Customer Segmentation Final Project

Virtual Internship

Ece Yavuzyılmaz

30-July-2023

Project	Customer Segmentation
Submitted By	Ece Yavuzyılmaz
Email	eceyavuzyilmaz@gmail.com
Country	Turkey
Specialization	Data Analyst
Internship Batch	LISUM21
Date	30 July 2023

Agenda

- Problem Statement
- Data Information
- Data Understanding
- Exploratory Data Analysis (EDA)
- Modeling Technique
- Recommendations
- Model Selection and Model Building
- Conclusion

Problem Statement

Data Glacier

Your Deep Learning Partner

1.Problem Description

XYZ bank wants to roll out Christmas offers to their customers. But Bank does not want to roll out same offer to all customers instead they want to roll out personalized offer to particular set of customers. If they manually start understanding the category of customer then this will be not efficient and also they will not be able to uncover the hidden pattern in the data (pattern which group certain kind of customer in one category). Bank approached ABC analytics company to solve their problem. Bank also shared information with ABC analytics that they don't want more than 5 group as this will be inefficient for their campaign.

2.Business Understanding

ABC analytics proposed customer segmentation approach to Bank. ABC analytics assigned this talk to their analytics team and instructed their team to come up with the approach and feature which group similar behavior customer in one category and others in different category.

Data Information

Total number of observations	**1000000**
Total number of files	1
Total number of features	48
Base format of the file	csv
Size of the data	154 MB
Total number of observations	1000000
Total number of files	1

Data Information

fecha_dato: The table is partitioned for this column

ncodpers: Customer code

ind_empleado: Employee index: A active, B ex employed, F filial, N not employee, P pasive

pais_residencia: Customer's Country residence

sexo: Customer's sex

age: Age

fecha_alta: The date in which the customer became as the first holder of a contract in the bank

ind_nuevo: New customer Index. 1 if the customer registered in the last 6 months.

antiguedad: Customer seniority (in months)

indrel: 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)

ult_fec_cli_1t: Last date as primary customer (if he isn't at the end of the month)

indrel_1mes: Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)

tiprel_1mes: Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)

Data Information

indresi: Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)

indext: Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)

conyuemp: Spouse index. 1 if the customer is spouse of an employee

canal_entrada: channel used by the customer to join

indfall: Deceased index. N/S

tipodom: Addres type. 1, primary address

cod_prov: Province code (customer's address)

nomprov: Province name

ind_actividad_cliente: Activity index (1, active customer; 0, inactive customer)

renta: Gross income of the household

ind_ahor_fin_ult1: Saving Account

ind_aval_fin_ult1: Guarantees

ind_cco_fin_ult1: Current Accounts

ind_cder_fin_ult1: Derivada Account

ind_cno_fin_ult1: Payroll Account

ind_ctju_fin_ult1: Junior Account

ind_ctma_fin_ult1: Más particular Account

ind_ctop_fin_ult1: particular Account

Data Information

ind_ctpp_fin_ult1: Particular Plus Account
ind_deco_fin_ult1: Short-term deposits
ind_deme_fin_ult1: Medium-term deposits
ind_dela_fin_ult1: Long-term deposits
ind_ecue_fin_ult1: e-account
ind_fond_fin_ult1: Funds
ind_hip_fin_ult1: Mortgage
ind_plan_fin_ult1: Pensions
ind_pres_fin_ult1: Loans
ind_reca_fin_ult1: Taxes
ind_tjcr_fin_ult1: Credit Card
ind_valo_fin_ult1: Securities
ind_viv_fin_ult1: Home Account
ind_nomina_ult1: Payroll
ind_nom_pens_ult1: Pensions
ind_recibo_ult1: Direct Debit

Task

1. Business Understanding
2. Data Understanding
3. EDA
4. Feature Engineering
4. Model Building
5. Model Evaluation
6. Presentation (Recommendation slide is must)
7. Document the challenges

Data Understanding

The dataset consists of 48 columns and 1000000 rows.

```
data.head()
```

	id	fecha_dato	cust_code	employee_index	custom_country_residence	sexo	age	date_regist	new_cust_index	cust_senior
0	0	2015-01-28	1375586	N	ES	H	35	2015-01-28	0.0	
1	1	2015-01-28	1050611	N	ES	V	23	2015-01-28	0.0	
2	2	2015-01-28	1050612	N	ES	V	23	2015-01-28	0.0	
3	3	2015-01-28	1050613	N	ES	H	22	2015-01-28	0.0	
4	4	2015-01-28	1050614	N	ES	V	23	2015-01-28	0.0	

5 rows × 45 columns

```
[4] data.shape
```

```
(1000000, 48)
```

Data Glacier

Your Deep Learning Partner

Data Understanding

All data types are seen as 'objects'. These data types have been changed as datetime64[ns] (2), float64 (3), int64 (29), object (11).

```
data['cust_code']=data['cust_code'].astype("int64")
data['cust_seniority']=data['cust_seniority'].astype("int64")
data['activity_index']=data['activity_index'].astype("int64")
data['save_account']=data['save_account'].astype("int64")
data['current_account']=data['current_account'].astype("int64")
data['derivada_account']=data['derivada_account'].astype("int64")
data['payroll_account']=data['payroll_account'].astype("int64")
data['junior_account']=data['junior_account'].astype("int64")
data['mas_particu_account']=data['mas_particu_account'].astype("int64")
data['particu_account']=data['particu_account'].astype("int64")
data['particu_plus_account']=data['particu_plus_account'].astype("int64")
data['short_term_deposit']=data['short_term_deposit'].astype("int64")
data['medium_term_deposit']=data['medium_term_deposit'].astype("int64")
data['long_term_deposit']=data['long_term_deposit'].astype("int64")
data['e_account']=data['e_account'].astype("int64")
data['e_account']=data['e_account'].astype("int64")
data['funds']=data['funds'].astype("int64")
data['mortgage']=data['mortgage'].astype("int64")
data['pensions']=data['pensions'].astype("int64")
data['loans']=data['loans'].astype("int64")
data['taxes']=data['taxes'].astype("int64")
data['credit_card']=data['credit_card'].astype("int64")
data['securities']=data['securities'].astype("int64")
data['home_account']=data['home_account'].astype("int64")
data['direct_debit']=data['direct_debit'].astype("int64")
```

Data Understanding

Data Glacier

Your Deep Learning Partner

Missing Values

Two different methods were used to fill in the missing value. These methods are; fillna() and interpolate().

id	0
fecha_dato	0
cust_code	0
employee_index	10782
custom_country_residence	10782
sexo	10786
age	10782
date_regist	0
new_cust_index	10782
cust_seniority	0
primary_cust	10782
cust_type	10782
cust_relation_type	10782
residence_index	10782
foreigner_index	10782
channel	10861
deceased_index	10782
addres_type	10782
cod_prov	17734
province_name	17734
activity_index	10782
gross_income	175183
save_account	0
current_account	0
derivada_account	0
payroll_account	0
junior_account	0
mas_particu_account	0
particu_account	0
particu_plus_account	0
short_term_deposit	0
medium_term_deposit	0
long_term_deposit	0
e_account	0
funds	0
mortgage	0
pensions	0
loans	0
taxes	0
credit_card	0
securities	0
home_account	0
payroll	5402

Data Understanding

Data Glacier

Your Deep Learning Partner

- **fillna():**

  *We can use **fillna()** function to fill NaN values. In this here, we can use different methods such 'backfill', 'bfill','pad','ffill' and mean/median/mode based approach to fill the missing values.*

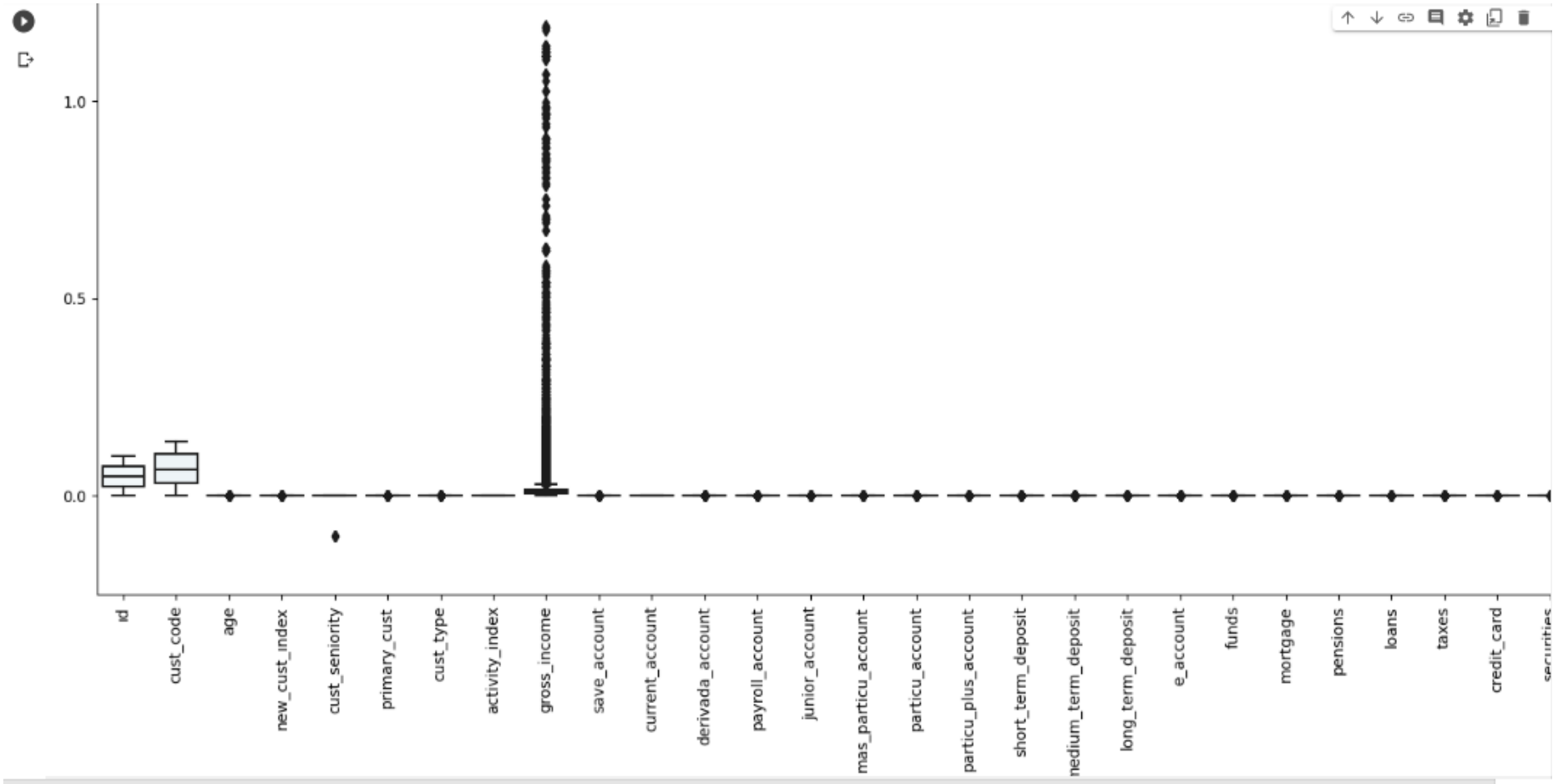
```
data['age']=data['age'].fillna(data['age'].mean()).astype("int64")
data['cust_type']=data['cust_type'].fillna(method ='ffill').astype("int64")
data['primary_cust']=data['primary_cust'].fillna(method ='ffill')
data['cust_relation_type']=data['cust_relation_type'].fillna(method ='ffill')
data['residence_index']=data['residence_index'].fillna(method='ffill')
data['foreigner_index']=data['foreigner_index'].fillna(method='ffill')
data['channel']=data['channel'].fillna(method='ffill')
data['deceased_index']=data['deceased_index'].fillna(method='ffill')
data['addres_type']=data['addres_type'].fillna(method='ffill')
data['province_name']=data['province_name'].fillna(method='ffill')
data['activity_index']=data['activity_index'].fillna(method='ffill')
data['payroll']=data['payroll'].fillna(method='ffill').astype("int64")
data['a_pension']=data['a_pension'].fillna(method='ffill').astype("int64")
data['employee_index']=data['employee_index'].fillna(method='ffill')
data['custom_country_residence']=data['custom_country_residence'].fillna(method='ffill')
data['sexo']=data['sexo'].fillna(method='ffill')
data['new_cust_index']=data['new_cust_index'].fillna(method='ffill')
data['cod_prov']=data['cod_prov'].fillna(method='ffill')
```

