Tugas Grup Project #1 (TGP #1) - Eksplorasi dan Praproses Data

Disusun oleh :

Kelompok 4

Sulis Avandhy - 05211940000084

Aflah Aditya - 05211942000001

M Zuhdi Afi A - 05211940000135

INSTITUT TEKNOLOGI SEPULUH NOPEMBER SURABAYA

SEMESTER GENAP 2021/2022

1. Analisis Univariate

Pada analisis univariate, digunakan bantuan python dan library python khusus yaitu pandas_profiling. Library ini dapat membantu mempercepat dalam analisis univarite karena dapat membuat sebuah laporan secara lengkap untuk masing masing variabel.

Dataset statistics

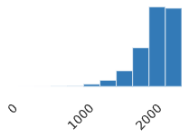
Number of variables	19
Number of observations	28382
Missing cells	3871
Missing cells (%)	0.7%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	7.2 MiB
Average record size in memory	265.1 B

Variable types

NUM	15
CAT	3
BOOL	1

1.1. vintage

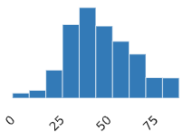
Distinct count	1459	Mean	2091.1441054189277
Unique (%)	5.1%	Minimum	73
Missing	0	Maximum	2476
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom “vintage” dapat ditarik beberapa kesimpulan diantaranya adalah distribusi data pada kurva diatas menunjukkan negative skew yang berarti berdasarkan data yang diberikan banyak diantaranya yang merupakan customer lama dan rata-rata pelanggan telah menggunakan layanan selama 2091 hari.

1.2. age

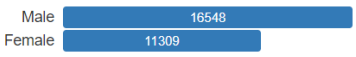
Distinct count	90	Mean	48.208336269466564
Unique (%)	0.3%	Minimum	1
Missing	0	Maximum	90
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom “age” dapat ditarik beberapa kesimpulan diantaranya adalah rentang umur customer yang menggunakan layanan pada bank tersebut mulai dari 1-90 tahun dan rata-rata umur customer adalah 48 tahun.

1.3. gender

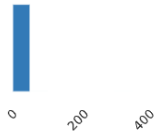
Distinct count	2
Unique (%)	< 0.1%
Missing	525
Missing (%)	1.8%
Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom “gender” dapat ditarik beberapa kesimpulan diantaranya adalah mayoritas customer yang menggunakan layanan pada bank ini adalah laki-laki dengan jumlah 16.548 orang sedangkan jumlah perempuan adalah 11.309 orang. Lalu terdapat 525 missing value atau sebesar 1.8% dari keseluruhan data.

1.4. dependents

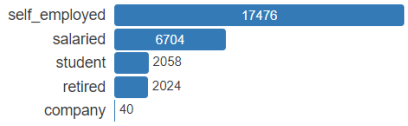
Distinct count	15	Mean	3.472356186581272
Unique (%)	0.1%	Minimum	0.0
Missing	2463	Maximum	520.0
Missing (%)	8.7%	Zeros	21435
Infinite	0	Zeros (%)	75.5%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom "dependents" dapat ditarik beberapa kesimpulan diantaranya adalah jumlah tanggungan customer yang menggunakan layanan pada bank tersebut mulai dari 0 hingga 520 orang dan rata-rata jumlah tanggungan customer adalah 3-4 orang. Variabel ini memiliki jumlah missing value terbanyak yaitu 8.7%.

1.5. occupation

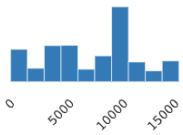
Distinct count	5
Unique (%)	< 0.1%
Missing	80
Missing (%)	0.3%
Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom "occupation" dapat ditarik beberapa kesimpulan diantaranya adalah terdapat 5 jenis pekerjaan dari seluruh customer yang ada yaitu *self employee*, *salaried*, *student*, *retired*, dan *company*. Pekerjaan customer terbanyak adalah *self employed* dengan jumlah 17.476 dan pekerjaan customer tersedikit adalah *retired*. Sedangkan company adalah perusahaan yang bermitra dengan bank tersebut bukan sebuah jenis pekerjaan.

1.6. city

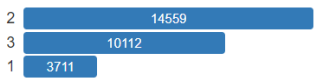
Distinct count	1604	Mean	7961.095761267631
Unique (%)	5.8%	Minimum	0.0
Missing	803	Maximum	16490.0
Missing (%)	2.8%	Zeros	16
Infinite	0	Zeros (%)	0.1%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom "city" dapat ditarik beberapa kesimpulan diantaranya adalah terdapat cukup banyak jumlah missing value yaitu berjumlah 803 kolom dengan persentase sebesar 2.8%. Variabel ini menggunakan tipe data real number dengan frekuensi terbanyak dimiliki oleh kota 10200 yaitu sebesar 12.3%.

1.7. customer_nw_category

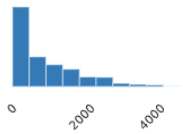
Distinct count	3
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom "customer_nw_category" dapat ditarik beberapa kesimpulan diantaranya adalah net worth customer diklasifikasikan menjadi 3 (1:High, 2:Medium, 3:Low). Berdasarkan data yang diberikan terdapat 3711 customer yang diklasifikasikan ke High, 14.559 orang yang diklasifikasikan ke Medium, dan 10.112 orang yang diklasifikasikan ke Low.

