# “Eksplorasi Data Customer Churn Berdasarkan Segmentasi Pelanggan, Kualitas Layanan, dan Sensitivitas Harga pada Dataset Telekomunikasi”


**Oleh:**

**Kelompok 2**

Alya Dhamierea Islamay Putri (255150700111019)
Aisyah Berlianti Afriandi (255150700111022)
Salma Ninda Syahputri (255150701111024)
I Dewa Ayu Puspa Dewi (255150700111045)

**Dosen:**

**Bayu Rahayudi, S.T., M.T., M.M.**

**Program Studi Teknologi Informasi**

**Fakultas Ilmu Komputer**

**Universitas Brawijaya**


**2026**

# DAFTAR ISI

**DAFTAR ISI** ..... **2**
**BAB I** ..... **5**
**PENDAHULUAN** ..... **5**
1.1 Latar Belakang ..... 5
**1.2 Rumusan Masalah** ..... **5**
1.3 Tujuan Penelitian ..... 5
1.4 Manfaat Penelitian ..... 5
**BAB II** ..... **6**
**DASAR TEORI** ..... **6**
2.1 Konsep Dasar Keilmuan ..... 6
2.1.1 Data Science ..... 6
2.1.2 Machine Learning ..... 6
2.2 Konsep Domain Bisnis ..... 7
2.2.1 Customer Churn Prediction ..... 7
2.2.2 Customer Segmentation ..... 7
2.2.3 Kualitas Layanan & Sensitivitas Harga ..... 8
2.2.4 Customer Retention ..... 10
2.2.5 Customer Churn ..... 11
2.2.6 Pricing ..... 11
2.2.7 Customer Value ..... 12
2.2.8 customer lifetime value ..... 12
2.3 Metodologi Data & Analysis ..... 12
2.3.1 Data Preprocessing ..... 12
2.3.2 Exploratory Data Analysis (EDA) ..... 13
2.3.2 Data Visualization ..... 13
2.4 Teknologi ..... 14
2.4.1 Bahasa Pemrograman Python ..... 14
2.4.2 Library Pandas ..... 14
2.4.3 Library Numpy ..... 15
2.4.4 Library Matplotlib ..... 15
2.4.5 Library Seaborn ..... 16
**BAB III** ..... **17**
**METODOLOGI PENELITIAN** ..... **17**
3.1 Jenis Penelitian ..... 17
3.2 Dataset ..... 17
1. Pilar Analisis Segmentasi Aktivitas & Perilaku Pelanggan (Segmentation) ..... 18
2. Pilar Analisis Layanan dan Kendala Pelanggan (Service & Pain Points) ..... 19
3. Pilar Analisis Sensitivitas Harga dan Kondisi Finansial (Financial Sensitivity) ..... 19
3.3 Tahapan Penelitian ..... 20
3.4 Data Preprocessing ..... 21
3.5 General EDA ..... 24

3.6 Specific EDA..... 25
3.6.1 Segmentasi..... 25
3.6.2 Service Quality..... 28
3.6.2.1 Pengorganisasian Dataset..... 28
3.6.2.2 Statistik Deskriptif..... 29
3.6.2.3 Distribusi Variabel..... 30
3.6.2.4 Analisis Hubungan dengan Churn..... 31
3.6.2.5 Analisis Multivariat (Tabel Pivot)..... 31
3.6.2.6 Validasi Outlier..... 32
3.6.2.7 Matriks Korelasi..... 32
3.6.3 Analisis Finansial dan Sensitivitas Harga..... 32
**BAB IV..... 36**
**HASIL DAN PEMBAHASAN..... 36**
4.1 Hasil Data Preprocessing..... 36
4.1.1. Hasil Setelah Preprocessing..... 36
4.1.2 Perubahan Struktur Dataset (Kolom & Baris)..... 36
4.1.3 Rekayasa Fitur Baru (Feature Engineering)..... 37
a. Fitur Berbasis Durasi dan Finansial..... 37
b. Fitur Indikator Anomali Perilaku..... 37
c. Encoding Variabel Kategorikal..... 38
d. Ekstraksi Fitur Teks (Text Preprocessing)..... 38
4.1.4 Karakteristik Dataset Final..... 38
4.2 Hasil General EDA..... 38
4.3 Hasil Analisis Customer Segmentation..... 40
4.3.1 Behavioral Matrix Berbasis Median..... 41
4.4 Hasil Analisis Service Quality..... 42
4.4.1 Persiapan dan Validasi Dataset..... 42
4.4.2 Statistik Deskriptif Variabel Kunci..... 43
4.4.3 Distribusi Variabel Kunci..... 44
4.4.3.1 Distribusi Support Tickets..... 44
4.4.3.2 Distribusi Jenis Kontrak..... 45
4.4.3.3 Distribusi Kota..... 46
4.4.3.4 Distribusi Kategori Feedback..... 47
4.4.4 Analisis Hubungan Dengan Churn..... 47
4.4.4.1 Support Tickets vs Churn..... 48
4.4.4.2 Jenis Kontrak vs Churn..... 49
4.4.4.3 Kota vs Churn..... 49
4.4.5 Analisis Multivariat..... 50
4.4.6 Validasi Outlier..... 50
4.4.7 Korelasi Antar Variabel..... 51
4.4.8 Ringkasan Insight dan Rekomendasi..... 51
4.5 Hasil Analisis Financial Sensitivity..... 52
4.5.1 Hubungan Beban Finansial dengan Tingkat Pendidikan..... 52
4.5.2 Hubungan Beban Finansial dengan Usia Pelanggan..... 53

4.5.3 Pembagian Geografis yang Terdampak Risiko Tertinggi.....53
4.5.4 Rekomendasi Bisnis.....53
4.6 Insight Bisnis.....54
4.6.1 Segmentasi Pelanggan Mengungkap Fenomena Wasted Value.....54
4.6.2 Frekuensi Keluhan dan Jenis Kontrak Sebagai Sinyal Churn Terukur.....54
4.6.3 Tekanan Finansial Bersifat Universal Lintas Demografi.....55
4.6.4 Integrasi Insight: Efek Berlipat Antar Dimensi.....55
4.7 Rekomendasi Bisnis.....56
4.7.1 Rekomendasi Berdasarkan Analisis Segmentasi (Notebook A).....56
4.7.2 Rekomendasi Berdasarkan Analisis Kualitas Layanan (Notebook B).....56
4.7.3 Rekomendasi Berdasarkan Analisis Finansial (Notebook C).....57
4.7.4 Matriks Strategi Retensi Terintegrasi.....57
**BAB V.....59**
**PENUTUP.....59**
5.1 Kesimpulan.....59
5.2 Saran.....60
5.2.1 Saran untuk Pengembangan Proyek.....60
5.2.2 Saran untuk Perusahaan Telekomunikasi.....60
**DAFTAR PUSTAKA.....61**
**LAMPIRAN.....64**

# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Churn pelanggan merupakan tantangan penting dalam industri telekomunikasi karena tingginya persaingan antaroperator dan kemudahan pelanggan untuk berpindah layanan (*operator switching*) (Latief et al., 2021). Dari sisi perilaku pelanggan, segmentasi berdasarkan pola penggunaan layanan diperlukan untuk mengenali kelompok pelanggan yang rentan churn dan menyusun strategi retensi yang lebih tepat sasaran. Dari sisi kualitas layanan, masalah seperti kualitas panggilan yang buruk, jaringan tidak stabil, dan gangguan konektivitas internet diketahui menjadi pendorong utama churn (Ribeiro dkk., 2023), sementara kualitas layanan juga berpengaruh terhadap loyalitas melalui kepuasan pelanggan (Yum & Yoo, 2023). Di sisi lain, kondisi finansial pelanggan turut memengaruhi keputusan churn, terutama ketika beban tagihan meningkat di tengah tekanan ekonomi dan sensitivitas harga konsumen (Pakistan Bureau of Statistics, 2026; Khan dkk., 2019). Oleh karena itu, analisis churn pada proyek ini dilakukan melalui tiga perspektif utama, yaitu **segmentasi pelanggan**, **kualitas layanan**, dan **sensitivitas harga**, agar diperoleh gambaran yang lebih komprehensif mengenai faktor-faktor yang memengaruhi churn pelanggan telekomunikasi.

## 1.2 Rumusan Masalah

1. Bagaimana pola penggunaan layanan pelanggan dapat digunakan untuk mengidentifikasi segmen pelanggan yang rentan mengalami churn?
2. Bagaimana kualitas layanan dan pain points pelanggan berkontribusi terhadap keputusan churn?
3. Bagaimana kondisi finansial pelanggan memengaruhi sensitivitas harga dan keputusan churn?

## 1.3 Tujuan Penelitian

1. Mengidentifikasi segmen pelanggan berdasarkan pola penggunaan layanan dan menganalisis kecenderungan churn pada tiap segmen.
2. Menganalisis hubungan kualitas layanan dan pain points pelanggan terhadap churn.
3. Menganalisis pengaruh kondisi finansial pelanggan terhadap sensitivitas harga dan keputusan churn.

## 1.4 Manfaat Penelitian

1. Bagi Perusahaan (Manfaat Praktis)
   Membantu pihak manajemen telekomunikasi dan tim CRM dalam memetakan karakteristik serta perilaku penggunaan layanan pelanggan secara akurat, sehingga perusahaan dapat merancang strategi retensi yang lebih efektif dan personal untuk mencegah *customer churn*.
2. Bagi Kualitas Layanan & Operasional

Memberikan rekomendasi konkret mengenai aspek operasional apa saja yang menjadi kendala utama pelanggan (*pain points*) serta bagaimana peningkatan kualitas layanan dapat memitigasi keluhan pelanggan sebelum mereka memutuskan untuk *churn*.

3. Bagi Pengambilan Kebijakan Finansial
   Menjadi landasan strategis bagi perusahaan dalam menyusun kebijakan harga (pricing) dan promosi yang adaptif, dengan mempertimbangkan tingkat stabilitas finansial serta sensitivitas harga dari masing-masing segmen pelanggan.

# BAB II
# DASAR TEORI

## 2.1 Konsep Dasar Keilmuan

### 2.1.1 Data Science

Sains data adalah kumpulan teknik interdisipliner yang digunakan untuk mengekstraksi nilai atau wawasan bermakna dari data (Kotu dan Deshpande, 2019). Bidang ini menggabungkan aplikasi bisnis dari pembelajaran mesin (*machine learning*), statistika, matematika, dan visualisasi untuk menemukan pola dan hubungan tersembunyi guna mendukung pengambilan keputusan yang krusial (Kotu dan Deshpande, 2019). Dalam perspektif praktis, sains data dipandang sebagai sebuah proses berkelanjutan, bukan sekadar peristiwa tunggal, di mana data digunakan sebagai instrumen untuk memahami fenomena dunia nyata melalui validasi hipotesis dan model (Ratnawati, 2024). Secara sistematis, alur kerja sains data umumnya mengikuti enam tahapan utama, yang dimulai dari penetapan tujuan penelitian (*research goal*), pengumpulan data, persiapan data, eksplorasi data, pemodelan, hingga tahap presentasi dan otomatisasi hasil (Ratnawati, 2020).

### 2.1.2 Machine Learning

Pembelajaran mesin (*machine learning*) merupakan sub-bidang dari kecerdasan buatan (*artificial intelligence*) yang memberikan kemampuan pada sistem komputer untuk belajar dari pengalaman dalam bentuk data (Kotu dan Deshpande, 2019). Definisi klasik menyebutkan bahwa *machine learning* memungkinkan komputer untuk belajar tanpa harus diprogram secara eksplisit untuk setiap detail instruksinya (Ratnawati, 2020). Berbeda dengan pemrograman tradisional, algoritme ini mengambil data pelatihan (*training data*) untuk secara otomatis membangun model perwakilan yang dapat memetakan masukan (*input*) menjadi keluaran (*output*) yang diinginkan (Kotu dan Deshpande, 2019).

Secara metodologis, *machine learning* dibagi menjadi beberapa kategori utama:

1. Supervised Learning: Teknik yang membangun model dengan mendeskripsikan hubungan antara variabel independen dan variabel target yang sudah memiliki label sebelumnya, seperti pada tugas klasifikasi atau regresi (Provost dan Fawcett, 2013).
2. Unsupervised Learning: Teknik yang bertujuan menemukan pengelompokan alami atau regularitas dalam data tanpa didorong oleh variabel target tertentu, yang umumnya diimplementasikan dalam tugas klasterisasi (Provost dan Fawcett, 2013).

Dalam konteks sains data, pembelajaran mesin menjadi alat krusial untuk melakukan prediksi perilaku masa depan, seperti dalam kasus prediksi *churn* pelanggan atau analisis sensitivitas harga (Kotu dan Deshpande, 2019)

## 2.2 Konsep Domain Bisnis

### 2.2.1 Customer Churn Prediction

*Customer churn*, atau sering disebut sebagai *attrition*, didefinisikan sebagai fenomena di mana pelanggan berhenti berlangganan atau berpindah dari satu penyedia layanan ke penyedia layanan lainnya (Provost dan Fawcett, 2013). Di industri telekomunikasi yang pasarnya sudah jenuh, mempertahankan pelanggan yang sudah ada jauh lebih efisien secara biaya dibandingkan strategi akuisisi pelanggan baru (Provost dan Fawcett, 2013). Oleh karena itu, Churn Prediction menjadi instrumen krusial bagi perusahaan untuk mengidentifikasi individu yang memiliki risiko tinggi untuk meninggalkan layanan sebelum kontrak mereka berakhir (Kotu dan Deshpande, 2019).

Secara metodologis, prediksi churn umumnya dirumuskan sebagai masalah klasifikasi biner dalam *supervised learning* (Provost dan Fawcett, 2013). Proses ini melibatkan penggunaan data historis (seperti pola penggunaan, riwayat interaksi bantuan, dan profil tagihan) untuk membangun model yang mampu membedakan antara pelanggan yang bertahan (*stay*) dan pelanggan yang akan keluar (*leave*) (Kotu dan Deshpande, 2019). Output dari pemodelan ini berupa skor probabilitas, yang memungkinkan departemen pemasaran untuk melakukan intervensi proaktif, seperti menawarkan program loyalitas atau paket khusus untuk meminimalkan potensi kerugian finansial akibat hilangnya pelanggan bernilai tinggi (Kotu dan Deshpande, 2019; Provost dan Fawcett, 2013).

### 2.2.2 Customer Segmentation

*Customer segmentation* adalah proses krusial dalam sains data untuk mengeksplorasi dan menemukan kelompok-kelompok bermakna (klaster) di mana individu dalam satu kelompok memiliki kemiripan ciri yang tinggi, namun sangat berbeda dengan individu di kelompok lain (Kotu dan Deshpande, 2019). Di industri telekomunikasi, segmentasi dilakukan untuk mengidentifikasi "pelanggan prototipe" (deskripsi pelanggan tipikal) berdasarkan pola penggunaan produk (*usage patterns*) agar strategi retensi dan pesan pemasaran dapat dipersonalisasi secara efektif (Kotu dan Deshpande, 2019).

**Pendekatan Sains Data dalam Segmentasi**

Secara metodologis, segmentasi dalam sains data terbagi menjadi dua pendekatan utama:

1. **Unsupervised Segmentation (Clustering):** Menemukan pengelompokan alami dalam data tanpa didorong oleh variabel target tertentu untuk memahami struktur dasar populasi (Provost dan Fawcett, 2013).
2. **Supervised Segmentation:** Membagi populasi berdasarkan karakteristik target spesifik yang ingin diprediksi, seperti risiko pembatalan layanan atau *churn* (Provost dan Fawcett, 2013).

**Matriks Perilaku dan Ukuran Tendensi Sentral**

Dalam membangun **Behavioral Matrix** (Matriks Perilaku), variabel numerik seperti durasi panggilan, penggunaan data, dan beban tagihan digunakan untuk membedakan interaksi pelanggan secara aktual (Kotu dan Deshpande, 2019). Pembagian kuadran dalam matriks ini memerlukan ambang batas (*threshold*) yang ditentukan melalui ukuran tendensi sentral.

Penggunaan nilai **Median** sangat direkomendasikan sebagai titik potong dibandingkan rata-rata (*mean*) karena sifatnya yang **tangguh (*robust*)** (Ratnawati, 2019). Median tidak sensitif terhadap data pencilan (*outlier*), sehingga mampu memberikan representasi titik pusat yang lebih akurat pada distribusi data yang miring atau tidak merata (Ratnawati, 2019).

**Package Mismatch dan Hubungannya dengan Churn**

Pemetaan pelanggan ke dalam persona (seperti *Internet Enthusiast* atau *Power User*) sangat penting untuk mendeteksi fenomena ***package mismatch*** atau ***wasted value*** (Provost dan Fawcett, 2013). Kondisi ini terjadi ketika pelanggan membayar fasilitas yang tidak mereka gunakan secara maksimal, yang sering kali menjadi pemicu utama ketidakpuasan dan keputusan pelanggan untuk berhenti berlangganan atau melakukan ***churn*** (Provost dan Fawcett, 2013)

### 2.2.3 Kualitas Layanan & Sensitivitas Harga

Kualitas layanan merupakan faktor determinan dalam retensi pelanggan yang diukur melalui berbagai sinyal interaksi, seperti jumlah tiket bantuan (*support tickets*) dan kategori umpan balik (*feedback*) pelanggan (Kotu dan Deshpande, 2019). Dalam perspektif sains data, pemantauan terhadap indikator kualitas ini sangat penting untuk mengidentifikasi titik kritis (*pain points*) yang dapat memicu ketidakpuasan sebelum pelanggan memutuskan untuk berhenti berlangganan (Ratnawati, 2024; Provost dan Fawcett, 2013). Analisis yang mendalam terhadap pola data interaksi membantu perusahaan menemukan efisiensi operasional dan meningkatkan nilai layanan yang ditawarkan (Ratnawati, 2024).

Sensitivitas harga mengacu pada sejauh mana beban finansial, yang direpresentasikan oleh tagihan bulanan (*monthly bill*), memengaruhi loyalitas pelanggan (Provost dan Fawcett, 2013). Di industri telekomunikasi, fenomena *wasted value* terjadi saat pelanggan merasa membayar mahal untuk fasilitas yang tidak mereka gunakan secara optimal, yang secara signifikan meningkatkan sensitivitas mereka terhadap harga (Kotu dan Deshpande, 2019). Teori ekonomi data menunjukkan bahwa pelanggan yang memiliki loyalitas rendah cenderung sangat sensitif terhadap fluktuasi harga atau beban tagihan yang tinggi (Provost dan Fawcett, 2013). Oleh karena itu, pemahaman terhadap elastisitas harga pada setiap segmen memungkinkan perusahaan untuk mengoptimalkan strategi penentuan harga (*pricing strategy*) guna meminimalkan angka *churn* dan memaksimalkan nilai seumur hidup pelanggan (Provost dan Fawcett, 2013; Kotu dan Deshpande, 2019).

### 2.3.3 Customer Satisfication

Customer satisfaction atau kepuasan pelanggan merupakan salah satu konsep utama dalam pemasaran yang menggambarkan tingkat kesesuaian antara harapan pelanggan sebelum menggunakan suatu produk atau layanan dengan pengalaman aktual yang diperoleh setelah penggunaan. Apabila kinerja layanan memenuhi atau bahkan melampaui harapan pelanggan, maka pelanggan akan merasa puas. Sebaliknya, apabila kinerja layanan berada di bawah harapan, pelanggan akan mengalami ketidakpuasan yang berpotensi memengaruhi perilaku mereka terhadap perusahaan, seperti mengurangi penggunaan layanan atau berpindah ke penyedia lain. Konsep ini telah menjadi salah satu indikator penting dalam mengevaluasi kualitas layanan serta keberhasilan perusahaan dalam mempertahankan pelanggan.

Menurut Kotler dan Keller (2016), customer satisfaction adalah perasaan senang atau kecewa yang muncul setelah pelanggan membandingkan persepsi terhadap kinerja suatu produk atau layanan dengan harapan yang dimilikinya. Walaupun definisi tersebut merupakan teori klasik yang masih banyak dijadikan acuan, penelitian terbaru menunjukkan bahwa customer satisfaction tidak lagi hanya dipengaruhi oleh kualitas produk, tetapi juga dipengaruhi oleh pengalaman pelanggan secara menyeluruh (customer experience), kemudahan memperoleh layanan, responsivitas perusahaan, kualitas komunikasi, hingga kemampuan perusahaan dalam menyelesaikan keluhan pelanggan.

Meta-analisis yang dilakukan oleh Vikas Mittal dan rekan-rekannya menganalisis 245 penelitian dengan lebih dari satu juta responden dan menemukan bahwa customer satisfaction memiliki hubungan positif yang signifikan terhadap loyalitas pelanggan, retensi pelanggan, pembelian ulang (repurchase intention), komunikasi positif dari pelanggan (word of mouth), peningkatan pengeluaran pelanggan, serta kinerja keuangan perusahaan. Temuan tersebut memperlihatkan bahwa perusahaan yang berhasil menjaga kepuasan pelanggan memiliki peluang lebih besar untuk mempertahankan pelanggan dalam jangka panjang sekaligus meningkatkan profitabilitas perusahaan.

Dalam industri telekomunikasi, kepuasan pelanggan menjadi salah satu faktor yang paling menentukan keberhasilan perusahaan dalam mempertahankan pelanggan. Pelanggan umumnya mengevaluasi kualitas jaringan, kestabilan koneksi internet, kualitas panggilan telepon, kecepatan penanganan komplain, transparansi sistem penagihan, serta pelayanan customer service sebelum memutuskan untuk tetap menggunakan layanan atau berpindah ke operator lain. Oleh karena itu, perusahaan telekomunikasi harus mampu menjaga kualitas layanan secara konsisten agar tingkat kepuasan pelanggan tetap tinggi.

Pada penelitian ini, customer satisfaction menjadi landasan teori dalam menganalisis kualitas layanan perusahaan telekomunikasi. Variabel seperti **support_tickets**, **customer_feedback**, dan **contract_type** digunakan untuk menggambarkan pengalaman pelanggan terhadap layanan perusahaan. Banyaknya keluhan pelanggan maupun tingginya frekuensi tiket dukungan diperkirakan menunjukkan tingkat kepuasan yang lebih rendah sehingga dapat meningkatkan kemungkinan pelanggan melakukan churn.