- **interpolate()**

```
# Interpolate backwardly across the column:
data['gross_income'].interpolate(method ='linear', limit_direction ='backward', inplace=True)
#data['gross_income']=data['gross_income'].fillna(data['gross_income'].mean())
```

Data Understanding

Outlier Detection



Exploratory Data Analysis (EDA)

Data Describe

	count	mean	std	min	25%	50%	75%	max
id	1000000.0	499999.500000	288675.278932	0.0	249999.75	499999.5	749999.25	999999.0
cust_code	1000000.0	690596.670395	404408.432011	15889.0	336411.00	664476.0	1074511.25	1379131.0
age	1000000.0	43.266717	17.065626	2.0	27.00	43.0	53.00	116.0
new_cust_index	1000000.0	0.000496	0.022266	0.0	0.00	0.0	0.00	1.0
cust_seniority	1000000.0	93.092962	2001.260509	-999999.0	33.00	95.0	156.00	246.0
primary_cust	1000000.0	1.109172	3.269090	1.0	1.00	1.0	1.00	99.0
cust_type	1000000.0	1.000088	0.013191	1.0	1.00	1.0	1.00	3.0
cod_prov	1000000.0	26.858072	12.416646	1.0	18.00	28.0	33.00	52.0
activity_index	1000000.0	0.565418	0.495702	0.0	0.00	1.0	1.00	1.0
gross_income	1000000.0	139104.125364	229305.904318	1202.0	73622.75	108048.0	162324.00	28894395.0
save_account	1000000.0	0.000177	0.013303	0.0	0.00	0.0	0.00	1.0
current_account	1000000.0	0.749626	0.433229	0.0	0.00	1.0	1.00	1.0
derivada_account	1000000.0	0.000591	0.024303	0.0	0.00	0.0	0.00	1.0
payroll_account	1000000.0	0.105296	0.306935	0.0	0.00	0.0	0.00	1.0

Exploratory Data Analysis (EDA)

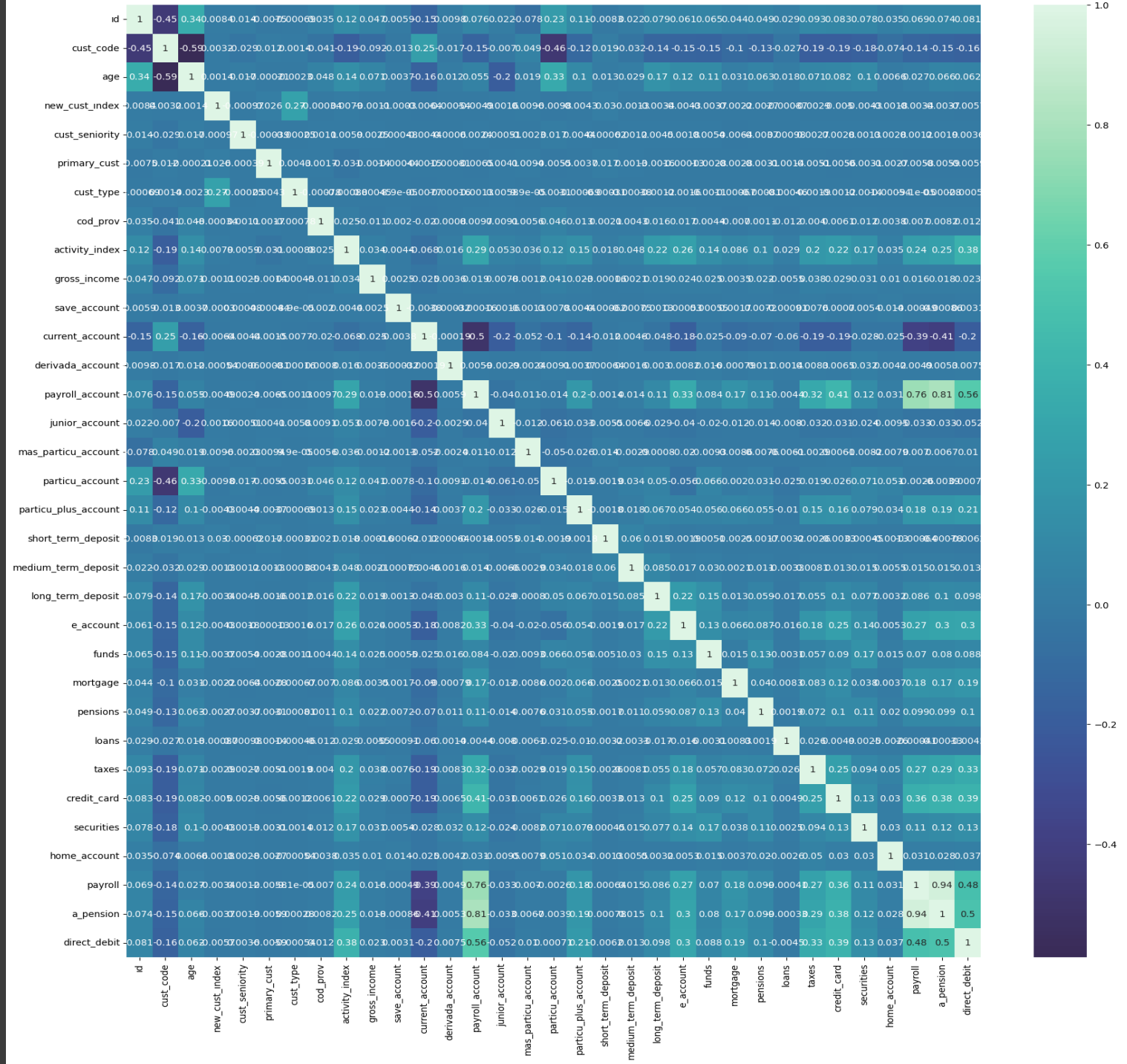
Data Describe

particu_plus_account	1000000.0	0.072079	0.258619	0.0	0.00	0.0	0.00	1.0
short_term_deposit	1000000.0	0.002158	0.046404	0.0	0.00	0.0	0.00	1.0
medium_term_deposit	1000000.0	0.003150	0.056036	0.0	0.00	0.0	0.00	1.0
long_term_deposit	1000000.0	0.066881	0.249816	0.0	0.00	0.0	0.00	1.0
e_account	1000000.0	0.106267	0.308179	0.0	0.00	0.0	0.00	1.0
funds	1000000.0	0.027182	0.162614	0.0	0.00	0.0	0.00	1.0
mortgage	1000000.0	0.009982	0.099410	0.0	0.00	0.0	0.00	1.0
pensions	1000000.0	0.014553	0.119755	0.0	0.00	0.0	0.00	1.0
loans	1000000.0	0.004661	0.068112	0.0	0.00	0.0	0.00	1.0
taxes	1000000.0	0.072581	0.259448	0.0	0.00	0.0	0.00	1.0
credit_card	1000000.0	0.066084	0.248429	0.0	0.00	0.0	0.00	1.0
securities	1000000.0	0.039378	0.194493	0.0	0.00	0.0	0.00	1.0
home_account	1000000.0	0.006442	0.080003	0.0	0.00	0.0	0.00	1.0
payroll	1000000.0	0.071693	0.257979	0.0	0.00	0.0	0.00	1.0
a_pension	1000000.0	0.079606	0.270682	0.0	0.00	0.0	0.00	1.0

Exploratory Data Analysis (EDA)

- The two features with the strongest correlation in the dataset are ‘Payroll' and ‘Pensions’ with 0.94.

- It is seen that there is a strong positive correlation with 0.81 correlation between ‘Pensions' and ‘Payroll Account' features.

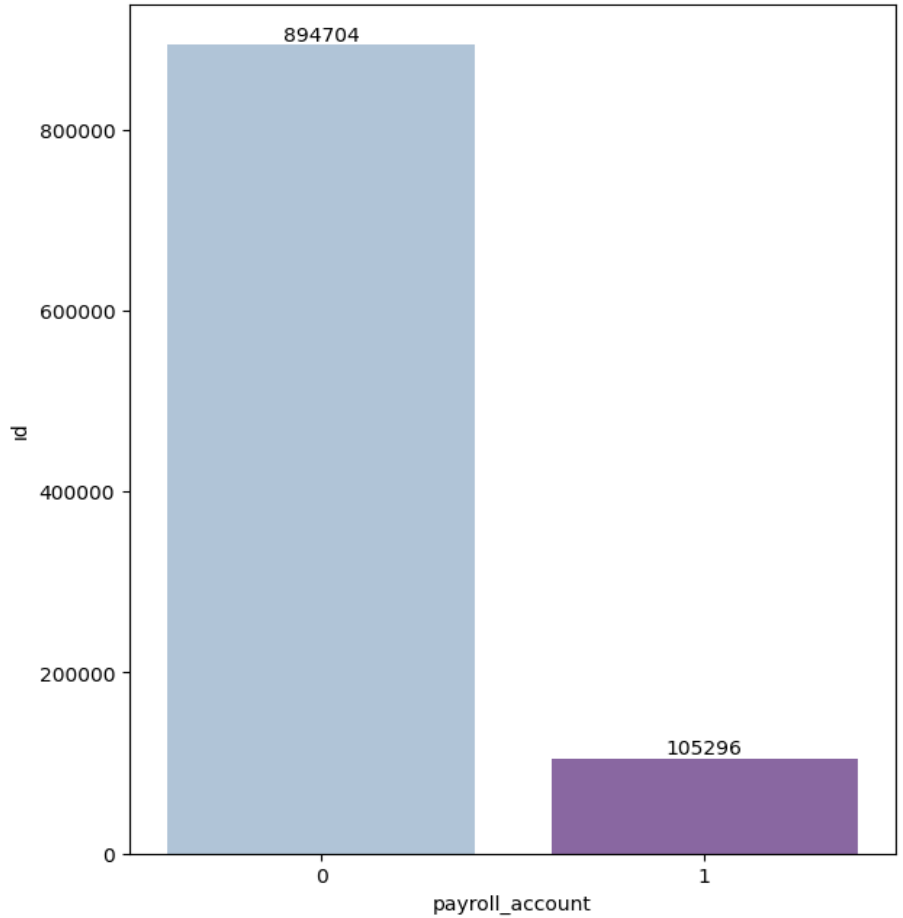


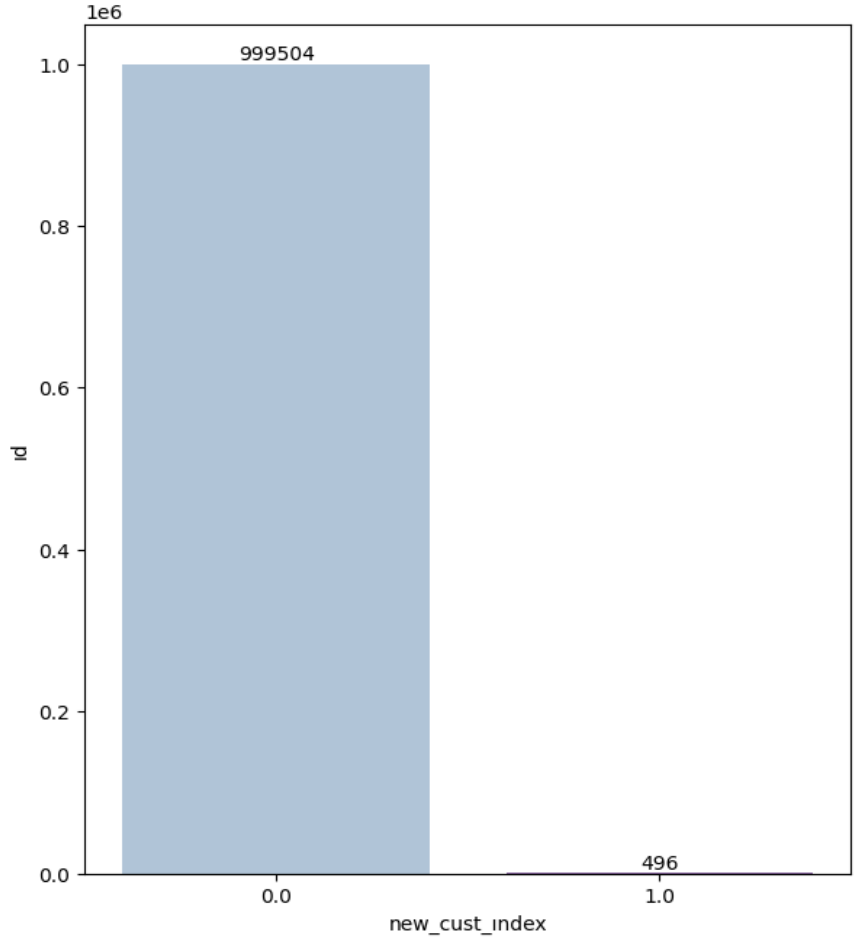
Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis (EDA)