1.8. branch_code

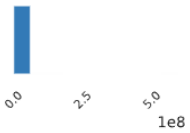
Distinct count	3185	Mean	925.9750193784794
Unique (%)	11.2%	Minimum	1
Missing	0	Maximum	4782
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom “branch_code” dapat ditarik beberapa kesimpulan diantaranya bahwa setiap akun kastemer memiliki kode cabang mereka masing-masing tetapi satu kode dapat dimiliki oleh banyak akun customer. Contohnya yaitu kode 19 dimiliki oleh 145 akun customer.

1.9. current_balance

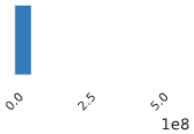
Distinct count	27814	Mean	664336.6531604538
Unique (%)	98.0%	Minimum	-550396
Missing	0	Maximum	590590403
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom “current_balance” dapat ditarik beberapa kesimpulan diantaranya adalah variabel ini memiliki tingkat korelasi yang tinggi dengan 3 variabel lainnya yaitu; previous_month_end_balance, average_monthly_balance_prevQ, dan current_month_balance sebagai input data untuk mencari jumlah saldo customer per hari ini. Dapat dilihat bahwa jumlah saldo per hari ini terbanyak yang dimiliki oleh seorang customer dapat mencapai 590.590.403.

1.10. previous_month_end_balance

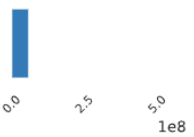
Distinct count	27853	Mean	678960.4222042139
Unique (%)	98.1%	Minimum	-314957
Missing	0	Maximum	574043863
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom “previous_month_end_balance” dapat ditarik beberapa kesimpulan diantaranya adalah variabel ini memiliki korelasi dengan beberapa variabel lainnya seperti average_monthly_balance_prevQ dan average_monthly_balance_prevQ2 untuk mencari rata-rata saldo customer pada previous or previous to previous quarter. Lalu didapatkan informasi customer lainnya seperti saldo akhir tertinggi pada bulan sebelumnya adalah 574.043.863 dan saldo akhir terendah pada bulan sebelumnya adalah -314.957 serta rata-rata saldo pada akhir bulan sebelumnya adalah 678.960

1.11. average_monthly_balance_prevQ

Distinct count	27790	Mean	684547.9812204918
Unique (%)	97.9%	Minimum	14290
Missing	0	Maximum	570028957
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom “average_monthly_balance_prevQ” dapat ditarik beberapa kesimpulan diantaranya adalah variabel ini memiliki korelasi dengan beberapa variabel lainnya

seperti current_balance dan previous_month_end_balance sebagai input data untuk mencari rata-rata saldo customer pada previous quarter. Lalu didapatkan informasi customer lainnya seperti saldo rata-rata bulanan tertinggi pada kuartal sebelumnya adalah 570.028.957 dan saldo rata-rata bulanan terendah pada kuartal sebelumnya adalah 14.290 serta nilai mean dari saldo rata-rata bulanan pada kuartal sebelumnya adalah 684.547,98

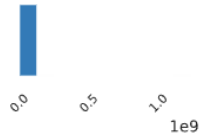
1.12. average_monthly_balance_prevQ2

Distinct count	27879	Mean	641136.308752026
Unique (%)	98.2%	Minimum	-1069193
Missing	0	Maximum	452604901
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB

Berdasarkan hasil analisis univariate pada kolom "average_monthly_balance_prevQ2" dapat ditarik beberapa kesimpulan diantaranya adalah variabel ini memiliki korelasi dengan beberapa variabel lainnya seperti current_balance dan previous_month_end_balance sebagai input data untuk mencari rata-rata saldo customer pada previous to previous quarter. Lalu didapatkan informasi customer lainnya seperti saldo rata-rata bulanan tertinggi pada dua kuartal sebelumnya adalah 452.604.901 dan saldo rata-rata bulanan terendah pada dua kuartal sebelumnya adalah -1.069.193 serta nilai mean dari saldo rata-rata bulanan pada dua kuartal sebelumnya adalah 641.136,3

1.13. current_month_credit

Distinct count	10203	Mean	310965.749383412
Unique (%)	35.9%	Minimum	1
Missing	0	Maximum	1226984539
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom "current_month_credit" dapat ditarik beberapa kesimpulan diantaranya adalah variabel ini memiliki korelasi dengan variabel current_month_debit sebagai perbandingan total jumlah kredit dan debit pada bulan ini. Total jumlah kredit terbanyak bulan ini yang dimiliki seorang customer yaitu 1.226.984.539. Rata-rata total jumlah kredit bulan ini yaitu 310.965.

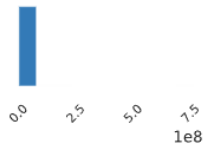
1.14. previous_month_credit

Distinct count	10491	Mean	296944.94246353325
Unique (%)	37.0%	Minimum	1
Missing	0	Maximum	236180829
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB

Berdasarkan hasil analisis univariate pada kolom "previous_month_credit" dapat ditarik beberapa kesimpulan diantaranya adalah variabel ini memiliki korelasi dengan variabel previous_month_debit sebagai perbandingan total jumlah kredit dan debit pada bulan lalu. Selain itu, jumlah kredit terbanyak pada bulan lalu yang dimiliki seorang customer yaitu 236..180.829 sedangkan jumlah kredit terkecil pada bulan lalu yang dimiliki seorang customer yaitu 1 serta rata-rata total jumlah kredit bulan lalu yaitu 296.944.