### 2.2.4 Customer Retention

Customer retention merupakan kemampuan perusahaan dalam mempertahankan pelanggan agar tetap menggunakan produk atau layanan dalam jangka waktu yang panjang. Berbeda dengan customer acquisition yang berfokus pada memperoleh pelanggan baru, customer retention lebih menitikberatkan pada upaya membangun hubungan jangka panjang sehingga pelanggan tetap loyal terhadap perusahaan. Strategi retensi pelanggan menjadi semakin penting karena mempertahankan pelanggan yang sudah ada umumnya membutuhkan biaya yang lebih rendah dibandingkan memperoleh pelanggan baru.

Customer retention dipengaruhi oleh berbagai faktor, antara lain kepuasan pelanggan, kualitas layanan, kepercayaan pelanggan, komunikasi perusahaan, pengalaman pelanggan, serta hubungan emosional yang terbentuk antara pelanggan dengan perusahaan. Penelitian systematic literature review oleh Albérico Rosário menunjukkan bahwa relationship marketing berperan penting dalam meningkatkan customer retention melalui peningkatan kepuasan pelanggan, komitmen, kepercayaan, dan kualitas komunikasi antara perusahaan dengan pelanggan.

Penelitian lain menunjukkan bahwa pengalaman pelanggan (customer experience) yang positif memiliki pengaruh yang signifikan terhadap retensi pelanggan. Selain itu, perusahaan yang mampu menciptakan hubungan emosional dengan pelanggan memiliki kemungkinan lebih besar untuk mempertahankan pelanggan dibandingkan perusahaan yang hanya mengandalkan kontrak atau biaya perpindahan (switching cost). Dengan demikian, customer retention tidak hanya dipengaruhi oleh faktor ekonomi, tetapi juga dipengaruhi oleh pengalaman dan hubungan yang dibangun selama pelanggan menggunakan layanan perusahaan.

Dalam industri telekomunikasi, customer retention menjadi salah satu indikator utama keberhasilan perusahaan karena tingkat persaingan yang tinggi membuat pelanggan dapat dengan mudah berpindah ke operator lain. Oleh sebab itu, perusahaan perlu memahami faktor-faktor yang menyebabkan pelanggan tetap bertahan maupun meninggalkan perusahaan sehingga strategi retensi dapat disusun secara lebih efektif.

Pada penelitian ini, customer retention menjadi dasar dalam mengevaluasi bagaimana kualitas layanan, jumlah komplain pelanggan, jenis kontrak, serta umpan balik pelanggan dapat memengaruhi keputusan pelanggan untuk tetap menggunakan layanan perusahaan. Analisis tersebut diharapkan dapat memberikan rekomendasi bagi perusahaan dalam meningkatkan strategi retensi pelanggan dan menekan tingkat customer churn.

### 2.2.5 Customer Churn

Customer churn merupakan kondisi ketika pelanggan memutuskan untuk menghentikan penggunaan suatu produk atau layanan dalam periode tertentu. Dalam industri telekomunikasi, churn biasanya ditandai dengan pelanggan yang menutup layanan, berpindah ke operator lain, atau tidak lagi melakukan penggunaan layanan dalam jangka waktu tertentu.

Tingginya tingkat churn menjadi salah satu tantangan utama perusahaan karena berdampak langsung terhadap penurunan pendapatan, meningkatnya biaya akuisisi pelanggan baru, serta berkurangnya profitabilitas perusahaan.

Systematic literature review yang dilakukan oleh Hugo Ribeiro mengidentifikasi bahwa customer satisfaction, service quality, switching cost, usia pelanggan, jenis kelamin, serta berbagai karakteristik perilaku pelanggan merupakan determinan utama yang memengaruhi customer churn pada industri telekomunikasi. Penelitian tersebut juga menjelaskan bahwa churn dipengaruhi oleh kombinasi faktor perilaku pelanggan dan karakteristik yang tersimpan dalam sistem manajemen pelanggan perusahaan, sehingga analisis data menjadi sangat penting dalam memahami penyebab pelanggan berpindah layanan.

Perkembangan teknologi data science dan machine learning memungkinkan perusahaan melakukan prediksi customer churn secara lebih akurat. Berbagai penelitian menunjukkan bahwa analisis terhadap pola penggunaan internet, durasi panggilan, besarnya tagihan bulanan, jumlah tiket dukungan, maupun karakteristik pelanggan mampu membantu perusahaan mengidentifikasi pelanggan yang memiliki risiko tinggi untuk melakukan churn sehingga tindakan retensi dapat dilakukan lebih awal.

Dalam konteks penelitian ini, customer churn menjadi variabel target yang dianalisis menggunakan pendekatan exploratory data analysis (EDA). Variabel-variabel seperti **support_tickets**, **customer_feedback**, **contract_type**, dan **city** dianalisis untuk mengetahui hubungan antara kualitas layanan dengan kemungkinan pelanggan melakukan churn. Hasil analisis diharapkan dapat menjadi dasar dalam memberikan rekomendasi strategis kepada perusahaan telekomunikasi untuk meningkatkan kualitas layanan, memperkuat retensi pelanggan, serta menurunkan tingkat customer churn.

### 2.2.6 Pricing

Pricing atau penetapan harga menentukan pendapatan (*revenue*) dan profitabilitas suatu perusahaan. Harga bukan sekadar angka nominal yang harus dibayar oleh konsumen, melainkan representasi nilai yang dipertukarkan untuk mendapatkan suatu manfaat layanan. Harga berfungsi sebagai indikator kualitas dan beban keuangan dalam pandangan konsumen. Ketika situasi makroekonomi suatu negara bergejolak, seperti yang terjadi di Pakistan ketika tingkat inflasi mencapai 11,7%, daya beli riil masyarakat akan terganggu. Kondisi ini menyebabkan sensitivitas harga yang lebih tinggi. Sensitivitas harga mengacu pada seberapa besar pengaruh perubahan harga atau tarif pengeluaran pada keputusan pembelian pelanggan. Pelanggan yang sangat sensitif terhadap harga akan sangat sensitif terhadap kenaikan biaya bulanan yang lebih besar daripada rasio bill-to-income mereka. Akibatnya, penyesuaian tarif yang tidak rasional akan menyebabkan pelanggan berhenti berlangganan dengan cepat.

### 2.2.7 Customer Value

Nilai pelanggan, juga dikenal sebagai "nilai pelanggan", didefinisikan sebagai penilaian menyeluruh konsumen tentang seberapa bermanfaat atau bermanfaat suatu produk atau layanan berdasarkan persepsi mereka tentang keuntungan yang mereka terima (manfaat) dibandingkan dengan biaya yang mereka korbankan (biaya). Secara matematis, nilai pelanggan dirumuskan sebagai berikut:

$$\text{Customer Value} = \frac{\text{Perceived Benefits}}{\text{Perceived Costs}}$$

### 2.2.8 Customer Lifetime Value

Jumlah nilai tunai atau proyeksi keuntungan bersih yang akan dihasilkan oleh seorang pelanggan untuk perusahaan selama jangka waktu hubungan kontraktual (jangka waktu hubungan). Pelanggan dilihat oleh CLV sebagai aset jangka panjang, bukan komoditas transaksi sekali beli. Nilai CLV dalam industri telekomunikasi yang berbasis langganan (yang dikenal sebagai "industri berbasis langganan") sangat bergantung pada seberapa baik suatu perusahaan mempertahankan pelanggan dan mengurangi laju kehilangan pelanggan. Formula dasar untuk menemukan nilai CLV biasanya terdiri dari komponen berikut:

CLV = Average Revenue Per User x Customer Lifetime - Customer Acquisition Cost

## 2.3 Metodologi Data & Analysis

### 2.3.1 Data Preprocessing

Pra-pemprosesan data (*data preprocessing*) adalah fase krusial dalam siklus sains data yang bertujuan untuk menyaring dan menyiapkan data mentah agar siap digunakan dalam tahap pemodelan (Kotu dan Deshpande, 2019). Data dalam dunia nyata sering kali bersifat "kotor", yang ditandai dengan ketidaklengkapan (*missing values*), adanya gangguan (*noisy data*), serta ketidakkonsistenan format (Ratnawati, 2020). Kegagalan dalam menangani kualitas data ini akan berdampak langsung pada akurasi model, sesuai dengan prinsip *garbage in, garbage out* (Provost dan Fawcett, 2013).

Tugas utama dalam pemrosesan awal meliputi empat tahap (Ratnawati, 2020; Kotu dan Deshpande, 2019):

1. Pembersihan Data (*Data Cleansing*): Menangani nilai yang hilang, menghaluskan data yang bising, serta mengidentifikasi dan menangani pencilan (*outlier*).
2. Integrasi Data: Menggabungkan data dari berbagai sumber atau basis data yang berbeda menjadi satu kesatuan.

3. Transformasi Data: Melakukan normalisasi, standarisasi, atau pengkodean (*encoding*) variabel kategori menjadi numerik agar sesuai dengan kebutuhan algoritma.
4. Reduksi Data: Mengurangi volume data tanpa menghilangkan informasi penting guna meningkatkan efisiensi komputasi.

### 2.3.2 Exploratory Data Analysis (EDA)

*Exploratory Data Analysis* (EDA) adalah pendekatan analisis data yang bertujuan untuk memahami karakteristik dasar, struktur, dan pola tersembunyi di dalam dataset (Kotu dan Deshpande, 2019). Proses ini melibatkan penggunaan statistik deskriptif dan teknik visualisasi untuk mendeteksi anomali, menguji hipotesis, dan memeriksa asumsi sebelum dilakukan analisis yang lebih kompleks (Provost dan Fawcett, 2013). Secara praktis, EDA membantu peneliti menjawab pertanyaan mengenai nilai tipikal suatu atribut (tendensi sentral) dan tingkat persebarannya (*variabilitas*) (Kotu dan Deshpande, 2019).

Alur kerja EDA umumnya mengikuti sebuah Roadmap sembilan langkah, yang mencakup pengorganisasian data, penghitungan titik pusat (median/mean), analisis variabilitas, visualisasi distribusi, hingga pemahaman hubungan antar-atribut melalui matriks korelasi (Ratnawati, 2025; Kotu dan Deshpande, 2019). Melalui EDA, peneliti dapat memvalidasi apakah data telah memenuhi persyaratan untuk menjawab tujuan bisnis yang telah ditetapkan di awal proyek (Ratnawati, 2024).

### 2.3.2 Data Visualization

Visualisasi data didefinisikan sebagai metode untuk mengekspresikan data dalam bentuk visual abstrak guna mempermudah pemahaman terhadap hubungan antar-atribut yang kompleks (Kotu dan Deshpande, 2019). Penglihatan manusia memiliki kemampuan kognitif yang kuat untuk mengenali pola dan tren dengan cepat ketika informasi disajikan secara visual dibandingkan dalam bentuk tabel numerik mentah (Kotu dan Deshpande, 2019).

Beberapa teknik visualisasi utama yang umum digunakan dalam sains data meliputi (Ratnawati, 2020; Kotu dan Deshpande, 2019):

- Histogram: Digunakan untuk menunjukkan frekuensi kemunculan nilai dan bentuk distribusi suatu variabel tunggal (*univariate*).
- Boxplot: Berguna untuk melihat ringkasan statistik (median, kuartil) serta mendeteksi keberadaan *outlier* secara visual.
- Scatterplot: Digunakan untuk mengeksplorasi hubungan atau korelasi antara dua variabel numerik.

Pentingnya visualisasi ditegaskan melalui fenomena Anscombe's Quartet, yang membuktikan bahwa kumpulan data dengan properti statistik deskriptif yang identik dapat memiliki distribusi yang sangat berbeda ketika ditampilkan dalam bentuk grafik (Kotu dan Deshpande, 2019). Oleh karena itu, visualisasi merupakan alat eksplorasi yang tidak terpisahkan dalam memastikan integritas analisis data.

## 2.4 Teknologi

### 2.4.1 Bahasa Pemrograman Python

Python merupakan bahasa pemrograman tingkat tinggi (*high-level programming language*) yang dikembangkan dengan tujuan memberikan sintaks yang sederhana, mudah dipahami, serta mendukung berbagai paradigma pemrograman, seperti *procedural programming*, *object-oriented programming*, dan *functional programming*. Python banyak digunakan dalam berbagai bidang, termasuk pengembangan perangkat lunak, otomatisasi sistem, kecerdasan buatan (*Artificial Intelligence*), *Machine Learning*, hingga *Data Science* karena memiliki ekosistem pustaka (*library*) yang sangat lengkap dan komunitas pengguna yang besar (Python Software Foundation, 2024).

Dalam bidang *Data Science*, Python menjadi salah satu bahasa pemrograman yang paling populer karena mampu mengolah data dalam jumlah besar secara efisien serta menyediakan berbagai pustaka yang mendukung proses analisis data, visualisasi, hingga pembangunan model *Machine Learning*. Menurut Van Rossum dan Drake (2021), keunggulan Python terletak pada keterbacaan sintaks (*readability*), kemudahan integrasi dengan berbagai teknologi, serta fleksibilitasnya dalam menangani berbagai kebutuhan komputasi ilmiah.

Pada penelitian ini, Python digunakan sebagai bahasa pemrograman utama untuk melakukan seluruh proses pengolahan data, mulai dari pembacaan dataset (*data loading*), *data preprocessing*, *feature engineering*, *Exploratory Data Analysis (EDA)*, visualisasi data, hingga analisis hubungan antar variabel. Penggunaan Python memungkinkan seluruh tahapan analisis dilakukan secara terstruktur, terdokumentasi dengan baik, serta mudah direproduksi (*reproducible research*), sehingga mendukung kualitas penelitian berbasis data (Python Software Foundation, 2024).

### 2.4.2 Library Pandas

Pandas merupakan pustaka (library) Python yang dirancang khusus untuk melakukan manipulasi dan analisis data terstruktur. Pandas menyediakan dua struktur data utama, yaitu Series untuk data satu dimensi dan DataFrame untuk data dua dimensi yang menyerupai tabel pada basis data maupun spreadsheet. Struktur tersebut memungkinkan pengguna melakukan proses pengolahan data secara efisien, seperti membaca data, membersihkan data, melakukan transformasi, agregasi, penggabungan (merge), hingga analisis statistik deskriptif (McKinney, 2022).

Dalam proses Data Science, Pandas menjadi salah satu library yang paling penting karena hampir seluruh tahapan awal analisis data dilakukan menggunakan DataFrame. Berbagai fungsi seperti read_csv(), info(), describe(), groupby(), merge(), dan pivot_table() memungkinkan proses eksplorasi data dilakukan secara sistematis dan efisien (Pandas Development Team, 2025).

Pada penelitian ini, Pandas digunakan untuk memuat dataset pelanggan telekomunikasi, melakukan pemeriksaan struktur data, mendeteksi nilai yang hilang (missing values), menghapus data duplikat, melakukan feature engineering, mengelompokkan data berdasarkan kategori tertentu, hingga menghasilkan tabel ringkasan yang digunakan dalam proses Exploratory Data Analysis (EDA).

### 2.4.3 Library Numpy

NumPy (Numerical Python) merupakan pustaka komputasi numerik yang menyediakan struktur data berupa ndarray (N-dimensional array) untuk menyimpan dan memproses data numerik secara efisien. NumPy dirancang agar mampu melakukan operasi matematika dan aljabar linear dengan performa tinggi dibandingkan struktur data bawaan Python karena sebagian besar implementasinya menggunakan bahasa C (Harris et al., 2020).

Selain menyediakan operasi matematika dasar, NumPy juga mendukung berbagai fungsi statistik, manipulasi matriks, pembangkitan bilangan acak (random number generation), serta operasi logika yang banyak digunakan dalam analisis data dan Machine Learning. Oleh karena itu, sebagian besar library analisis data di Python, termasuk Pandas, Scikit-learn, dan Matplotlib, dibangun di atas NumPy (NumPy Developers, 2025).

Dalam penelitian ini, NumPy digunakan untuk mendukung proses perhitungan numerik selama tahap preprocessing, seperti perhitungan median, kuartil, Interquartile Range (IQR), penanganan outlier, serta berbagai operasi matematika yang diperlukan dalam proses feature engineering.

### 2.4.4 Library Matplotlib

Matplotlib merupakan library visualisasi data dua dimensi yang paling banyak digunakan dalam ekosistem Python. Library ini menyediakan berbagai fungsi untuk menghasilkan grafik seperti line chart, bar chart, scatter plot, histogram, pie chart, hingga box plot. Visualisasi tersebut membantu peneliti memahami pola distribusi data, mengidentifikasi hubungan antar variabel, serta mendeteksi keberadaan outlier secara visual (Hunter, 2007).

Salah satu keunggulan Matplotlib adalah fleksibilitasnya dalam mengatur hampir seluruh aspek visualisasi, seperti warna, ukuran gambar, label sumbu, judul grafik, legenda, hingga anotasi. Kemampuan tersebut menjadikan Matplotlib sebagai fondasi utama berbagai library visualisasi lain di Python, termasuk Seaborn (Matplotlib Development Team, 2025).

Pada penelitian ini, Matplotlib digunakan untuk membuat berbagai visualisasi dasar selama proses Exploratory Data Analysis, seperti histogram, box plot, bar chart, dan pie chart guna memberikan gambaran mengenai karakteristik pelanggan serta distribusi variabel yang digunakan dalam analisis.

### 2.4.5 Library Seaborn

Seaborn merupakan library visualisasi statistik berbasis Matplotlib yang dikembangkan untuk menghasilkan grafik dengan tampilan yang lebih informatif dan estetis. Seaborn menyediakan fungsi visualisasi tingkat tinggi (high-level interface) yang memudahkan pengguna dalam membuat grafik statistik tanpa perlu melakukan banyak konfigurasi secara manual (Waskom, 2021).

Library ini mendukung berbagai jenis visualisasi yang umum digunakan dalam analisis data, seperti countplot, barplot, boxplot, violin plot, pairplot, heatmap, dan scatterplot. Selain itu, Seaborn memiliki kemampuan untuk secara otomatis menghitung agregasi statistik, seperti rata-rata dan interval kepercayaan (confidence interval), sehingga mempermudah proses eksplorasi data (Seaborn Development Team, 2025).

Dalam penelitian ini, Seaborn digunakan untuk menghasilkan berbagai visualisasi statistik pada tahap General Exploratory Data Analysis (EDA) maupun Specific EDA. Visualisasi tersebut meliputi distribusi pelanggan berdasarkan karakteristik demografi, hubungan antara variabel terhadap customer churn, analisis support tickets, jenis kontrak pelanggan, hingga heatmap korelasi antar variabel. Penggunaan Seaborn membantu proses interpretasi data menjadi lebih mudah dipahami dan mendukung penyusunan rekomendasi bisnis berdasarkan hasil analisis.

# BAB III
# METODOLOGI PENELITIAN

## 3.1 Jenis Penelitian

Penelitian ini menggunakan jenis penelitian kuantitatif berbasis eksperimental terapan di bidang data sains. Eksperimen difokuskan pada pemodelan prediktif klasifikasi biner, di mana algoritma dilatih untuk mengenali pola perilaku historis, atribut demografi, kondisi finansial, tingkat pemanfaatan layanan, serta sentimen tertulis dari pelanggan guna mengklasifikasikan status akhir pelanggan ke dalam dua kelas diskrit: apakah pelanggan akan mempertahankan layanan (retained) atau berhenti berlangganan (churn).

## 3.2 Dataset

Dataset yang melandasi seluruh rangkaian eksperimen ini bertajuk Telecom Customer Churn Feature Engineering Dataset yang diperoleh dari platform kompetisi data sains Kaggle dan dimuat secara langsung melalui repositori GitHub resmi Kelompok 2. Dataset mentah ini mencakup total basis data sebanyak **1.200 baris observasi** dengan **15 kolom atribut awal**. Karakteristik data merupakan kombinasi dari tipe numerik kontinu, numerik diskrit, kategorikal nominal, kategorikal ordinal, serta data tekstual tidak terstruktur (*unstructured text*).

Tabel 3.1 Struktur Spesifikasi Atribut Dataset Awal

| No | Nama Fitur | Tipe Data | Deskripsi Akademis & Fungsional |
|---|---|---|---|
| 1 | **customer_id** | Integer / Object | Identitas unik numerik setiap individu pelanggan (rentang 10001–11200). Berfungsi murni sebagai penanda (*identifier*) dan tidak memiliki nilai prediktif sejalan prinsip kausalitas, sehingga dieksklusi dalam pemodelan. |
| 2 | **signup_date** | String / Date | Tanggal pencatatan registrasi awal pelanggan (rentang waktu 1 Januari 2019 hingga 14 April 2022). Digunakan sebagai variabel basis untuk menghitung durasi retensi lewat rekayasa fitur temporal. |
| 3 | **age** | Integer | Representasi usia biologis pelanggan dalam satuan tahun (rentang 18–69 tahun, rata-rata 43,5 tahun). Berguna untuk pemetaan segmentasi demografi. |
| 4 | **gender** | Categorical | Jenis kelamin pelanggan dengan variasi nilai binomial (*Male* dan *Female*). |
| 5 | **city** | Categorical | Variabel spasial yang menunjukkan kota domisili operasional pelanggan (contoh: Multan, Peshawar, Gujranwala, Lahore, Hyderabad). |

| No | Nama Fitur | Tipe Data | Deskripsi Akademis & Fungsional |
|---|---|---|---|
| 6 | **education_level** | Categorical | Tingkat pendidikan formal terakhir pelanggan (*Primary*, *Secondary*, *Bachelor*, *Master*) yang mencerminkan profil latar belakang sosial. |
| 7 | **employment_status** | Categorical | Status kekaryaan atau okupasi aktif dari pelanggan (*Employed*, *Self-employed*, *Student*, *Unemployed*). |
| 8 | **monthly_income** | Float | Besaran pendapatan finansial bulanan riil dari pelanggan. Berfungsi sebagai proksi daya beli dan kapasitas ekonomi. |
| 9 | **monthly_bill** | Float | Total nilai tagihan finansial yang dibebankan atas pemakaian layanan telekomunikasi setiap bulannya. |
| 10 | **internet_usage_gb** | Float | Volume konsumsi kuota data internet pelanggan dalam satuan Gigabyte (GB) sepanjang siklus penagihan berjalan. |
| 11 | **call_minutes** | Float | Akumulasi durasi komunikasi suara (panggilan telepon) yang dihabiskan pelanggan dalam satuan menit. |
| 12 | **contract_type** | Categorical | Metode ikatan kontrak hukum penagihan layanan yang dipilih (*Monthly*, *6-Month*, *12-Month*). |
| 13 | **support_tickets** | Integer | Kuantitas berkas keluhan resmi atau tiket aduan yang diajukan pelanggan ke bagian pusat bantuan pelanggan (*customer support*). |
| 14 | **customer_feedback** | Text | Ulasan kualitatif naratif teks bebas yang ditulis langsung oleh pelanggan mengenai impresi atau kendala operasional layanan. |
| 15 | **churn** | Binary | **Variabel Target** (Label Klasifikasi) di mana nilai 1 merepresentasikan kejadian pelanggan berhenti berlangganan (*Churn*), sedangkan nilai 0 merepresentasikan pelanggan tetap aktif (*Retained*). |

1. Pilar Analisis Segmentasi Aktivitas & Perilaku Pelanggan (*Segmentation*)

Pilar ini memetakan karakteristik perilaku pelanggan berdasarkan tingkat penggunaan atau konsumsi riil terhadap layanan.