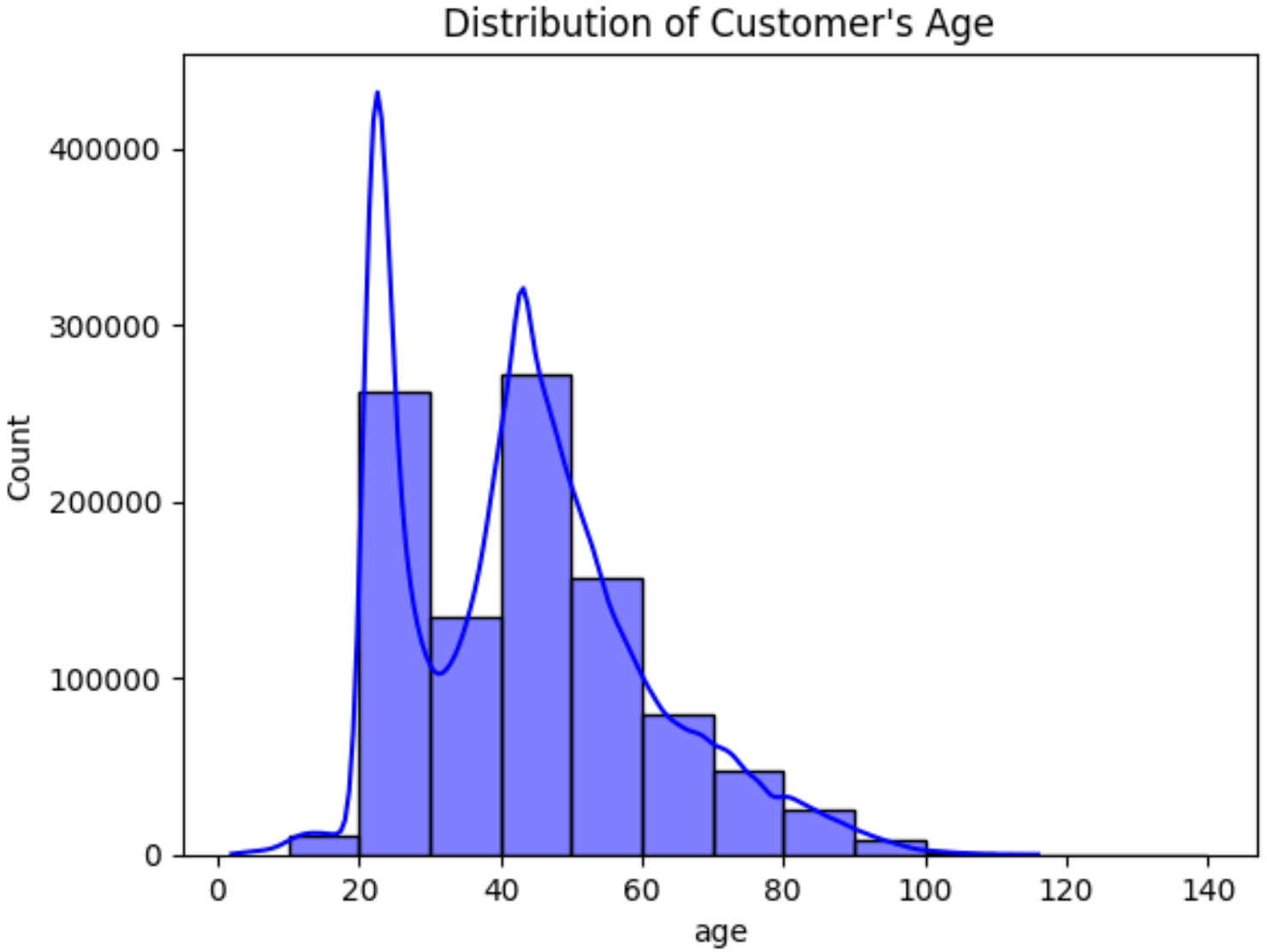
Here, payroll account shows the number of payroll accounts of customers registered with the bank, and the new customer index shows the number of customers registered in the last 6 months.





Exploratory Data Analysis (EDA)

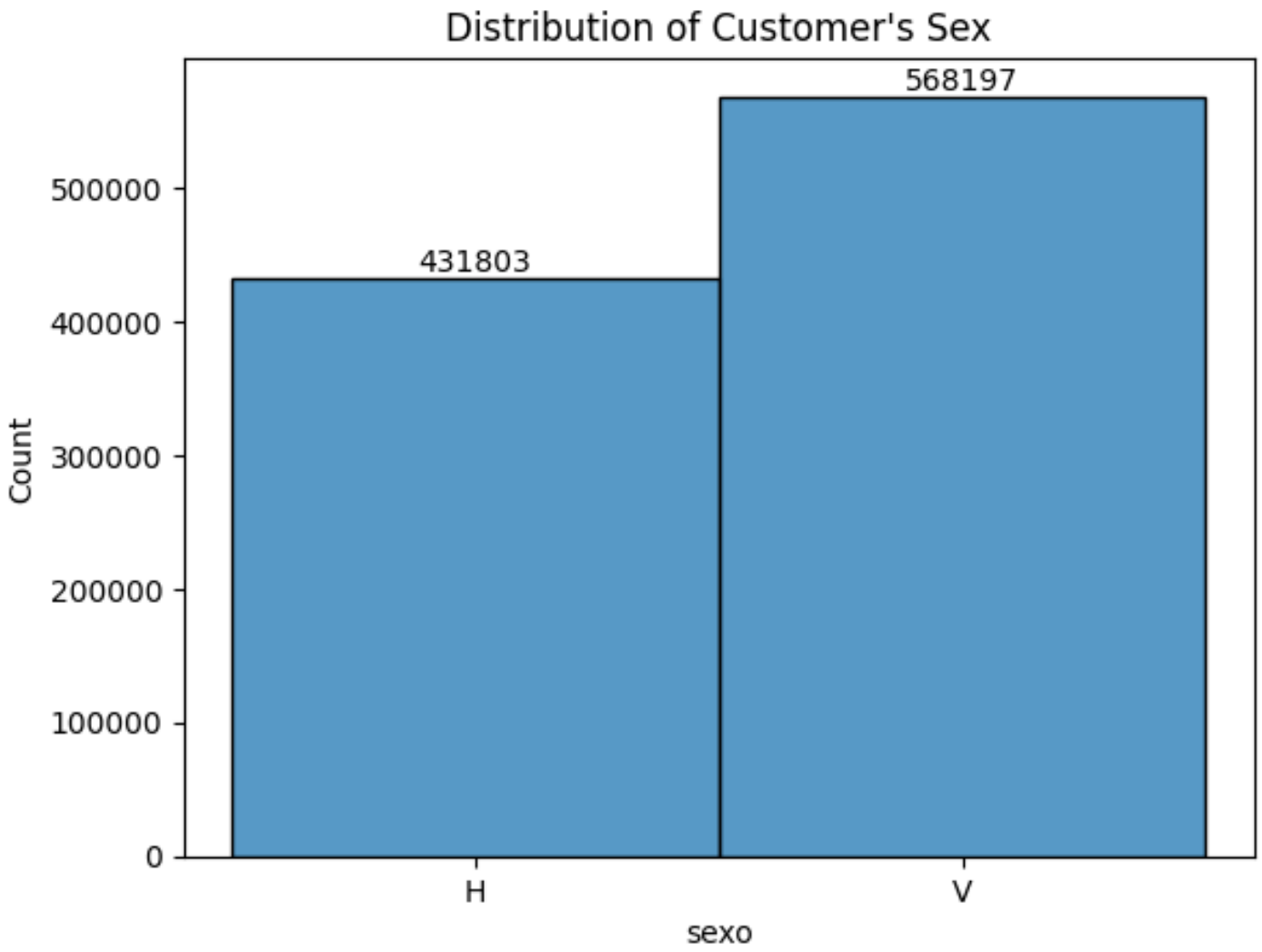
- In this here, we is seen that most of the customers belong in the age range of 20-60.



Data Glacier
Your Deep Learning Partner

Exploratory Data Analysis (EDA)

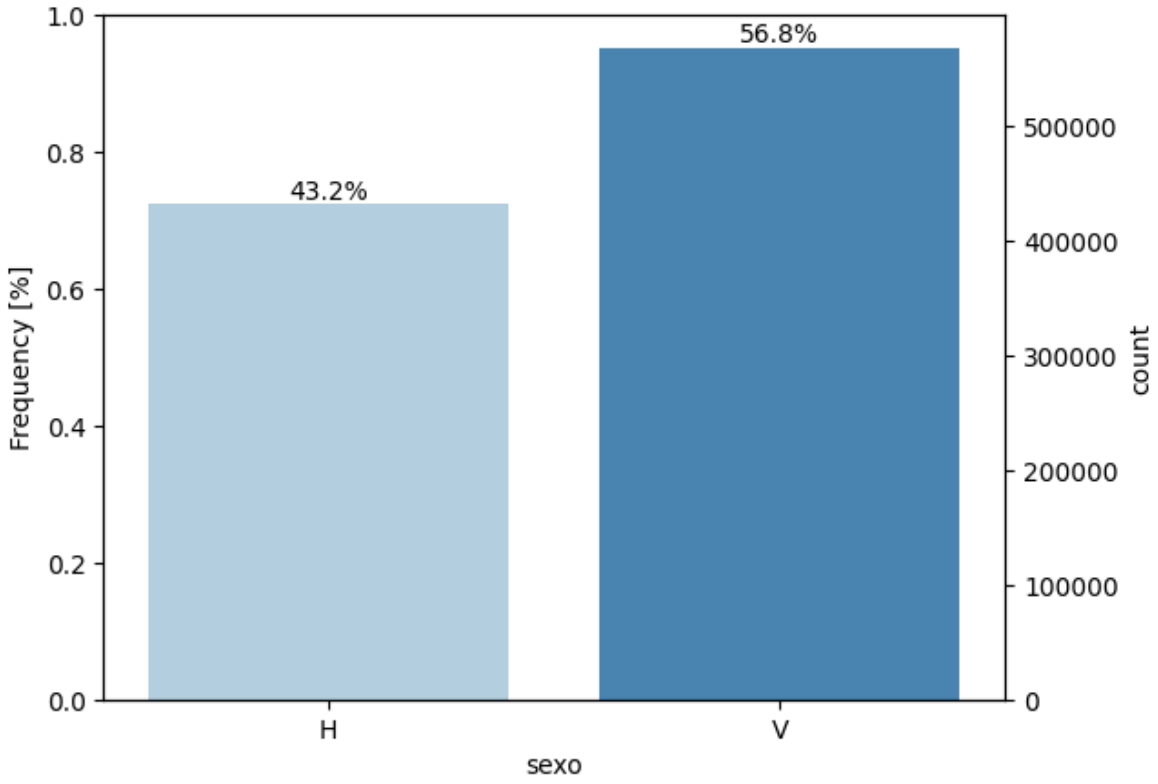
- This chart shows the sex distribution of customers. 431.803 of the customers are female, and 568.197 are male.