1.15. current_month_debit

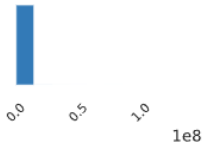
Distinct count	13437	Mean	328485.44207596366
Unique (%)	47.3%	Minimum	1
Missing	0	Maximum	763785736
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom “current_month_debit” dapat ditarik beberapa kesimpulan diantaranya adalah variabel ini memiliki korelasi dengan variabel current_month_credit sebagai perbandingan total jumlah kredit dan debit pada bulan ini. Total jumlah debit terbanyak bulan ini yang dimiliki seorang customer yaitu 763.785.736 sedangkan total jumlah debit tersedikit bulan ini yang dimiliki seorang customer yaitu 1. Lalu, rata-rata total jumlah kredit bulan ini yaitu 328.485.

1.16. previous_month_debit

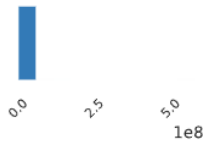
Distinct count	13753	Mean	295026.3651610176
Unique (%)	48.5%	Minimum	1
Missing	0	Maximum	141416806
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom “previous_month_debit” dapat ditarik beberapa kesimpulan diantaranya adalah variabel ini memiliki korelasi dengan variabel previous_month_credit sebagai perbandingan total jumlah kredit dan debit pada bulan lalu. Selain itu, jumlah debit terbanyak pada bulan lalu yang dimiliki seorang customer yaitu 141.416.806 sedangkan total jumlah debit tersedikit pada bulan lalu yang dimiliki seorang customer yaitu 1 serta rata-rata jumlah kredit bulan lalu yaitu 295.026.

1.17. current_month_balance

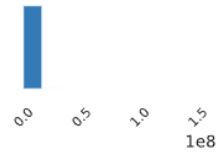
Distinct count	27900	Mean	675884.0763159749
Unique (%)	98.3%	Minimum	-337418
Missing	0	Maximum	577818477
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom “current_month_balance” dapat ditarik beberapa kesimpulan diantaranya adalah variabel ini memiliki korelasi dengan 3 variabel lainnya yaitu; current_balance, average_monthly_balance_prevQ, dan previous_month_end_balance sebagai input data untuk mencari saldo rata-rata customer bulan ini. Rata-rata saldo bulan yaitu 675.884, dengan saldo akhir tertinggi yaitu 577.818.477 dan saldo akhir terendah yaitu -337.418.

1.18. previous_month_balance

Distinct count	27885	Mean	663884.1024240716
Unique (%)	98.2%	Minimum	-517192
Missing	0	Maximum	157283293
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	221.9 KiB



Berdasarkan hasil analisis univariate pada kolom “previous_month_balance” dapat ditarik beberapa kesimpulan diantaranya adalah rata-rata saldo pada akhir bulan sebelumnya yaitu 663.884, dengan saldo akhir tertinggi pada bulan sebelumnya yaitu 157.283.293 dan saldo akhir terendah pada bulan sebelumnya yaitu -517.192.

1.19. churn

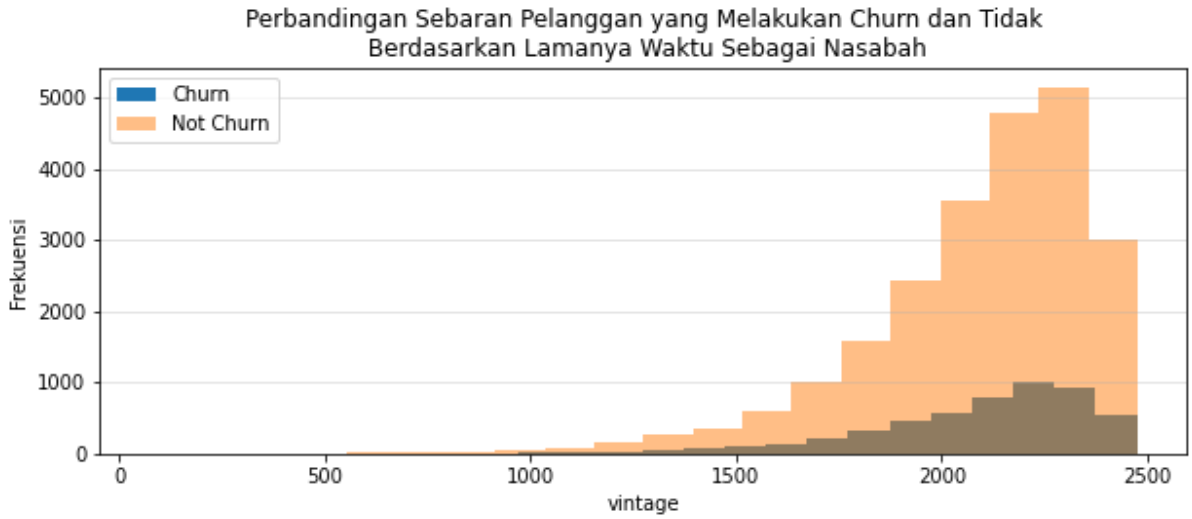
Distinct count	2
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	221.9 KiB