Fokus Atribut:

- internet_usage_gb_capped: Volume konsumsi kuota data internet (setelah pembersihan *outlier*).
- call_minutes_capped: Akumulasi durasi panggilan telepon (setelah pembersihan *outlier*).
- monthly_bill_capped: Total nilai tagihan finansial bulanan (setelah pembersihan *outlier*).

- churn: Variabel target biner (indikator kehilangan atau bertahannya pelanggan).

2. Pilar Analisis Layanan dan Kendala Pelanggan (*Service & Pain Points*)

Pilar ini bertujuan untuk mengidentifikasi tingkat kepuasan pelanggan, hambatan operasional, aspek kontrak, serta titik kritis (*pain points*) pemicu *churn*.

Fokus Atribut:

- support_tickets: Frekuensi aduan resmi pelanggan sebagai indikator kendala operasional (*pain points*).
- contract_type: Bentuk ikatan komitmen waktu pelanggan (*Monthly, 6-Month, 12-Month*).
- city: Lokasi geografis pelanggan untuk mendeteksi kendala layanan berbasis regional.
- customer_feedback: Ulasan naratif jujur pelanggan untuk menangkap kata kunci kritis masalah.
- churn: Variabel target biner (indikator kehilangan atau bertahannya pelanggan).

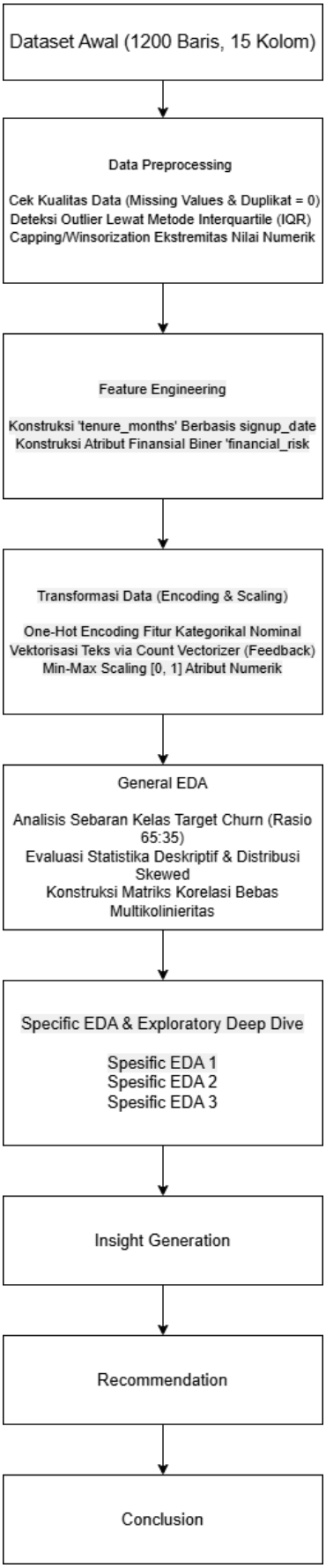
3. Pilar Analisis Sensitivitas Harga dan Kondisi Finansial (*Financial Sensitivity*)

Pilar ini dirancang untuk menggali profil ekonomi pelanggan serta sensitivitas mereka terhadap struktur tarif layanan di berbagai wilayah geografis.

Fokus Atribut:

- monthly_bill_capped: Beban pengeluaran tagihan bulanan.
- monthly_income_scaled: Kapasitas pendapatan bulanan pelanggan.
- financial_risk: Indikator biner tekanan finansial jika tagihan melebihi pendapatan.
- age_scaled: Korelasi rentang usia dengan kestabilan finansial.
- city_Faisalabad, city_Gujranwala, city_Hyderabad, city_Islamabad, city_Karachi, city_Lahore, city_Multan, city_Peshawar, city_Quetta, city_Rawalpindi, dan city_Sialkot: Atribut spasial kota hasil *One-Hot Encoding*.
- churn: Variabel target biner

## 3.3 Tahapan Penelitian



### 3.4 Data Preprocessing

*Preprocessing* data difungsikan untuk mentransformasikan data mentah yang rawan distorsi menjadi representasi yang optimal bagi model komputer. Berdasarkan implementasi pemrograman pada repositori notebook kelompok, rincian operasional yang diterapkan mencakup:

a. **Missing Value Audit**

Pengecekan menyeluruh dijalankan menggunakan fungsi pemrograman terintegrasi untuk mendeteksi keberadaan nilai kosong di seluruh baris data. Hasil audit mengonfirmasi bahwa dataset ini bersih sempurna dari data kosong (0 baris *missing values*), menghilangkan kebutuhan untuk imputasi statistik (seperti rata-rata atau modus).

```
missing = df.isnull().sum()
missing_pct = (missing / len(df) * 100).round(2)
missing_df = pd.DataFrame({
    'Jumlah Missing': missing,
    'Persentase (%)': missing_pct
})

print("Pemeriksaan Missing Values per Kolom:")
print("-" * 45)
if missing_df['Jumlah Missing'].sum() == 0:
    print("Tidak ditemukan missing values pada seluruh kolom dataset.")
    print("Dataset dalam kondisi lengkap dan tidak memerlukan imputasi.")
else:
    display(missing_df[missing_df['Jumlah Missing'] > 0].sort_values('Jumlah Missing', ascending=False))
```

b. **Duplicate Record Check**

Evaluasi baris berulang dijalankan untuk mengeliminasi potensi bias replikasi sampel. Hasil pengujian menunjukkan tidak terdapat data duplikat sama sekali (0 baris duplikat), menegaskan keselarasan data observasi secara individual. Berikut kodenya:

```
n_dup = df.duplicated().sum()
print(f"Jumlah baris duplikat: {n_dup}")

if n_dup > 0:
    df.drop_duplicates(inplace=True)
    print(f"{n_dup} baris duplikat berhasil dihapus.")
    print(f"Shape setelah pembersihan: {df.shape}")
else:
    print("Tidak ditemukan baris duplikat. Tidak ada tindakan yang diperlukan.")

print(f"Jumlah baris setelah pengecekan duplikat: {len(df)}")
```

```
Jumlah baris duplikat: 0
Tidak ditemukan baris duplikat. Tidak ada tindakan yang diperlukan.
Jumlah baris setelah pengecekan duplikat: 1200
```

c. **Penanganan Outlier Eksstrem (Capping/Winsorization)**

Analisis distribusi menunjukkan adanya ekor panjang ke kanan (*right-skewed distribution*) pada variabel finansial dan pemakaian, seperti *monthly_income*, *monthly_bill*, *internet_usage_gb*, *call_minutes*, dan *support_tickets*. Hal ini mengonfirmasi adanya pencilan (outlier) yang valid namun ekstrem. Guna mengatasinya tanpa kehilangan baris data berharga, diterapkan metode **Interquartile Range (IQR)** dengan formulasi:

Batas Atas = Q3 + (1.5 × IQR)

Nilai-nilai ekstrim yang melampaui ambang Batas Atas tersebut disesuaikan secara proaktif ke nilai ambang batas maksimumnya (**Capping/Winsorization**) agar deviasi kuadratik tidak mengacaukan konvergensi fungsi loss algoritma klasifikasi.

```python
# Fungsi untuk menerapkan capping outlier dengan IQR
def cap_outliers_iqr(df_col):
    Q1 = df_col.quantile(0.25)
    Q3 = df_col.quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return df_col.clip(lower=lower_bound, upper=upper_bound)

# Kolom-kolom yang akan ditangani dengan capping
columns_to_cap = ['monthly_bill', 'internet_usage_gb', 'call_minutes', 'support_tickets']

for col in columns_to_cap:
    if col in df.columns:
        df[f'{col}_capped'] = cap_outliers_iqr(df[col])
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        upper = Q3 + 1.5 * IQR
        print(f"Outlier '{col}' berhasil di-capping. Batas atas: {upper:.1f} (max asli: {df[col].max():.1f})")

print("\nRingkasan Statistik Setelah Capping:")
display(df[columns_to_cap + [f'{c}_capped' for c in columns_to_cap]].describe())
```

```
Outlier 'monthly_bill' berhasil di-capping. Batas atas: 13776.1 (max asli: 36361.0)
Outlier 'internet_usage_gb' berhasil di-capping. Batas atas: 50.1 (max asli: 82.2)
Outlier 'call_minutes' berhasil di-capping. Batas atas: 428.5 (max asli: 865.0)
Outlier 'support_tickets' berhasil di-capping. Batas atas: 5.0 (max asli: 6.0)

Ringkasan Statistik Setelah Capping:
```

| | monthly_bill | internet_usage_gb | call_minutes | support_tickets | monthly_bill_capped | internet_usage_gb_capped | call_minutes_capped | support_tickets_capped |
|---|---|---|---|---|---|---|---|---|
| count | 1200.00 | 1200.00 | 1200.00 | 1200.00 | 1200.00 | 1200.00 | 1200.00 | 1200.00 |
| mean | 5939.15 | 20.15 | 182.83 | 1.19 | 5709.50 | 19.96 | 179.85 | 1.19 |
| std | 4046.47 | 12.17 | 106.57 | 1.08 | 3226.47 | 11.56 | 96.84 | 1.08 |
| min | 710.00 | 1.10 | 14.00 | 0.00 | 710.00 | 1.10 | 14.00 | 0.00 |
| 25% | 3291.75 | 11.10 | 106.00 | 0.00 | 3291.75 | 11.10 | 106.00 | 0.00 |
| 50% | 4871.00 | 17.80 | 164.00 | 1.00 | 4871.00 | 17.80 | 164.00 | 1.00 |
| 75% | 7485.50 | 26.70 | 235.00 | 2.00 | 7485.50 | 26.70 | 235.00 | 2.00 |
| max | 36361.00 | 82.20 | 865.00 | 6.00 | 13776.12 | 50.10 | 428.50 | 5.00 |

d. **Feature Engineering**

Guna meningkatkan derajat informatif data terhadap fenomena churn, dibentuk dua fitur turunan baru yang bersifat strategis:

- **tenure_months:** Fitur temporal ini diekstrak dengan memparsing string tanggal *signup_date*, lalu menghitung jarak masa berlangganan dalam satuan bulan terhadap titik waktu referensi proyek. Atribut ini sangat krusial mengingat pelanggan baru secara historis lebih rentan mengalami churn dibandingkan pelanggan yang telah melewati fase loyalitas awal. Setelah ekstraksi sukses, kolom asli *signup_date* dihapus demi efisiensi dimensi.
- **financial_risk:** Sebuah variabel biner indikator tekanan finansial (nilai 1 jika komponen pengeluaran *monthly_bill* bernilai lebih besar daripada kapasitas pendapatan *monthly_income*; nilai 0 jika pengeluaran berada dalam batas aman pendapatan), menangkap korelasi langsung antara beban keuangan dengan keputusan pelanggan mengakhiri kontrak.

```python
# Fitur tenure_months
if 'signup_date' in df.columns:
    df['signup_date'] = pd.to_datetime(df['signup_date'])
    tanggal_referensi = pd.to_datetime('2026-06-25')
    df['tenure_months'] = ((tanggal_referensi - df['signup_date']).dt.days // 30).astype(int)
    df.drop(columns=['signup_date'], inplace=True)
    print("Fitur 'tenure_months' berhasil dibuat dan kolom 'signup_date' dihapus.")
    print(f"  Rentang tenure : {df['tenure_months'].min()} - {df['tenure_months'].max()} bulan")
    print(f"  Rata-rata      : {df['tenure_months'].mean():.1f} bulan")

# Fitur financial_risk
if 'monthly_bill' in df.columns and 'monthly_income' in df.columns:
    df['financial_risk'] = np.where(df['monthly_bill'] > df['monthly_income'], 1, 0)
    n_risk = df['financial_risk'].sum()
    print(f"\nFitur 'financial_risk' berhasil dibuat.")
    print(f"  Pelanggan berisiko finansial : {n_risk} ({n_risk/len(df)*100:.1f}%)")

print("\n5 baris pertama setelah pembuatan fitur baru:")
display(df.head())
```

```
Fitur 'tenure_months' berhasil dibuat dan kolom 'signup_date' dihapus.
  Rentang tenure : 51 - 91 bulan
  Rata-rata      : 70.6 bulan

Fitur 'financial_risk' berhasil dibuat.
  Pelanggan berisiko finansial : 23 (1.9%)

5 baris pertama setelah pembuatan fitur baru:
```

| | age | gender | city | education_level | employment_status | monthly_income | monthly_bill | internet_usage_gb | call_minutes | contract_type | support_tickets | customer_feedback | churn | tenure_months | financial_risk |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 65 | Male | Multan | Bachelor | Employed | 57643.00 | 10049.00 | 5.80 | 52.00 | 6-Month | 1 | Coverage is poor in my location | 0 | 91 | 0 |
| 1 | 22 | Female | Peshawar | Master | Employed | 18207.00 | 2752.00 | 11.90 | 22.00 | 6-Month | 2 | Billing issues occurred multiple times | 0 | 91 | 0 |
| 2 | 43 | Male | Gujranwala | Secondary | Employed | 13075.00 | 3155.00 | 40.00 | 87.00 | Monthly | 0 | Billing issues occurred multiple times | 0 | 91 | 0 |
| 3 | 21 | Male | Lahore | Secondary | Employed | 38890.00 | 5859.00 | 8.50 | 147.00 | 6-Month | 2 | Customer support was helpful | 0 | 90 | 0 |
| 4 | 37 | Female | Hyderabad | Primary | Employed | 17475.00 | 5106.00 | 3.30 | 183.00 | 12-Month | 0 | Coverage is poor in my location | 0 | 90 | 0 |

e. **Encoding Variabel Kategorikal dan Vektorisasi Teks**

Transformasi data kualitatif dijalankan berdasarkan klasifikasi fiturnya:

- **One-Hot Encoding:** Diterapkan pada fitur kategorikal nominal dengan variasi nilai majemuk (*gender*, *city*, dan *employment_status*) menggunakan pustaka fungsi pemrograman. Langkah ini memecah kolom kategorikal menjadi kolom biner baru (0 atau 1) untuk mencegah model mengasumsikan adanya tingkatan besaran urutan (*ordinal hierarchy bias*) yang sebenarnya tidak ada pada variabel nominal.
- **Text Feature Extraction (Count Vectorizer):** Atribut naratif ulasan *customer_feedback* dibersihkan terlebih dahulu dari spasi kosong sekunder (*whitespace trimming*). Selanjutnya, teks tidak terstruktur ditransformasikan ke dalam representasi matriks kuantitatif frekuensi kemunculan kata kunci menggunakan metode **Count Vectorizer**. Teknik ini berhasil memetakan klaster kata kunci utama penyulut ketidakpuasan pelanggan ke dalam dimensi numerik terukur (seperti pendeteksian isu tingginya biaya penagihan atau keluhan buruknya sinyal jaringan).

```python
# 1. Gender - binary encoding
if 'gender' in df.columns:
    df['gender'] = df['gender'].map({'Female': 0, 'Male': 1, 'female': 0, 'male': 1})
    print("Encoding 'gender': Female=0, Male=1")
    print(df['gender'].value_counts().to_string())

# 2. Education Level - ordinal encoding
edu_mapping = {'Primary': 0, 'Secondary': 1, 'Bachelor': 2, 'Master': 3, 'PhD': 4}
if 'education_level' in df.columns:
    df['education_level_encoded'] = df['education_level'].map(edu_mapping)
    print("\nEncoding 'education_level' (ordinal):")
    print(df['education_level_encoded'].value_counts().sort_index().to_string())

# 3. Contract Type - ordinal encoding
contract_mapping = {'Monthly': 0, '6-Month': 1, '12-Month': 2}
if 'contract_type' in df.columns:
    df['contract_type_encoded'] = df['contract_type'].map(contract_mapping)
    print("\nEncoding 'contract_type' (ordinal):")
    print(df['contract_type_encoded'].value_counts().sort_index().to_string())

# 4. Employment Status - one-hot encoding
if 'employment_status' in df.columns:
    df_emp_dummies = pd.get_dummies(df['employment_status'], prefix='emp', drop_first=False, dtype=int)
    df = pd.concat([df, df_emp_dummies], axis=1)
    print(f"\nOne-Hot Encoding 'employment_status': {list(df_emp_dummies.columns)}")

print("\nProses encoding selesai.")
display(df.head())
```

```
Encoding 'gender': Female=0, Male=1
gender
1    603
0    597

Encoding 'education_level' (ordinal):
education_level_encoded
0    165
1    304
2    422
3    249
4     60

Encoding 'contract_type' (ordinal):
contract_type_encoded
0    569
1    286
2    345

One-Hot Encoding 'employment_status': ['emp_Employed', 'emp_Retired', 'emp_Self-Employed', 'emp_Student', 'emp_Unemployed']

Proses encoding selesai.
```

| | age | gender | city | education_level | employment_status | monthly_income | monthly_bill | internet_usage_gb | call_minutes | contract_type | support_tickets | customer_feedback | churn |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 65 | 1 | Multan | Bachelor | Employed | 57643.00 | 10049.00 | 5.80 | 52.00 | 6-Month | 1 | Coverage is poor in my location | 0 |
| 1 | 22 | 0 | Peshawar | Master | Employed | 18207.00 | 2752.00 | 11.90 | 22.00 | 6-Month | 2 | Billing issues occurred multiple times | 0 |
| 2 | 43 | 1 | Gujranwala | Secondary | Employed | 13075.00 | 3155.00 | 40.00 | 87.00 | Monthly | 0 | Billing issues occurred multiple times | 0 |
| 3 | 21 | 1 | Lahore | Secondary | Employed | 38890.00 | 5859.00 | 8.50 | 147.00 | 6-Month | 2 | Customer support was helpful | 0 |

f. **Feature Scaling (Normalisasi Numerik)**

Guna menghindari dominasi variabel berskala besar (seperti pendapatan bernilai puluhan ribu) terhadap variabel berskala kecil (seperti kuantitas tiket aduan berdomain satuan) saat pelatihan model, seluruh fitur numerik non-kontinu seperti *age* ditransformasikan menggunakan metode Min-Max Scaling ke dalam domain rentang nilai seragam antara [0, 1].

### 3.5 General EDA

Tahap EDA Umum difungsikan untuk memetakan struktur dasar serta memeriksa karakteristik distribusi variabel di dalam dataset setelah melalui rangkaian proses preprocessing. Fokus utama tahapan ini berorientasi pada validasi kesiapan data sebelum didistribusikan ke tahapan pemodelan prediktif. Implementasi teknis pada tahapan ini meliputi:

**a. Analisis Proporsi Variabel Target (Target Class Distribution Analysis)**

Pemeriksaan tingkat keseimbangan kelas pada variabel target churn dilakukan secara visual. Metode ini diimplementasikan menggunakan fungsi sns.countplot() atau diagram batang frekuensi guna mendeteksi adanya indikasi class imbalance. Langkah ini penting untuk menentukan apakah model klasifikasi membutuhkan penanganan berbasis data khusus (seperti SMOTE) atau cukup mengandalkan penyesuaian metrik evaluasi pada algoritma.

**b. Analisis Hubungan Variabel Kunci (Bivariate Key Features Analysis)**

Eksplorasi hubungan antara variabel-variabel prediktor kunci terhadap variabel target churn dilakukan dengan teknik analisis bivariat. Secara teknis, proses ini memanfaatkan matriks visualisasi bertingkat (subplots) menggunakan kombinasi fungsi sns.barplot() atau sns.boxplot(). Metode ini digunakan untuk mengevaluasi secara visual tren sebaran data, tingkat skewness (kemencengan) pasca-capping, serta pola interaksi awal fitur-fitur penting (seperti support_tickets, contract_type, dan volume penggunaan layanan) terhadap kecenderungan churn.

**c. Uji Multikolinieritas Melalui Matriks Korelasi (Correlation Matrix Mapping)**

Pengujian multikolinieritas dilakukan untuk mendeteksi adanya korelasi linear yang terlalu tinggi (redundansi) antar variabel independen numerik yang dapat mengganggu stabilitas bobot model prediktif. Metode yang digunakan adalah menghitung koefisien korelasi menggunakan fungsi .corr(method='pearson') pada Pandas DataFrame. Hasil perhitungan koefisien tersebut kemudian dipetakan secara visual ke dalam bentuk grafik Heatmap interaktif menggunakan pustaka sns.heatmap() dengan parameter annot=True untuk mempermudah audit angka korelasi secara menyeluruh.