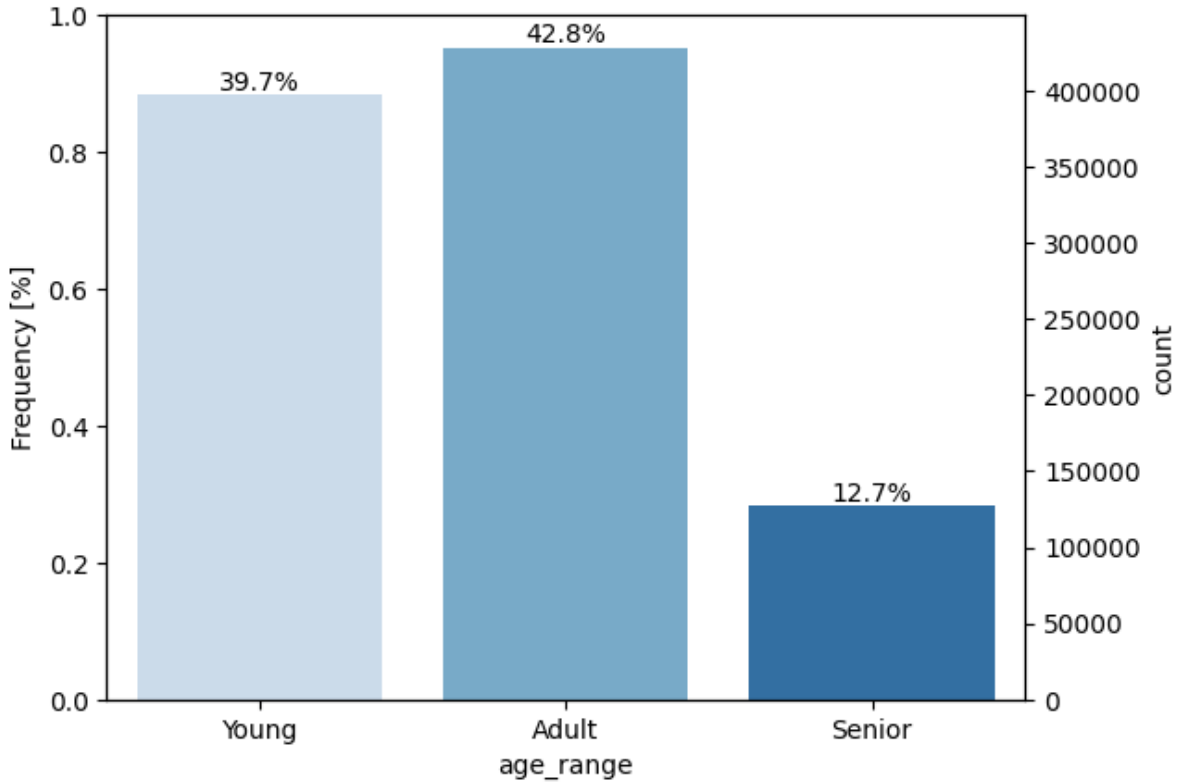


Data Glacier
Your Deep Learning Partner

Exploratory Data Analysis (EDA)

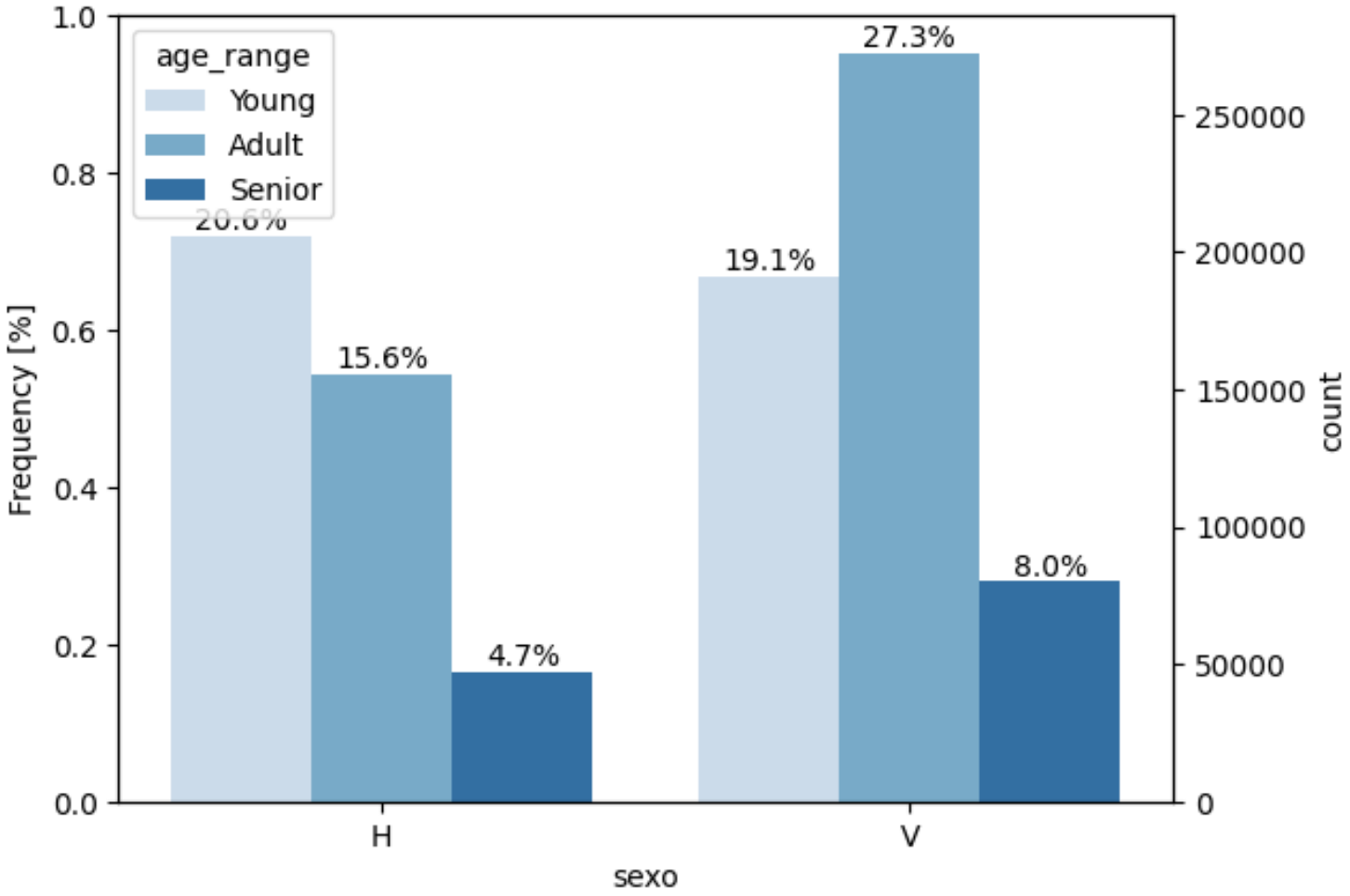
Here, the frequency distributions of customers by gender and age are shown. The number of adult customers is higher.





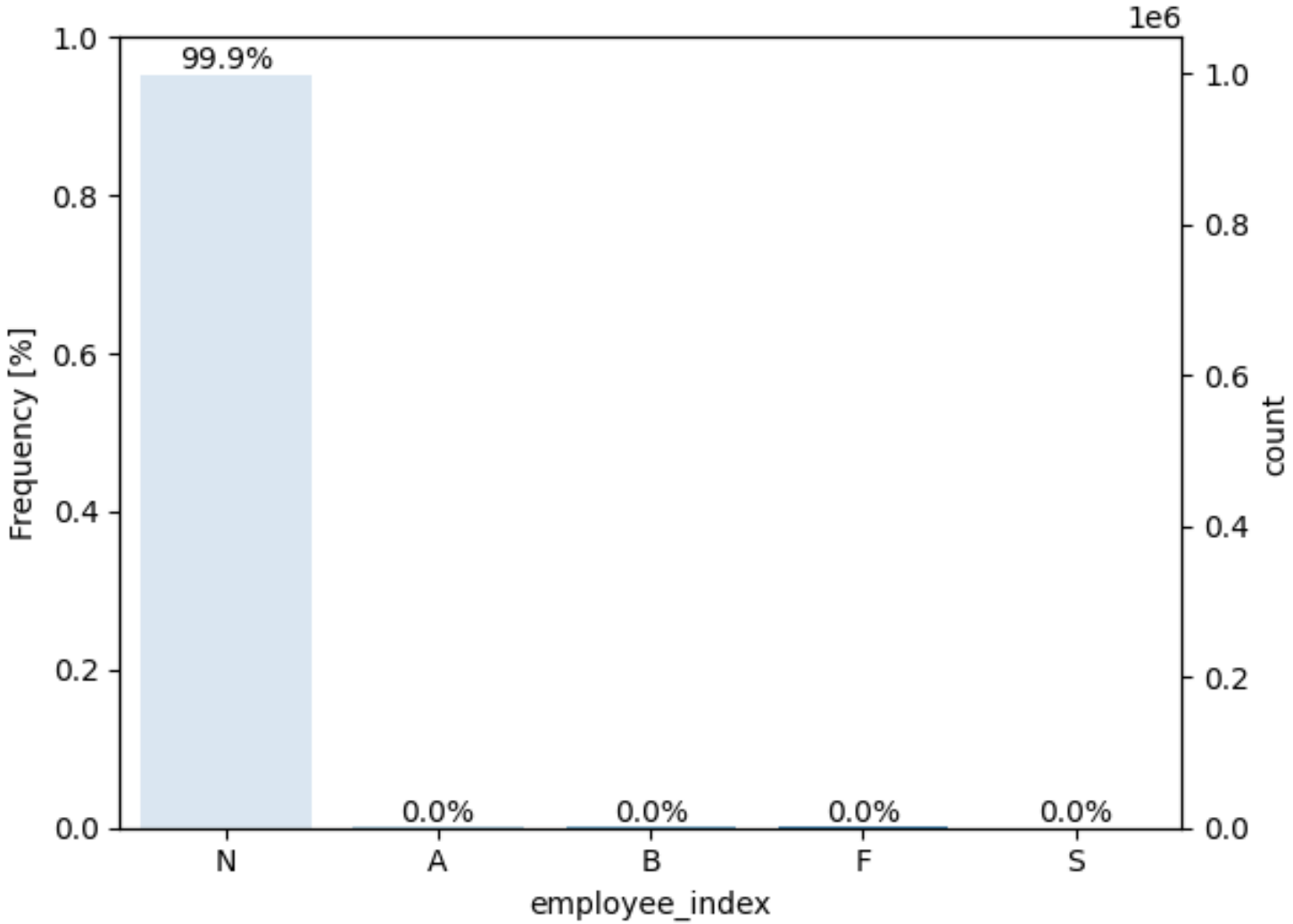
Exploratory Data Analysis (EDA)

While the rate of young customers is higher in females, the rate of adult customers is higher in males.