Berdasarkan hasil analisis univariate pada kolom “churn” dapat ditarik beberapa kesimpulan diantaranya adalah variabel ini merupakan variabel dependen yang akan dilakukan komparasi dengan variabel lain untuk mengetahui penyebab atau tanda-tanda customer akan melakukan churn. Apabila dilihat dari data statistik diatas maka dapat ditarik kesimpulan bahwa dari data tersebut ada 5.260 customer atau 18,5% yang melakukan churn dari seluruh data yang disediakan

2. **Analisis Bivariate**

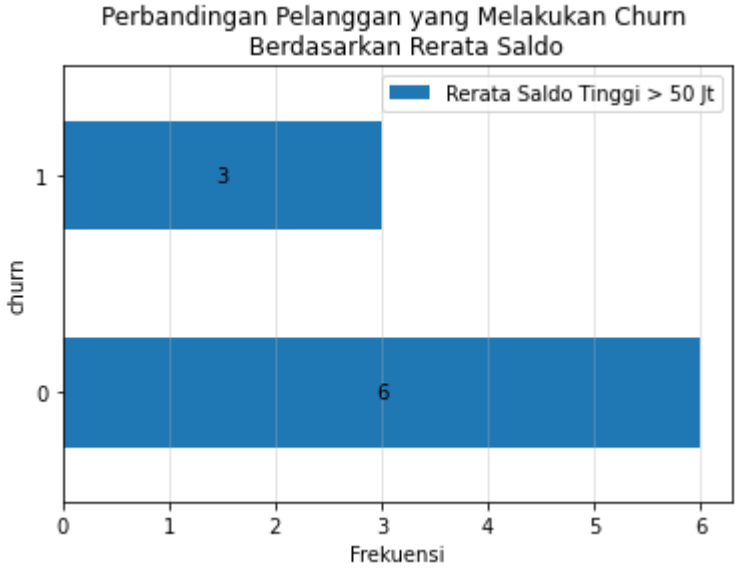
2.1. **Apakah kastemer lama (vintage customer) cenderung untuk melakukan churn?**

Berdasarkan hasil exploratory dari data yang diberikan kami memberi kesimpulan bahwa customer lama cenderung melakukan churn karena apabila dilihat dari gambar histogram dibawah, semakin lama pelanggan menggunakan layanan pada bank ini maka semakin banyak juga pelanggan yang melakukan churn. Walaupun akan terlihat jauh lebih sedikit pelanggan yang melakukan churn apabila dibandingkan dengan grafik pelanggan yang tidak melakukan churn berdasarkan durasi penggunaan layanan pada bank ini.



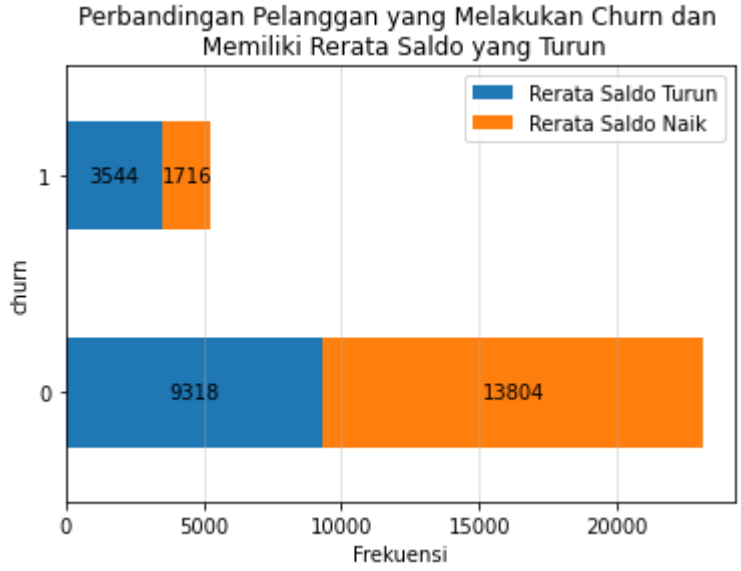
2.2. **Apakah kastemer dengan rata-rata saldo tertinggi akan cenderung untuk melakukan churn?**

Berdasarkan hasil exploratory dari data yang diberikan kami mengasumsikan saldo tertinggi yaitu customer dengan saldo lebih dari 50 juta. Kesimpulan yang didapat yaitu bahwa customer dengan rerata saldo tinggi > 50 jt tidak cenderung untuk melakukan churn. Hal ini dapat ditunjukkan dengan grafik dibawah yang mana apabila dihitung persentase antara customer yang melakukan dan tidak melakukan churn adalah 33,33%.



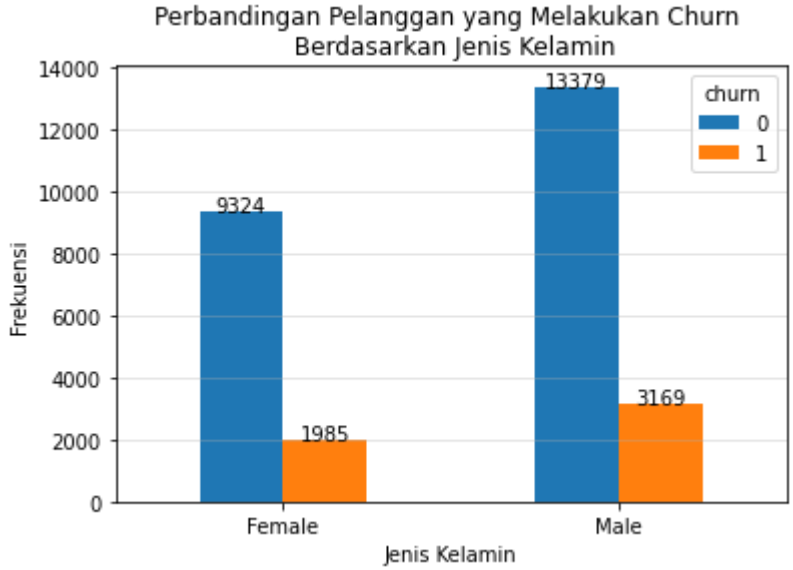
2.3. **Apakah kastemer yang mengalami penurunan saldo bulanan akan cenderung untuk melakukan churn?**

