aceptable "na" as answer is "estado_civil" and using a similar code the result finded in table and graphic style are:

```
estado = pd.read_csv("/kaggle/input/finalproj/query_estadocivil.csv")
display(estado)
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/query_e

display(Image("/kaggle/input/finalproj/graphic/gra_estado.png"))
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/graphic
```

	estado_civil	_col1
0	na	116
1	solteiro	766
2	divorciado	161
3	casado	1521

link



link

Equaly as the las variable we can notice that "na" answer is not a espressive one, so is possible discosidering this one.

```
print("About the possibles cards we have this possibles answers:")
display(Code("SELECT tipo_cartao, COUNT(*) FROM credito GROUP BY tipo_cartao", language =
card = pd.read_csv("/kaggle/input/finalproj/query_tipocartao.csv")
card
```

```
About the possibles cards we have this possibles answers:
SELECT tipo_cartao, COUNT(*) FROM credito GROUP BY tipo_cartao
```

	tipo_cartao	_col1
0	platinum	2
1	blue	2453
2	gold	16
3	silver	93

Considering variable "sexo" was used:

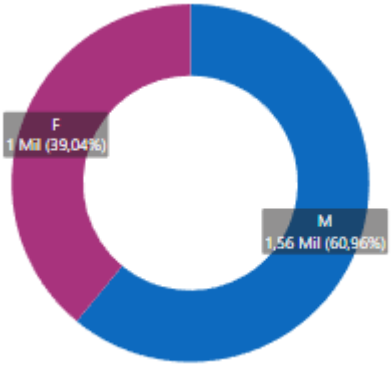
```
SELECT sexo, COUNT() FROM credito GROUP BY sexo
```

and the result are:

```
sexo = pd.read_csv("/kaggle/input/finalproj/query_sexo.csv", sep = ",")
display(sexo)
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/query_s
display(Image("/kaggle/input/finalproj/graphic/gra_sexo.png"))
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/graphic
```

	sexo	_col1
0	F	1001
1	M	1563

link



link

We can notice that a bigger part of our base is of male group.

```
print("Using the code")
display(Code("SELECT MAX(limite_credito) AS limite_credito, escolaridade, tipo_cartao, se
lim_max = pd.read_csv("/kaggle/input/finalproj/query_limitecredito_max.csv")
display(lim_max)
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/query_l
print("is possible analyse the maximo value of the credit limit considering the schoolars

print("The followed graphic shows the limit of the credit ")
display(Image("/kaggle/input/finalproj/graphic/gra_lim.png"))
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/graphic
```

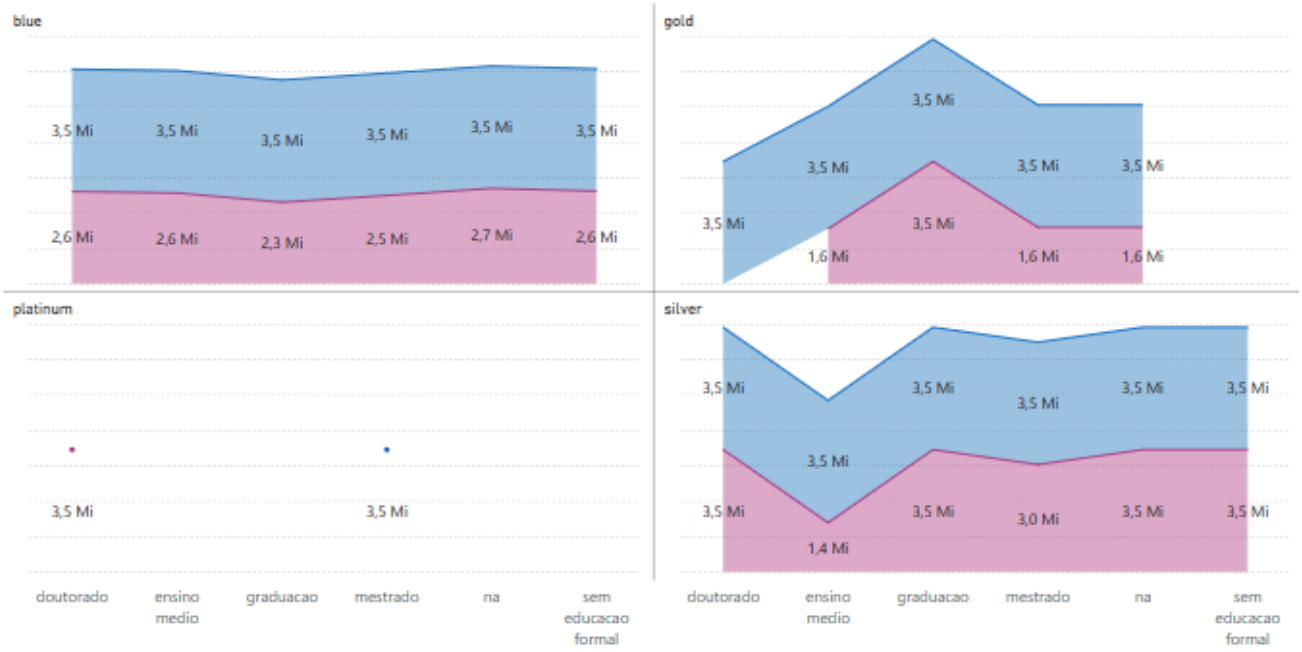
Using the code

```
SELECT MAX(limite_credito) AS limite_credito, escolaridade, tipo_cartao, sexo FROM cr
```