Exploratory Data Analysis (EDA)

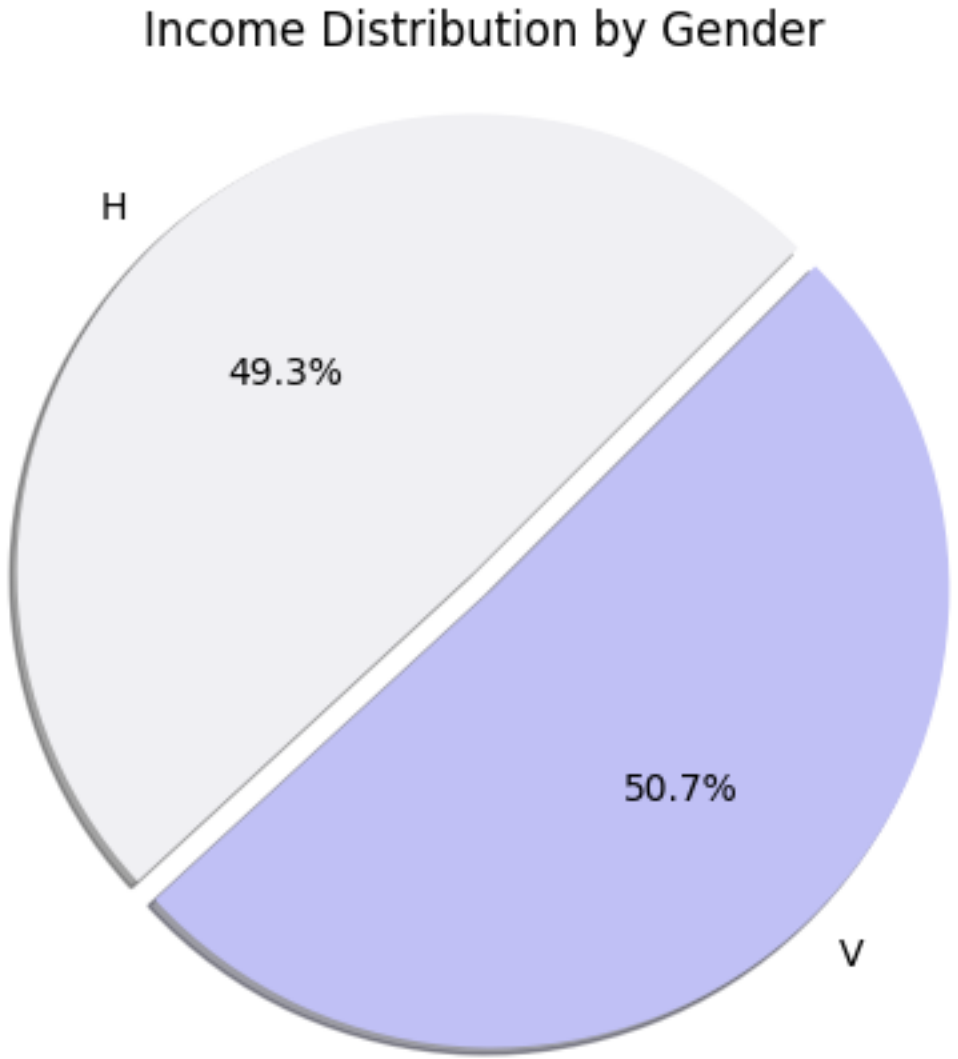
- 99% of customers are not bank employees.(employee index: A active, B ex employed, F filial, N not employee, P passive)



Data Glacier
Your Deep Learning Partner

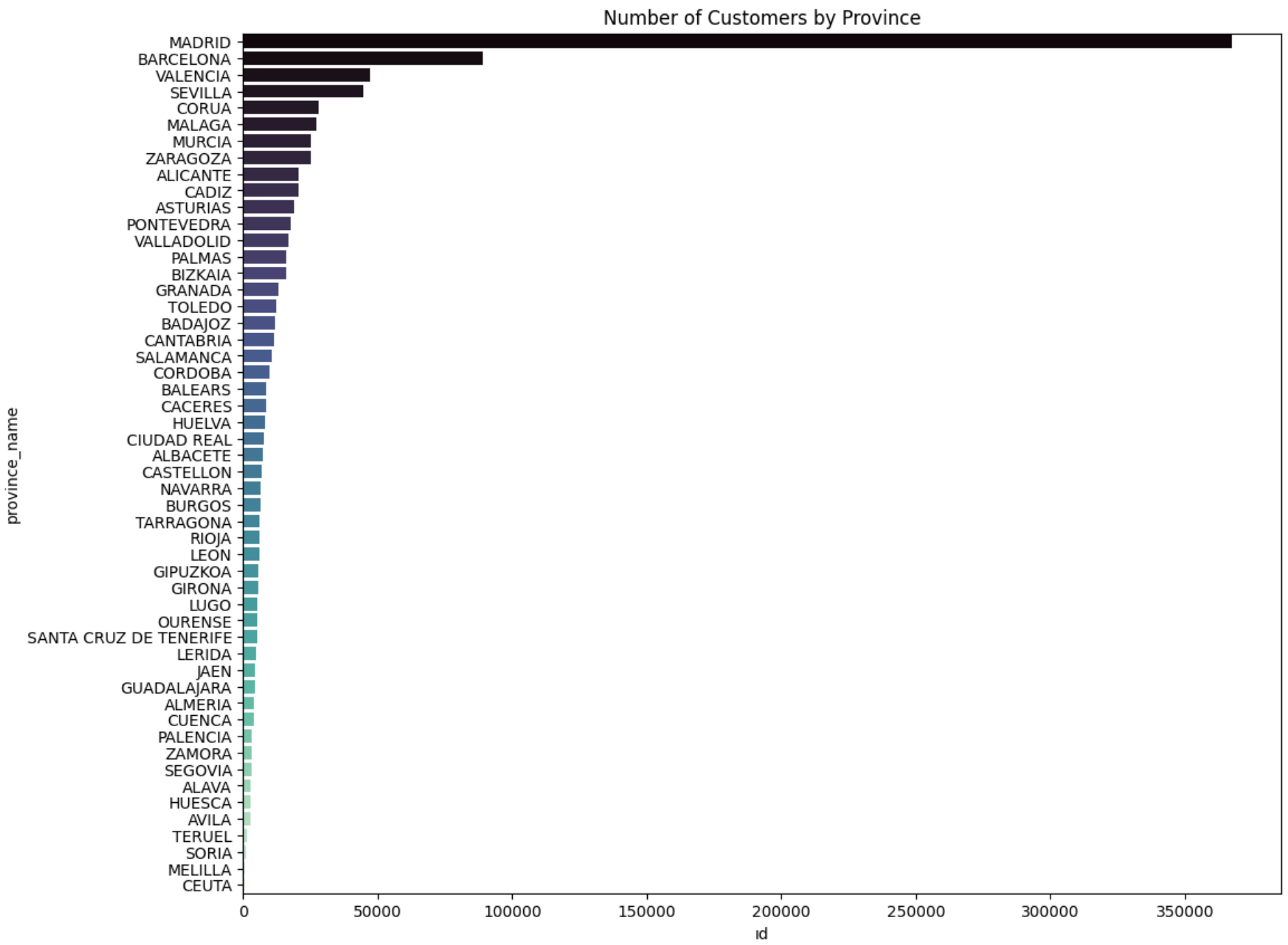
Exploratory Data Analysis (EDA)

- When we look at the average income distribution, it is seen that this rate is slightly higher in males.

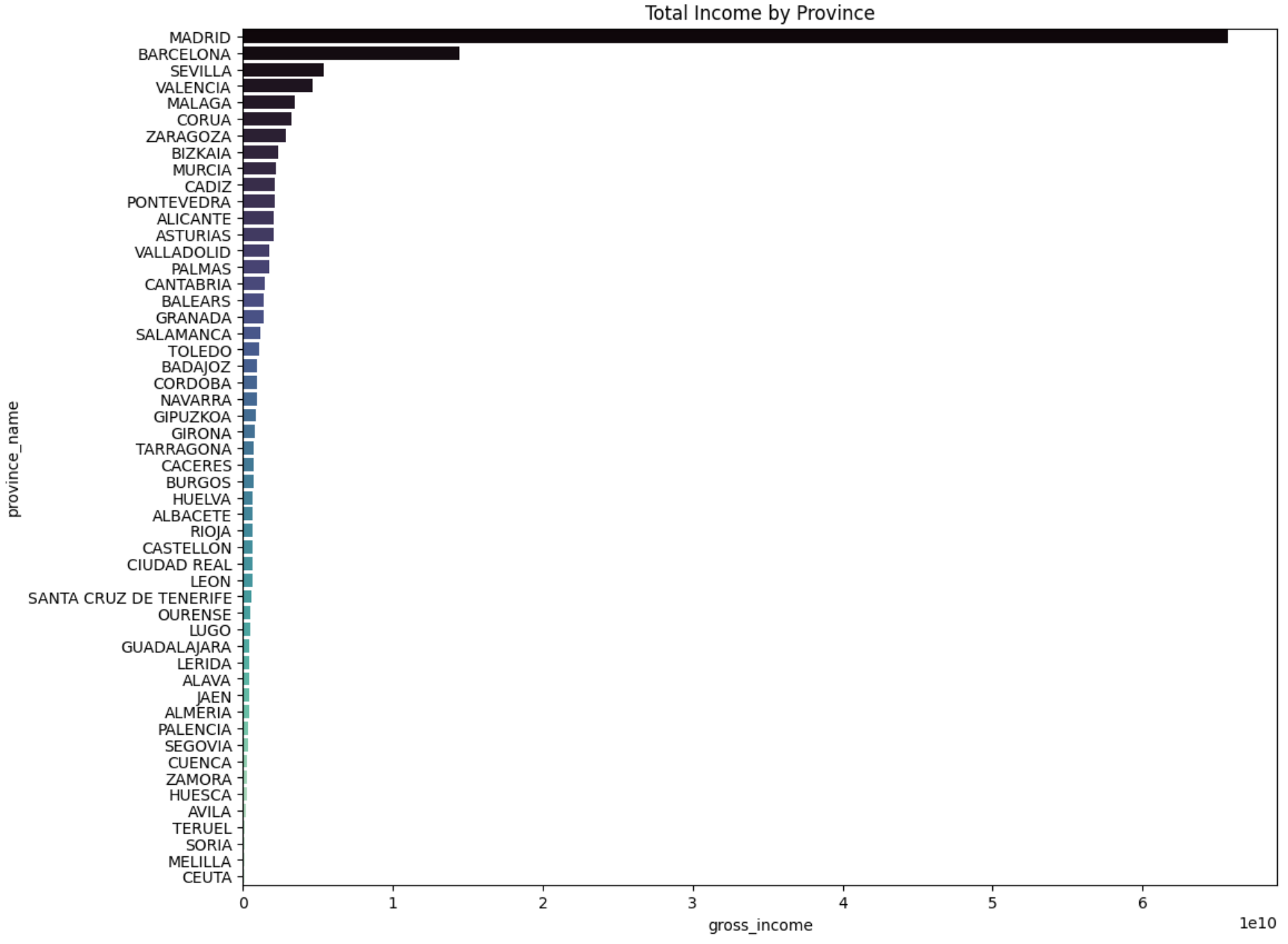


Data Glacier
Your Deep Learning Partner

Exploratory Data Analysis (EDA)

