1.0-hans

July 16, 2025

1 Improving Customer Churn Rate in Telecom: A Machine Learning Approach

by JCDS2602 - Alpha Team - Abe, Alfi, Hans ## Section 1. Business Understanding

1.0.1 1.1 Background

Perusahaan XYZ merupakan penyedia layanan telekomunikasi terkemuka yang dikenal karena pendekatannya yang inovatif dan berfokus pada pelanggan. Beroperasi di pasar yang sangat kompetitif, XYZ menawarkan beragam layanan seperti telepon seluler, internet broadband, dan layanan digital, yang ditujukan untuk pelanggan individu maupun bisnis. Meskipun memiliki posisi yang kuat di pasar, XYZ menghadapi tantangan besar terkait churn pelanggan, yaitu hilangnya pelanggan yang dapat berdampak langsung pada pendapatan dan pertumbuhan perusahaan. Dengan memanfaatkan analisis data dan wawasan pelanggan, XYZ berupaya mengurangi tingkat churn serta meningkatkan retensi pelanggan, guna menjaga keberlanjutan dan profitabilitas jangka panjang di industri telekomunikasi yang dinamis.

Churn pelanggan di industri telekomunikasi mengacu pada tingkat pelanggan yang berhenti menggunakan layanan dalam periode waktu tertentu. Fenomena ini sangat krusial karena berdampak langsung terhadap pendapatan dan laba perusahaan. Alasan pelanggan berhenti bisa bermacam-macam, mulai dari ketidakpuasan terhadap kualitas layanan, harga yang tidak kompetitif, layanan pelanggan yang buruk, hingga penawaran menarik dari pesaing. Memahami dan mengelola churn sangat penting agar perusahaan dapat menjaga basis pelanggan yang stabil serta kesehatan finansialnya. Dalam konteks proyek ini, seorang pelanggan didefinisikan sebagai “churn” jika mereka tidak memperpanjang atau menghentikan layanan berlangganan mereka dalam periode tertentu.

1.0.2 1.2 Gap Analysis

Fokus utama dari proyek ini adalah mengubah pendekatan perusahaan dalam menangani churn pelanggan, dari yang sebelumnya bersifat reaktif menjadi strategi mitigasi risiko yang proaktif. Saat ini, perusahaan belum memiliki mekanisme yang efektif untuk mengidentifikasi pelanggan mana yang berisiko tinggi berhenti berlangganan. Akibatnya, upaya retensi sering kali dilakukan terlambat atau tidak tepat sasaran, seperti memberikan penawaran secara massal yang justru menyebabkan pemborosan anggaran.

Solusi yang diusulkan adalah pengembangan strategi mitigasi churn yang cerdas, dimulai dengan identifikasi dini terhadap pelanggan berisiko tinggi. Setelah pelanggan ini dikenali, tindakan preventif seperti pemberian penawaran khusus senilai $139 dapat dialokasikan secara lebih efisien hanya kepada mereka yang memang berpotensi churn. Pendekatan ini diharapkan mampu menekan

angka churn secara signifikan, sekaligus menghindari biaya akuisisi pelanggan baru yang jauh lebih mahal.

Berdasarkan data industri, rata-rata **Customer Acquisition Cost (CAC)** di sektor telekomunikasi adalah sekitar **$694 per pelanggan baru**, mencakup biaya pemasaran dan penjualan hingga pelanggan resmi menggunakan layanan (inbeat.agency, Investopedia). Sementara itu, menurut riset dari **Simon-Kucher & Partners**, biaya untuk mempertahankan pelanggan (**Customer Retention Cost/CRC**) diperkirakan **lima kali lebih murah** dibandingkan biaya akuisisi, sehingga asumsi CRC ditetapkan sebesar **$139 per pelanggan (dibulatkan)** (Simon-Kucher).

Untuk mendukung strategi ini, proyek bertujuan membangun alat prediksi churn berbasis data. Alat ini akan berfungsi sebagai sistem peringatan dini (*early warning system*) yang memungkinkan perusahaan untuk mengidentifikasi risiko churn secara akurat dan melakukan intervensi sebelum pelanggan benar-benar berhenti. Dengan demikian, perusahaan dapat meningkatkan efisiensi anggaran dan mempertahankan pendapatan secara lebih berkelanjutan.

1.0.3 1.3 Problem Statements

Perusahaan belum memiliki kemampuan yang andal untuk mengidentifikasi pelanggan yang berisiko churn maupun memahami faktor-faktor utama yang menyebabkan churn. Akibatnya, strategi retensi menjadi tidak efisien dan sering kali terlambat.

Proyek ini berupaya menjawab dua pertanyaan kunci berikut:

- **Bagaimana cara mengidentifikasi pelanggan yang berisiko churn secara akurat sebelum mereka berhenti berlangganan?**

- **Faktor-faktor apa saja yang paling berkontribusi terhadap churn, baik dari sisi demografi maupun perilaku penggunaan layanan?**

1.0.4 1.4 Goals

Tujuan dari proyek ini adalah membantu perusahaan mengurangi jumlah pelanggan yang berhenti berlangganan dengan cara yang lebih tepat dan efisien. Fokus utamanya adalah:

- **Membuat model prediksi** yang bisa mengenali pelanggan yang berisiko churn sebelum mereka benar-benar pergi.
- **Menemukan faktor-faktor utama** yang membuat pelanggan berhenti, seperti lama berlangganan, jumlah tagihan, jenis layanan, atau data demografi.
- **Memberikan rekomendasi strategis** berdasarkan data, agar perusahaan bisa mengambil tindakan retensi yang lebih terarah dan hemat biaya.

1.0.5 1.5 Analytical Approach

Pendekatan analisis akan dilakukan dalam dua tahap utama. Pertama, akan dilakukan Analisis Data Eksploratif (EDA). Tahap ini seperti "menggali" data yang sudah ada untuk mencari tahu lebih dalam tentang perilaku pelanggan. Tujuannya adalah untuk menemukan pola atau ciri-ciri menarik dari pelanggan yang cenderung berhenti berlangganan. Hasil dari analisis ini bisa langsung memberikan rekomendasi awal untuk bisnis.

Kedua, wawasan dari tahap pertama akan digunakan untuk membangun sebuah model klasifikasi. Sederhananya, ini adalah sistem cerdas yang dilatih untuk memprediksi pelanggan mana yang kemungkinan besar akan churn. Untuk membuktikan kegunaannya, kinerja model ini akan diukur dan dibandingkan dengan skenario "tanpa model", yaitu kondisi perusahaan saat ini yang tidak memiliki sistem prediksi. Perbandingan ini akan menunjukkan secara jelas keuntungan dari penerapan pendekatan berbasis data.

1.0.6 1.6 Metrics Evaluation: F2-Score

F2-Score adalah versi modifikasi dari F1-Score yang dirancang khusus untuk situasi di mana **Recall** dianggap lebih penting daripada Presisi. Metrik ini menggabungkan Presisi dan Recall, namun memberikan bobot empat kali lebih besar pada Recall. Hal ini menjadikannya pilihan ideal untuk skenario bisnis di mana biaya akibat gagal mendeteksi sebuah kasus (**False Negative**) jauh lebih merugikan daripada biaya akibat salah menandai (**False Positive**).

Dalam konteks prediksi *churn*, kesalahan prediksi memiliki implikasi biaya yang signifikan:

- **False Negative (FN):** Terjadi ketika model memprediksi pelanggan akan **tetap setia**, padahal **kenyataannya pelanggan tersebut *churn***. Ini adalah kesalahan paling merugikan karena perusahaan kehilangan pelanggan tanpa sempat melakukan tindakan pencegahan. Setiap kali terjadi FN, perusahaan kehilangan seorang pelanggan dan harus mengeluarkan biaya untuk mencari pelanggan baru, yang dikenal sebagai **Customer Acquisition Cost (CAC)**. Kami mengasumsikan biaya per kesalahan FN adalah **$694**.

$$\text{Total Biaya FN} = \text{Jumlah FN} \times \$694$$

- **False Positive (FP):** Terjadi ketika model memprediksi pelanggan akan ***churn***, padahal **kenyataannya pelanggan tersebut tetap setia**. Kesalahan ini tidak separah FN, namun tetap menimbulkan biaya yang tidak perlu. Ketika terjadi FP, perusahaan akan mengeluarkan biaya untuk tindakan retensi (misalnya, memberikan diskon atau bonus) kepada pelanggan yang sebenarnya tidak berniat pergi. Biaya ini disebut **Customer Retention Cost (CRC)**. Kami mengasumsikan biaya per kesalahan FP adalah **$139**.

$$\text{Total Biaya FP} = \text{Jumlah FP} \times \$139$$

Mengingat bahwa biaya akibat **False Negative (CAC)** ($694) jauh lebih tinggi daripada biaya akibat **False Positive (CRC)** ($139), sangat penting bagi model untuk meminimalkan False Negatives. F2-Score secara khusus mendorong model untuk memaksimalkan penemuan pelanggan yang berisiko *churn* (yaitu, meningkatkan Recall), sehingga secara efektif mengurangi kerugian finansial yang terkait dengan kehilangan pelanggan. Ini menjadikan F2-Score sebagai alat ukur yang paling tepat karena ia mendorong model untuk memaksimalkan deteksi pelanggan yang berisiko *churn*.

$$\text{F2-Score} = 5 \times \frac{\text{Precision} \times \text{Recall}}{4 \times \text{Precision} + \text{Recall}}$$

1.0.7 1.7 Success Criteria

Proyek ini dianggap berhasil jika:

- **Model prediksi churn cukup akurat**, dengan F2-Score minimal 70%.
- **Pelanggan yang berisiko churn bisa dikenali lebih awal** sebelum mereka berhenti.
- **Faktor-faktor utama churn bisa dipahami** oleh tim bisnis, sehingga dapat digunakan untuk pengambilan keputusan.