Berdasarkan hasil exploratory dari data yang diberikan kami memberi kesimpulan bahwa customer yang mengalami penurunan saldo bulanan tidak cenderung untuk melakukan churn. Hal ini dapat dibuktikan dengan grafik "perbandingan pelanggan yang melakukan churn dan memiliki rerata saldo yang turun" pada gambar dibawah ini, yang mana terdapat 3544 customer yang melakukan churn karena mengalami penurunan saldo bulanan sedangkan total orang yang mengalami penurunan saldo bulanan adalah 12862 customer. Dan apabila dihitung nilai persentase antara customer yang melakukan churn dan tidak, maka terdapat 73,5% customer yang tidak melakukan churn dari seluruh customer yang mengalami penurunan saldo bulanan



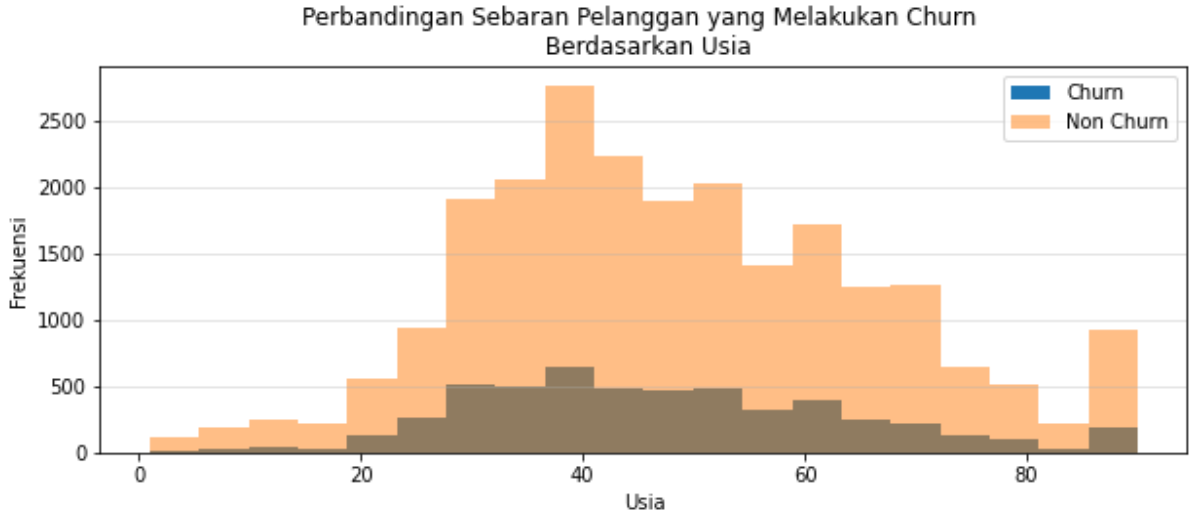
2.4. **Apakah kastemer perempuan mempunyai kecenderangan yang rendah untuk melakukan churn?**

Berdasarkan hasil exploratory dari data yang diberikan kami memberi kesimpulan bahwa customer perempuan mempunyai kecenderungan yang rendah untuk melakukan churn dengan asumsi dapat dikatakan tingkat kecenderungannya rendah apabila persentase jumlah customer yang melakukan churn pada suatu kondisi tersebut kurang dari 20%. Sedangkan dari hasil grafik tersebut didapatkan ada 1.985 customer perempuan yang melakukan churn dari 11.309 customer perempuan dan apabila dihitung persentasenya didapatkan angka 17.5% atau kurang dari 20%.



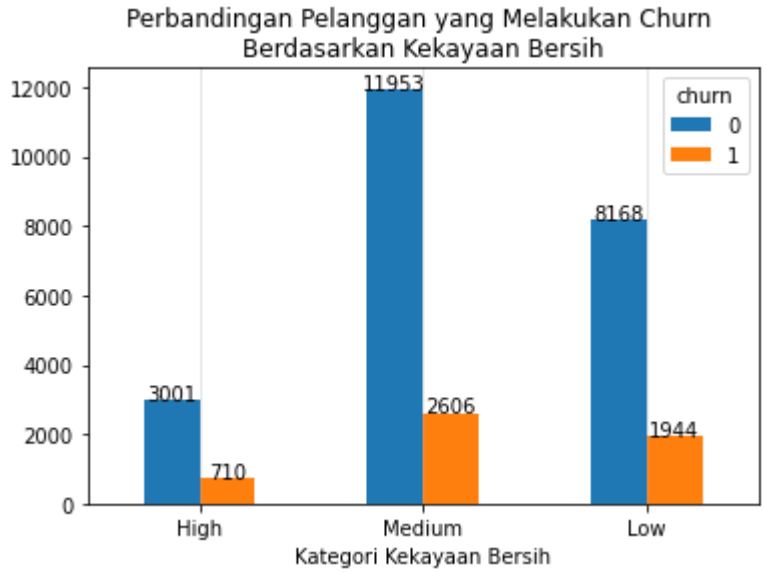
2.5. **Apakah kastemer muda akan mempunyai kecenderungan melakukan churn?**