### 3.6 Specific EDA

#### 3.6.1 Segmentasi

**Sumber dan Akuisisi Data**

Import library dan clean data  dari github

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Tarik data bersih langsung dari GitHub
url = 'https://raw.githubusercontent.com/AIU7AgrataEleigh/Telecom-Churn-Rate_Kelompok-2/main/
df = pd.read_csv(url)

# Melihat  5 baris data teratas
df.head()
```

| | gender | churn | monthly_bill_capped | internet_usage_gb_capped | call_minutes_capped | support_ |
|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 10049.0 | 5.8 | 52.0 | |
| **1** | 0 | 0 | 2752.0 | 11.9 | 22.0 | |
| **2** | 1 | 0 | 3155.0 | 40.0 | 87.0 | |
| **3** | 1 | 0 | 5859.0 | 8.5 | 147.0 | |
| **4** | 0 | 0 | 5106.0 | 3.3 | 183.0 | |

5 rows × 42 columns

**Visualisasi Data**

(1) pengecekan dimensi data menggunakan df.shape,

(2) pengecekan tipe data setiap kolom menggunakan df.info(),

```python
print(df.shape)

df.info()

df.describe()
```

```
(1200, 42)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1200 entries, 0 to 1199
Data columns (total 42 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   gender                      1200 non-null   int64
```

(3) pengecekan missing value menggunakan df.isna().sum().

```python
missing = pd.DataFrame({
    "Missing Values": df.isna().sum(),
    "Persentase (%)": (df.isna().mean()*100).round(2)
})

missing
```

| | | |
|---|---|---|
| emp_Employed | 0 | 0.0 |
| emp_Retired | 0 | 0.0 |

**Pemilihan Variabel**

Variabel yang digunakan difokuskan pada tiga indikator utama yaitu aktivitas data (*internet_usage_gb_capped*), aktivitas telepon (*call_minutes_capped*), dan beban tagihan (*monthly_bill_capped*) untuk membentuk profil pelanggan, dengan variabel *churn* sebagai indikator target.

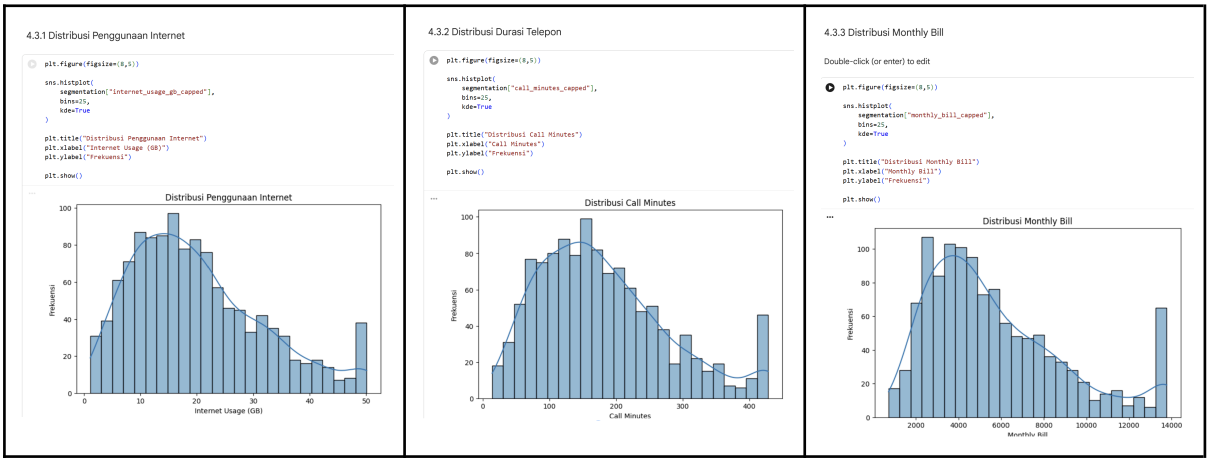
| | internet_usage_gb_capped | call_minutes_capped | monthly_bill_capped | churn |
|---|---|---|---|---|
| 0 | 5.8 | 52.0 | 10049.0 | 0 |
| 1 | 11.9 | 22.0 | 2752.0 | 0 |
| 2 | 40.0 | 87.0 | 3155.0 | 0 |
| 3 | 8.5 | 147.0 | 5859.0 | 0 |
| 4 | 3.3 | 183.0 | 5106.0 | 0 |

**Profil Statistik**

| | Mean | Median | Mode | Minimum | Maximum | Range | Variance | Std Dev | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|
| internet_usage_gb_capped | 19.961 | 17.8 | 50.100 | 1.1 | 50.100 | 49.000 | 1.336680e+02 | 11.561 | 0.750 | -0.010 |
| call_minutes_capped | 179.851 | 164.0 | 428.500 | 14.0 | 428.500 | 414.500 | 9.377901e+03 | 96.840 | 0.761 | 0.089 |
| monthly_bill_capped | 5709.496 | 4871.0 | 13776.125 | 710.0 | 13776.125 | 13066.125 | 1.041013e+07 | 3226.474 | 0.971 | 0.258 |
| churn | 0.350 | 0.0 | 0.000 | 0.0 | 1.000 | 1.000 | 2.280000e-01 | 0.477 | 0.630 | -1.606 |

Statistik deskriptif digunakan untuk mengetahui karakteristik pusat data, dengan penggunaan nilai median sebagai batas segmentasi yang *robust* terhadap *outlie*r.

## Distribusi




Visualisasi distribusi menunjukkan keragaman perilaku pelanggan, yang mengindikasikan bahwa strategi pemasaran tidak bisa disamaratakan bagi seluruh pelanggan

## Metode Analisis Statistik Deskriptif & Visualisasi

| | Mean | Median | Mode | Minimum | Maximum | Range | Variance | Std Dev | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|
| **internet_usage_gb_capped** | 19.961 | 17.8 | 50.100 | 1.1 | 50.100 | 49.000 | 1.336680e+02 | 11.561 | 0.750 | -0.010 |
| **call_minutes_capped** | 179.851 | 164.0 | 428.500 | 14.0 | 428.500 | 414.500 | 9.377901e+03 | 96.840 | 0.761 | 0.089 |
| **monthly_bill_capped** | 5709.496 | 4871.0 | 13776.125 | 710.0 | 13776.125 | 13066.125 | 1.041013e+07 | 3226.474 | 0.971 | 0.258 |
| **churn** | 0.350 | 0.0 | 0.000 | 0.0 | 1.000 | 1.000 | 2.280000e-01 | 0.477 | 0.630 | -1.606 |

Karakteristik pusat digunakan untuk mengetahui nilai tipikal dari masing-masing variabel numerik sebelum dilakukan visualisasi data. Ukuran yang digunakan meliputi mean, median, dan mode

### 4.4.3 Pivot Analysis

```
pivot_summary = segmentation.groupby("churn").agg({
    "internet_usage_gb_capped":"mean",
    "call_minutes_capped":"mean",
    "monthly_bill_capped":"mean"
}).round(2)

pivot_summary.index = ["Non-Churn","Churn"]

pivot_summary
```

| | internet_usage_gb_capped | call_minutes_capped | monthly_bill_capped |
|---|---|---|---|
| **Non-Churn** | 20.75 | 176.09 | 5270.75 |
| **Churn** | 18.49 | 186.83 | 6524.31 |

| customer_segment | monthly_bill_capped |
|---|---|
| Traditional Communicator | 5781.568182 |
| Power User | 5709.486431 |
| Internet Enthusiast | 5677.368289 |
| Casual User | 5670.199751 |

Analisis dilakukan menggunakan rata-rata monthly_bill sebagai indikator kontribusi pendapatan dari user

| customer_segment | internet_usage_gb_capped | call_minutes_capped | monthly_bill_capped |
|---|---|---|---|
| Casual User | 10.82 | 102.07 | 5670.20 |
| Internet Enthusiast | 29.34 | 105.42 | 5677.37 |
| Power User | 28.78 | 258.61 | 5709.49 |
| Traditional Communicator | 10.79 | 252.76 | 5781.57 |

Kemudian, membagi segmentasi kembali berdasarkan penggunaan internet usage dan call minute

| Customer Persona | Karakteristik Perilaku | Rata-rata Tagihan | Churn Rate |
|---|---|---|---|
| Power User | Internet Tinggi, Telepon Tinggi | 5.709,49 | 33,22% |
| Internet Enthusiast | Internet Tinggi, Telepon Rendah | 5.677,37 | 30,87% |
| Traditional Communicator | Internet Rendah, Telepon Tinggi | 5.781,57 | 39,39% |
| Casual User | Internet Rendah, Telepon Rendah | 5.670,20 | 36,54% |

Segmentasi dilakukan berbasis *Behavioral Matrix* dengan metode median (Internet: 17,8 GB; Telepon: 164 menit), menghasilkan empat persona: *Power User*, *Internet Enthusiast*, *Traditional Communicator*, dan *Casual User*

### 3.6.2 Service Quality

3.6.2.1 Pengorganisasian Dataset

Sebelum analisis dilakukan, dataset dipersiapkan melalui tiga langkah sebagai berikut.

a. Rename Kolom

```
df = df.rename(columns={'churn': 'churn_status'})
```

Kolom churn diganti namanya menjadi churn_status untuk meningkatkan kejelasan semantik variabel target.

b. Pembuatan Label Kontrak

```
contract_mapping = {0: 'Monthly', 1: '6 Month', 2: '12 Month'}
df['contract_type_label'] = df['contract_type_encoded'].map(contract_mapping)
```

Nilai numerik pada kolom contract_type_encoded dipetakan kembali menjadi label teks yang lebih mudah dibaca untuk keperluan visualisasi dan analisis.

c. Rekonstruksi Kolom Kota

```
city_columns = [col for col in df.columns if col.startswith('city_')]
df['city_name'] = df[city_columns].idxmax(axis=1).str.replace('city_', '')
```

Kolom city_name di rekonstruksi dari kolom-kolom one-hot encoded kota menggunakan fungsi idxmax() untuk mendapatkan nama kota aktual setiap pelanggan.

Statistik deskriptif dihitung menggunakan empat metode untuk memahami karakteristik dasar data sebagai berikut.

a. describe()

```
display(df.describe())
```

Menghasilkan ringkasan statistik: count, mean, std, min, kuartil (25%, 50%, 75%), dan max untuk seluruh kolom numerik.

b. mean ()

```
display(df.mean(numeric_only=True))
```

Menghitung rata-rata setiap kolom numerik sebagai ukuran tendensi sentral.

c. median ()

```
display(df.median(numeric_only=True))
```

Menghitung median setiap kolom numerik, lebih robust terhadap outlier dibanding mean.

d. mode()

```
for col in df.select_dtypes(include=['int64', 'float64', 'object', 'bool']).columns:
    modes = df[col].mode()
    print(f"{col}: {modes.iloc[0]}")
```

Menghitung modus untuk setiap kolom, baik numerik maupun kategorikal.

3.6.2.3 Distribusi Variabel

Distribusi empat variabel kunci divisualisasikan untuk memahami pola individual masing-masing variabel.

a. Distribusi Support Tickets

```
plt.subplot(1, 2, 1)
sns.histplot(df['support_tickets_capped'], bins=10, kde=True)

plt.subplot(1, 2, 2)
sns.boxplot(y=df['support_tickets_capped'])
```

Histogram dengan KDE (Kernel Density Estimation) digunakan untuk menampilkan frekuensi distribusi, sedangkan boxplot digunakan untuk mengidentifikasi sebaran data dan outlier.

b. Distribusi Jenis Kontrak

```
sns.countplot(x='contract_type_label', data=df, palette='viridis',
              order=['Monthly', '6 Month', '12 Month'],
              hue='contract_type_label', legend=False)
```

Countplot digunakan untuk memvisualisasikan frekuensi setiap kategori jenis kontrak.

c. Distribusi Kota

```
sns.countplot(x='city_name', data=df, ...)
```

Countplot digunakan untuk menampilkan sebaran jumlah pelanggan per kota.

d. Distribusi Kategori Feedback

```
sns.countplot(x='feedback_category', data=df, ...)
```

Countplot digunakan untuk menampilkan frekuensi kemunculan setiap kategori umpan balik pelanggan.

3.6.2.4 Analisis Hubungan dengan Churn

Hubungan antara variabel kualitas layanan dan churn_status dianalisis menggunakan visualisasi dan perhitungan statistik.

a. Support Ticket vs Churn

```
sns.boxplot(x='churn_status', y='support_tickets_capped', data=df)

churn_tickets = df.groupby('churn_status')['support_tickets_capped'].mean()
print(churn_tickets)
```

Boxplot digunakan untuk membandingkan distribusi jumlah tiket antara pelanggan churn dan tidak churn. Rata-rata tiket per kelompok dihitung untuk memperkuat interpretasi.

b. Jenis Kontrak vs Churn

```
churn_contract = df.groupby('contract_type_label')['churn_status'].mean()
sns.barplot(x='contract_type_label', y='churn_status', data=df, ...)
```

Tingkat churn rate dihitung per kategori jenis kontrak dan divisualisasikan dengan barplot.

c. Kota vs Churn

```
churn_city = df.groupby('city_name')['churn_status'].mean()
sns.barplot(x='city_name', y='churn_status', data=df, ...)
```

Churn rate dihitung dan divisualisasikan per kota untuk mengidentifikasi pola geografis.

3.6.2.5 Analisis Multivariat (Tabel Pivot)

Tabel pivot digunakan untuk menganalisis hubungan multivariat yang tidak terlihat dari analisis univariat.

a. Churn berdasarkan Jenis Kontrak dan Kota

```
pivot_churn = df.pivot_table(
    values='churn_status',
    index='city_name',
    columns='contract_type_label',
```

```
    aggfunc='mean'
)
```

b. Rata-rata Support Tickets per Jenis Kontrak

```
pivot_tickets = df.pivot_table(
    values='support_tickets_capped',
    index='contract_type_label',
    aggfunc='mean'
)
```

3.6.2.6 Validasi Outlier

Outlier divalidasi menggunakan boxplot untuk variabel numerik kunci.

```
fig, axes = plt.subplots(1, 3, figsize=(15, 5))
sns.boxplot(y=df['support_tickets_capped'], ax=axes[0])
sns.boxplot(y=df['monthly_bill_capped'], ax=axes[1])
sns.boxplot(y=df['tenure_months'], ax=axes[2])
```

Visualisasi ini memastikan bahwa proses capping pada tahap preprocessing telah efektif membatasi nilai ekstrim.

3.6.2.7 Matriks Korelasi

Visualisasi ini memastikan bahwa proses capping pada tahap preprocessing telah efektif membatasi nilai ekstrim.

```
numeric_cols = ['support_tickets_capped', 'contract_type_encoded',
                'feedback_category', 'churn_status', ...]
corr_matrix = df[numeric_cols].corr()

sns.heatmap(corr_matrix, annot=True, fmt='.2f',
            cmap='coolwarm', center=0)
```

Nilai korelasi mendekati +1 atau -1 mengindikasikan hubungan linear yang kuat antar variabel.

### 3.6.3 Analisis Finansial dan Sensitivitas Harga

**Variabel yang digunakan:** monthly_bill_capped, monthly_income_scaled, financial_risk, age_scaled, variabel city(city_Faisalabad, city_Gujranwala, city_Hyderabad, city_Islamabad, city_Karachi, city_Lahore, city_Multan, city_Peshawar, city_Quetta, city_Rawalpindi, city_Sialkot)

**Statistik Deskriptif**

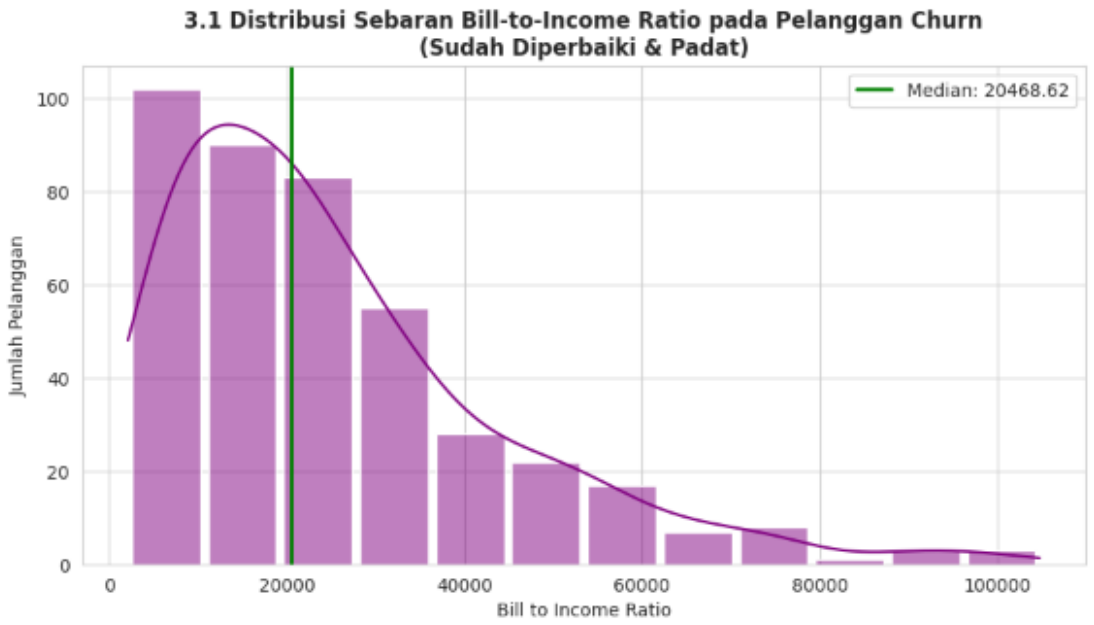
analisis menggunakan fungsi .describe() berfungsi untuk memberikan bukti angka riil yang akurat secara statistik

| | monthly_bill_capped | monthly_income_scaled | bill_to_income_ratio |
|---|---|---|---|
| count | 420.000000 | 420.000000 | 4.200000e+02 |
| mean | 6524.313690 | 0.223742 | inf |
| std | 3601.822682 | 0.161462 | NaN |
| min | 1116.000000 | 0.000000 | 2.648293e+03 |
| 25% | 3651.500000 | 0.106445 | 1.638283e+04 |
| 50% | 5616.500000 | 0.181842 | 3.271497e+04 |
| 75% | 8676.250000 | 0.300958 | 6.654770e+04 |
| max | 13776.125000 | 0.716033 | inf |

Feature engineering digunakan untuk membuat variabel interaksi baru dengan kode df_churn['bill_to_income_ratio'] = df_churn['monthly_bill_capped'] / df_churn['monthly_income_scaled']. Ini dilakukan untuk mengukur tingkat tekanan ekonomi yang dialami populasi churn secara keseluruhan. Penggabungan dua dimensi keuangan—biaya tagihan dan kapasitas pendapatan—menjadi nilai rasio ini didasarkan pada teori sensitivitas harga. Teori ini menyatakan bahwa tingkat stres finansial konsumen tidak dapat diukur hanya dari besarnya tagihan, tetapi juga dari seberapa agresif biaya bulanan tersebut menggerus ruang saku ekonomi rumah tangga.

**Distribusi Rasio beban Finansial**

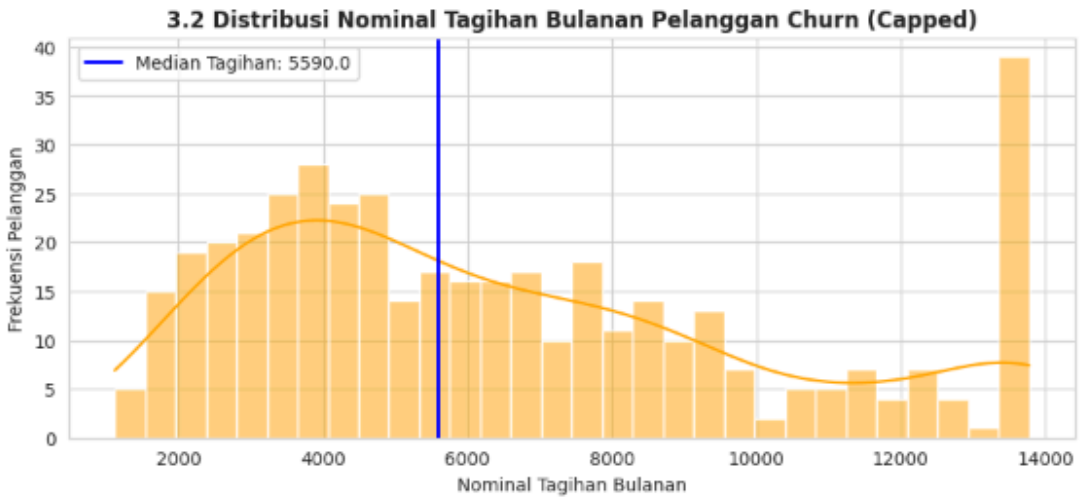
```
# --- COPIED & FIXED FOR HISTOGRAM RATIO ---
plt.figure(figsize=(10, 5))

# PERBAIKAN UTAMA: Nilai bins diturunkan ke 12, ditambah 'shrink=0.9' agar grafik padat dan rapi
sns.histplot(data=df_churn, x='bill_to_income_ratio', kde=True, color='purple', bins=12, shrink=0.9)

# Menambahkan garis penanda nilai tengah (Median)
plt.axvline(df_churn['bill_to_income_ratio'].median(), color='green', linestyle='-', linewidth=2,
            label=f"Median: {df_churn['bill_to_income_ratio'].median():.2f}")

plt.title('3.1 Distribusi Sebaran Bill-to-Income Ratio pada Pelanggan Churn\n(Sudah Diperbaiki & Padat)', fontsize=12, fontweight='bold')
plt.xlabel('Bill to Income Ratio')
plt.ylabel('Jumlah Pelanggan')
plt.legend()
plt.show()
```




Grafik ini menunjukkan bahwa mayoritas pelanggan yang churn ternyata merasakan tekanan ekonomi atau beban tagihan yang seragam dan sama-sama berat. Kondisi finansial yang terjepit inilah yang diduga kuat memicu mereka untuk berhenti berlangganan.

**Distribusi Tagihan Bulanan**

```
plt.figure(figsize=(10, 4))
sns.histplot(data=df_churn, x='monthly_bill_capped', kde=True, color='orange', bins=30)
plt.axvline(df_churn['monthly_bill_capped'].median(), color='blue', linestyle='-', linewidth=2,
            label=f"Median Tagihan: {df_churn['monthly_bill_capped'].median():.1f}")
plt.title('3.2 Distribusi Nominal Tagihan Bulanan Pelanggan Churn (Capped)', fontweight='bold', fontsize=12)
plt.xlabel('Nominal Tagihan Bulanan')
plt.ylabel('Frekuensi Pelanggan')
plt.legend()
plt.show()
```



ini adalah nominal Tagihan Bulanan dari pelanggan yang churn. Grafiknya cenderung tinggi di sebelah kanan, artinya banyak sekali pelanggan yang keluar justru berada di kelompok yang tagihan bulanannya mahal atau tinggi (mendekati batas maksimal/capped).

**Distribusi pendapatan Bulanan**