1.1 Section 2. Data Understanding

1.1.1 2.1 Dataset Information

[1]:
```
import os
import warnings
from pathlib import Path

# Suppress all warnings
warnings.filterwarnings("ignore")

# Data handling
import pandas as pd
import numpy as np

# Visualization
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as msno

# Statistics
from scipy.stats import spearmanr, normaltest, mannwhitneyu, kruskal

# Pandas display settings
pd.set_option("display.max_columns", None)

# Seaborn aesthetics
sns.set_theme()

# File path
data_path = Path("../data/WA_Fn-UseC_-Telco-Customer-Churn.csv")

# Load dataset
if data_path.exists():
    real_df = pd.read_csv(data_path)
    df = real_df.copy()
    print(df.info())
    display(df.head())
else:
    print(f"File not found: {data_path}")
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
```

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
None
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	\
0	7590-VHVEG	Female	0	Yes	No	1	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	
3	7795-CFOCW	Male	0	No	No	45	No	
4	9237-HQITU	Female	0	No	No	2	Yes	

	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	\
0	No phone service	DSL	No	Yes	
1	No	DSL	Yes	No	
2	No	DSL	Yes	Yes	
3	No phone service	DSL	Yes	No	
4	No	Fiber optic	No	No	

	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	\
0	No	No	No	No	Month-to-month	
1	Yes	No	No	No	One year	
2	No	No	No	No	Month-to-month	
3	Yes	Yes	No	No	One year	
4	No	No	No	No	Month-to-month	

```
  PaperlessBilling              PaymentMethod  MonthlyCharges TotalCharges  \
0              Yes           Electronic check           29.85        29.85
1               No               Mailed check           56.95       1889.5
2              Yes               Mailed check           53.85       108.15
3               No  Bank transfer (automatic)           42.30      1840.75
4              Yes           Electronic check           70.70       151.65

  Churn
0    No
1    No
2   Yes
3    No
4   Yes
```

Berikut adalah penjelasan untuk masing-masing kolom:

Kolom

Definisi

Value dan Penjelasan

customerID

ID unik untuk setiap pelanggan.

Teks alfanumerik: Kode unik pengenal pelanggan.

gender

Jenis kelamin pelanggan.

Male: Laki-laki.Female: Perempuan.

SeniorCitizen

Bagaimanakah pelanggan seorang warga senior.

1: Ya, warga senior.0: Bukan warga senior.

Partner

Bagaimanakah pelanggan memiliki pasangan.

Yes: Punya pasangan.No: Tidak punya pasangan.

Dependents

Bagaimanakah pelanggan memiliki tanggungan.

Yes: Punya tanggungan.No: Tidak punya tanggungan.

tenure

Lama berlangganan dalam bulan.

Numerik: Jumlah bulan pelanggan bersama perusahaan.

Contract

Jenis kontrak berlangganan.

Month-to-month: Kontrak bulanan.One year: Kontrak satu tahun.Two year: Kontrak dua tahun.

PaperlessBilling

Menggunakan tagihan elektronik.

Yes: Tagihan via email.No: Tagihan fisik.

PaymentMethod

Metode pembayaran yang digunakan.

Electronic check: Pembayaran dengan cek elektronik.Mailed check: Pembayaran dengan mengirimkan cek fisik.Bank transfer (automatic): Pembayaran melalui transfer bank otomatis.Credit card (automatic): Pembayaran melalui kartu kredit otomatis.

MonthlyCharges

Tagihan setiap bulan.

Numerik Desimal: Total biaya bulanan untuk semua layanan.

TotalCharges

Total tagihan selama berlangganan.

Numerik Desimal: Akumulasi semua tagihan bulanan.

PhoneService

Bagaimanakah pelanggan punya layanan telepon.

Yes: Punya layanan telepon.No: Tidak punya.

MultipleLines

Bagaimanakah punya lebih dari satu saluran telepon.

Yes: Ya.No: Punya telepon, tapi hanya 1 saluran.No phone service: Tidak punya layanan telepon.

InternetService

Tipe layanan internet pelanggan.

DSL: Layanan internet melalui jalur telepon standar.Fiber optic: Layanan internet berkecepatan tinggi melalui kabel fiber optik.No: Tidak berlangganan internet.

OnlineSecurity

Bagaimanakah punya layanan keamanan online.

Yes: Ya.No: Pelanggan berlangganan internet namun tidak menggunakan layanan ini.No internet service: Tidak punya layanan internet.

OnlineBackup

Bagaimanakah punya layanan backup online.

Yes: Ya.No: Pelanggan berlangganan internet namun tidak menggunakan layanan ini.No internet service: Tidak punya layanan internet.

DeviceProtection

Bagaimanakah punya proteksi perangkat.

Yes: Ya.No: Pelanggan berlangganan internet namun tidak menggunakan layanan ini.No internet service: Tidak punya layanan internet.

TechSupport

Bagaimanakah punya dukungan teknis premium.

Yes: Ya.No: Pelanggan berlangganan internet namun tidak menggunakan layanan ini.No internet service: Tidak punya layanan internet.

StreamingTV

Bagaimanakah streaming TV dari Telco.

Yes: Ya.No: Pelanggan berlangganan internet namun tidak menggunakan layanan ini.No internet service: Tidak punya layanan internet.

StreamingMovies

Bagaimanakah streaming film dari Telco.

Yes: Ya.No: Pelanggan berlangganan internet namun tidak menggunakan layanan ini.No internet service: Tidak punya layanan internet.

Churn

(Target) Bagaimanakah pelanggan berhenti.

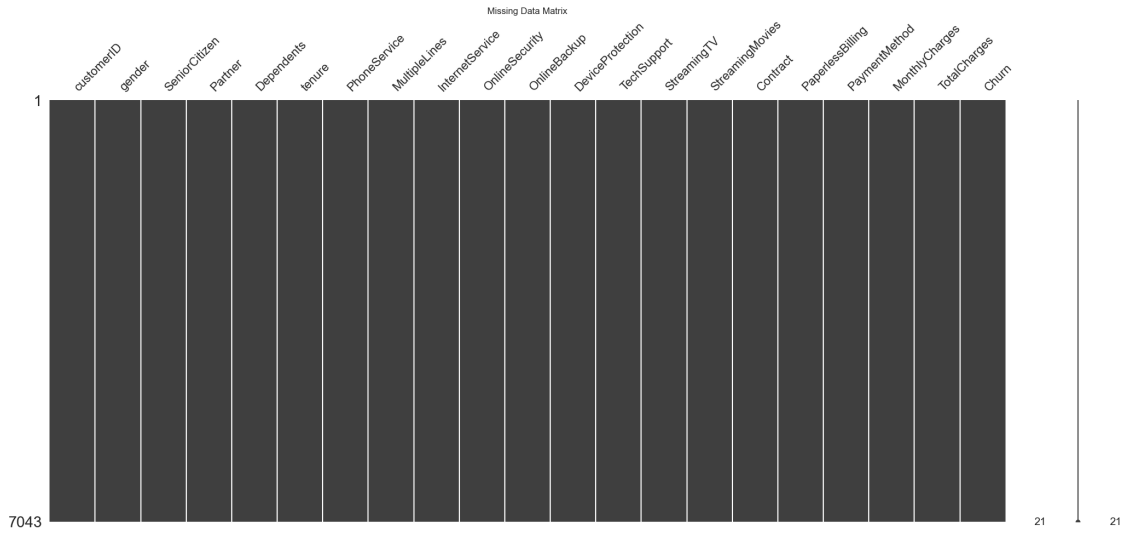
Yes: Pelanggan berhenti berlangganan.No: Pelanggan masih aktif.

1.1.2 2.2 Missing Values Checking

[2]:
```
# Check missing values count per column
missing_counts = df.isnull().sum()
print(missing_counts)

# Visualize missing data pattern
msno.matrix(df)
plt.title('Missing Data Matrix')
plt.show()
```

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
```

```
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```



Meskipun tidak terdapat missing values secara eksplisit dalam dataset, keberadaan nilai `tenure = 0` perlu diperhatikan karena dapat merepresentasikan pelanggan yang baru mendaftar dan belum benar-benar aktif, atau langsung churn pada bulan pertama. Hal ini penting karena bisa mempengaruhi kualitas data dan performa model machine learning, terutama jika jumlahnya kecil namun berdampak besar dalam proses pembelajaran model. Dari sisi bisnis, pelanggan dengan `tenure = 0` mungkin belum sempat menggunakan layanan secara penuh, sehingga berpotensi mencerminkan signup palsu atau pembatalan instan. Oleh karena itu, baris dengan nilai ini sebaiknya dianalisis lebih lanjut, dipertimbangkan untuk dipisahkan, atau bahkan dibuang jika tidak relevan.

[3]:
```
df[df['tenure']==0]
```

[3]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	\
488	4472-LVYGI	Female	0	Yes	Yes	0	
753	3115-CZMZD	Male	0	No	Yes	0	
936	5709-LVOEQ	Female	0	Yes	Yes	0	
1082	4367-NUYAO	Male	0	Yes	Yes	0	
1340	1371-DWPAZ	Female	0	Yes	Yes	0	

```
3331  7644-OMVMY    Male              0     Yes        Yes       0
3826  3213-VVOLG    Male              0     Yes        Yes       0
4380  2520-SGTTA  Female              0     Yes        Yes       0
5218  2923-ARZLG    Male              0     Yes        Yes       0
6670  4075-WKNIU  Female              0     Yes        Yes       0
6754  2775-SEFEE    Male              0      No        Yes       0

     PhoneService     MultipleLines InternetService       OnlineSecurity  \
488            No  No phone service             DSL                  Yes
753           Yes                No              No  No internet service
936           Yes                No             DSL                  Yes
1082          Yes               Yes              No  No internet service
1340           No  No phone service             DSL                  Yes
3331          Yes                No              No  No internet service
3826          Yes               Yes              No  No internet service
4380          Yes                No              No  No internet service
5218          Yes                No              No  No internet service
6670          Yes               Yes             DSL                   No
6754          Yes               Yes             DSL                  Yes

            OnlineBackup     DeviceProtection          TechSupport  \
488                   No                  Yes                  Yes
753   No internet service  No internet service  No internet service
936                  Yes                  Yes                   No
1082  No internet service  No internet service  No internet service
1340                 Yes                  Yes                  Yes
3331  No internet service  No internet service  No internet service
3826  No internet service  No internet service  No internet service
4380  No internet service  No internet service  No internet service
5218  No internet service  No internet service  No internet service
6670                 Yes                  Yes                  Yes
6754                 Yes                   No                  Yes

             StreamingTV      StreamingMovies  Contract PaperlessBilling  \
488                  Yes                   No  Two year              Yes
753   No internet service  No internet service  Two year               No
936                  Yes                  Yes  Two year               No
1082  No internet service  No internet service  Two year               No
1340                 Yes                   No  Two year               No
3331  No internet service  No internet service  Two year               No
3826  No internet service  No internet service  Two year               No
4380  No internet service  No internet service  Two year               No
5218  No internet service  No internet service  One year              Yes
6670                 Yes                   No  Two year               No
6754                  No                   No  Two year              Yes

                 PaymentMethod  MonthlyCharges TotalCharges Churn
```

```
488   Bank transfer (automatic)          52.55                 No
753                Mailed check          20.25                 No
936                Mailed check          80.85                 No
1082               Mailed check          25.75                 No
1340    Credit card (automatic)          56.05                 No
3331               Mailed check          19.85                 No
3826               Mailed check          25.35                 No
4380               Mailed check          20.00                 No
5218               Mailed check          19.70                 No
6670               Mailed check          73.35                 No
6754  Bank transfer (automatic)          61.90                 No
```