Berdasarkan hasil exploratory dari data yang diberikan kami memberi kesimpulan bahwa customer muda tidak mempunyai kecenderungan untuk melakukan churn dengan asumsi customer muda ini berusia pada rentang umur 1-20 tahun. Hal ini dapat dibuktikan dari grafik "Perbandingan sebaran pelanggan yang Melakukan Churn dan tidak, Berdasarkan Usia" pada gambar dibawah ini, yang mana apabila dilihat pada rentang umur 1-20 tahun ini cenderung sedikit bahkan hampir tidak ada yang melakukan churn dan ini bertolak belakang dengan customer yang berusia diatas 20 atau 30 tahun keatas.



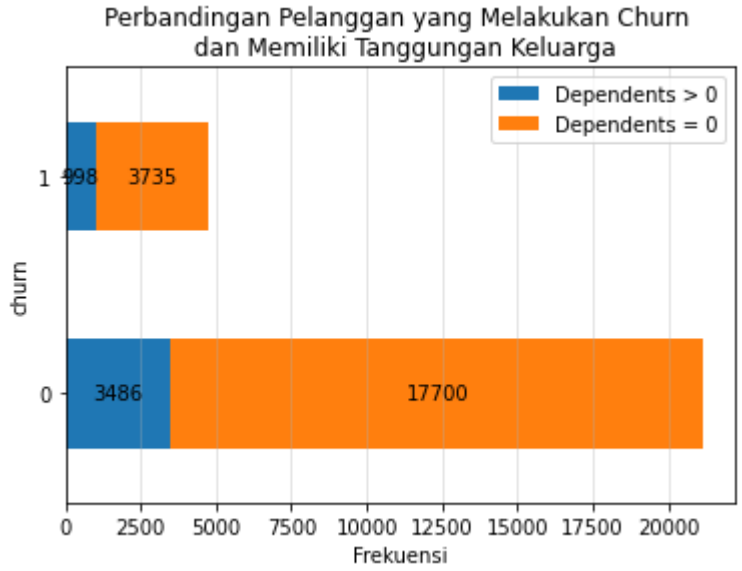
2.6. **Apakah kastemer yang dengan penghasilan kecil akan cenderung melakukan churn?**

Berdasarkan hasil exploratory dari data yang diberikan kami memberi kesimpulan bahwa variabel penghasilan customer kurang berpengaruh terhadap variabel churn atau dengan kata lain customer dengan penghasilan kecil cenderung tidak melakukan churn. Kami berasumsi bahwa jumlah penghasilan akan berbanding lurus dengan jumlah kekayaan bersih, sehingga kami membuktikannya dengan menggunakan grafik histogram dibawah ini dengan judul "Perbandingan Pelanggan yang Melakukan Churn Berdasarkan Kekayaan Bersih". Kami mengasumsikan bahwa variabel ini tidak mempengaruhi churn karena apabila dihitung persentase antara yang melakukan dan tidak melakukan churn pada setiap kategori (High, Medium, Low) diperoleh hasil yang hampir sama antar ketiganya. High diperoleh 18.9%, Medium diperoleh 17.9%, dan High diperoleh 19.2%.



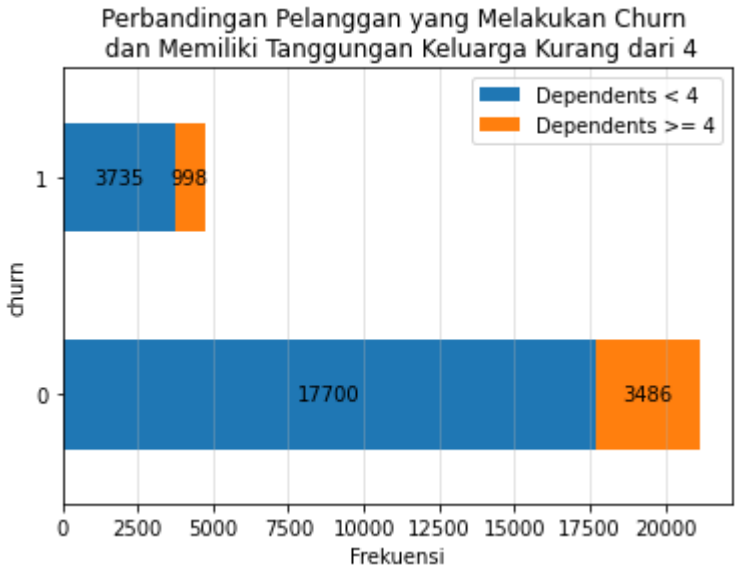
2.7. **Apakah kastemer yang memiliki tanggungan keluarga cenderung untuk melakukan churn?**

Berdasarkan grafik “Perbandingan Pelanggan yang Melakukan Churn dan Memiliki Tanggungan Keluarga” dibawah ini, customer yang memiliki tanggungan keluarga tidak cenderung untuk melakukan churn. Pada grafik tersebut terdapat keterangan “Dependents > 0, yaitu customer yang memiliki tanggungan keluarga sedangkan Dependents = 0, yaitu customer tanpa tanggungan keluarga”. Jumlah customer yang melakukan churn yaitu 998 sedangkan jumlah customer yang tidak melakukan churn yaitu 3486 dari total 4484 customer yang memiliki tanggungan keluarga. Apabila dihitung nilai persentase antara customer yang melakukan churn dan tidak, maka terdapat 22,3% customer yang melakukan churn dari seluruh customer yang memiliki tanggungan keluarga.