```
plt.figure(figsize=(10, 4))
sns.histplot(data=df_churn, x='monthly_income_scaled', kde=True, color='purple', bins=30)
plt.axvline(df_churn['monthly_income_scaled'].median(), color='green', linestyle='-', linewidth=2,
            label=f"Median Pendapatan: {df_churn['monthly_income_scaled'].median():.3f}")
plt.title('3.3 Distribusi Sebaran Tingkat Pendapatan Pelanggan Churn (Scaled)', fontweight='bold', fontsize=12)
plt.xlabel('Nilai Pendapatan Bersih Terskala')
plt.ylabel('Frekuensi Pelanggan')
plt.legend()
plt.show()
```

Grafik ini menunjukkan profil Pendapatan Bulanan dari pelanggan yang churn. Grafik ini menunjukkan bahwa pelanggan yang churn tidak hanya berasal dari kelompok berpendapatan rendah, melainkan tersebar di berbagai tingkat pendapatan.

# BAB IV
# HASIL DAN PEMBAHASAN

## 4.1 Hasil Data Preprocessing

### 4.1.1. Hasil Setelah Preprocessing

Sebelum dilakukan pemrosesan, dataset mentah dievaluasi untuk mendeteksi permasalahan kualitas data seperti missing values, data duplikat, nilai tidak masuk akal (impossible values), inkonsistensi penulisan, dan keberadaan data pencilan (outlier). Berikut adalah rangkuman tindakan serta hasil setelah preprocessing dilakukan:

a. **Penanganan Missing Values & Duplikat**
   Setelah melalui proses verifikasi, dipastikan tidak terdapat nilai yang hilang (missing values = 0) maupun baris data yang ganda (duplikat = 0) di dalam dataset final, sehingga integritas data tabular tetap terjaga.
b. **Pembersihan Data Kategorikal**
   Dilakukan penghapusan spasi kosong (whitespace trimming) pada seluruh fitur kategorikal untuk mencegah redundansi kategori akibat kesalahan input teks.
c. **Penanganan Outlier (Data Pencilan)**
   Fitur-fitur numerik utama seperti monthly_income, monthly_bill, internet_usage_gb, dan call_minutes diidentifikasi menggunakan metode Interquartile Range (IQR). Untuk meminimalkan bias akibat nilai ekstrem tanpa menghapus baris data, diterapkan metode pencatatan batas atas/bawah (capping atau Winsorization).
d. **Transformasi Skala (Scaling)**
   Variabel numerik yang telah dibersihkan kemudian ditransformasikan menggunakan Min-Max Scaling guna menyamakan rentang nilai (domain skala 0 hingga 1), sehingga algoritma pemodelan dapat bekerja lebih optimal dan stabil tanpa terpengaruh oleh perbedaan satuan besaran fitur.

### 4.1.2 Perubahan Struktur Dataset (Kolom & Baris)

Proses transformasi data, encoding variabel kategorikal, serta ekstraksi fitur teks menyebabkan perubahan signifikan pada struktur dimensi dataset. Perubahan struktur ini dirangkum dalam tabel berikut:

| **Karakteristik Dimensi** | **Dataset Awal (Mentah)** | **Dataset Final (Hasil Preprocessing)** | **Keterangan Perubahan** |
|---|---|---|---|
| Jumlah Baris | 1,200 baris | 1,200 baris | Jumlah baris tetap dipertahankan guna |

| Karakteristik Dimensi | Dataset Awal (Mentah) | Dataset Final (Hasil Preprocessing) | Keterangan Perubahan |
|---|---|---|---|
| | | | menjaga volume sampel eksperimen. |
| Jumlah Kolom | 15 kolom | 42 kolom | Mengalami penambahan sebanyak 27 kolom baru akibat *feature engineering* dan encoding. |
| Kolom Numerik | 7 kolom | 42 kolom | Seluruh kolom dalam dataset final telah berbentuk angka (int64 / float64). |
| Kolom Non-Numerik (Object/Text) | 8 kolom | 0 kolom | Semua fitur kategorikal dan teks bebas telah berhasil dikonversi sepenuhnya. |

### 4.1.3 Rekayasa Fitur Baru (Feature Engineering)

Untuk memperkaya informasi perilaku pelanggan, dibentuk beberapa kelompok fitur baru yang diekstrak dari variabel demografi, keuangan, serta umpan balik pelanggan:

**a. Fitur Berbasis Durasi dan Finansial**

- **tenure_months:** Fitur ini dihitung berdasarkan selisih antara tanggal referensi analisis (25 Juni 2026) dengan tanggal registrasi pelanggan (signup_date). Fitur ini memberikan gambaran durasi loyalitas pelanggan dalam satuan bulan (berkisar antara 51 hingga 91 bulan).
- **financial_risk:** Fitur indikator biner (1 jika total tagihan bulanan monthly_bill lebih besar daripada pendapatan bulanan monthly_income, dan 0 jika sebaliknya). Fitur ini membantu mendeteksi risiko gagal bayar yang berpotensi memicu churn.

**b. Fitur Indikator Anomali Perilaku**

- **anomaly_pay_not_use:** Menandakan pelanggan yang membayar tagihan bulanan tinggi namun memiliki tingkat penggunaan internet atau menit panggilan yang sangat rendah.
- **anomaly_happy_leaver:** Menandakan segmen pelanggan khusus yang melakukan churn tanpa memberikan sinyal keluhan atau tiket dukungan (*support tickets* = 0) sebelumnya.
- **anomaly_genz_call_enthusiast:** Menandakan karakteristik unik dari kelompok usia muda (Gen-Z) yang memiliki durasi panggilan suara (*call minutes*) sangat tinggi di luar

rata-rata kelompoknya.

**c. Encoding Variabel Kategorikal**

Variabel kategorikal nominal (seperti jenis kelamin, kota domisili, tingkat pendidikan, status pekerjaan, dan jenis kontrak) ditransformasikan menggunakan metode **One-Hot Encoding**. Proses ini menghasilkan kolom-kolom biner baru (seperti city_Hyderabad, city_Lahore, dsb.) sehingga dapat langsung diproses oleh algoritma klasifikasi.

**d. Ekstraksi Fitur Teks (Text Preprocessing)**

Kolom umpan balik bebas customer_feedback diekstrak menjadi fitur numerik menggunakan teknik binerisasi teks atau **Count Vectorizer** berbasis kata kunci spesifik untuk menangkap sentimen dan keluhan utama pelanggan, meliputi:

- **Masalah Biaya/Tagihan:** Berdasarkan kemunculan kata kunci seperti *'billing'*, *'expensive'*.
- **Masalah Jaringan/Sinyal:** Berdasarkan kemunculan kata kunci seperti *'poor'*, *'unreliable'*, *'drops'*, *'slow'*.
- **Layanan Customer Support:** Berdasarkan kemunculan kata kunci seperti *'support'*.
- **Sentimen Puas/Positif:** Berdasarkan kemunculan kata kunci seperti *'happy'*, *'satisfied'*, *'good'*.

### 4.1.4 Karakteristik Dataset Final

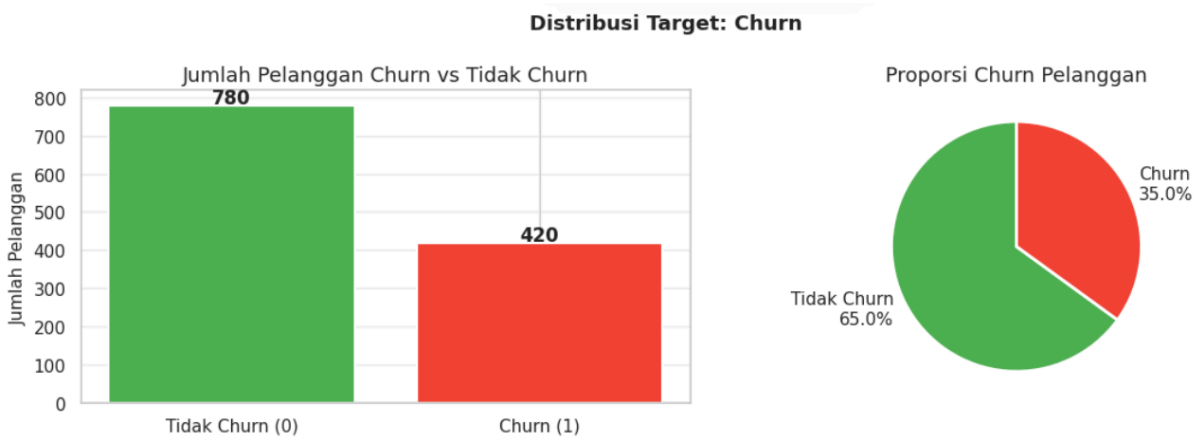
Dataset final yang dihasilkan memiliki spesifikasi teknis sebagai berikut:

- **Ukuran Matriks Data:** 1,200 baris × 42 kolom.
- **Kebersihan Data:** 100% bersih dari nilai kosong, data duplikat, maupun teks mentah.
- **Kesiapan Model:** Seluruh data telah berformat numerik dan memiliki sebaran fitur yang stabil setelah proses pembatasan outlier (*capping*), meminimalkan risiko bias saat model melakukan proses training.

Dengan struktur data yang komprehensif ini, dataset siap didistribusikan ke dalam analisis lanjutan pada masing-masing sub-tim (Analisis Demografi, Analisis Penggunaan Layanan, dan Pemodelan Klasifikasi Prediktif Churn).

## 4.2 Hasil General EDA

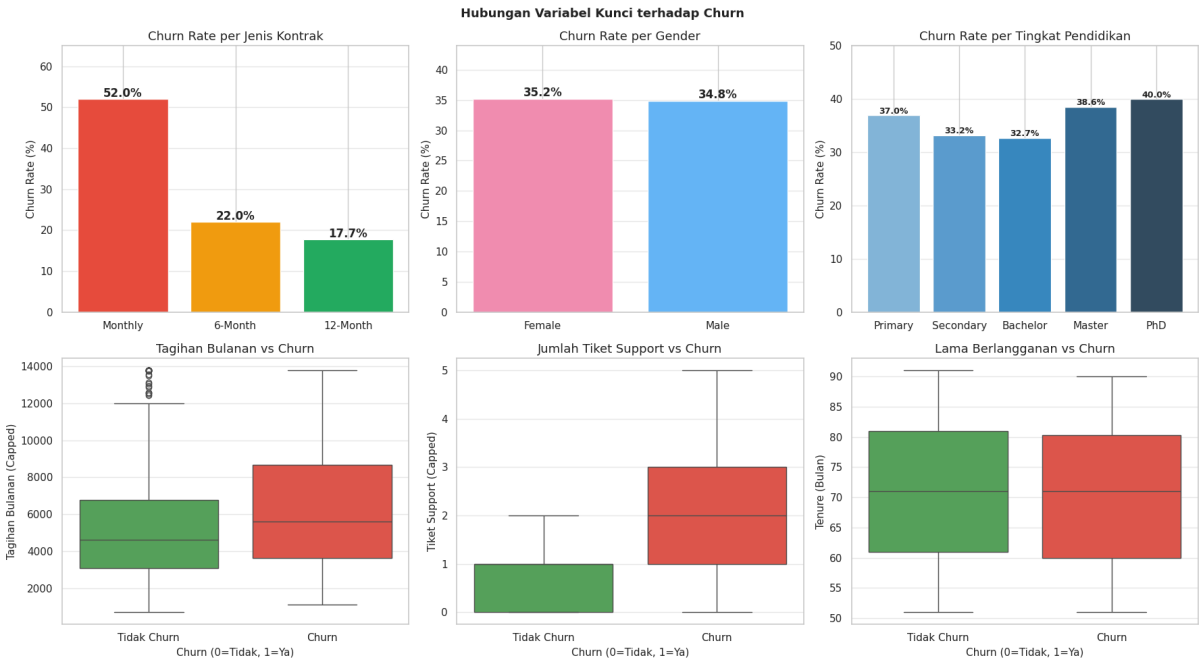
- **Evaluasi Distribusi Proporsi Target Churn**




*Gambar 4.2.1 Distribusi Proporsi Variabel Target Churn*

Berdasarkan visualisasi frekuensi variabel target, didapatkan rasio pembagian kelas sebesar 65% untuk kelas pelanggan aktif bertahan (Retained / 0) dan 35% untuk kelas pelanggan keluar (Churn / 1). Distribusi ini menandakan bahwa dataset berada dalam kondisi ketidakseimbangan kelas tingkat ringan (mildly imbalanced). Implikasinya, model machine learning klasifikasi biner nantinya dapat dilatih secara langsung tanpa keharusan manipulasi sintetik radikal seperti aplikasi algoritma SMOTE, namun proses pengujian harus tetap menggunakan metrik evaluasi yang sensitif (seperti F1-Score atau AUC-ROC) alih-alih murni mengandalkan metrik akurasi umum.

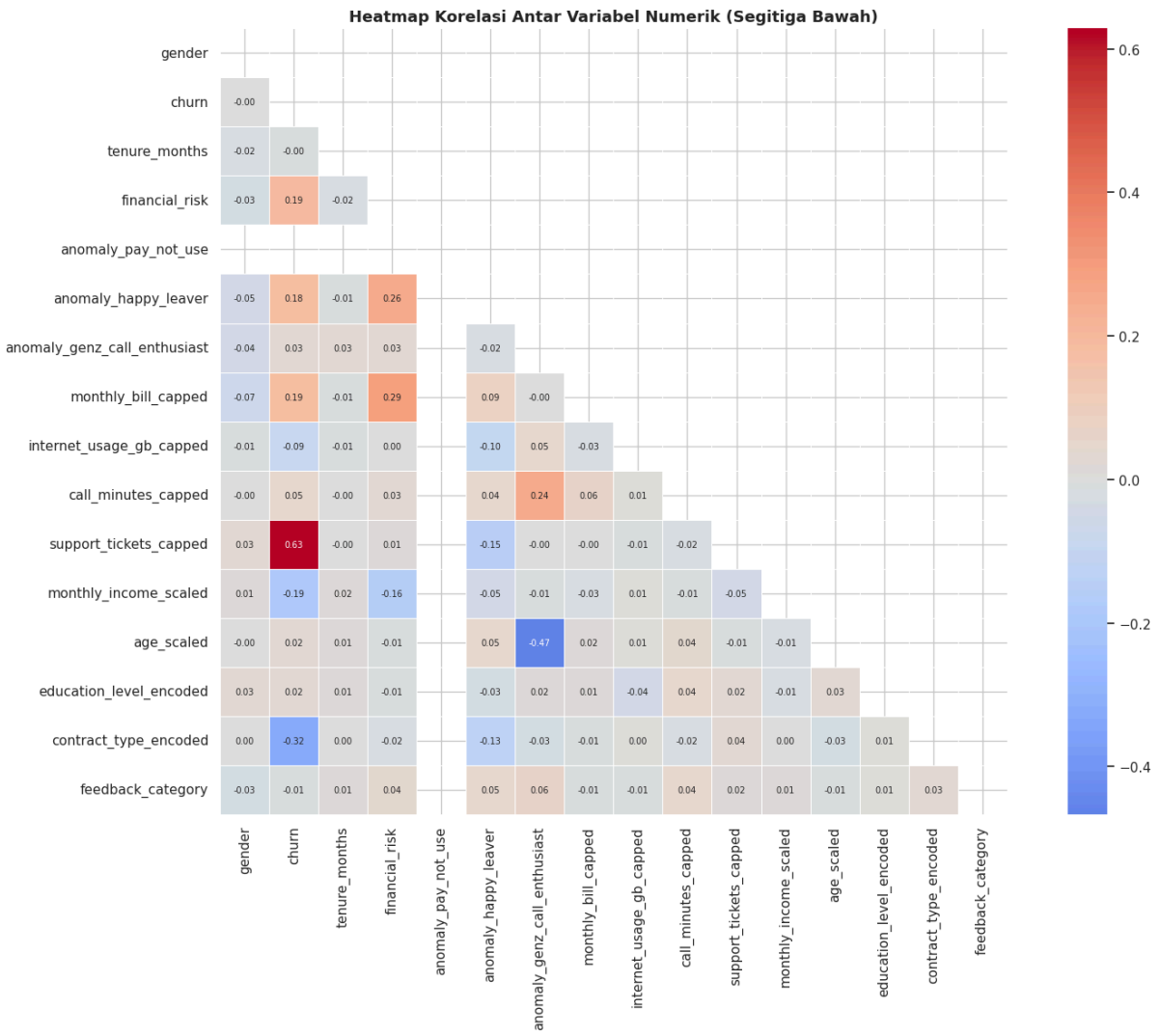
- **Identifikasi Hubungan Awal & Skewness**




*Gambar 4.2.2 Analisis Visual Hubungan Atribut Kunci terhadap Churn Rate*

Analisis visual melalui visualisasi sebaran histogram membuktikan bahwa rekayasa pembatasan nilai (*capping outlier*) berhasil mereduksi deviasi ekstrim sebaran data tanpa merusak bentuk alami distribusi populasi sampel. Hubungan awal menunjukkan adanya kecenderungan korelasi linear positif antara tingginya frekuensi keluhan pada fitur *support_tickets* dengan kecenderungan nilai target churn.