	limite_credito	escolaridade	tipo_cartao	sexo
0	34516.99	na	silver	F
1	34516.99	sem educacao formal	blue	M
2	34516.98	mestrado	gold	M
3	34516.97	mestrado	blue	M
4	34516.96	sem educacao formal	silver	M
5	34516.96	doutorado	platinum	F
6	34516.96	na	blue	M
7	34516.95	ensino medio	gold	M
8	34516.94	graduacao	gold	F
9	34516.94	graduacao	gold	M

link

is possible analyse the maximo value of the credit limit considering the schoolarship
The followed graphic shows the limit of the credit



link

Using this code

```
SELECT MAX(valor_transacoes_12m) AS maior_valor_gasto, AVG(valor_transacoes_12m) AS medi
FROM credito
GROUP BY sexo
```

was founded this table as resoults:

```
val_ = pd.read_csv("/kaggle/input/finalproj/query_transacoes_max.csv")
display(val_)
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/query_t
```

	maior_valor_gasto	media_valor_gasto	min_valor_gasto	sexo
0	4686.93	1807.9680	530.36	M
1	4776.58	1839.6226	510.16	F

link

Using the code

```
SELECT AVG(qtd_produtos) AS qts_produtos, AVG(valor_transacoes_12m) AS media_valor_trans
FROM credito
WHERE salario_anual != 'na'
GROUP BY sexo, salario_anual
ORDER BY avg(valor_transacoes_12m) DESC
```

we could find the result:

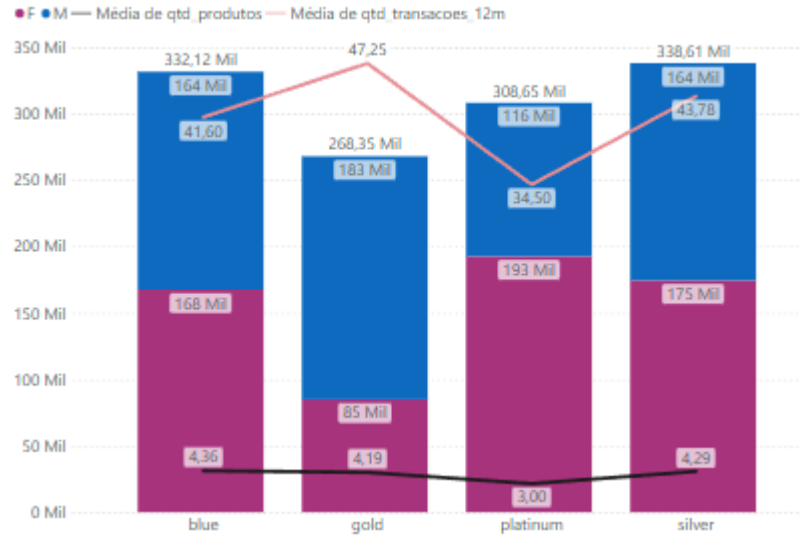
```
qtd_ = pd.read_csv("/kaggle/input/finalproj/query_compras.csv")
display(qtd_)
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/query_c
```

	qts_produtos	media_valor_transacoes	qtd_transacoes	sexo	salario_anual
0	4.394231	1963.6719	44.653846	M	menos que $40K
1	4.291971	1878.2974	42.602190	M	$40K - $60K
2	4.408710	1845.1332	42.507538	F	menos que $40K
3	4.352550	1818.6364	41.097561	M	$60K - $80K
4	4.440415	1781.4299	41.300518	F	$40K - $60K
5	4.329918	1755.2499	40.446721	M	$80K - $120K
6	4.328829	1701.4652	39.842342	M	$120K +

link

```
print("The followed graphic shows the quantified of products buyed")
display(Image("/kaggle/input/finalproj/graphic/gra_compras.png"))
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/graphic
```

The followed graphic shows the quantified of products buyed

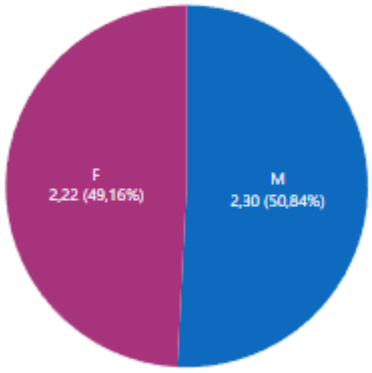


link

Conclusion

We can study another combinatiosn of dates, like the number of inactive months of tyhe acount oer mounth:

```
display(Image("/kaggle/input/finalproj/graphic/gra_sex.png"))
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/graphic
print("where we can undertand taht are no differences between the time that the account o
```

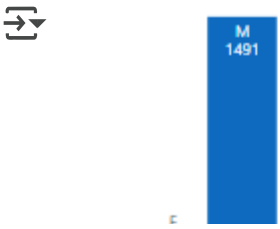


link

```
where we can undertand taht are no differences between the time that the account oof
```

Or the difference between the type pof cards, like the number of each type per sex:

```
display(Image("/kaggle/input/finalproj/graphic/gra_sexo_cartao.png"))
display(Markdown("[link](https://github.com/MahFr115/EBAC_SQL_FinalProj/blob/main/graphic
print("looking at this graphic we can notice a little difference betwen the numbers of ma
```


M
1491