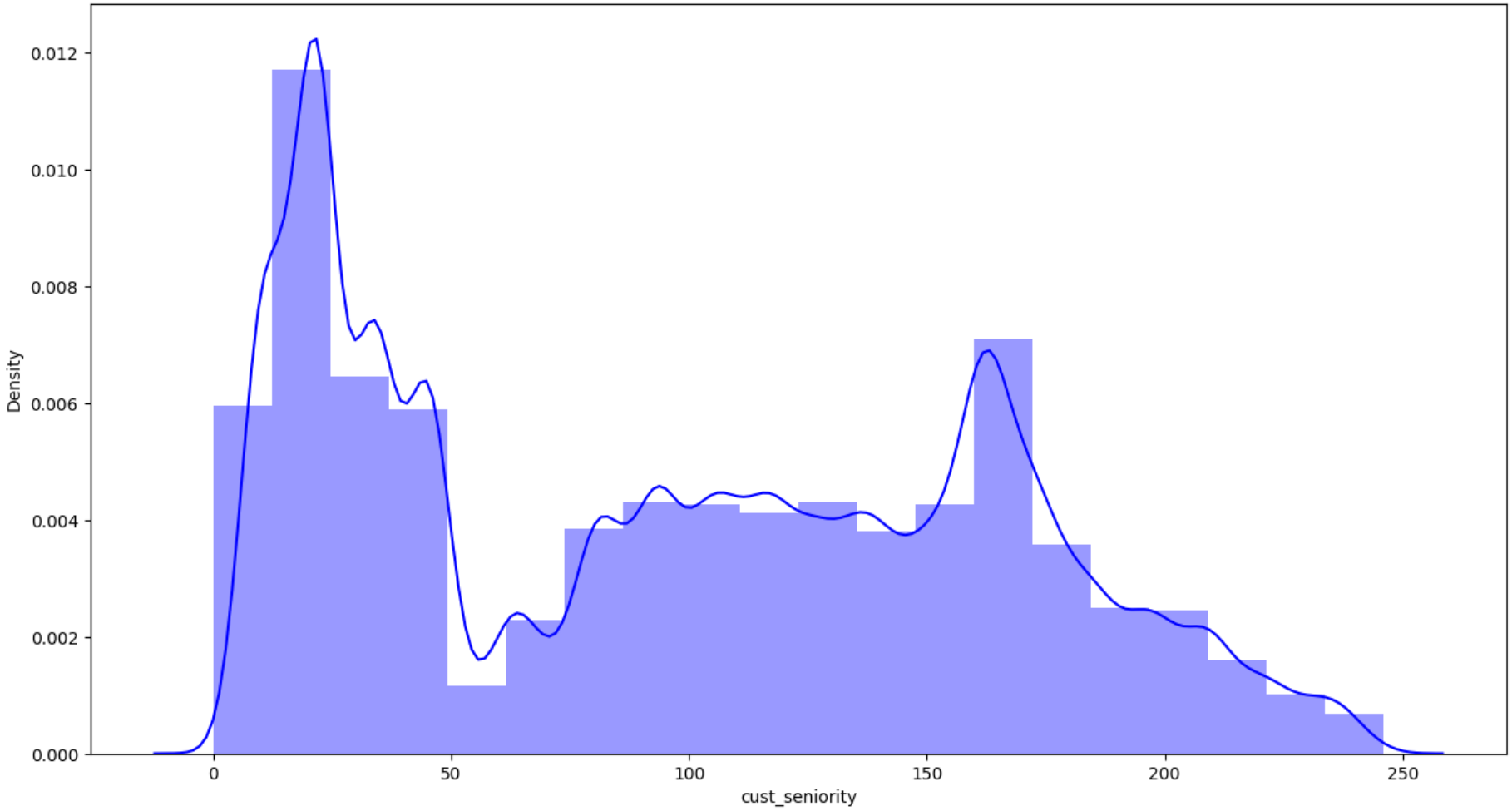


Exploratory Data Analysis (EDA)



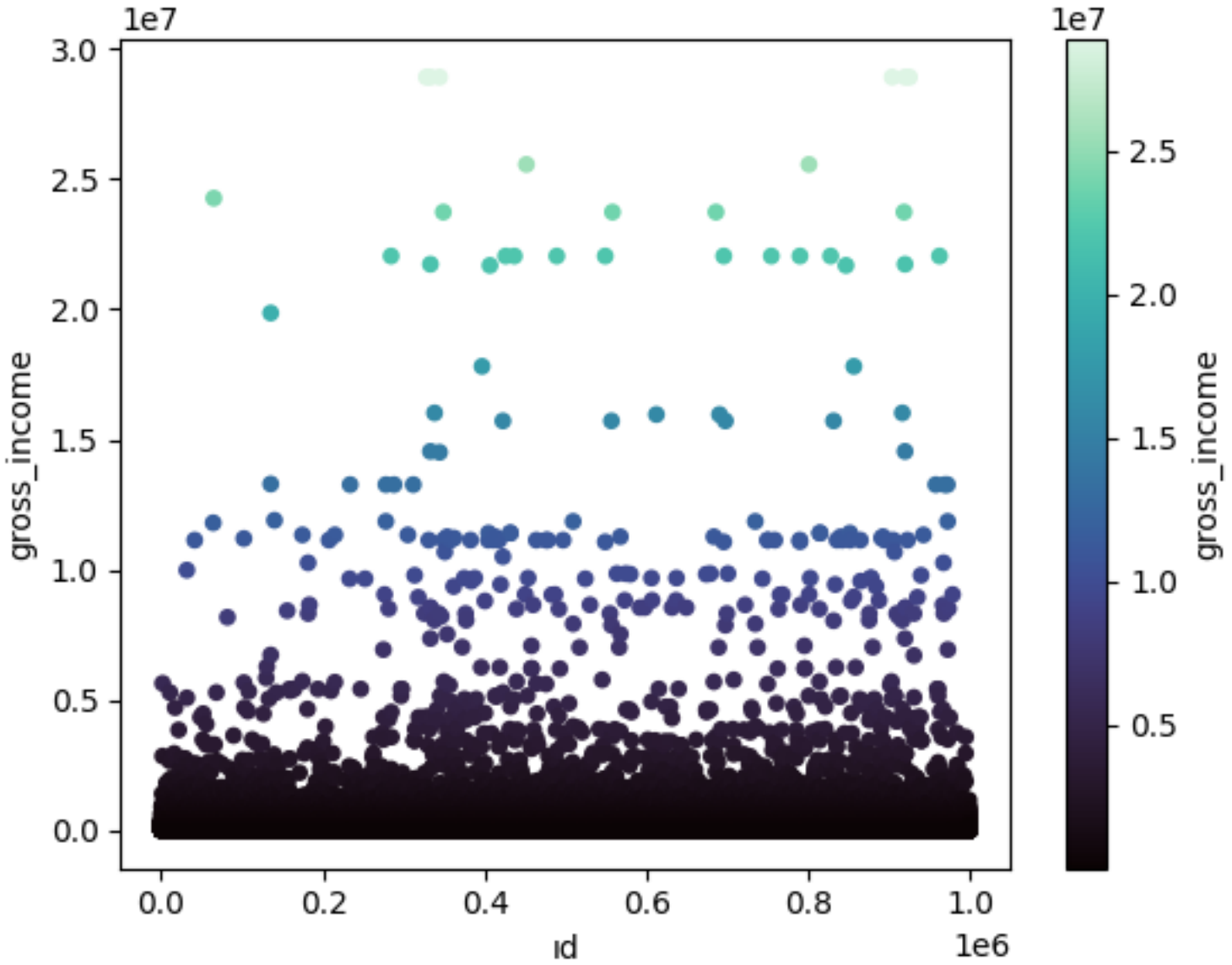
Exploratory Data Analysis (EDA)

Customer seniority (in months)

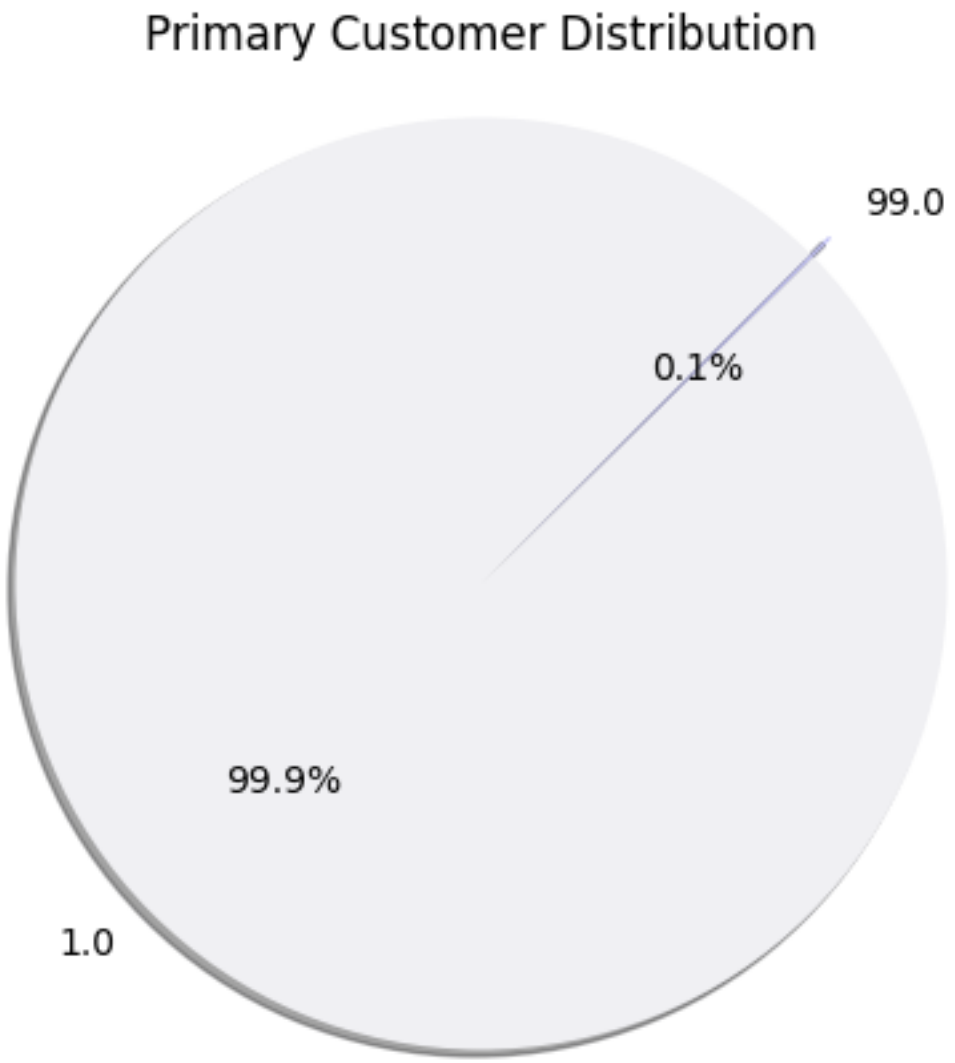


Exploratory Data Analysis (EDA)

gross income distribution



Exploratory Data Analysis (EDA)