- **Uji Multikolinieritas Melalui Matriks Korelasi**




*Gambar 4.2.3 Heatmap Matriks Korelasi Antar Variabel Pasca-Preprocessing*

Evaluasi koefisien matriks korelasi linier antar variabel numerik pasca-preprocessing dijalankan secara menyeluruh. Pengujian ini memastikan tidak terdapat pasangan variabel independen yang memiliki korelasi linear terlalu tinggi (multikolinieritas ekstrem), sehingga menjamin stabilitas estimasi parameter bobot pada tahapan pemodelan prediktif klasifikasi selanjutnya. Rangkaian temuan umum dari EDA ini memberikan pondasi data yang kokoh bagi tim pengerjaan sub-analisis berikutnya.

## 4.3 Hasil Analisis Customer Segmentation

Segmentasi pelanggan (Customer Segmentation) dilakukan untuk memetakan perilaku penggunaan layanan dan dampaknya terhadap tingkat churn. Analisis ini difokuskan pada tiga variabel numerik utama yang telah melalui proses penanganan outlier (Winsorization/Capping), yaitu internet_usage_gb_capped, call_minutes_capped, dan monthly_bill_capped.

### 4.3.1 Behavioral Matrix Berbasis Median

Segmentasi dilakukan menggunakan metode **Behavioral Matrix** dengan membagi pelanggan ke dalam empat kuadran berdasarkan nilai **Median** penggunaan layanan (Internet: 17,8 GB; Telepon: 164 menit).

| Customer Persona | Karakteristik Perilaku | Rata-rata Tagihan | Churn Rate |
|---|---|---|---|
| **Power User** | Internet Tinggi, Telepon Tinggi | 5.709,49 | 33,22% |
| **Internet Enthusiast** | Internet Tinggi, Telepon Rendah | 5.677,37 | 30,87% |
| **Traditional Communicator** | Internet Rendah, Telepon Tinggi | 5.781,57 | 39,39% |
| **Casual User** | Internet Rendah, Telepon Rendah | 5.670,20 | 36,54% |

*(Sumber: Sintesis Bagian 5.6 Jupyter Notebook )*

**Analisis Utama:**

1. **Identifikasi "Digital Natives":** Segmen *Internet Enthusiast* memiliki rata-rata penggunaan data tertinggi (**29,34 GB**) namun jarang menelepon (**105,42 menit**). Mereka memiliki *Package Usage Ratio* tertinggi (**0,34**), menunjukkan efisiensi pemanfaatan kuota internet yang baik.
2. **Konfirmasi Fenomena "Wasted Value":** Ditemukan inefisiensi pada segmen *Internet Enthusiast*. Meskipun mereka jarang menelepon, tagihan rata-rata mereka (**5.677,37**) hampir setara dengan *Traditional Communicator* yang menelepon dua kali lipat lebih banyak. Ini membuktikan pelanggan terpaksa membayar fasilitas telepon yang tidak mereka gunakan.
3. **Segmen Risiko Tinggi:** *Traditional Communicator* memberikan kontribusi pendapatan tertinggi, namun juga memiliki **Churn Rate tertinggi (39,39%)**. Kehilangan pelanggan pada segmen ini akan memberikan dampak finansial terbesar bagi perusahaan

**Rekomendasi Bisnis:**

Berdasarkan hasil pemetaan perilaku dan risiko *churn*, perusahaan disarankan mengimplementasikan strategi berbasis persona sebagai berikut:

- **Internet Enthusiast (Solusi *Wasted Value*):** Meluncurkan **paket "Add-on Data Only"** dan *bundling* layanan hiburan (*streaming/gaming*) untuk pelanggan dengan konsumsi data tinggi (rata-rata **29,34 GB**) guna menghilangkan persepsi beban biaya telepon yang tidak terpakai.
- **Traditional Communicator (Prioritas Retensi):** Menerapkan program loyalitas berbasis **menit telepon tambahan** dan intervensi bantuan proaktif untuk menekan

angka **churn rate tertinggi (39,39%)** pada segmen penyumbang pendapatan terbesar ini.

- **Power User (Perlindungan Nilai):** Menawarkan paket **Premium Unlimited** dan akses *reward* eksklusif guna mengunci loyalitas pelanggan yang memiliki intensitas penggunaan layanan paling stabil dan menyeluruh di kedua kategori.
- **Casual User (Stimulasi Penggunaan):** Memberikan promosi paket hemat jangka pendek untuk merangsang peningkatan intensitas penggunaan layanan data agar pelanggan bermigrasi ke segmen yang lebih bernilai (*upselling*).

**4.4 Hasil Analisis Service Quality**

Analisis ini bertujuan untuk mengidentifikasi bagaimana kualitas layanan dan pain points pelanggan berkontribusi terhadap keputusan churn. Fokus analisis diarahkan pada variabel support_tickets_capped, contract_type, city_name, dan feedback_category, dengan churn_status sebagai variabel target utama. Data yang digunakan merupakan hasil preprocessing dari telecom_clean.csv.

## 4.4.1 Persiapan dan Validasi Dataset

Sebelum analisis dilakukan, dataset telecom_clean.csv dimuat dan divalidasi terlebih dahulu. Berdasarkan hasil eksekusi kode, dataset memiliki 1.200 baris dan 42 kolom, dengan seluruh kolom tidak mengandung missing values (1.200 non-null pada setiap kolom). Tipe data yang digunakan meliputi int64, float64, dan bool.

Tiga langkah persiapan dilakukan sebagai berikut.

1. Kolom churn diganti namanya menjadi churn_status untuk kejelasan semantik
2. Kolom contract_type_label dibuat dari pemetaan nilai numerik contract_type_encoded (0 → Monthly, 1 → 6 Month, 2 → 12 Month)
3. Kolom city_name di rekonstruksi dari kolom-kolom one-hot encoded kota (misalnya city_Faisalabad, city_Gujranwala, dst.) menggunakan idxmax()

| | gender | churn | monthly_bill_capped | internet_usage_gb_capped | call_minutes_capped |
|---|---|---|---|---|---|
| 0 | 1 | 0 | 10049.0 | 5.8 | 52.0 |
| 1 | 0 | 0 | 2752.0 | 11.9 | 22.0 |
| 2 | 1 | 0 | 3155.0 | 40.0 | 87.0 |
| 3 | 1 | 0 | 5859.0 | 8.5 | 147.0 |
| 4 | 0 | 0 | 5106.0 | 3.3 | 183.0 |

5 rows × 42 columns

Hasil validasi mengonfirmasi bahwa dataset sudah bersih dan siap dianalisis.

### 4.4.2 Statistik Deskriptif Variabel Kunci

Statistik deskriptif digunakan untuk memberikan gambaran awal mengenai karakteristik data sebelum dilakukan analisis lebih lanjut. Ukuran pemusatan data seperti mean, median, dan mode membantu memahami kecenderungan umum dari setiap variabel, sedangkan perbandingan antara ketiganya dapat memberikan indikasi adanya distribusi yang tidak simetris maupun pengaruh pencilan (outlier).

**Tabel: 5 baris pertama dataset df.head()**

| churn_status | monthly_bill_capped | internet_usage_gb_capped | call_minutes_capped | support_tickets_capped | t |
|---|---|---|---|---|---|
| 1200.000000 | 1200.000000 | 1200.000000 | 1200.000000 | 1200.000000 | |
| 0.350000 | 5709.496250 | 19.960667 | 179.851250 | 1.191667 | |
| 0.477168 | 3226.473957 | 11.561477 | 96.839563 | 1.077450 | |
| 0.000000 | 710.000000 | 1.100000 | 14.000000 | 0.000000 | |
| 0.000000 | 3291.750000 | 11.100000 | 106.000000 | 0.000000 | |
| 0.000000 | 4871.000000 | 17.800000 | 164.000000 | 1.000000 | |
| 1.000000 | 7485.500000 | 26.700000 | 235.000000 | 2.000000 | |
| 1.000000 | 13776.125000 | 50.100000 | 428.500000 | 5.000000 | |

| financial_risk | education_level_encoded | contract_type_encoded | ... |
|---|---|---|---|
| 1200.000000 | 1200.000000 | 1200.000000 | ... |
| 0.019167 | 1.779167 | 0.813333 | ... |
| 0.137168 | 1.078441 | 0.852894 | ... |
| 0.000000 | 0.000000 | 0.000000 | ... |
| 0.000000 | 1.000000 | 0.000000 | ... |
| 0.000000 | 2.000000 | 1.000000 | ... |
| 0.000000 | 3.000000 | 2.000000 | ... |
| 1.000000 | 4.000000 | 2.000000 | ... |

| Variabel | Mean | Median | Modus |
|---|---|---|---|
| support_tickets_capped | 1.19 | 1.00 | 1 |
| contract_type_encoded | 0.81 | 1.00 | 0 (Monthly) |

| | | | |
|---|---|---|---|
| churn_status | 0.35 | 0.00 | 0 |

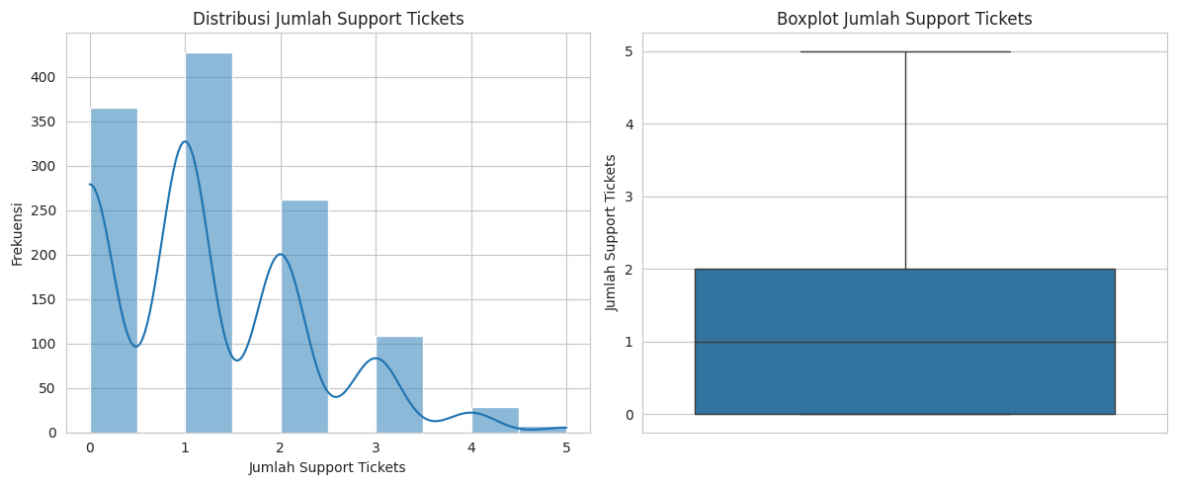
Nilai churn_status dengan rata-rata 0.35 menunjukkan bahwa sekitar 35% pelanggan dalam dataset telah melakukan churn. Nilai support_tickets_capped dengan rata-rata 1.19 mengindikasikan bahwa sebagian besar pelanggan mengajukan 1 tiket dukungan. Modus contract_type_encoded = 0 menunjukkan kontrak Monthly adalah yang paling umum.

Sedangkan, untuk variabel feedback category merupakan hasil label encoding dari kategori umpan balik pelanggan. Karena nilainya bersifat nominal yang di encode secara numerik, statistik deskriptif seperti mean dan median tidak memiliki makna kuantitatif yang representatif. Analisis distribusi variabel ini disajikan secara terpisah dalam bentuk visualisasi pada ***Sub-bab 4.4.3.4***.