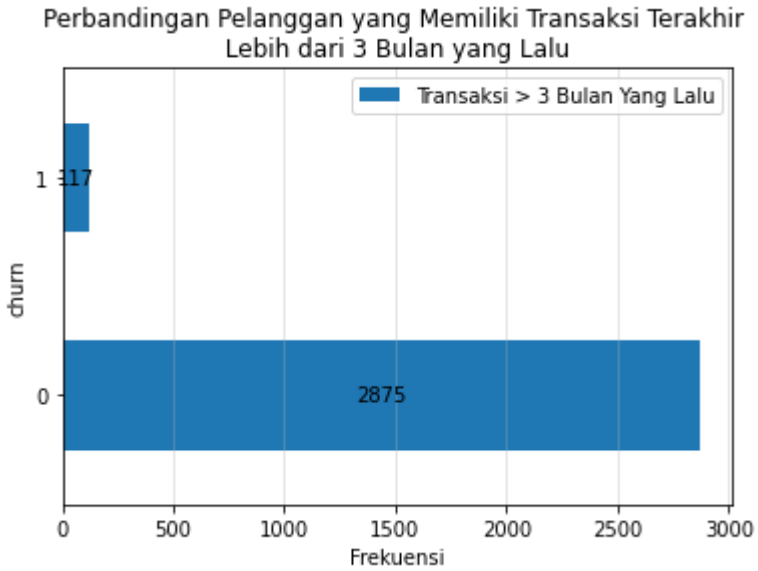
2.8. **Apakah kastemer dengan rata-rata jumlah tanggungan keluarga kurang dari 4 cenderung untuk melakukan churn?**

Berdasarkan hasil exploratory dari data serta grafik yang diberikan, kami memberi kesimpulan bahwa customer dengan rerata jumlah tanggungan keluarga kurang dari 4 tidak cenderung untuk melakukan churn. Pada grafik tersebut terdapat keterangan “Dependents < 4, yaitu customer yang memiliki tanggungan keluarga kurang dari 4 sedangkan Dependents >= 4, yaitu customer dengan tanggungan 4 keluarga atau lebih”. Jika dibandingkan dengan kriteria customer yang sama tetapi melakukan churn, terdapat 3735 customer yang melakukan churn sedangkan jumlah customer yang tidak melakukan churn yaitu 17700, dari total 21435 customer dengan rerata jumlah tanggungan keluarga kurang dari 4. Apabila dihitung nilai persentase antara customer yang melakukan churn dan tidak, maka terdapat 17,4% customer yang melakukan churn dari seluruh customer dengan rerata jumlah tanggungan keluarga kurang dari 4.



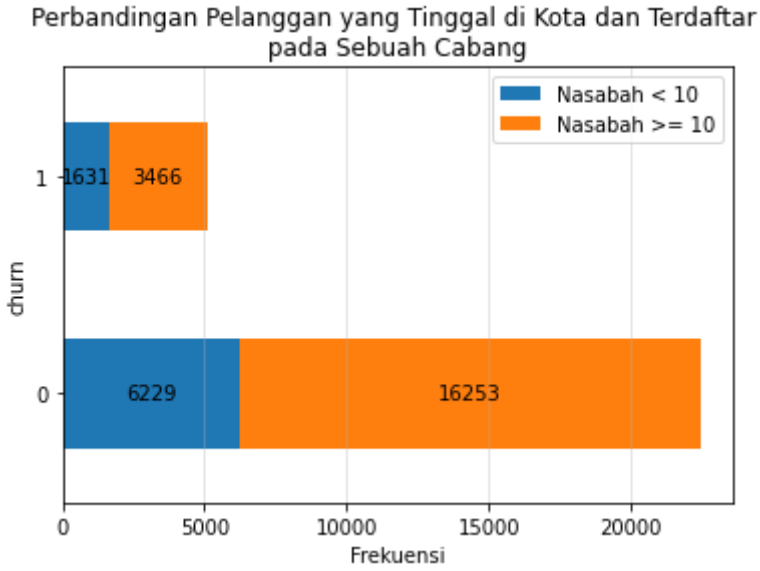
2.9. **Apakah kastemer yang melakukan transaksi terakhir lebih dari 6 bulan lalu mempunyai kecenderungan churn yang lebih tinggi?**

Berdasarkan hasil exploratory dari data serta grafik yang diberikan, kami memberi kesimpulan bahwa customer yang melakukan transaksi lebih dari 3 bulan yang lalu memiliki kecenderungan churn yang lebih rendah. Kami menggunakan data transaksi 3 bulan lalu menghitung selisih saldo dari dua kuartal sebelumnya, tetapi tidak ditemukan data yang memenuhi hal tersebut sehingga digunakan data selisih saldo dari kuartal sebelumnya. Kesimpulan yang diambil dapat dibuktikan dengan menggunakan grafik bar pada gambar di bawah. Pada gambar tersebut, dapat dilihat bahwa hanya 117 pelanggan atau 0,04% dari keseluruhan pelanggan yang bertransaksi 3 bulan terakhir. Sementara itu, dengan menggunakan kriteria yang sama, terdapat 2875 pelanggan yang tidak melakukan churn.