Dari tabel yang ditampilkan, terlihat bahwa seluruh baris dengan `tenure = 0` memiliki nilai `TotalCharges` yang kosong (missing). Ini memperkuat indikasi adanya **anomali data**, karena secara logika bisnis, jika seorang pelanggan sudah memiliki `MonthlyCharges` tetapi `TotalCharges` kosong, hal ini tidak konsisten. Seharusnya, meskipun pelanggan baru, `TotalCharges` minimal setara dengan `MonthlyCharges` jika telah berjalan satu bulan. Selain itu, sebagian besar pelanggan ini memiliki kontrak jangka panjang seperti "Two year", yang tidak wajar jika langsung berstatus tenure = 0 tanpa adanya tagihan. Temuan ini menunjukkan bahwa baris-baris tersebut kemungkinan merupakan data yang belum tereksekusi penuh dalam sistem atau input yang belum lengkap. Maka, penting untuk memperlakukan baris ini secara khusus—baik dengan membersihkan, memisahkan, atau mengecualikan dari pelatihan model tergantung pada tujuan analisisnya.

[4]:
```
df = df[df['tenure'] != 0]
df
```

[4]:
```
      customerID  gender  SeniorCitizen Partner Dependents  tenure  \
0     7590-VHVEG  Female              0     Yes         No       1
1     5575-GNVDE    Male              0      No         No      34
2     3668-QPYBK    Male              0      No         No       2
3     7795-CFOCW    Male              0      No         No      45
4     9237-HQITU  Female              0      No         No       2
…          …     …              …      …        …     …
7038  6840-RESVB    Male              0     Yes        Yes      24
7039  2234-XADUH  Female              0     Yes        Yes      72
7040  4801-JZAZL  Female              0     Yes        Yes      11
7041  8361-LTMKD    Male              1     Yes         No       4
7042  3186-AJIEK    Male              0      No         No      66

     PhoneService     MultipleLines InternetService OnlineSecurity  \
0              No  No phone service             DSL             No
1             Yes                No             DSL            Yes
2             Yes                No             DSL            Yes
3              No  No phone service             DSL            Yes
4             Yes                No     Fiber optic             No
…           …                …             …            …
7038          Yes               Yes             DSL            Yes
7039          Yes               Yes     Fiber optic             No
```

```
7040           No  No phone service             DSL            Yes
7041          Yes               Yes     Fiber optic             No
7042          Yes                No     Fiber optic            Yes

     OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies  \
0             Yes               No          No          No              No
1              No              Yes          No          No              No
2             Yes               No          No          No              No
3              No              Yes         Yes          No              No
4              No               No          No          No              No
…             …                …           …           …               …
7038           No              Yes         Yes         Yes             Yes
7039          Yes              Yes          No         Yes             Yes
7040           No               No          No          No              No
7041           No               No          No          No              No
7042           No              Yes         Yes         Yes             Yes

            Contract PaperlessBilling              PaymentMethod  \
0     Month-to-month              Yes           Electronic check
1           One year               No               Mailed check
2     Month-to-month              Yes               Mailed check
3           One year               No  Bank transfer (automatic)
4     Month-to-month              Yes           Electronic check
…                …               …                          …
7038        One year              Yes               Mailed check
7039        One year              Yes    Credit card (automatic)
7040  Month-to-month              Yes           Electronic check
7041  Month-to-month              Yes               Mailed check
7042        Two year              Yes  Bank transfer (automatic)

      MonthlyCharges TotalCharges Churn
0              29.85        29.85    No
1              56.95       1889.5    No
2              53.85       108.15   Yes
3              42.30      1840.75    No
4              70.70       151.65   Yes
…              …          …   …
7038           84.80       1990.5    No
7039          103.20       7362.9    No
7040           29.60       346.45    No
7041           74.40        306.6   Yes
7042          105.65       6844.5    No

[7032 rows x 21 columns]
```

Terdapat anomali pada data dengan tenure = 0, di mana nilai TotalCharges tidak terisi. Hal ini wajar karena pelanggan tersebut baru bergabung sehingga belum memiliki tagihan dan status churn-nya belum dapat diketahui dengan jelas. Sesuai kesepakatan, 11 baris dengan tenure =

0 dihapus dari dataset karena dianggap tidak merepresentasikan perilaku pelanggan aktif. Data ini hanya mencakup sekitar 0.15% dari total, sehingga dampaknya terhadap analisis sangat kecil. Keputusan ini juga didukung oleh literatur, di mana menurut Schafer (1999) dan Bennett (2001), data yang hilang kurang dari 5% umumnya tidak menyebabkan bias signifikan jika dihapus secara langsung. Oleh karena itu, penghapusan dilakukan untuk keperluan EDA maupun pelatihan model machine learning.

1.1.3 2.3 Duplicated Values Checking

[5]:
```
# Count duplicated rows
num_duplicates = df.duplicated().sum()
print(f"Number of duplicated rows: {num_duplicates}")
```

```
Number of duplicated rows: 0
```

Hasil pemeriksaan duplikasi dengan fungsi `df.duplicated().sum()` menunjukkan bahwa tidak terdapat baris yang sama persis (duplikat) dalam dataset, yaitu sejumlah 0 baris duplikat. Hal ini menandakan bahwa setiap entri dalam data bersifat unik setelah penghapusan baris dengan `tenure = 0`, sehingga tidak diperlukan tindakan lebih lanjut terkait duplikasi. Keberadaan data yang bebas duplikasi sangat penting untuk menjaga kualitas analisis dan mencegah bias yang dapat muncul pada tahap pelatihan model machine learning.

1.1.4 2.4 Dataset Restructuring for Better EDA

[6]:
```
# Convert object columns to category dtype
object_cols = df.select_dtypes(include=['object']).columns.tolist()
object_cols = [col for col in object_cols if col not in ('customerID',␣
 ↪'TotalCharges')]
df[object_cols] = df[object_cols].astype('category')
df['SeniorCitizen'] = df['SeniorCitizen'].map({0: 'No', 1: 'Yes'}).
 ↪astype('category')
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')

# Confirm changes
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 7032 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7032 non-null   object
 1   gender            7032 non-null   category
 2   SeniorCitizen     7032 non-null   category
 3   Partner           7032 non-null   category
 4   Dependents        7032 non-null   category
 5   tenure            7032 non-null   int64
 6   PhoneService      7032 non-null   category
 7   MultipleLines     7032 non-null   category
```

```
 8   InternetService   7032 non-null   category
 9   OnlineSecurity    7032 non-null   category
 10  OnlineBackup      7032 non-null   category
 11  DeviceProtection  7032 non-null   category
 12  TechSupport       7032 non-null   category
 13  StreamingTV       7032 non-null   category
 14  StreamingMovies   7032 non-null   category
 15  Contract          7032 non-null   category
 16  PaperlessBilling  7032 non-null   category
 17  PaymentMethod     7032 non-null   category
 18  MonthlyCharges    7032 non-null   float64
 19  TotalCharges      7032 non-null   float64
 20  Churn             7032 non-null   category
dtypes: category(17), float64(2), int64(1), object(1)
memory usage: 393.6+ KB
```

[7]:
```
df.head()
```

[7]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
0	7590-VHVEG	Female	No	Yes	No	1	No
1	5575-GNVDE	Male	No	No	No	34	Yes
2	3668-QPYBK	Male	No	No	No	2	Yes
3	7795-CFOCW	Male	No	No	No	45	No
4	9237-HQITU	Female	No	No	No	2	Yes

	MultipleLines	InternetService	OnlineSecurity	OnlineBackup
0	No phone service	DSL	No	Yes
1	No	DSL	Yes	No
2	No	DSL	Yes	Yes
3	No phone service	DSL	Yes	No
4	No	Fiber optic	No	No

	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract
0	No	No	No	No	Month-to-month
1	Yes	No	No	No	One year
2	No	No	No	No	Month-to-month
3	Yes	Yes	No	No	One year
4	No	No	No	No	Month-to-month

	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges
0	Yes	Electronic check	29.85	29.85
1	No	Mailed check	56.95	1889.50
2	Yes	Mailed check	53.85	108.15
3	No	Bank transfer (automatic)	42.30	1840.75
4	Yes	Electronic check	70.70	151.65

	Churn

```
0    No
1    No
2   Yes
3    No
4   Yes
```

Untuk menyederhanakan analisis dan mengoptimalkan penggunaan memori, seluruh kolom bertipe `object`—kecuali `customerID` dan `TotalCharges`—diubah menjadi tipe data `category`, karena kolom-kolom ini merepresentasikan data kategorikal. Selain itu, kolom `SeniorCitizen`, yang semula berupa numerik biner (0 dan 1), dipetakan menjadi kategori `"No"` dan `"Yes"` untuk meningkatkan interpretabilitas. Kolom `TotalCharges` juga dikonversi ke tipe numerik (`float64`) dengan `errors='coerce'` untuk memastikan konsistensi data. Setelah transformasi, dataset terdiri dari 21 kolom dengan 17 kolom bertipe kategori, 2 numerik kontinu (`MonthlyCharges`, `TotalCharges`), 1 numerik diskrit (`tenure`), dan 1 kolom identitas (`customerID`). Transformasi ini berhasil memperbaiki struktur data untuk keperluan analisis eksploratif dan pemodelan lebih lanjut.