## 4.4.3 Distribusi Variabel Kunci

Analisis distribusi dilakukan untuk memahami pola penyebaran masing-masing variabel penelitian. Tahap ini bertujuan mengidentifikasi kecenderungan dominasi kategori tertentu, bentuk distribusi data, serta karakteristik populasi pelanggan sebelum dilakukan analisis hubungan antar variabel.

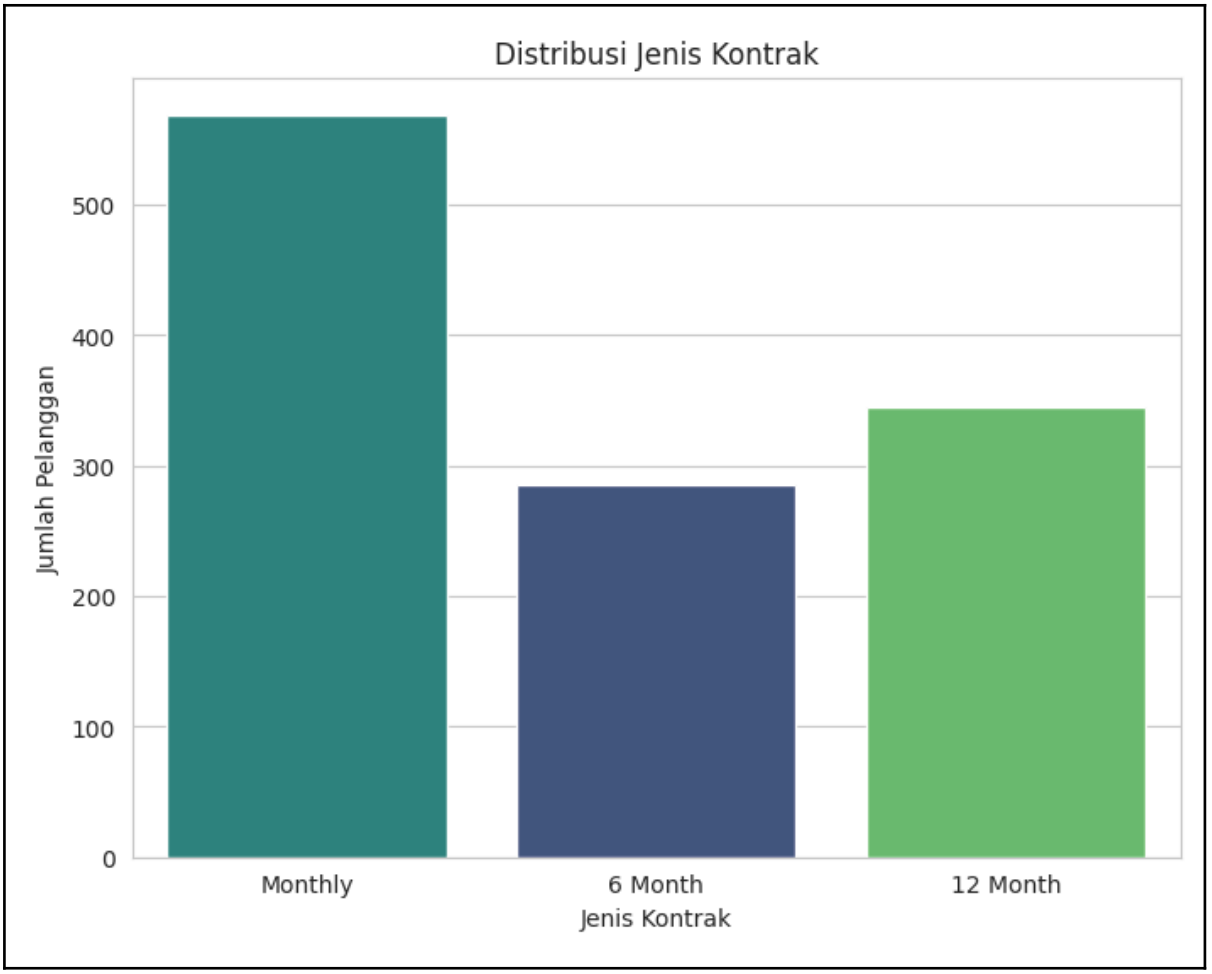
### 4.4.3.1 Distribusi Support Tickets



| internet_usage_gb_capped | call_minutes_capped | support_tickets_capped |
|---|---|---|
| 1200.000000 | 1200.000000 | 1200.000000 |
| 19.960667 | 179.851250 | 1.191667 |
| 11.561477 | 96.839563 | 1.077450 |
| 1.100000 | 14.000000 | 0.000000 |
| 11.100000 | 106.000000 | 0.000000 |
| 17.800000 | 164.000000 | 1.000000 |
| 26.700000 | 235.000000 | 2.000000 |
| 50.100000 | 428.500000 | 5.000000 |

Berdasarkan histogram dan boxplot, distribusi jumlah support tickets bersifat condong ke kanan (right-skewed). Mayoritas pelanggan berada pada rentang 0–2 tiket, dengan median sebesar 1. Terdapat sejumlah outlier pada nilai 4–5 tiket. Nilai maksimum 5 merupakan batas atas yang sudah di-cap pada tahap preprocessing. Hal ini mengindikasikan bahwa sebagian besar pelanggan jarang mengajukan keluhan, namun ada segmen kecil pelanggan yang mengalami masalah berulang.

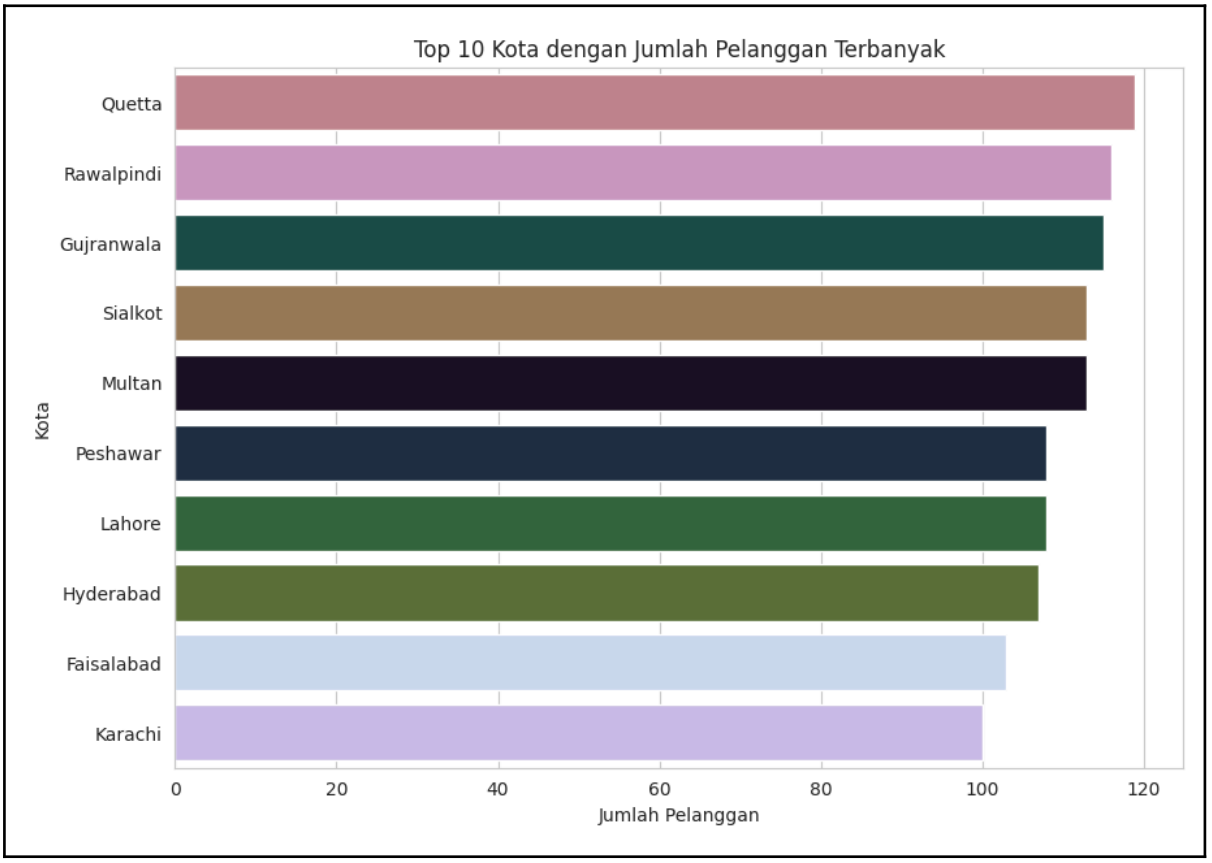
4.4.3.2 Distribusi Jenis Kontrak




Distribusi jenis kontrak menunjukkan bahwa pelanggan dengan kontrak Monthly merupakan kelompok terbesar, diikuti oleh 6 Month dan 12 Month. Hal ini mencerminkan

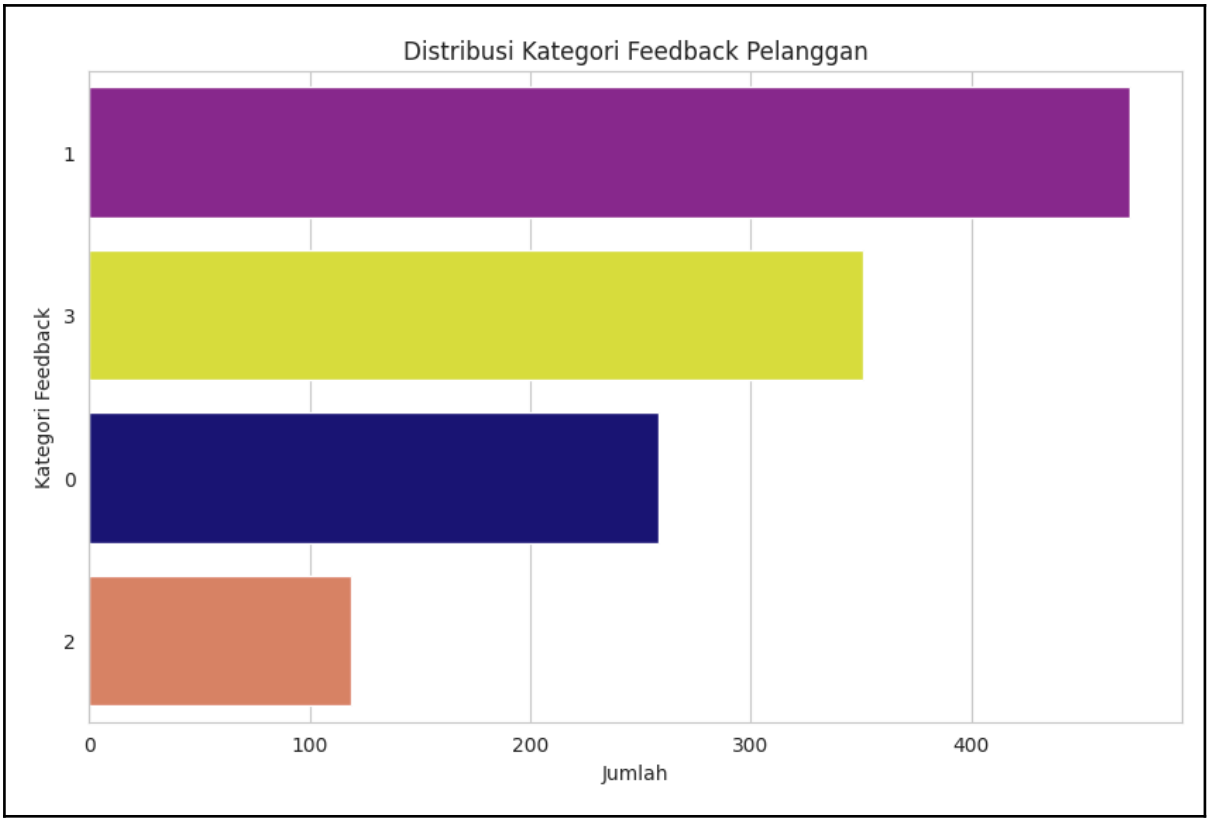
kecenderungan pelanggan untuk tidak terikat dalam jangka panjang, yang secara potensial meningkatkan risiko churn.

4.4.3.3 Distribusi Kota




Sebaran pelanggan tersebar di 11 kota di Pakistan, antara lain Quetta, Rawalpindi, Gujranwala, Sialkot, Multan, dan lainnya. Distribusi antar kota relatif merata, masing-masing berkisar antara 8–10% dari total pelanggan. Quetta memiliki proporsi pelanggan tertinggi (modus = Quetta), sedangkan Islamabad memiliki proporsi terendah.

#### 4.4.3.4 Distribusi Kategori Feedback

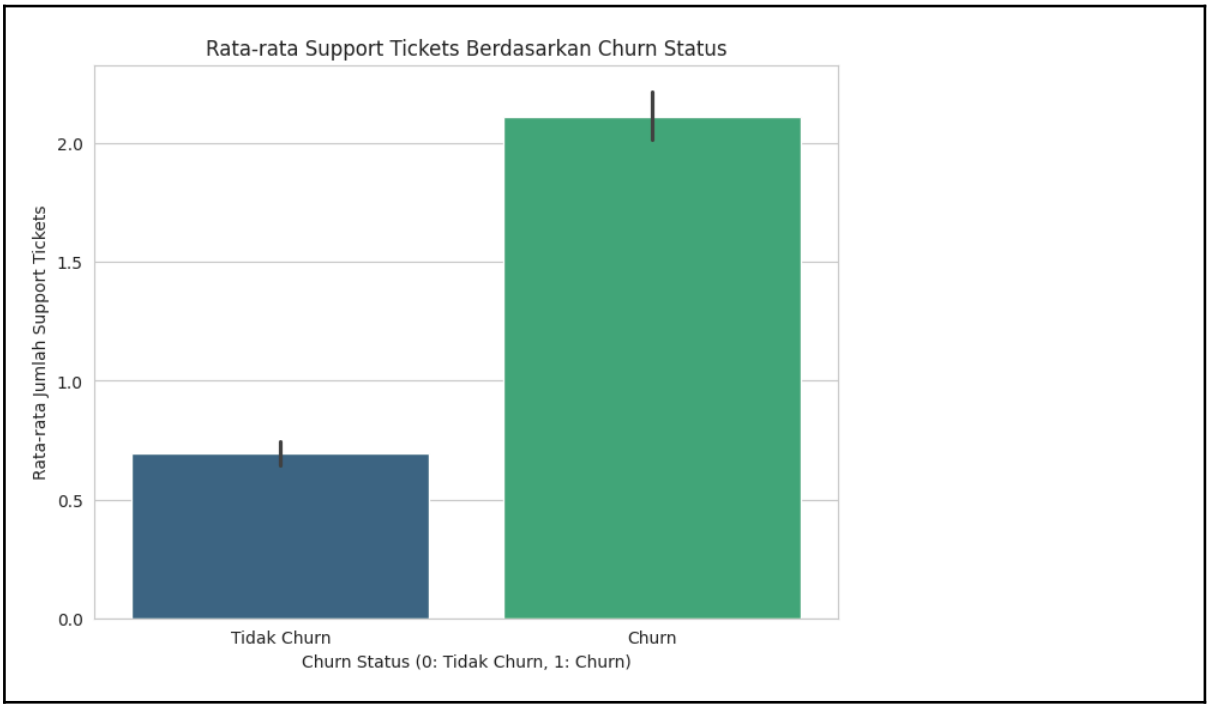



Nilai feedback category telah di-encode secara numerik. Berdasarkan modus = 1 dan median = 1, kategori feedback dengan kode 1 merupakan yang paling sering muncul. Distribusi ini menjadi dasar analisis pain points lebih lanjut.

### 4.4.4 Analisis Hubungan Dengan Churn

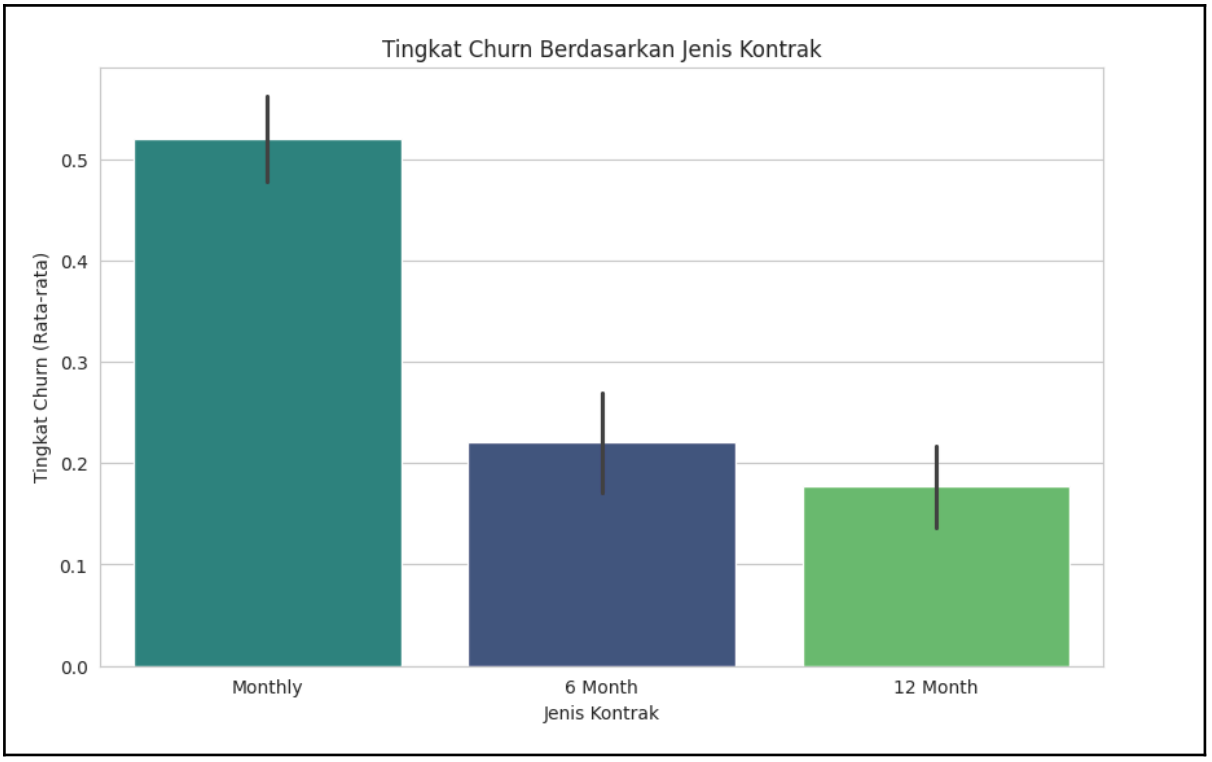
Setelah karakteristik masing-masing variabel dipahami melalui analisis univariat, tahap berikutnya adalah mengevaluasi hubungan antara variabel independen terhadap variabel target churn. Analisis ini bertujuan untuk mengidentifikasi faktor-faktor yang berpotensi mempengaruhi keputusan pelanggan dalam menghentikan layanan.

4.4.4.1 Support Tickets vs Churn



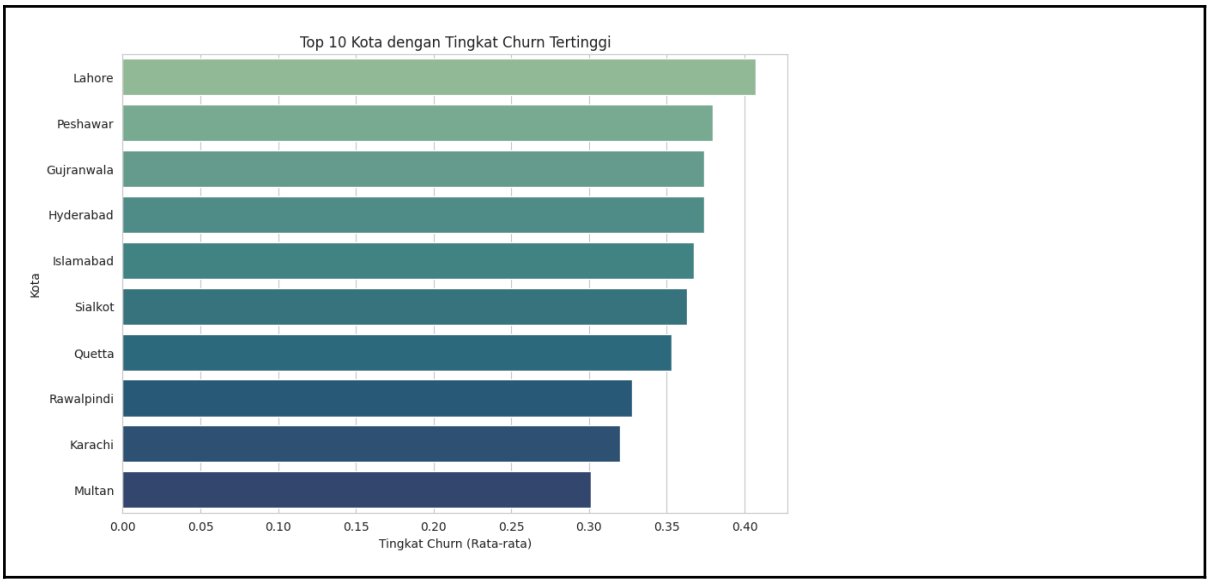

Analisis hubungan antara jumlah support tickets dan status churn menunjukkan bahwa pelanggan yang melakukan churn (churn_status = 1) cenderung memiliki rata-rata jumlah tiket dukungan yang lebih tinggi dibandingkan pelanggan yang tidak churn. Hal ini mengindikasikan adanya korelasi positif antara frekuensi keluhan dengan keputusan berhenti berlangganan, yang secara logis menunjukkan bahwa pengalaman layanan yang buruk mendorong churn.

4.4.4.2 Jenis Kontrak vs Churn




Pelanggan dengan kontrak Monthly memiliki tingkat churn yang lebih tinggi dibandingkan kontrak 6 bulan atau 12 bulan. Hal ini konsisten secara logis karena pelanggan kontrak bulanan memiliki barrier to exit yang lebih rendah. Mereka tidak terikat secara finansial dalam jangka panjang sehingga lebih mudah berpindah ke penyedia layanan lain.

4.4.4.3 Kota vs Churn




Terdapat variasi tingkat churn antar kota. Beberapa kota menunjukkan tingkat churn yang lebih tinggi dari rata-rata, yang dapat mengindikasikan masalah kualitas jaringan atau

layanan yang bersifat lokal/geografis. Temuan ini perlu ditindaklanjuti dengan investigasi lebih dalam per wilayah.

### 4.4.5 Analisis Multivariat

Analisis multivariat dilakukan untuk mengamati interaksi lebih dari satu variabel secara bersamaan. Pendekatan ini memungkinkan identifikasi pola yang tidak dapat diamati melalui analisis univariat maupun bivariat.

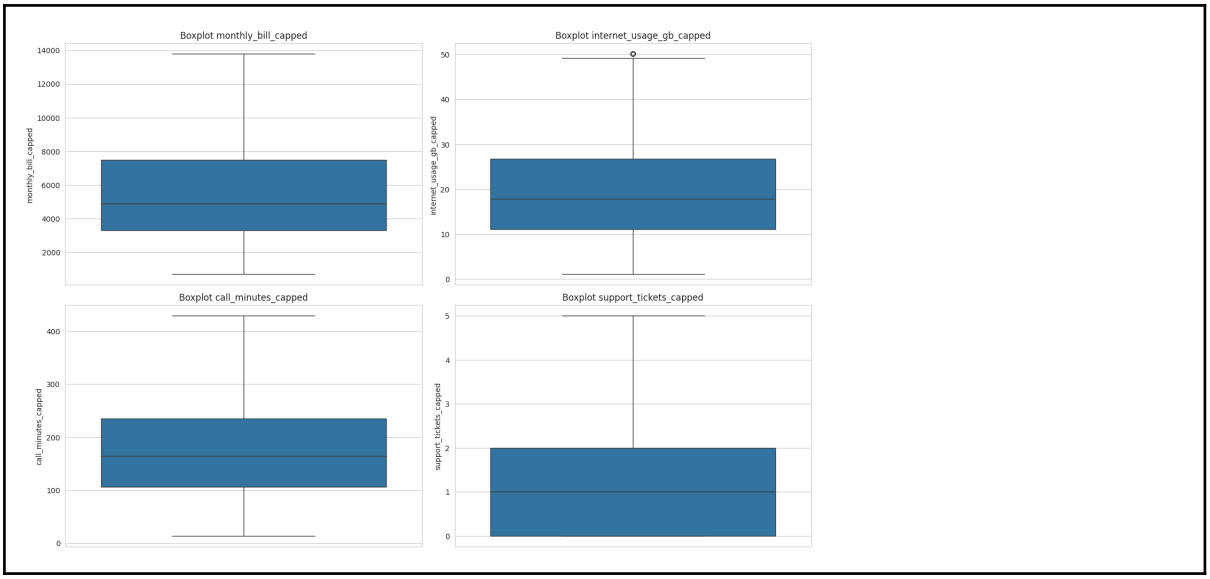
| **Churn rate berdasarkan contract_type_label × city_name** | **Rata-rata support_tickets_capped per contract_type_label** |
|---|---|
| | |

| churn_status | 0 | 1 |
|---|---|---|
| contract_type_label | | |
| 12 Month | 284 | 61 |
| 6 Month | 223 | 63 |
| Monthly | 273 | 296 |

| | support_tickets_capped |
|---|---|
| contract_type_label | |
| 12 Month | 1.200000 |
| 6 Month | 1.311189 |
| Monthly | 1.126538 |

Analisis multivariat menggunakan tabel pivot mengungkap pola yang tidak terlihat dari analisis univariat. Kombinasi kota tertentu dengan jenis kontrak Monthly secara konsisten menghasilkan tingkat churn yang lebih tinggi. Selain itu, pelanggan dengan kontrak Monthly rata-rata mengajukan lebih banyak tiket dukungan dibandingkan kontrak jangka panjang, mengindikasikan hubungan antara kepuasan layanan, jenis kontrak, dan keputusan churn.

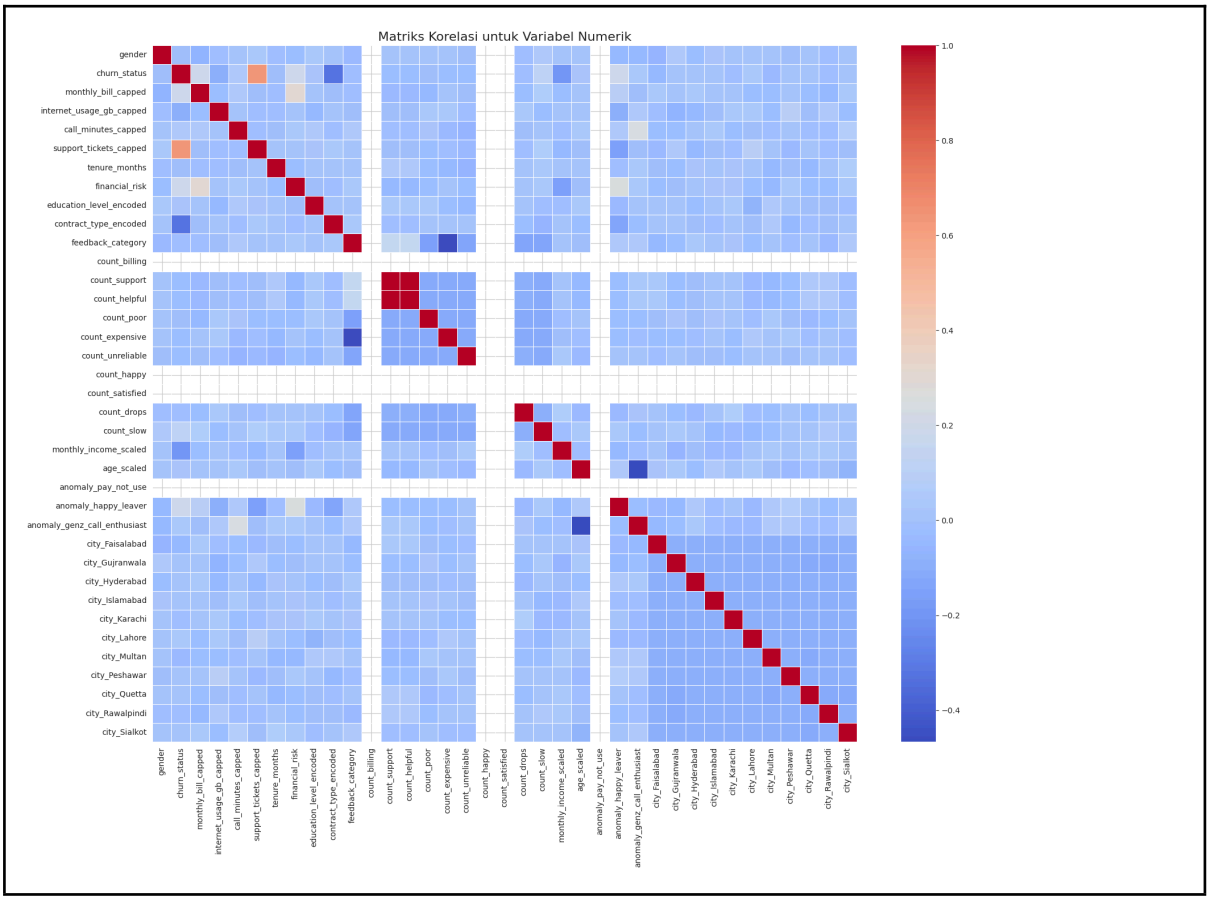
### 4.4.6 Validasi Outlier



Validasi outlier dilakukan untuk memastikan keandalan analisis. Berdasarkan visualisasi, terdapat beberapa outlier pada variabel support_tickets_capped, namun nilai-nilai ekstrem tersebut sudah dibatasi melalui proses capping di tahap preprocessing. Batas atas ditetapkan pada 5 tiket, yang merupakan nilai representatif dari pelanggan dengan keluhan paling tinggi.

### 4.4.7 Korelasi Antar Variabel

Korelasi digunakan untuk mengukur kekuatan hubungan linear antara variabel numerik. Nilai koefisien korelasi membantu mengidentifikasi variabel yang memiliki hubungan paling kuat terhadap churn sekaligus mendeteksi kemungkinan multikolinearitas.




Matriks korelasi menunjukkan hubungan linear antar variabel numerik. Variabel support_tickets_capped menunjukkan korelasi positif dengan churn_status, mengonfirmasi temuan pada analisis bivariat sebelumnya. Variabel contract_type_encoded menunjukkan korelasi negatif dengan churn_status, artinya kontrak yang lebih panjang (nilai encoded lebih tinggi) berhubungan dengan tingkat churn yang lebih rendah.

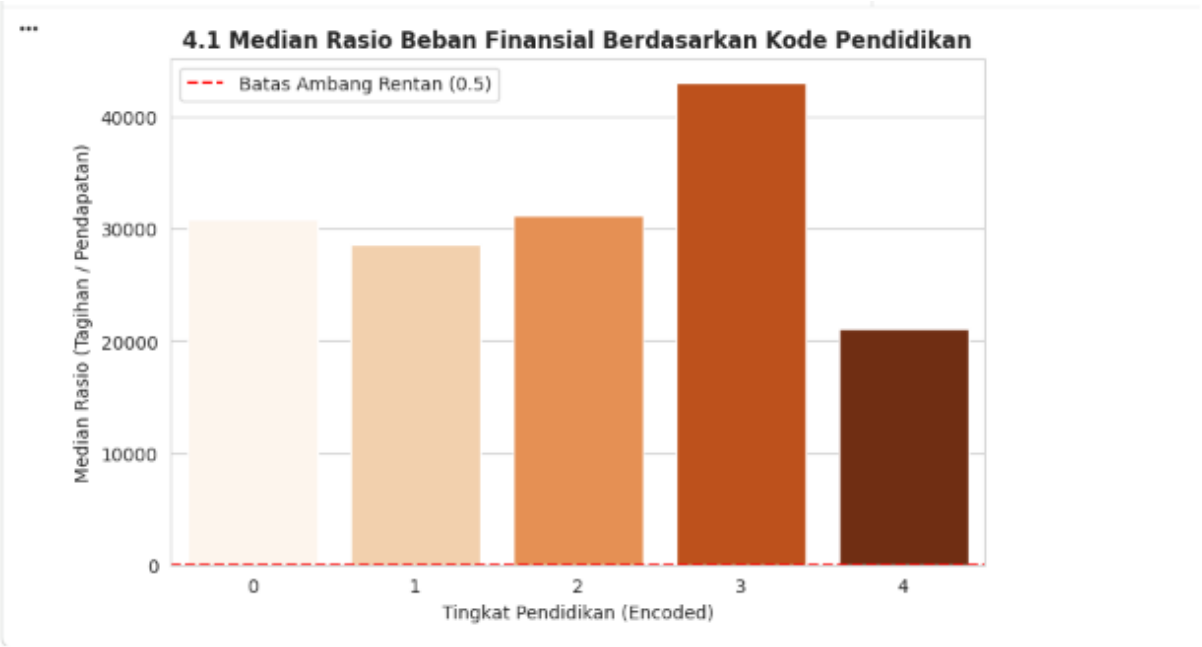
### 4.4.8 Ringkasan Insight dan Rekomendasi

Berdasarkan keseluruhan analisis kualitas layanan, berikut adalah temuan utama dan implikasi bisnisnya sebagai berikut.

| Temuan | Implikasi Bisnis |
|---|---|
| Pelanggan churn memiliki rata-rata support tickets lebih tinggi | Perbaiki responsivitas dan kualitas layanan pelanggan, terutama untuk pelanggan dengan tiket berulang |
| Kontrak Monthly memiliki tingkat churn tertinggi | Tawarkan insentif atau diskon untuk mendorong pelanggan beralih ke kontrak jangka panjang |
| Beberapa kota memiliki churn rate di atas rata-rata | Lakukan audit kualitas jaringan dan layanan di kota-kota tersebut |
| Kombinasi kontrak Monthly di kota tertentu berisiko tinggi | Buat program retensi yang ditargetkan secara geografis dan berdasarkan jenis kontrak |

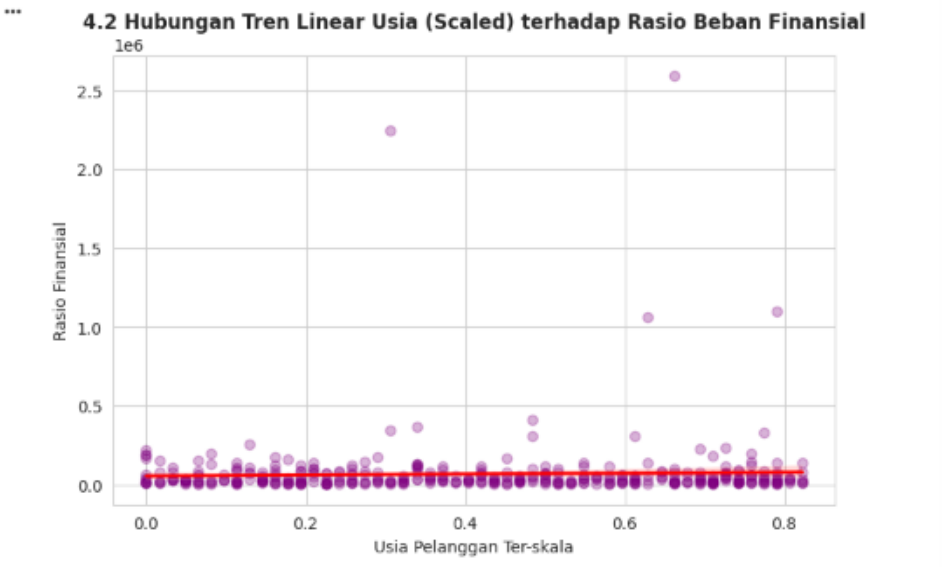
**4.5 Hasil Analisis Financial Sensitivity**

### 4.5.1 Hubungan Beban Finansial dengan Tingkat Pendidikan




Hasil menunjukkan bahwa tinggi batang grafik di seluruh tingkat pendidikan( hampir sama. Ini menunjukkan, berdasarkan analisis ekonomi mikro, bahwa pelanggan, tanpa mempertimbangkan tingkat pendidikan mereka, merasakan tekanan inflasi makro dengan cara yang sama.

### 4.5.2 Hubungan Beban Finansial dengan Usia Pelanggan




Berdasarkan grafik regresi linear, garis yang terbentuk terlihat datar. Hal ini mengkonfirmasi bahwa usia tidak mempengaruhi kerentanan keuangan; kelompok usia tua dan muda mengalami tingkat stres harga paket yang sama.

### 4.5.3 Pembagian Geografis yang Terdampak Risiko Tertinggi