2.10. **Apakah ada kemungkinan kastemer yang tinggal di kota dan terdaftar pada sebuah cabang dengan jumlah nasabah yang kecil akan cenderung melakukan churn?**

Berdasarkan hasil exploratory dari data yang diberikan kami memberi kesimpulan bahwa customer yang tinggal di kota dan terdaftar pada sebuah cabang dengan jumlah nasabah yang kecil akan cenderung melakukan churn. Kami berasumsi pelanggan cenderung melakukan churn apabila persentase pada kondisi tersebut lebih dari 20%. Kesimpulan yang telah kami buat dapat dibuktikan dengan menggunakan grafik histogram dibawah ini dengan judul "Perbandingan Pelanggan yang Tinggal di Kota dan terdaftar pada Sebuah Cabang", yang mana persentase pelanggan yang melakukan churn saat terdaftar pada sebuah cabang dengan jumlah nasabah yang kecil (X<10) adalah 20,75%. Sedangkan persentase pelanggan yang melakukan churn ketika terdaftar pada sebuah cabang dengan jumlah nasabah yang besar (X>=10) adalah 17,57%



3. Pra Proses Data

3.1. Missing Value

Pada tahapan ini, dilakukan analisis terhadap nilai yang kosong pada dataset. Analisis diawali dengan pengecekan jumlah *missing value* di setiap kolom pada dataset. Setelah dilakukan pengecekan, diketahui bahwa terdapat empat kolom yang memiliki missing value yaitu *gender* sebanyak 525 baris, *dependents* sebanyak 2463 baris, *occupation* sebanyak 80 baris, dan *customer_nw_category* sebanyak 803 baris. Pemrosesan *missing value* pada kolom gender dilakukan dengan mengisi nilai yang kosong dengan nilai modus dari kolom gender karena kolom gender merupakan kategori dan jumlah nilai yang kosong tidak terlalu besar jika dibandingkan dengan jumlah data keseluruhan.

Pemrosesan *missing value* pada kolom occupation juga dilakukan dengan mengisi nilai yang kosong dengan nilai modus dari kolom occupation karena kolom gender merupakan kategori dan jumlah nilai yang kosong tidak terlalu besar jika dibandingkan dengan jumlah data keseluruhan. Selanjutnya, untuk pemrosesan *missing value* pada kolom dependents, dilakukan dengan mengisi nilai yang hilang menggunakan nilai median dari kolom tersebut. Hal ini dilakukan karena data dependents merupakan data numerik dan sebaran dari data tersebut tidak berdistribusi normal sehingga lebih baik menggunakan median.

3.2. Duplicate Value

Tahapan selanjutnya adalah memeriksa nilai yang terduplikasi pada dataset. Berdasarkan hasil analisis menggunakan python, diketahui tidak terdapat data yang terduplikasi pada dataset.

3.3. Outlier

Tahapan selanjutnya adalah pemeriksaan outlier. Pada tahap ini, kolom yang akan dianalisis adalah seluruh kolom terkecuali kolom *customer_id, gender, occupation, city, customer_nw_category, dan branch_code.* Hal ini dikarenakan kolom tersebut merupakan data yang termasuk data kategorik. Analisis outlier menggunakan metode perhitungan batas atas dan batas bawah berdasar dari nilai interkuartil data. Batas bawah ditetapkan sebagai Q1 - 1.5 * IQR dan batas atas adalah Q3 + 1.5 * IQR dimana Q1 adalah kuartil 1, Q3 adalah kuartil 3 dan IQR adalah jarak interkuartil atau Q3 - Q1.

Analisis dilakukan menggunakan bantuan python dan library dengan membuat beberapa fungsi sehingga proses pendeteksian dan penghapusan outlier dapat dilakukan secara otomatis.

3.4. One Hot Encoding

Proses one hot encoding merupakan proses untuk mengganti suatu nilai kategorik menjadi nilai numerik. Proses ini akan membuat kolom sejumlah dengan banyaknya kategori secara unik dan mengisinya dengan nilai 1 atau 0. Pada analisis ini, proses one hot encoding akan diterapkan pada tiga kolom yaitu *gender, occupation, dan customer_nw_category*. Proses ini menggunakan bantuan python dan library pandas.

3.5. Drop Unnecessary Columns

Tahapan ini dilakukan untuk menghapus kolom yang tidak digunakan untuk analisis selanjutnya. Pada analisis ini dilakukan beberpa penghapusan kolom yaitu kolom *customer_id, gender, occupation, customer_nw_category,* dan *current_month_balance. customer_id* dihapus karena tidak terkait dengan proses prediksi. Sementara itu *gender, occupation, customer_nw_category* dihapus karena telah melalui proses one hot encoding dan kolom *current_month_balance* dihapus karena memiliki korelasi yang sangat tinggi dengan kolom *current_balance* dengan nilai korelasi 0.96.