1.1.5 2.5 Exploratory Data Analysis (EDA) - Univariat

[8]:
```python
def plot_box_and_kde(df, col, figsize=(15, 5)):
    # Visualisasi
    fig, axes = plt.subplots(nrows=1, ncols=2, figsize=figsize)
    plt.suptitle(f'Distribution of {col}', fontsize=16, y=1.02)
    sns.boxplot(data=df, x=col, orient='h', ax=axes[0])
    axes[0].set_title(f'Boxplot of {col}')

    sns.kdeplot(data=df, x=col, fill=True, ax=axes[1])
    axes[1].set_title(f'KDE Histplot of {col}')

    plt.tight_layout(rect=[0, 0, 1, 0.9])
    plt.show()

    # Statistik deskriptif + IQR
    stats = df[col].describe()
    q1 = stats['25%']
    q3 = stats['75%']
    iqr = q3 - q1
    stats['IQR'] = iqr

    # Tabel transpos
    stats_df = stats.to_frame(name=col).T
    display(stats_df)
```

Fungsi plot_box_and_kde digunakan untuk menganalisis variabel numerik seperti 'tenure'. Fungsi ini menyajikan dua plot berdampingan: sebuah boxplot untuk melihat ringkasan statistik (median, kuartil, jangkauan) dan sebuah KDE plot untuk melihat bentuk distribusi data. Selain visualisasi, fungsi ini juga menampilkan tabel statistik deskriptif yang mencakup mean, standar deviasi, dan IQR (Interquartile Range) untuk memberikan ringkasan kuantitatif.

[9]:
```
def plot_countplot_with_hue(df, col, figsize=(15, 5)):
    # Hitung Count dan Persentase
    counts = df[col].value_counts()
    percentages = counts / len(df) * 100
    total = len(df)

    # Buat urutan berdasarkan Count tertinggi ke terendah
    ordered_categories = counts.index.tolist()
    df[col] = pd.Categorical(df[col], categories=ordered_categories,␣
 ↪ordered=True)

    # Plot
    plt.figure(figsize=figsize)
    ax = sns.countplot(data=df, y=col, hue=col)
    plt.title(f'Distribution of {col}', fontsize=16)
    plt.xlabel('Count')
    plt.ylabel('')

    for p in ax.patches:
        width = p.get_width()
        y = p.get_y() + p.get_height() / 2
        label = f'{int(width)} ({width / total:.1%})'
        ax.text(width + total * 0.005, y, label, va='center', fontsize=9)

    plt.tight_layout()
    plt.show()

    # Tampilkan tabel count dan persentase (dengan reset index)
    table = pd.DataFrame({
        col: counts.index,
        'Count': counts.values,
        'Percentage': percentages.apply(lambda x: f'{x:.1f}%').values
    }).reset_index(drop=True)

    display(table)
```

Untuk variabel kategorikal, fungsi plot_countplot_with_hue digunakan. Fungsi ini menghitung frekuensi setiap kategori, mengurutkannya dari yang terbesar, lalu menampilkannya dalam bentuk countplot horizontal. Setiap bar pada plot diberi anotasi yang menunjukkan jumlah absolut dan persentase relatifnya, sehingga memberikan gambaran distribusi yang jelas dan informatif.

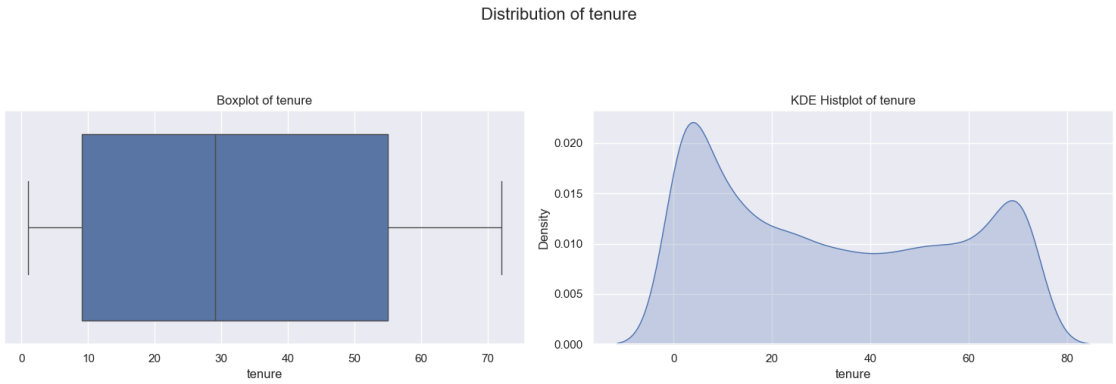
[10]:
```
def check_normality(df, col):
  # Ambil data dari kolom yang ditentukan dan hapus nilai yang hilang.
  data_to_test = df[col].dropna()

  # Jalankan uji D'Agostino-Pearson dan ambil p-value.
  statistic, p_value = normaltest(data_to_test)
```

```
  # Tampilkan nilai p-value-nya.
  print(f"P-value untuk kolom '{col}': {p_value:.2f}")

  # Cek p-value untuk mencetak konklusi hipotesis.
  if p_value < 0.05:
    print(f"Tolak Ho, terima Ha. Data '{col}' tidak terdistribusi normal.")
  else:
    print(f"Gagal tolak Ho. Data '{col}' terdistribusi normal.")
```

Fungsi ini bertujuan untuk menguji secara statistik Bagaimanakah data dalam sebuah kolom mengikuti distribusi normal atau tidak. Ia mengambil data dari kolom yang ditentukan, menjalankan uji normalitas D'Agostino-Pearson, lalu mencetak nilai p-value yang dihasilkan. Berdasarkan p-value tersebut, fungsi ini memberikan kesimpulan statistik Bagaimanakah kita harus menolak atau gagal menolak hipotesis nol, yang pada akhirnya menentukan Bagaimanakah data tersebut dianggap normal atau tidak. - Ho: Data terdistribusi normal. - Ha: Data tidak terdistribusi normal.

2.5.1 Berapa lama pelanggan biasanya tetap berlangganan layanan?

[11]:
```
plot_box_and_kde(df, 'tenure')
check_normality(df,'tenure')
```



	count	mean	std	min	25%	50%	75%	max	IQR
tenure	7032.0	32.421786	24.54526	1.0	9.0	29.0	55.0	72.0	46.0

```
P-value untuk kolom 'tenure': 0.00
Tolak Ho, terima Ha. Data 'tenure' tidak terdistribusi normal.
```

Business Insights

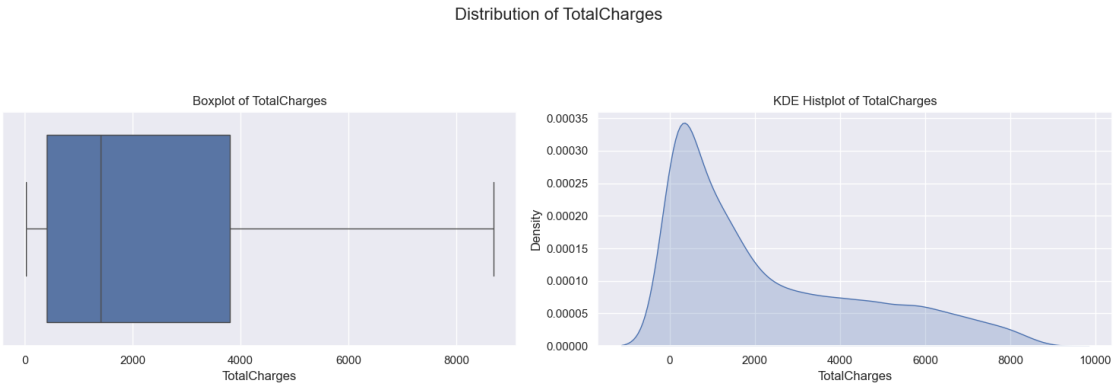
- Sebagian besar pelanggan memiliki lama berlangganan di bawah 3 tahun, dengan nilai tengah 29 bulan dan rata-rata 32 bulan. Hal ini menunjukkan pentingnya fokus pada retensi jangka menengah untuk menjaga basis pelanggan tetap stabil.
- Sebanyak 25% pelanggan berhenti dalam 9 bulan pertama, dan 75% pelanggan berhenti

sebelum bulan ke-55. Fakta ini menunjukkan bahwa risiko churn terbesar terjadi di awal masa langganan.

- Distribusi lama berlangganan menunjukkan dua puncak, yaitu pada masa awal dan mendekati 72 bulan. Ini menandakan adanya dua kelompok dominan: pelanggan yang cepat berhenti dan pelanggan sangat loyal.

2.5.2 Berapa besar total pengeluaran pelanggan selama mereka berlangganan?

[12]:
```
plot_box_and_kde(df, 'TotalCharges')
check_normality(df, 'TotalCharges')
```




```
                count         mean          std   min     25%      50%  \
TotalCharges   7032.0  2283.300441  2266.771362  18.8  401.45  1397.475

                    75%     max        IQR
TotalCharges  3794.7375  8684.8  3393.2875

P-value untuk kolom 'TotalCharges': 0.00
Tolak Ho, terima Ha. Data 'TotalCharges' tidak terdistribusi normal.
```

Business Insights

- Sebagian besar pelanggan memiliki total pengeluaran yang relatif rendah, dengan nilai tengah sebesar $1.397 dan hanya 25% pelanggan yang membayar lebih dari $3.794. Hal ini menunjukkan bahwa sebagian besar pelanggan belum menghasilkan nilai jangka panjang yang tinggi bagi perusahaan.
- Rata-rata total pengeluaran pelanggan adalah sekitar $2.283, namun sebaran sangat condong ke kanan hingga maksimum $8.684. Ini mengindikasikan adanya kelompok kecil pelanggan bernilai tinggi yang memberikan kontribusi signifikan terhadap pendapatan.
- IQR yang besar (±$3.393) menunjukkan variasi yang tinggi dalam nilai pelanggan. Artinya, kontribusi finansial dari setiap pelanggan sangat berbeda-beda.