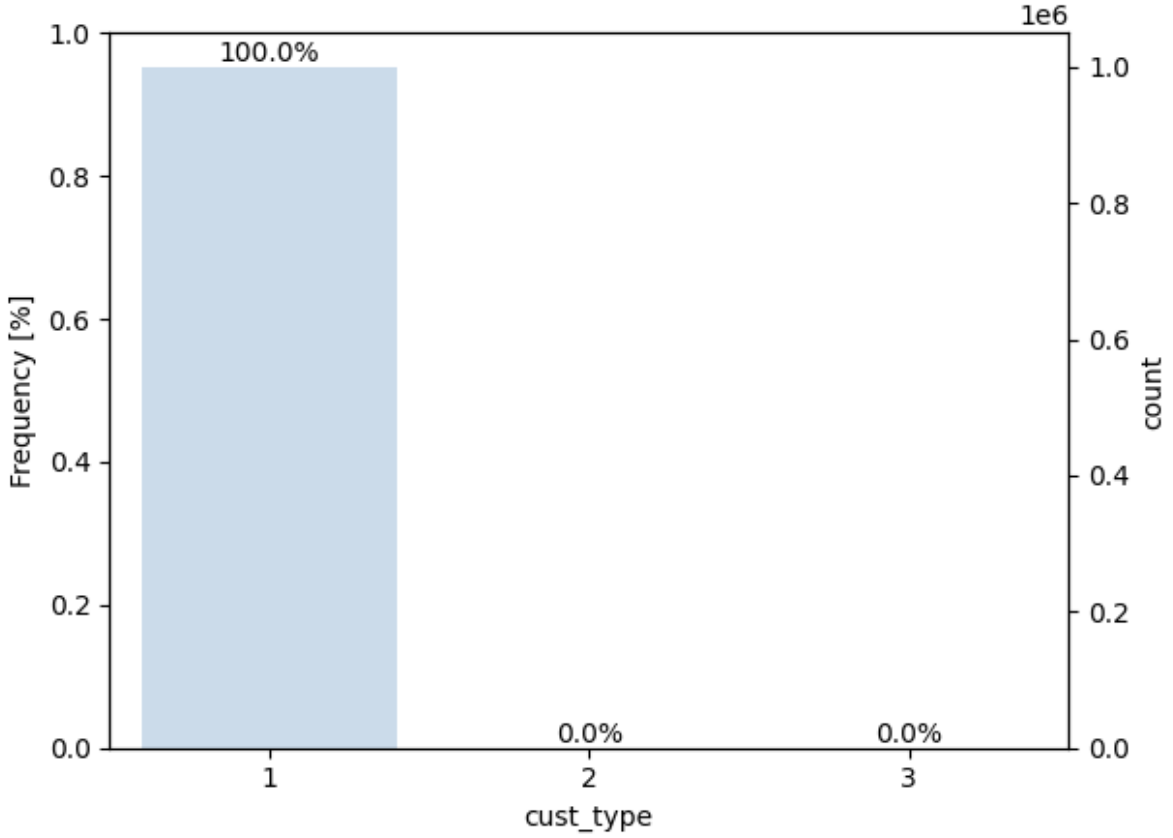


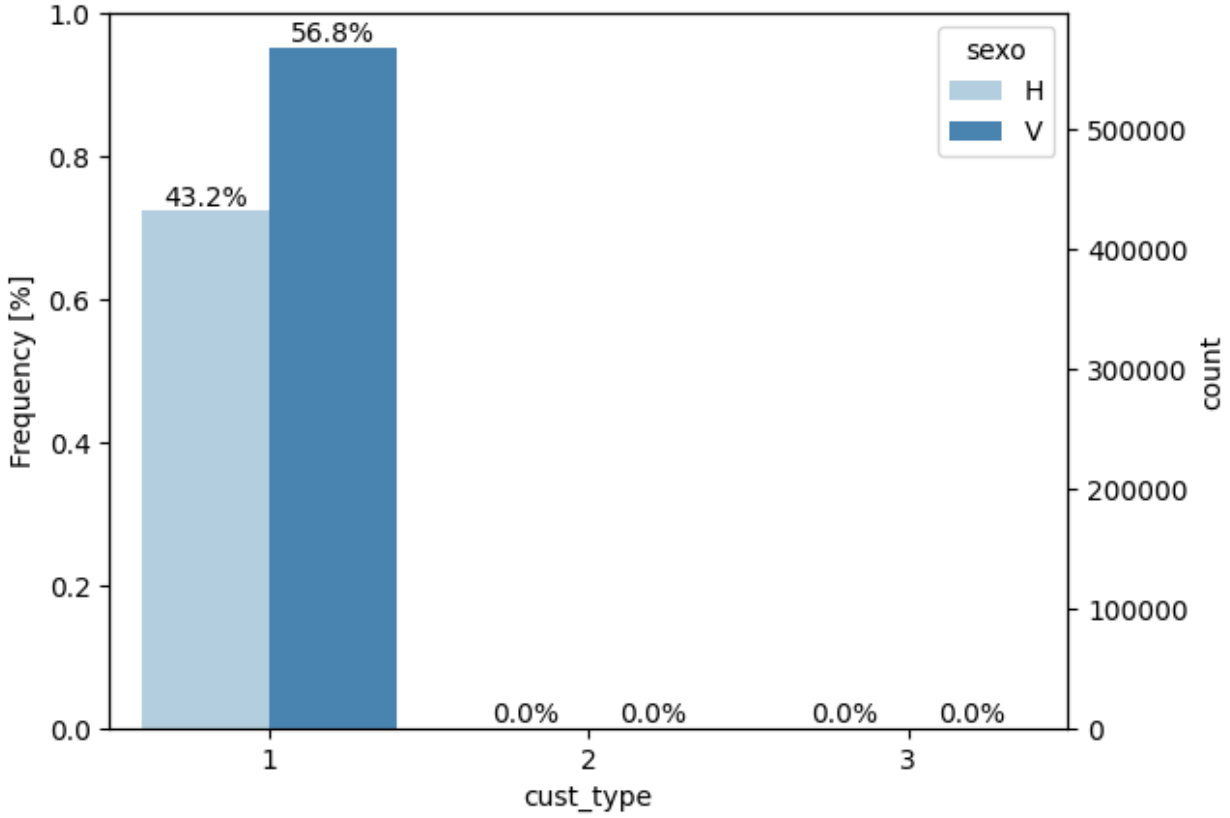
Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis (EDA)

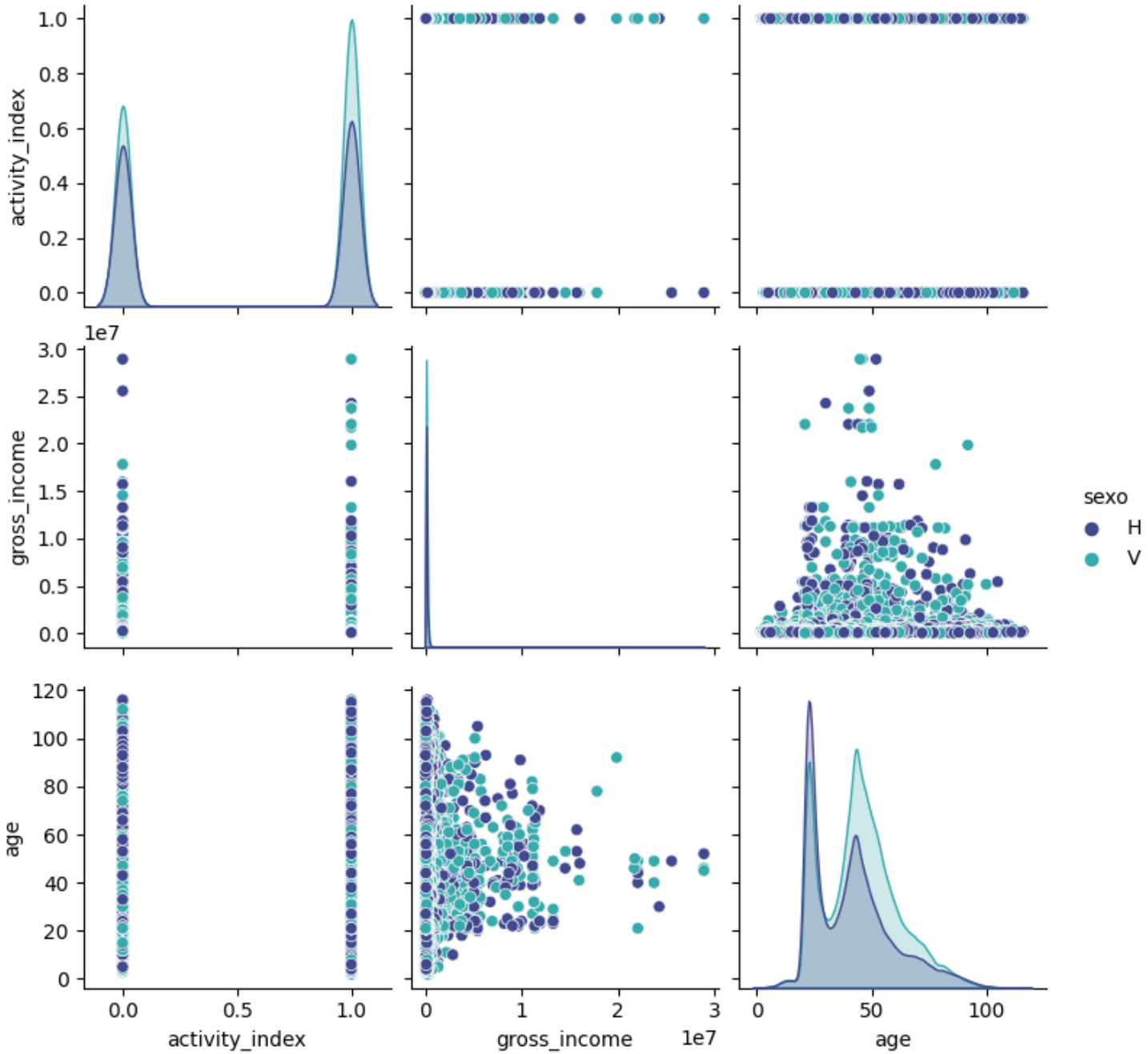
Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)





Exploratory Data Analysis (EDA)

Distribution of 'cost index, gross income and age' by gender



Data Glacier

Your Deep Learning Partner

Recommendations

Final Recommendation

A model can be built using the K-means method for customer segmentation and the Elbow method for determining the number of clusters.

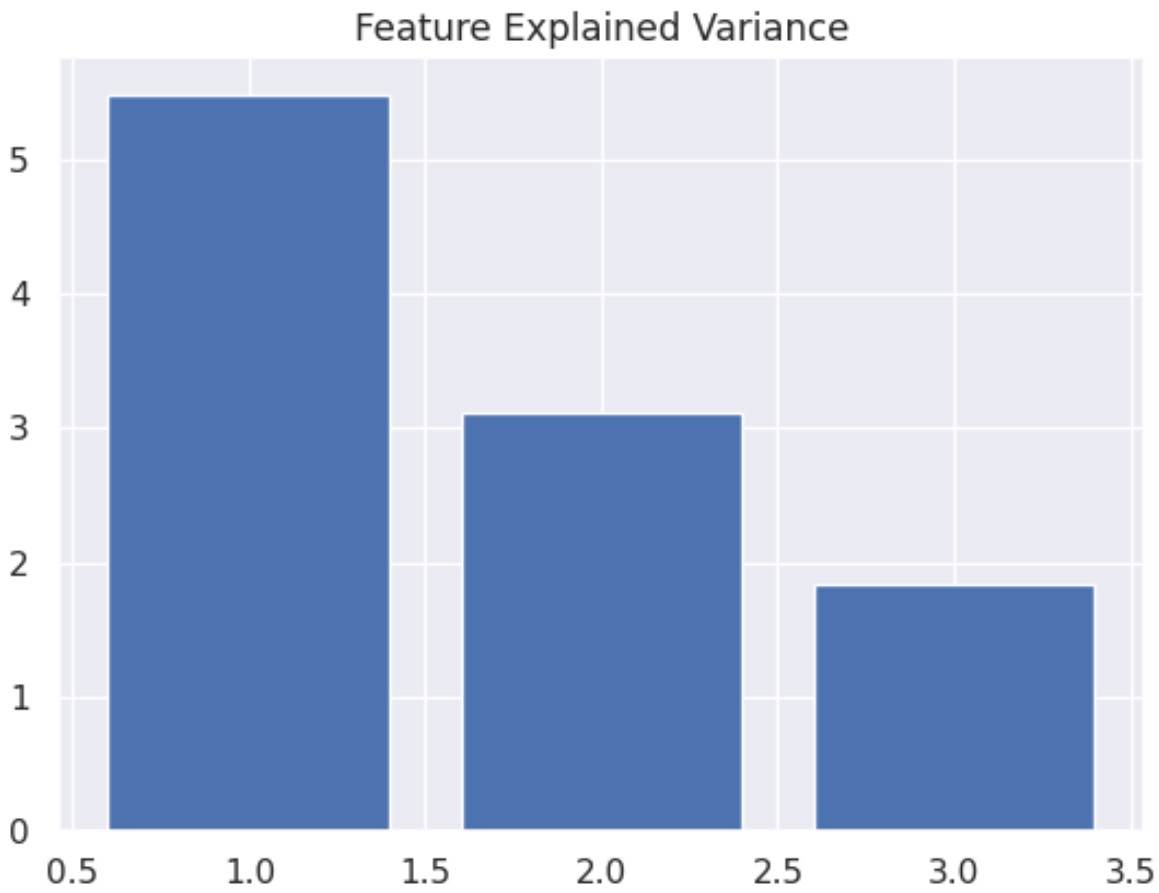
K-Means Clustering Algorithm: The KMeans model is an unsupervised machine learning model that works by simply splitting N observations into K numbers of clusters. The observations are grouped into these clusters based on how close they are to the mean of that cluster, which is commonly referred to as centroids.

- Specify number of clusters K.
- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.

Model Selection and Model Building

We will use the K-means method for customer segmentation. It is also quite effective model for customer segmentation. The K-Means model is an unsupervised machine learning model that works by simply splitting N observations into K numbers of clusters. The observations are grouped into these clusters based on how close they are to the mean of that cluster, which is commonly referred to as centroids.