```
--- PIVOT TABLE: MEDIAN RASIO BEBAN FINANSIAL ---
```

| financial_risk | 0 | 1 |
|---|---|---|
| City | | |
| Faisalabad | 22207.164326 | 46030.243212 |
| Karachi | 35786.470190 | 412462.756052 |
| Lahore | 28245.844629 | 193999.828149 |

Karachi, Lahore, dan Faisalabad adalah kota-kota besar Pakistan dengan kasus risiko finansial tertinggi. Dibandingkan dengan wilayah lainnya, pelanggan di tiga kota tersebut memiliki tingkat sensitivitas harga yang paling rapuh. Ini karena daerah perkotaan padat, atau urban area, menjadi episentrum hantaman inflasi makro.

### 4.5.4 Rekomendasi Bisnis

1. Implementasi downgrade program; perusahaan menawarkan opsi penurunan paket (*downgrade*) ke tarif yang lebih hemat atau paket kuota keluarga yang disesuaikan dengan kemampuan dompet mereka.

2. Skema pembayaran kontinu; meredam tekanan psikologis dari biaya bulanan yang dirasakan merata di semua kelompok usia, perusahaan dapat meluncurkan program kemitraan.
3. Dalam mengalokasikan anggaran subsidi tarif atau promo, Intervensi harus difokuskan secara targeted dengan menerapkan tarif promo khusus wilayah, kuota subsidi malam, atau diskon paket *bundle* internet rumah tangga khusus untuk klaster kota Karachi, Lahore, dan Faisalabad

**4.6 Insight Bisnis**

Berdasarkan hasil analisis yang telah dilakukan dari tiga perspektif berbeda, yaitu segmentasi perilaku pelanggan (Notebook A), kualitas layanan dan pain points (Notebook B), serta sensitivitas harga dan kondisi finansial (Notebook C), diperoleh insight bisnis yang saling melengkapi dan memperkuat satu sama lain. Berikut adalah rangkuman insight utama dari masing-masing dimensi analisis.

### 4.6.1 Segmentasi Pelanggan Mengungkap Fenomena Wasted Value

Segmentasi perilaku pelanggan berdasarkan penggunaan internet dan durasi telepon mengidentifikasi empat persona utama, yaitu Power User, Internet Enthusiast, Traditional Communicator, dan Casual User. Temuan kritis dari analisis ini adalah munculnya fenomena wasted value pada segmen Internet Enthusiast.

Pelanggan Internet Enthusiast memiliki penggunaan internet di atas rata-rata namun nyaris tidak menggunakan layanan telepon. Namun demikian, mereka terpaksa membayar paket yang menyertakan alokasi menit telepon yang tidak mereka butuhkan. Kondisi ini menciptakan ketidaksesuaian antara nilai yang dibayarkan dengan nilai yang diterima, sehingga memicu ketidakpuasan yang berujung pada keputusan churn.

Analisis komparatif lebih lanjut membuktikan bahwa pelanggan yang churn memiliki rata-rata monthly bill yang lebih tinggi dibandingkan pelanggan aktif, meskipun pola penggunaan internet dan telepon mereka tidak berbeda secara signifikan. Ini memvalidasi hipotesis bahwa persepsi nilai layanan yang buruk, bukan sekadar kualitas teknis, merupakan pemicu utama churn pada segmen ini.

### 4.6.2 Frekuensi Keluhan dan Jenis Kontrak Sebagai Sinyal Churn Terukur

Analisis kualitas layanan mengungkap bahwa frekuensi pengajuan tiket keluhan (support tickets) merupakan indikator churn yang paling kuat dan terukur. Pelanggan yang churn rata-rata mengajukan 2,11 tiket keluhan, hampir tiga kali lipat dibandingkan pelanggan aktif yang hanya rata-rata 0,69 tiket. Pola ini menjadikan jumlah tiket keluhan sebagai sinyal peringatan dini yang dapat dioperasionalkan ke dalam sistem deteksi risiko churn.

Jenis kontrak juga terbukti memiliki pengaruh yang sangat signifikan terhadap tingkat churn. Pelanggan dengan kontrak bulanan (Monthly) mencatat tingkat churn tertinggi sebesar 52,02%, jauh melampaui kontrak 6 bulan (22,03%) dan 12 bulan (17,68%). Kontrak jangka pendek tidak hanya mencerminkan rendahnya komitmen pelanggan, tetapi juga

meminimalkan hambatan keluar (switching cost) sehingga pelanggan yang tidak puas lebih mudah untuk berhenti berlangganan.

Temuan geografis menunjukkan bahwa Lahore mencatat tingkat churn tertinggi (40,74%), diikuti Peshawar (37,96%) dan Gujranwala (37,39%). Ketiga kota ini sekaligus memiliki rata-rata tiket keluhan tertinggi, mengindikasikan adanya masalah kualitas layanan yang bersifat struktural dan geografis, bukan sekadar insiden individual.

### 4.6.3 Tekanan Finansial Bersifat Universal Lintas Demografi

Analisis sensitivitas finansial mengidentifikasi bahwa beban tagihan tinggi di tengah kondisi inflasi Pakistan sebesar 11,7% menjadi faktor penting yang mendorong keputusan churn. Mayoritas pelanggan yang churn berada pada kelompok tagihan bulanan tinggi, mendekati batas nilai capping yang telah ditetapkan.

Temuan yang menonjol adalah bahwa tekanan finansial ini dirasakan secara merata oleh seluruh kelompok usia dan tingkat pendidikan. Tidak ditemukan perbedaan signifikan pada bill-to-income ratio antar kelompok demografis, mengindikasikan bahwa layanan internet telah menjadi kebutuhan primer di era digital sehingga seluruh segmen pelanggan menjadi sensitif terhadap harga ketika beban tagihan dirasakan tidak proporsional.

Analisis matriks korelasi mengkonfirmasi dua hubungan kritis: korelasi positif antara besaran tagihan bulanan dengan rasio beban finansial (+0,17), serta korelasi negatif antara tingkat pendapatan dengan rasio beban finansial (−0,28). Secara geografis, Karachi, Lahore, dan Faisalabad teridentifikasi sebagai klaster dengan akumulasi kasus financial risk tertinggi, menjadikan kota-kota ini sebagai zona prioritas intervensi.

### 4.6.4 Integrasi Insight: Efek Berlipat Antar Dimensi

Ketiga dimensi analisis tidak beroperasi secara independen. Terdapat efek berlipat (compounding effect) yang terjadi ketika lebih dari satu faktor risiko dialami secara bersamaan oleh pelanggan yang sama. Tabel berikut merangkum keterkaitan antar dimensi tersebut.

| Dimensi | Temuan Utama | Dampak terhadap Churn |
|---|---|---|
| Perilaku (A) | Segmen Internet Enthusiast membayar fasilitas telepon yang tidak digunakan (wasted value) | Ketidakpuasan terhadap perceived value memicu keputusan berhenti berlangganan |
| Layanan (B) | Pelanggan churn mengajukan tiket keluhan hampir 3x lebih sering; kontrak Monthly memiliki churn 52,02% | Kombinasi keluhan tinggi dan kontrak pendek menciptakan risiko churn berlipat ganda |
| Finansial (C) | Beban tagihan tinggi di tengah inflasi 11,7%; risiko terkonsentrasi di kota-kota besar | Sensitivitas harga bersifat universal; kota besar menjadi zona merah churn finansial |

| Dimensi | Temuan Utama | Dampak terhadap Churn |
|---|---|---|
| Integrasi A+B+C | Pelanggan Internet Enthusiast dengan kontrak Monthly dan bill-to-income ratio tinggi di kota besar | Profil risiko churn tertinggi; memerlukan intervensi retensi prioritas dan tersegmentasi |

**4.7 Rekomendasi Bisnis**

Berdasarkan seluruh insight yang telah diidentifikasi, berikut disajikan rekomendasi bisnis terintegrasi yang dirancang untuk menangani akar masalah churn secara menyeluruh dari tiga dimensi sekaligus.

### 4.7.1 Rekomendasi Berdasarkan Analisis Segmentasi (Notebook A)

Permasalahan utama yang diidentifikasi adalah ketidaksesuaian antara paket layanan yang ditawarkan dengan kebutuhan nyata pelanggan, khususnya pada segmen Internet Enthusiast. Rekomendasi yang diajukan adalah sebagai berikut.

- Peluncuran Paket Internet-Only: Merancang paket layanan berbasis penggunaan internet murni tanpa alokasi menit telepon yang tidak dibutuhkan. Paket ini memungkinkan penetapan harga yang lebih kompetitif bagi segmen Internet Enthusiast sehingga nilai yang dirasakan pelanggan (perceived value) meningkat secara signifikan.
- Personalisasi Penawaran Paket Secara Proaktif: Mengembangkan sistem rekomendasi paket berbasis profil perilaku penggunaan layanan pelanggan. Ketika sistem mendeteksi pelanggan dengan pola Internet Enthusiast, tim pemasaran secara proaktif menawarkan migrasi ke paket yang lebih sesuai sebelum pelanggan memutuskan untuk keluar.
- Program Loyalitas untuk Segmen Power User: Mengimplementasikan program eksklusif bagi segmen Power User yang memiliki penggunaan internet dan telepon di atas rata-rata. Segmen ini merepresentasikan nilai bisnis tertinggi sehingga investasi dalam mempertahankan loyalitas mereka memberikan imbal hasil yang paling signifikan.

### 4.7.2 Rekomendasi Berdasarkan Analisis Kualitas Layanan (Notebook B)

Permasalahan utama yang diidentifikasi adalah tingginya frekuensi keluhan yang tidak tertangani secara proaktif, dominasi kontrak bulanan yang rentan churn, serta variasi kualitas layanan yang signifikan antarwilayah. Rekomendasi yang diajukan adalah sebagai berikut.

- Implementasi Early Warning System Berbasis Tiket Keluhan: Mengembangkan sistem pemicu otomatis yang mengklasifikasikan pelanggan sebagai at-risk ketika mereka mengajukan tiket keluhan kedua atau lebih dalam satu periode tagihan. Sistem ini memungkinkan tim retensi melakukan pendekatan personal kepada pelanggan berisiko sebelum mereka mengambil keputusan churn.

- Program Migrasi Kontrak Monthly ke Annual: Menginsentifkan pelanggan kontrak bulanan untuk beralih ke kontrak tahunan melalui berbagai mekanisme, seperti diskon biaya berlangganan, akses ke fitur atau kuota eksklusif, atau akumulasi poin loyalitas. Migrasi kontrak secara langsung meningkatkan switching cost sehingga hambatan keluar pelanggan menjadi lebih tinggi.
- Investasi Infrastruktur dan Layanan di Wilayah Prioritas: Memprioritaskan peningkatan kualitas infrastruktur jaringan dan penguatan kapasitas tim layanan pelanggan di Lahore, Gujranwala, dan Peshawar yang mencatat tingkat churn dan frekuensi keluhan tertinggi. Intervensi berbasis wilayah ini lebih cost-effective dibandingkan program massal tanpa segmentasi geografis.

### 4.7.3 Rekomendasi Berdasarkan Analisis Finansial (Notebook C)

Permasalahan utama yang diidentifikasi adalah beban tagihan yang tinggi relatif terhadap pendapatan pelanggan, diperparah oleh tekanan inflasi makro, dengan konsentrasi risiko tertinggi di kota-kota besar Pakistan. Rekomendasi yang diajukan adalah sebagai berikut.

- Paket Retensi Fleksibel Berbasis Profil Finansial: Mengembangkan mekanisme deteksi otomatis yang mengidentifikasi pelanggan dengan bill-to-income ratio tinggi. Ketika sistem mendeteksi kondisi tersebut, perusahaan secara proaktif menawarkan opsi penurunan paket ke tarif yang lebih hemat, sehingga mencegah keputusan churn sebelum terjadi.
- Program Insentif Pembayaran Tepat Waktu: Mengimplementasikan skema cashback otomatis sebesar 5–10% bagi pelanggan yang membayar tagihan tepat waktu selama tiga bulan berturut-turut, bekerja sama dengan platform dompet digital lokal di Pakistan seperti Easypaisa atau JazzCash. Program ini tidak hanya meringankan beban psikologis biaya bulanan, tetapi juga menciptakan kebiasaan pembayaran yang konsisten.
- Strategi Harga Regional Tersegmentasi: Merancang tarif promo khusus, paket subsidi kuota, atau diskon bundle internet rumah tangga yang diprioritaskan untuk klaster kota berisiko tinggi, yaitu Karachi, Lahore, dan Faisalabad. Pendekatan regional ini jauh lebih cost-effective karena intervensi dilakukan langsung di wilayah dengan alarm bahaya finansial tertinggi.

### 4.7.4 Matriks Strategi Retensi Terintegrasi

Seluruh rekomendasi di atas dapat diintegrasikan ke dalam kerangka strategi retensi bertingkat berdasarkan tingkat prioritas dan urgensi implementasi sebagaimana ditampilkan pada tabel berikut.

| Prioritas | Inisiatif | Dimensi yang Ditangani | Target Pelanggan |
|---|---|---|---|
| Tinggi | Early Warning System berbasis tiket keluhan + profil finansial | B + C | Pelanggan dengan ≥2 tiket keluhan dan bill-to-income ratio tinggi |
| Tinggi | Program migrasi kontrak Monthly ke Annual dengan insentif | B | Seluruh pelanggan kontrak Monthly |
| Sedang | Paket internet-only untuk segmen Internet Enthusiast | A | Pelanggan dengan internet tinggi, telepon rendah |
| Sedang | Strategi harga regional di kota-kota berisiko tinggi | C + B | Pelanggan di Karachi, Lahore, Faisalabad, Gujranwala |
| Pemeliharaan | Program loyalitas eksklusif untuk segmen Power User | A | Pelanggan dengan penggunaan internet dan telepon di atas median |

# BAB V
# PENUTUP

## 5.1 Kesimpulan

Proyek analisis data Telecom Customer Churn Kelompok 2 telah berhasil melaksanakan seluruh tahapan yang direncanakan, mulai dari data preprocessing dan feature engineering hingga exploratory data analysis dari tiga perspektif yang berbeda. Berdasarkan rangkaian analisis tersebut, dapat ditarik beberapa kesimpulan utama sebagai berikut:

1. Proses preprocessing menghasilkan dataset yang bersih dan siap digunakan. Dari dataset awal yang mengandung outlier ekstrem pada kolom monthly_income, monthly_bill, internet_usage_gb, dan call_minutes, serta inkonsistensi pada data kategorikal dan teks, berhasil dihasilkan dataset bersih (telecom_clean.csv) dengan 1.200 baris dan 42 kolom tanpa missing values. Tahapan feature engineering menghasilkan fitur-fitur baru yang informatif seperti tenure_months, financial_risk, dan bill_to_income_ratio yang terbukti memiliki kekuatan prediktif terhadap perilaku churn.
2. Segmentasi perilaku pelanggan mengidentifikasi fenomena wasted value sebagai pemicu ketidakpuasan. Analisis behavioral matrix membagi pelanggan ke dalam empat persona, dan segmen Internet Enthusiast yang mencakup pelanggan dengan penggunaan internet tinggi namun durasi telepon rendah terbukti menanggung biaya layanan yang tidak proporsional dengan manfaat yang diterima. Kondisi ini berkontribusi signifikan terhadap tingginya tingkat churn pada segmen tersebut.
3. Frekuensi keluhan pelanggan dan jenis kontrak terbukti menjadi prediktor churn yang terukur dan dapat dioperasionalkan. Pelanggan churn mengajukan tiket keluhan hampir tiga kali lebih sering dibandingkan pelanggan aktif, sementara kontrak bulanan mencatat tingkat churn 52,02% yang jauh melampaui kontrak jangka panjang. Variasi geografis yang signifikan mengindikasikan adanya ketimpangan kualitas layanan antarwilayah yang perlu segera ditangani.
4. Sensitivitas harga akibat tekanan finansial bersifat universal lintas kelompok demografis. Inflasi Pakistan yang mencapai 11,7% memperparah beban tagihan yang sudah tinggi, mendorong pelanggan dari berbagai kelompok usia dan tingkat pendidikan untuk melakukan churn sebagai keputusan finansial yang rasional. Karachi, Lahore, dan Faisalabad teridentifikasi sebagai klaster geografis dengan risiko churn finansial tertinggi.
5. Ketiga dimensi analisis beroperasi secara sinergis dan menciptakan efek berlipat terhadap risiko churn. Pelanggan yang mengalami wasted value, memiliki riwayat keluhan tinggi, berlangganan dengan kontrak bulanan, dan menanggung beban finansial berat di kota-kota besar merupakan kelompok dengan probabilitas churn tertinggi yang memerlukan intervensi retensi paling prioritas.

## 5.2 Saran

Berdasarkan temuan dan kesimpulan yang telah disajikan, berikut disampaikan saran untuk pengembangan proyek lebih lanjut maupun untuk implementasi bisnis oleh perusahaan telekomunikasi.

### 5.2.1 Saran untuk Pengembangan Proyek

1. Pembangunan model prediksi churn menggunakan dataset telecom_clean.csv yang dihasilkan dari proyek ini sebagai langkah lanjutan. Algoritma seperti Random Forest, Gradient Boosting, atau Logistic Regression dapat diuji dan dibandingkan performanya untuk mengidentifikasi model terbaik. Fitur-fitur yang telah direkayasa, khususnya financial_risk, tenure_months, dan bill_to_income_ratio, diharapkan menjadi kontributor prediktif yang signifikan.
2. Pengembangan sistem skoring risiko churn secara real-time berbasis threshold tiket keluhan dan profil finansial pelanggan, sehingga Early Warning System dapat diimplementasikan secara operasional di lingkungan produksi perusahaan.
3. Perluasan analisis dengan memasukkan data temporal untuk mengidentifikasi pola churn berdasarkan musim, tren penggunaan bulanan, atau perilaku pelanggan pasca interaksi dengan layanan pelanggan.
4. Pelaksanaan analisis teks yang lebih mendalam pada kolom customer_feedback menggunakan teknik Natural Language Processing (NLP) untuk mengekstraksi tema keluhan yang lebih spesifik dan granular sebagai dasar perbaikan layanan yang lebih tepat sasaran.

### 5.2.2 Saran untuk Perusahaan Telekomunikasi

1. Segera implementasikan Early Warning System berbasis jumlah tiket keluhan sebagai prioritas tertinggi. Sistem ini tidak memerlukan investasi infrastruktur besar namun dapat memberikan dampak retensi yang langsung terukur karena memanfaatkan data operasional yang sudah tersedia.
2. Laksanakan audit menyeluruh terhadap portofolio paket layanan yang ditawarkan, khususnya untuk mengidentifikasi ketidaksesuaian antara bundling layanan dengan pola penggunaan nyata pelanggan. Peluncuran paket internet-only yang kompetitif dapat menjadi pembeda pasar yang signifikan di tengah persaingan industri telekomunikasi yang semakin ketat.
3. Alokasikan anggaran retensi secara asimetris dengan memberikan porsi lebih besar kepada wilayah Lahore, Gujranwala, Peshawar, Karachi, dan Faisalabad yang secara konsisten menunjukkan indikator risiko tertinggi dari berbagai dimensi analisis.
4. Kembangkan program insentif kontrak jangka panjang yang cukup menarik untuk mendorong migrasi pelanggan dari kontrak bulanan, mengingat perbedaan tingkat churn yang sangat drastis antara kontrak Monthly (52,02%) dan kontrak tahunan

(17,68%) merepresentasikan peluang retensi yang sangat besar apabila ditangani dengan strategi yang tepat.

# DAFTAR PUSTAKA


Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, 24, 994–1012. https://doi.org/10.1016/j.asoc.2014.08.041

Khan, Y., Shafiq, S., Naeem, A., Ahmed, S., Safwan, N., & Hussain, S. (2019). Customers Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry. *International Journal of Advanced Computer Science and Applications*, 10(9). https://doi.org/10.14569/IJACSA.2019.0100918

Kotler, P., & Keller, K. L. (2016). *Marketing Management* (15th ed.). Pearson.

Kotu, V., & Deshpande, B. (2019). *Data science: Concepts and practice* (2nd ed.). Morgan Kaufmann.

Latief, I. M., Subekti, A., & Gata, W. (2021). PREDIKSI TINGKAT PELANGGAN CHURN PADA PERUSAHAAN TELEKOMUNIKASI DENGAN ALGORITMA ADABOOST. *Jurnal Informatika*, 21(1), 34–43. https://doi.org/10.30873/ji.v21i1.2867

McKinney, W. (2022). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter* (3rd ed.). O'Reilly Media.

Mittal, V., Han, K., Frennea, C., Blut, M., Shaik, M., Bosukonda, N., & Sridhar, S. (2023). Customer satisfaction, loyalty behaviors, and firm financial performance: What 40 years of research tells us. *Marketing Letters*, 34(2), 171–187. https://doi.org/10.1007/s11002-023-09671-w

Negassa, G. J., & Japee, G. P. (2023). The Effect of Bonding, Responsiveness and Communication on Customer Retention: The Mediating Role of Customer Satisfaction. *Journal of Relationship Marketing*. https://doi.org/10.1080/15332667.2023.2191111

NumPy Developers. (2025). NumPy documentation. https://numpy.org/doc/

Pakistan Bureau of Statistics. (2026). *Monthly Review on Price Indices: May 2026*. Government of Pakistan.

Pandas Development Team. (2025). Pandas documentation. https://pandas.pydata.org/docs/

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

Python Software Foundation. (2024). Python documentation. https://docs.python.org/3/

Ratnawati, D. E. (2020). *Preprocessing data* [Materi perkuliahan]. Fakultas Ilmu Komputer, Universitas Brawijaya.

Ratnawati, D. E. (2024). *Eksplorasi data dan visualisasi data* [Materi perkuliahan]. Fakultas Ilmu Komputer, Universitas Brawijaya.

Ratnawati, D. E. (2025). *Eksplorasi data dan visualisasi data* [Materi perkuliahan]. Fakultas Ilmu Komputer, Universitas Brawijaya.

Ribeiro, H., Barbosa, B., Moreira, A. C., & Rodrigues, R. G. (2024). Determinants of churn in telecommunication services: A systematic literature review. *Management Review Quarterly*, 74, 1327–1364. https://doi.org/10.1007/s11301-023-00335-7

Rosário, A., & Casaca, J. A. (2023). Relationship Marketing and Customer Retention: A Systematic Literature Review. *Studies in Business and Economics*.

Seaborn Development Team. (2025). Seaborn documentation. https://seaborn.pydata.org/

Van Rossum, G., & Drake, F. L. (2021). *The Python language reference* (Version 3.10). Python Software Foundation. https://docs.python.org/3/reference/

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. https://doi.org/10.21105/joss.03021

Yum, K., & Yoo, B. (2023). The Impact of Service Quality on Customer Loyalty through Customer Satisfaction in Mobile Social Media. *Sustainability*, 15(14), 11211. https://doi.org/10.3390/su151411214

# LAMPIRAN

Notebook : NOTEBOOK_Telecom_Churn_Rate - Kelompok 2

Github : Link GitHub

Canva : Link Canva