2.5.3 Berapa biaya bulanan yang biasanya dibayarkan pelanggan?

[13]:
```
plot_box_and_kde(df, 'MonthlyCharges')
check_normality(df, 'MonthlyCharges')
```

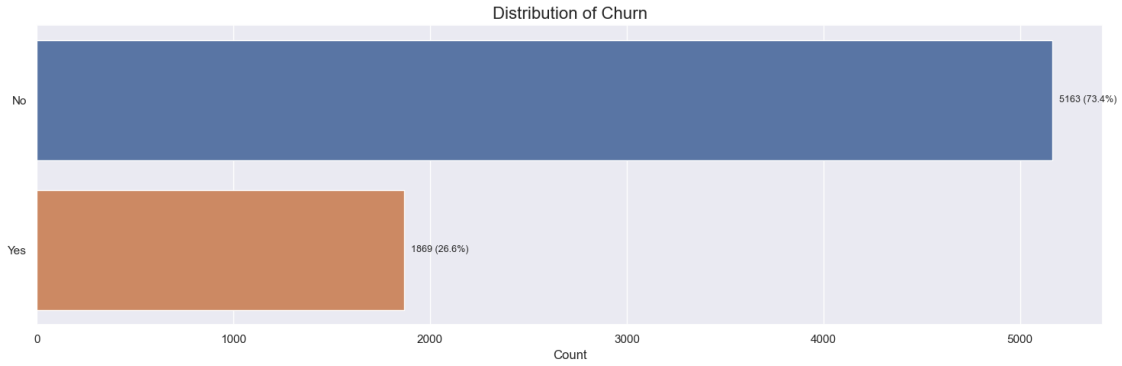

```
                  count       mean        std    min      25%    50%      75%  \
MonthlyCharges   7032.0  64.798208  30.085974  18.25  35.5875  70.35  89.8625

                    max     IQR
MonthlyCharges   118.75  54.275

P-value untuk kolom 'MonthlyCharges': 0.00
Tolak Ho, terima Ha. Data 'MonthlyCharges' tidak terdistribusi normal.
```

Business Insights

- Nilai tengah tagihan bulanan pelanggan adalah sekitar $70, dengan rata-rata sebesar $65. Ini menunjukkan bahwa sebagian besar pelanggan berada pada level harga menengah ke atas.
- Sebaran biaya bulanan bersifat bimodal, dengan konsentrasi pelanggan pada rentang harga rendah dan tinggi. Hal ini menunjukkan adanya dua kelompok utama pelanggan berdasarkan jenis layanan yang dipilih.
- IQR yang cukup besar ($54) menunjukkan variasi signifikan dalam struktur biaya antar pelanggan. Artinya, pelanggan membayar biaya yang sangat berbeda tergantung pada jenis dan kombinasi layanan yang diambil.

2.5.4 Bagaimana distribusi status churn pelanggan?

[14]:
```
plot_countplot_with_hue(df, 'Churn')
```



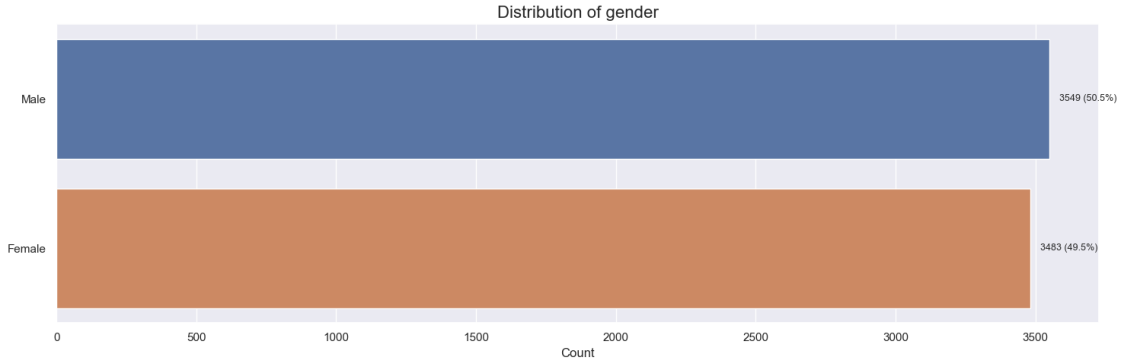
	Churn	Count	Percentage
0	No	5163	73.4%
1	Yes	1869	26.6%

Business Insights

- Sebanyak 26,6% pelanggan dalam dataset berhenti berlangganan (churn), sementara 73,4% masih aktif. Ini menunjukkan bahwa meskipun mayoritas pelanggan tetap bertahan, proporsi churn tetap signifikan dan tidak bisa diabaikan.
- Dataset bersifat tidak seimbang, dengan jumlah pelanggan yang tidak churn hampir tiga kali lebih banyak dibandingkan yang churn. Hal ini perlu diperhatikan dalam analisis lebih lanjut maupun pengembangan model prediktif.

2.5.5 Bagaimana karakteristik distribusi pelanggan berdasarkan jenis kelamin?

[15]:
```
plot_countplot_with_hue(df, 'gender')
```



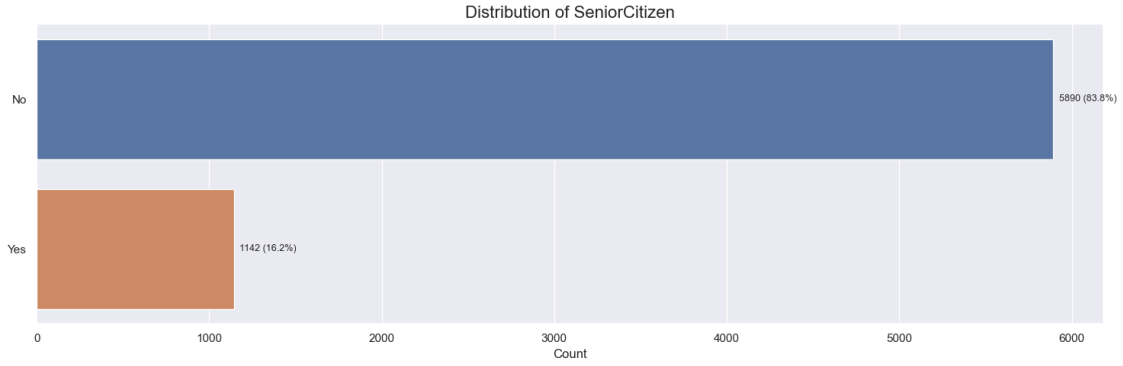
	gender	Count	Percentage
0	Male	3549	50.5%
1	Female	3483	49.5%

Business Insights

- Distribusi pelanggan berdasarkan gender sangat seimbang, dengan 50,5% laki-laki dan 49,5% perempuan. Tidak ada dominasi gender tertentu dalam populasi pelanggan.

2.5.6 Bagaimana karakteristik distribusi pelanggan berdasarkan status lansia?

[16]:
```
plot_countplot_with_hue(df, 'SeniorCitizen')
```



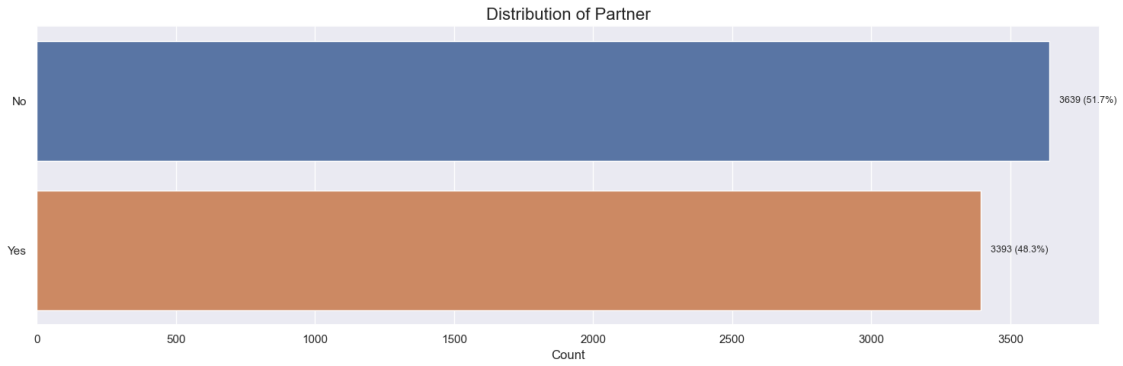
	SeniorCitizen	Count	Percentage
0	No	5890	83.8%
1	Yes	1142	16.2%

Business Insights

- Sebagian besar pelanggan (83,8%) bukan merupakan warga senior, sementara hanya 16,2% yang termasuk kategori senior. Artinya, kelompok senior merupakan minoritas dalam basis pelanggan saat ini.