Principal Component Analysis (PCA)

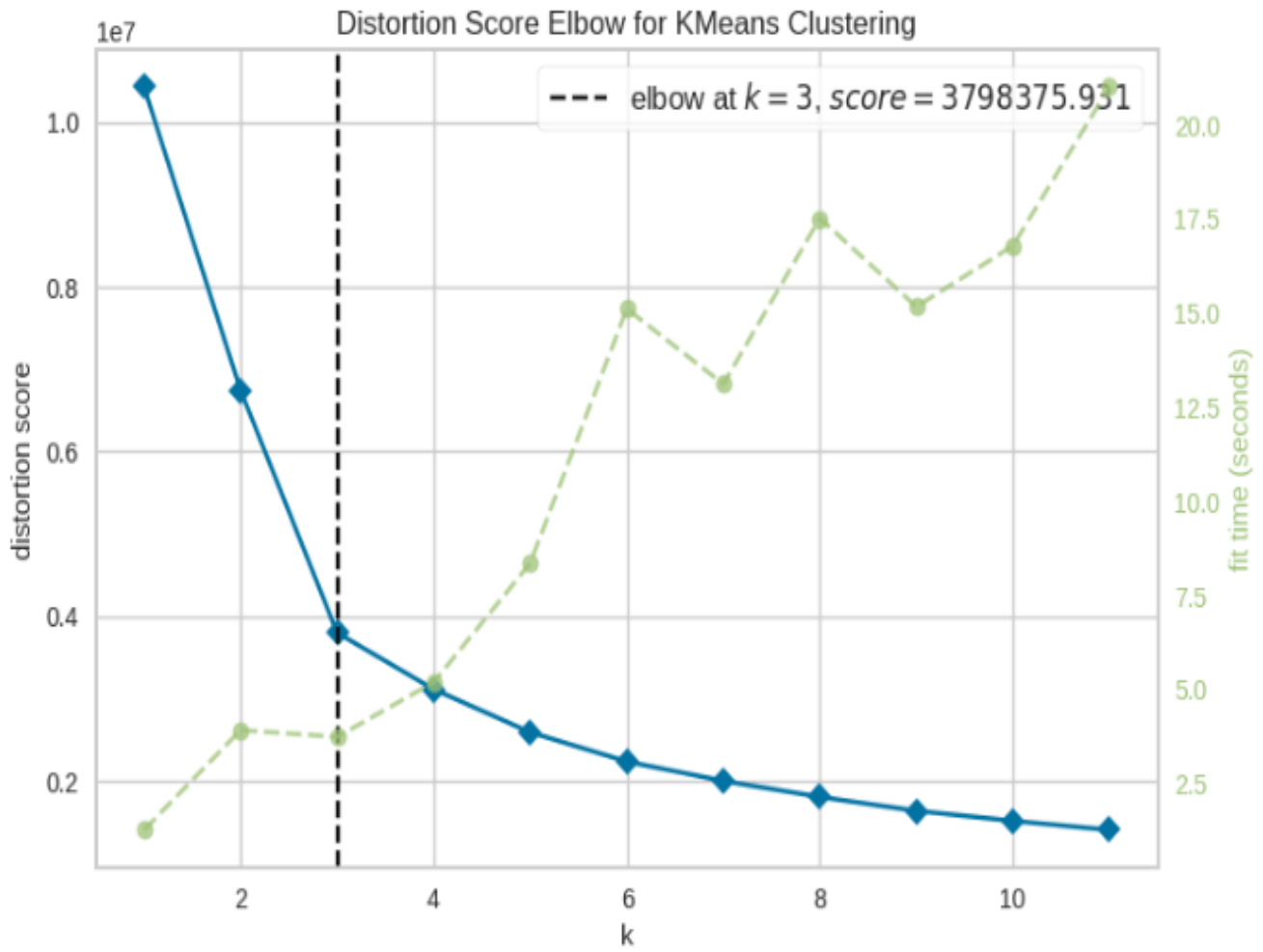


Data Glacier

Your Deep Learning Partner

Elbow Method

The graph shows that the optimal number of clusters (k) for our model is three.



Data Glacier

Your Deep Learning Partner

K_Means Clustering

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3, init='k-means++', random_state=0).fit(data_pca)
```

```
kmeans.labels_
```

```
array([1, 1, 1, ..., 1, 1, 1], dtype=int32)
```

```
kmeans.inertia_
```

```
3798378.5084681325
```

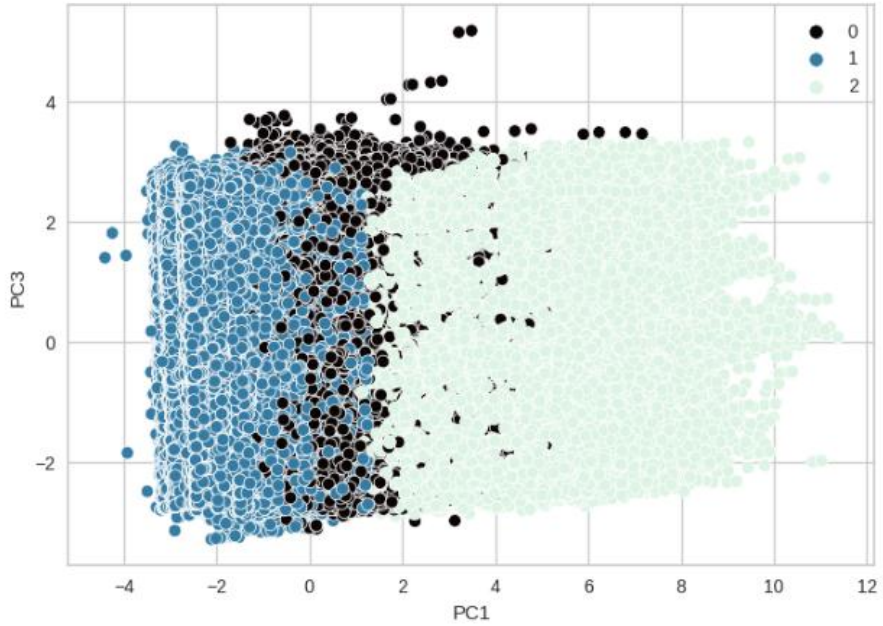
```
kmeans.cluster_centers_
```

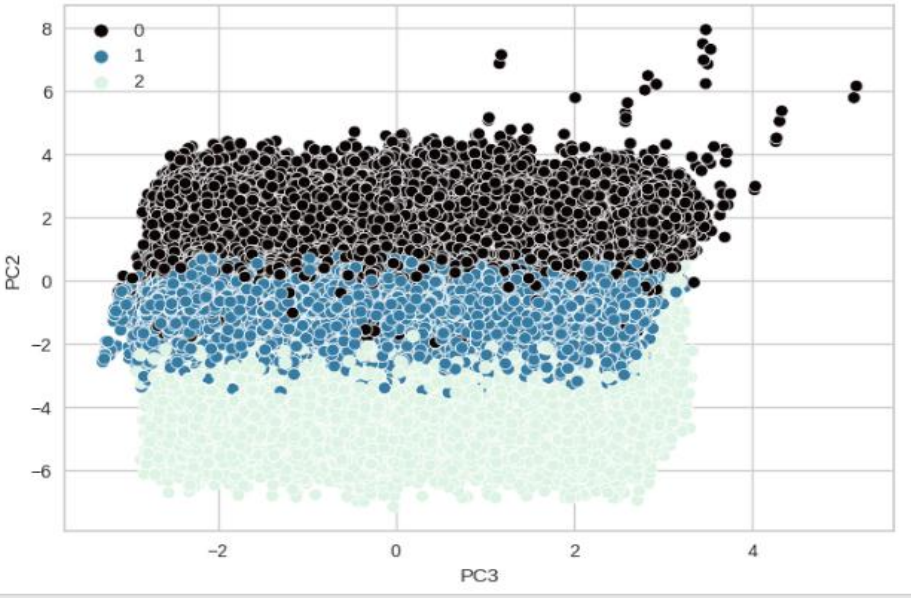
```
array([[ 0.61158024,  1.38439195,  0.02500715],
       [-2.07863506, -1.1083183 , -0.04310727],
       [ 5.03639032, -2.94844787,  0.03883934]])
```

```
from collections import Counter
Counter(kmeans.labels_)
```

```
Counter({1: 387087, 2: 96975, 0: 515938})
```

K_Means Clustering



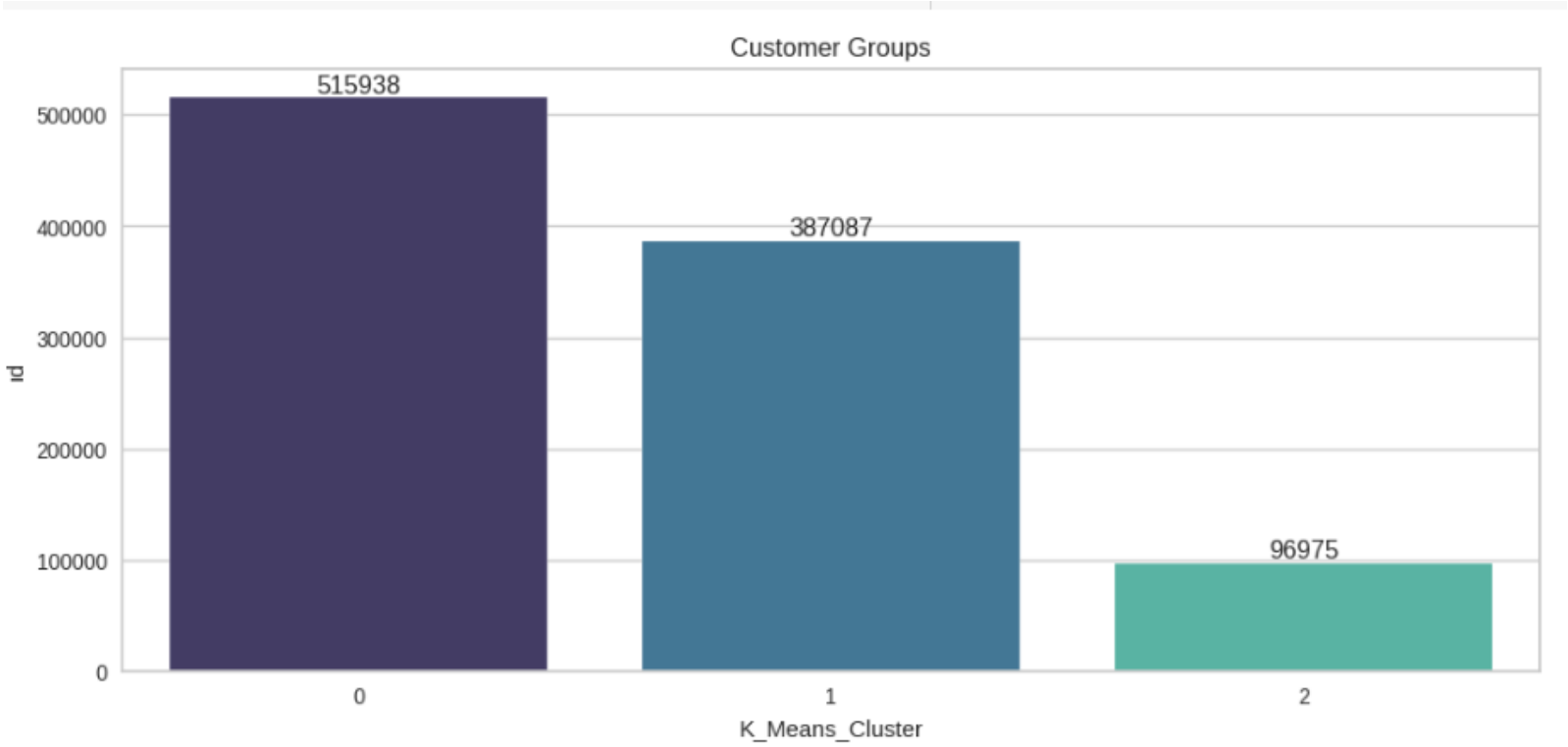


Data Glacier

Your Deep Learning Partner

K_Means Clustering

0 group is 52% of all customers, any improvements achieved in this customer group will dramatically benefit the campaign.



Conclusion

For an unsupervised machine learning task, the dataset was well-suited. To tackle this project, we used an unsupervised machine learning approach with the K Means clustering method using the PCA method as the dataset contained many features. After applying the Elbow method, it was determined that three clusters would be optimal. Using K-means clustering, patterns in the data were identified and used to create groups, paving the way for strategies tailored to these groups. In the future, customer groups can be created using specific features from the dataset, allowing for personalized offers to be extended.

Your Deep Learning Partner

Thank You