2.5.7 Bagaimana karakteristik distribusi pelanggan berdasarkan status memiliki pasangan?

[17]:
```
plot_countplot_with_hue(df, 'Partner')
```



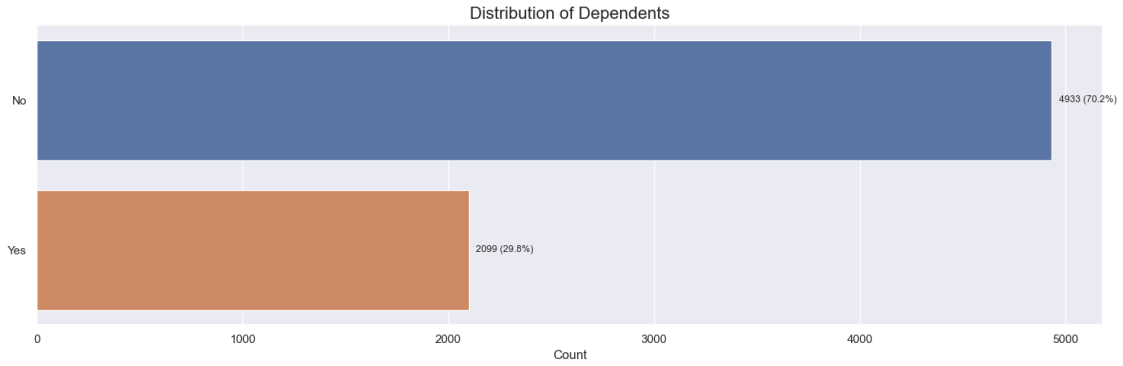
	Partner	Count	Percentage
0	No	3639	51.7%
1	Yes	3393	48.3%

Business Insights

- Distribusi pelanggan berdasarkan status memiliki pasangan cukup seimbang, dengan 51,7% tidak memiliki pasangan dan 48,3% memiliki pasangan. Tidak ada dominasi dari salah satu kelompok.

2.5.8 Bagaimana karakteristik distribusi pelanggan berdasarkan tanggungan keluarga?

[18]:
```
plot_countplot_with_hue(df, 'Dependents')
```



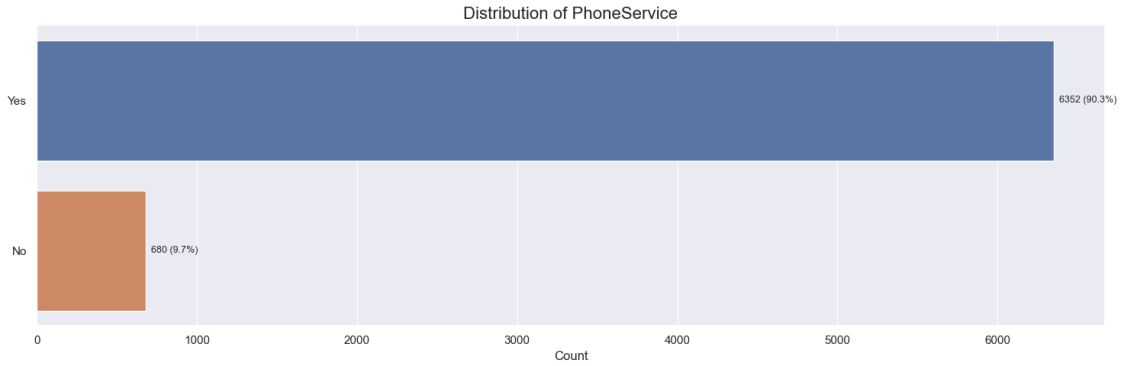
	Dependents	Count	Percentage
0	No	4933	70.2%
1	Yes	2099	29.8%

Business Insights

- Sebagian besar pelanggan (70,2%) tidak memiliki tanggungan, sementara hanya 29,8% yang memiliki tanggungan. Artinya, mayoritas pelanggan merupakan individu tanpa ketergantungan keluarga langsung.

2.5.9 Bagaimana karakteristik distribusi pelanggan berdasarkan kepemilikan layanan telepon ?

[19]:
```
plot_countplot_with_hue(df, 'PhoneService')
```



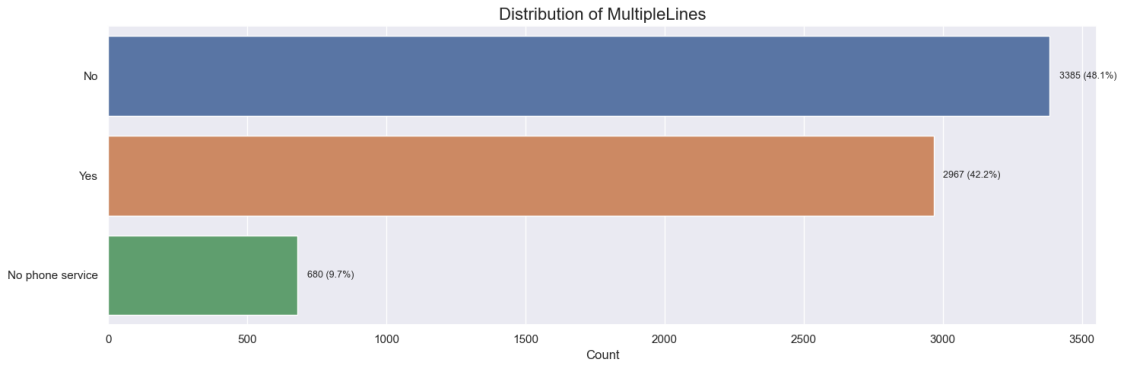
	PhoneService	Count	Percentage
0	Yes	6352	90.3%
1	No	680	9.7%

Business Insights

- Sebagian besar pelanggan (90,3%) menggunakan layanan telepon, sementara hanya 9,7% yang tidak. Ini menunjukkan bahwa layanan telepon merupakan fitur yang sangat umum di antara pelanggan.

2.5.10 Bagaimana karakteristik distribusi pelanggan berdasarkan kepemilikan beberapa saluran telepon?

[20]:
```
plot_countplot_with_hue(df, 'MultipleLines')
```



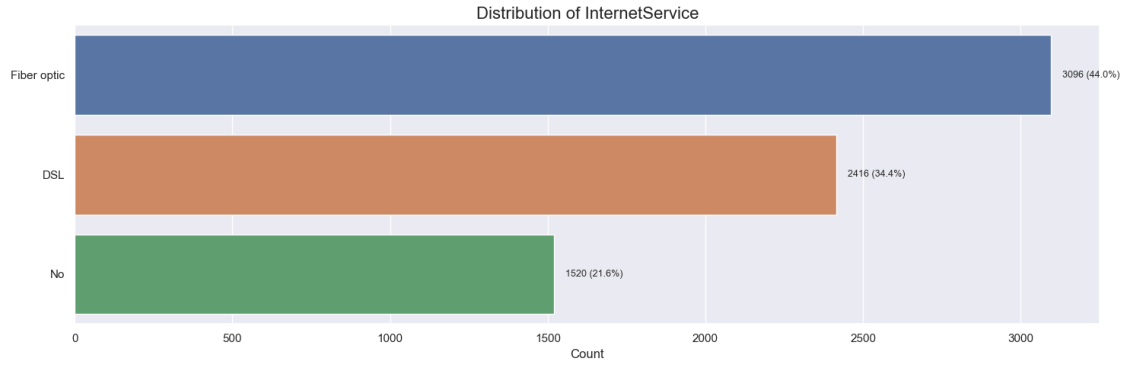
	MultipleLines	Count	Percentage
0	No	3385	48.1%
1	Yes	2967	42.2%
2	No phone service	680	9.7%

Business Insights

- Sebanyak 48,1% pelanggan memiliki satu jalur telepon, 42,2% memiliki lebih dari satu jalur, dan 9,7% tidak menggunakan layanan telepon sama sekali. Ini menunjukkan bahwa layanan tambahan berupa saluran telepon ganda cukup diminati oleh hampir separuh pengguna layanan telepon.

2.5.11 Bagaimana karakteristik distribusi pelanggan berdasarkan jenis layanan internet?

[21]:
```
plot_countplot_with_hue(df, 'InternetService')
```



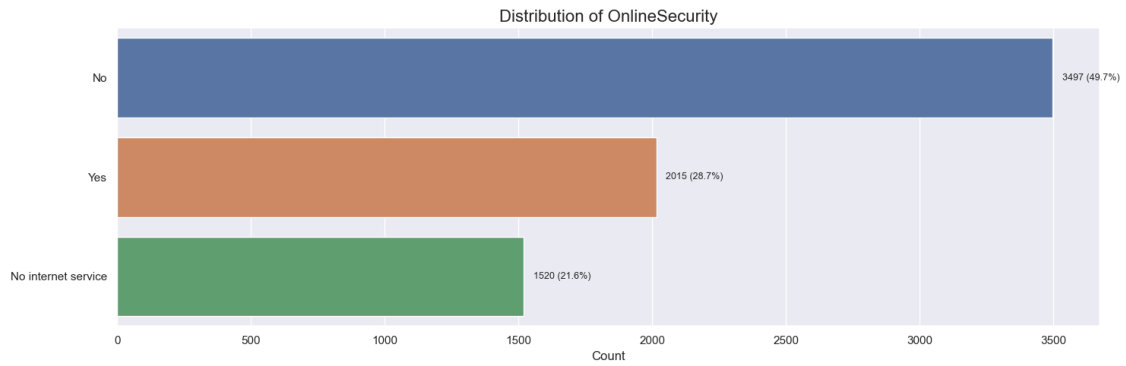
	InternetService	Count	Percentage
0	Fiber optic	3096	44.0%
1	DSL	2416	34.4%
2	No	1520	21.6%

Business Insights

- Sebagian besar pelanggan menggunakan layanan internet, dengan 44,0% memilih Fiber optic dan 34,4% menggunakan DSL. Hanya 21,6% yang tidak berlangganan internet. Ini menunjukkan bahwa layanan internet adalah produk utama dalam portofolio perusahaan.

2.5.12 Bagaimana karakteristik distribusi pelanggan berdasarkan status perlindungan keamanan online?

[22]:
```
plot_countplot_with_hue(df, 'OnlineSecurity')
```



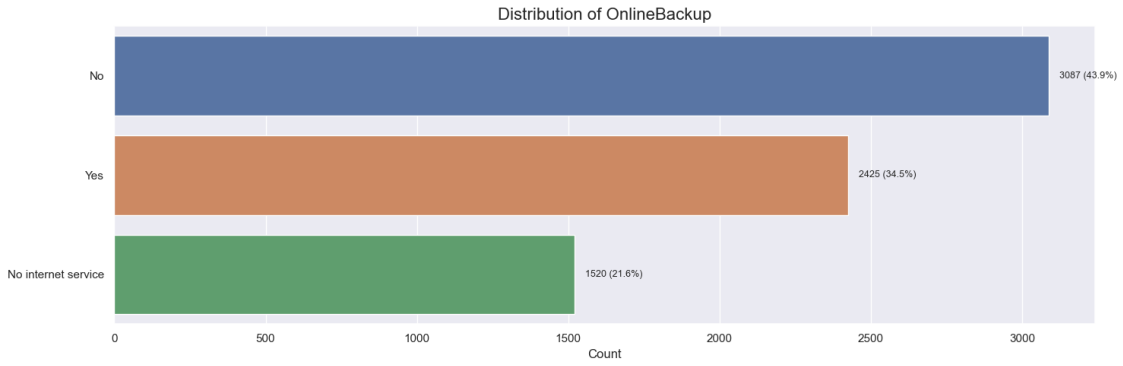
	OnlineSecurity	Count	Percentage
0	No	3497	49.7%
1	Yes	2015	28.7%
2	No internet service	1520	21.6%

Business Insights

- Sebanyak 49,7% pelanggan tidak menggunakan layanan Online Security, sementara hanya 28,7% yang menggunakannya. Sisanya 21,6% tidak memiliki layanan internet. Artinya, dari pelanggan internet aktif, mayoritas tidak memanfaatkan layanan keamanan online.

2.5.13 Bagaimana karakteristik distribusi pelanggan berdasarkan status cadangan data online?

[23]:
```
plot_countplot_with_hue(df, 'OnlineBackup')
```



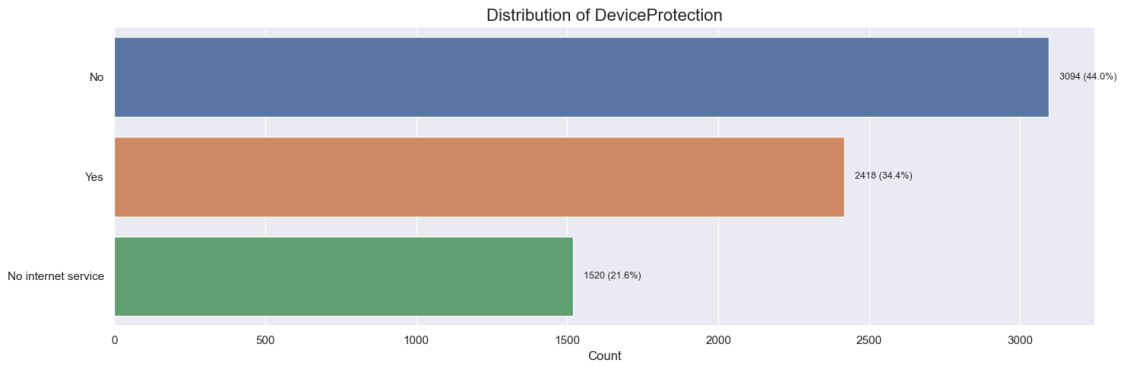
	OnlineBackup	Count	Percentage
0	No	3087	43.9%
1	Yes	2425	34.5%
2	No internet service	1520	21.6%

Business Insights

- Sebanyak 43,9% pelanggan tidak menggunakan layanan Online Backup, dan hanya 34,5% yang menggunakannya, sementara 21,6% tidak memiliki layanan internet. Artinya, mayoritas pelanggan internet tidak memanfaatkan layanan backup daring yang tersedia.

2.5.14 Bagaimana karakteristik distribusi pelanggan berdasarkan perlindungan perangkat?

[24]:
```
plot_countplot_with_hue(df, 'DeviceProtection')
```



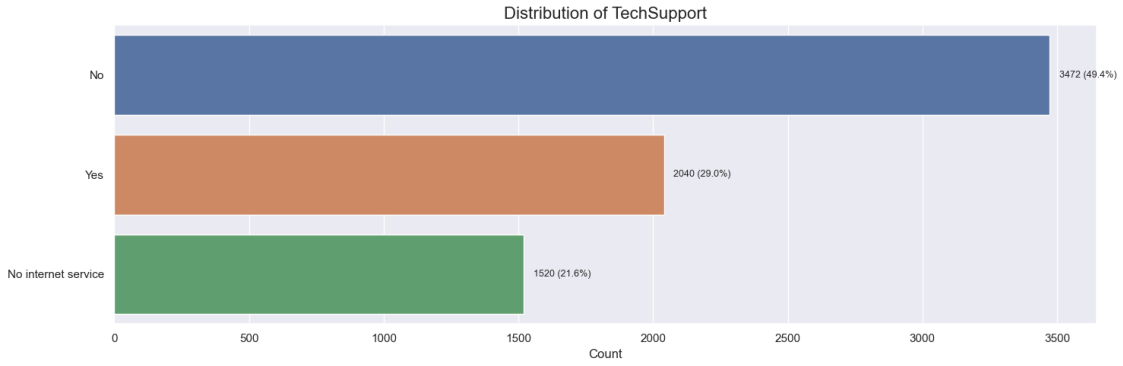
	DeviceProtection	Count	Percentage
0	No	3094	44.0%
1	Yes	2418	34.4%
2	No internet service	1520	21.6%

Business Insights

- Sebanyak 44,0% pelanggan tidak menggunakan layanan Device Protection, sementara 34,4% sudah memanfaatkannya. Sementara itu, 21,6% pelanggan tidak memiliki layanan internet. Ini menunjukkan bahwa perlindungan perangkat belum menjadi layanan yang dominan, meskipun masih memiliki potensi pasar yang cukup besar.

2.5.15 Bagaimana karakteristik distribusi pelanggan berdasarkan dukungan teknis?

[25]:
```
plot_countplot_with_hue(df, 'TechSupport')
```



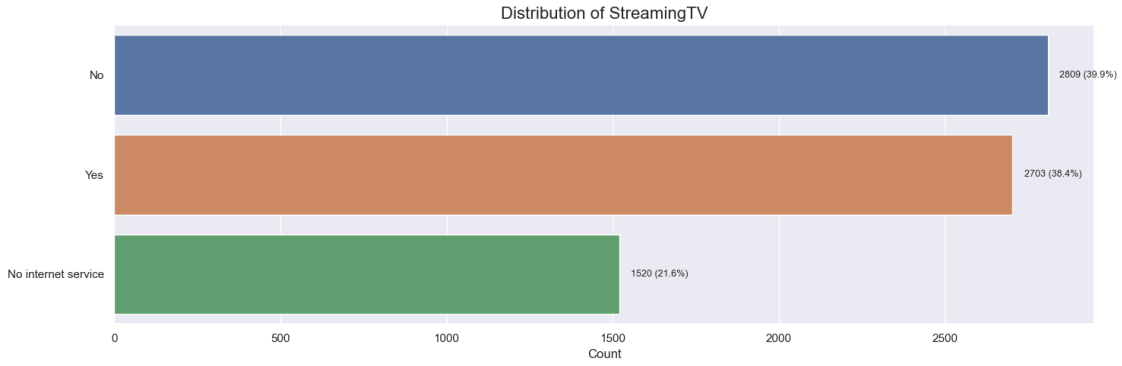
	TechSupport	Count	Percentage
0	No	3472	49.4%
1	Yes	2040	29.0%
2	No internet service	1520	21.6%

Business Insights

- Sebanyak 49,4% pelanggan tidak menggunakan layanan Tech Support, 29,0% menggunakannya, dan 21,6% tidak memiliki layanan internet. Ini menunjukkan bahwa mayoritas pelanggan internet belum memanfaatkan dukungan teknis sebagai bagian dari layanan mereka.

2.5.16 Bagaimana karakteristik distribusi pelanggan berdasarkan layanan streaming TV?

[26]:
```
plot_countplot_with_hue(df, 'StreamingTV')
```



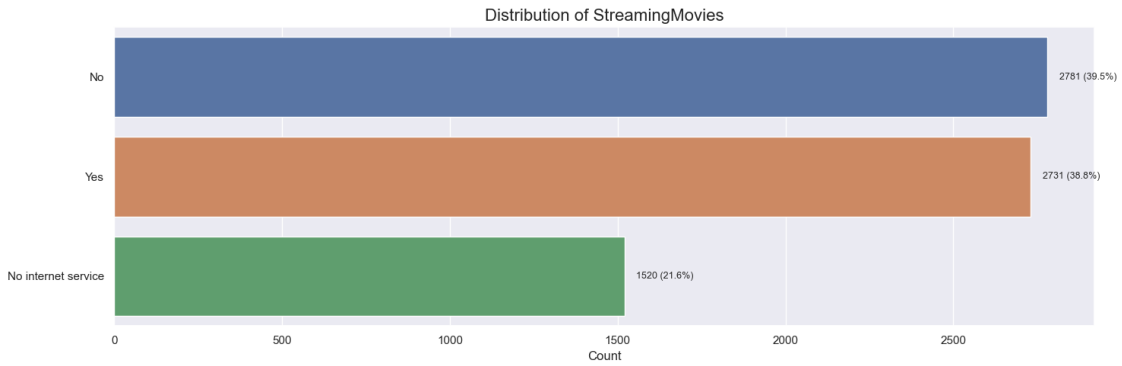
	StreamingTV	Count	Percentage
0	No	2809	39.9%
1	Yes	2703	38.4%
2	No internet service	1520	21.6%

Business Insights

- Penggunaan layanan Streaming TV terbagi cukup merata, dengan 39,9% pelanggan tidak menggunakannya dan 38,4% menggunakannya. Sebanyak 21,6% pelanggan tidak memiliki layanan internet. Ini menunjukkan bahwa layanan ini memiliki potensi pertumbuhan lebih lanjut di segmen pengguna internet.

2.5.17 Bagaimana karakteristik distribusi pelanggan berdasarkan layanan streaming film?

[27]:
```
plot_countplot_with_hue(df, 'StreamingMovies')
```



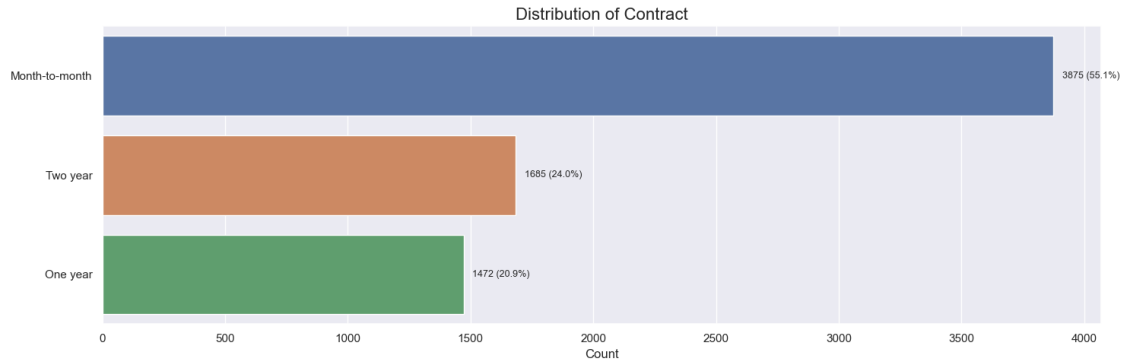
```
       StreamingMovies  Count Percentage
0                   No   2781      39.5%
1                  Yes   2731      38.8%
2  No internet service   1520      21.6%
```

Business Insights

- Penggunaan layanan Streaming Movies cukup seimbang, dengan 39,5% pelanggan tidak menggunakannya dan 38,8% menggunakannya. Sebanyak 21,6% pelanggan tidak memiliki layanan internet. Hal ini menunjukkan bahwa layanan ini memiliki tingkat adopsi yang hampir setara di kalangan pelanggan internet.

2.5.18 Bagaimana karakteristik distribusi pelanggan berdasarkan jenis kontrak langganan?

[28]:
```
plot_countplot_with_hue(df, 'Contract')
```



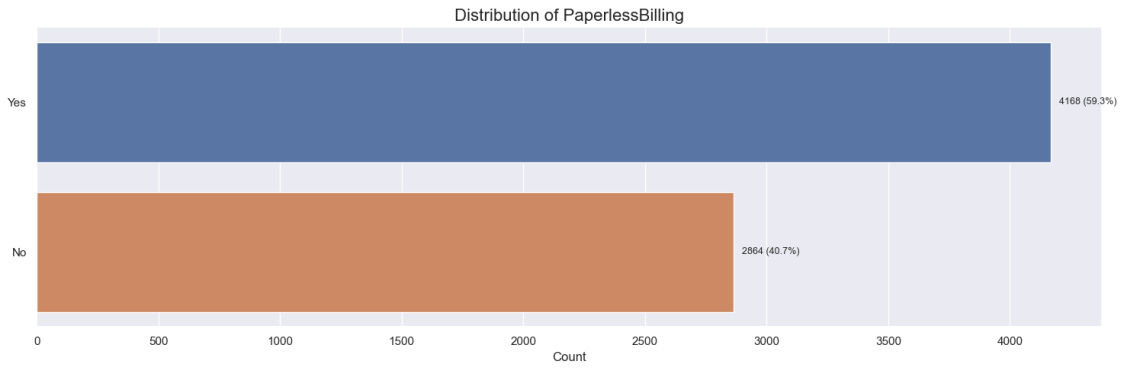

```
        Contract  Count Percentage
0  Month-to-month   3875      55.1%
1        Two year   1685      24.0%
2        One year   1472      20.9%
```

Business Insights

- Mayoritas pelanggan (55,1%) menggunakan kontrak bulanan (Month-to-month), sementara hanya 24,0% memilih kontrak dua tahun dan 20,9% kontrak satu tahun. Ini menunjukkan bahwa lebih dari separuh pelanggan memilih fleksibilitas tanpa komitmen jangka panjang.

2.5.19 Bagaimana karakteristik distribusi pelanggan berdasarkan metode paperless?

[29]:
```
plot_countplot_with_hue(df, 'PaperlessBilling')
```



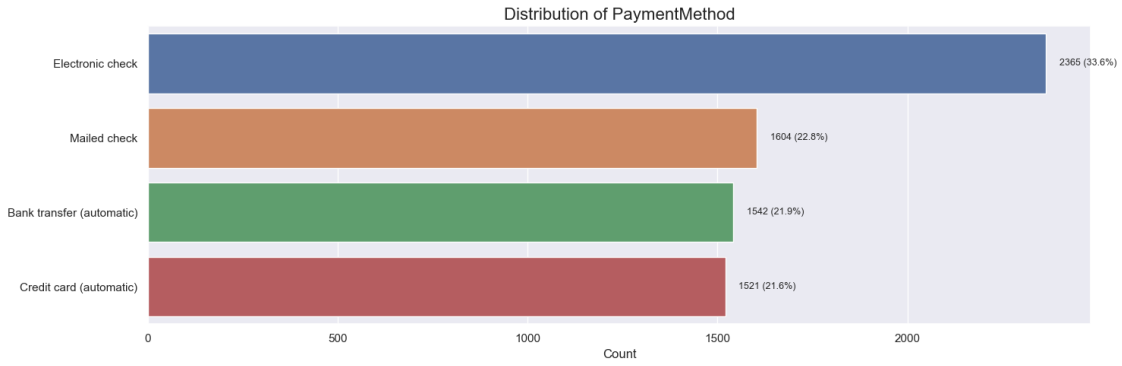
	PaperlessBilling	Count	Percentage
0	Yes	4168	59.3%
1	No	2864	40.7%

Business Insights

- **Jumlah pelanggan yang menggunakan *Paperless Billing* lebih banyak daripada yang tidak.** Hal ini menunjukkan bahwa mayoritas pelanggan lebih memilih untuk menerima tagihan secara digital daripada melalui surat fisik.

2.5.20 Bagaimana karakteristik distribusi pelanggan berdasarkan metode pembayaran?

[30]:
```
plot_countplot_with_hue(df, 'PaymentMethod')
```



	PaymentMethod	Count	Percentage
0	Electronic check	2365	33.6%
1	Mailed check	1604	22.8%
2	Bank transfer (automatic)	1542	21.9%
3	Credit card (automatic)	1521	21.6%

Business Insights

- Sebagian besar pelanggan (59,3%) menggunakan metode tagihan tanpa kertas (Paperless Billing), sementara 40,7% masih memilih tagihan fisik. Ini menunjukkan bahwa mayoritas pelanggan sudah terbiasa dengan pendekatan digital, meskipun sebagian masih memilih metode tradisional.

1.1.6 2.6 Exploratory Data Analysis (EDA) - Bivariat

[31]:
```
def plot_box_and_kde_churn(df, col, target='Churn', figsize=(15, 6)):
    # Visualisasi
    fig, axes = plt.subplots(nrows=1, ncols=2, figsize=figsize)
    plt.suptitle(f'Distribution of {col} Based on {target}', fontsize=16, y=1.
↪02)

    # Boxplot berdasarkan Churn
    # Menggunakan y=target untuk membuat boxplot terpisah untuk setiap kategori␣
↪Churn
    sns.boxplot(data=df, x=col, y=target, orient='h', ax=axes[0], hue=target)
    axes[0].set_title(f'Boxplot of {col}')

    # KDE plot berdasarkan Churn
    # Menggunakan hue=target untuk membuat kurva distribusi terpisah
    sns.kdeplot(data=df, x=col, hue=target, fill=True, common_norm=False,␣
↪alpha=0.5, ax=axes[1])
    axes[1].set_title(f'Distribution of {col}')

    plt.tight_layout(rect=[0, 0, 1, 0.95])
    plt.show()

    # Statistik deskriptif yang dikelompokkan berdasarkan Churn
    stats = df.groupby(target)[col].describe()
    q1 = stats['25%']
    q3 = stats['75%']
    iqr = q3 - q1
    stats['IQR'] = iqr

    # Tampilkan tabel statistik
    display(stats)
```

Fungsi `plot_box_and_kde_churn` digunakan untuk menganalisis distribusi variabel numerik berdasarkan kategori target seperti 'Churn'. Fungsi ini menghasilkan dua visualisasi berdampingan: boxplot horizontal yang memisahkan data berdasarkan nilai target untuk mengamati sebaran dan potensi outlier, serta KDE plot (Kernel Density Estimate) untuk melihat pola distribusi tiap kategori secara halus. Selain visualisasi, fungsi ini menghitung dan menampilkan tabel statistik deskriptif (count, mean, std, min, kuartil, dan IQR) yang dikelompokkan berdasarkan nilai target, sehingga mempermudah perbandingan karakteristik numerik antar kategori.

[32]:
```
def plot_stacked_barh_churn(df, col, target='Churn', figsize=(15, 5)):
    # Hitung count & percentage
    counts = df.groupby([col, target], observed=True).size().
 ↪unstack(fill_value=0)
    percentages = counts.div(counts.sum(axis=1), axis=0) * 100

    # Pastikan kolom urut: No, Yes
    desired_order = ['No', 'Yes']
    actual_order = [c for c in desired_order if c in percentages.columns]
    counts = counts[actual_order]
    percentages = percentages[actual_order]

    # Urutkan berdasarkan persentase churn ascending
    sort_order = percentages['No'].sort_values(ascending=True).index
    counts = counts.loc[sort_order]
    percentages = percentages.loc[sort_order]

    # Visualisasi manual
    fig, ax = plt.subplots(figsize=figsize)
    left = [0] * len(percentages)

    for status in actual_order:
        values = percentages[status]
        bar = ax.barh(percentages.index, values, left=left, label=status)

        for i, (pct, cnt) in enumerate(zip(values, counts[status])):
            if pct > 0:
                ax.text(left[i] + pct / 2, i, f'{pct:.1f}%\n({cnt})',
                        ha='center', va='center', fontsize=8,
                        color='white' if pct > 15 else 'black')
        left = [l + v for l, v in zip(left, values)]

    plt.suptitle(f'Distribution of {col} Based on {target}', fontsize=16, y=1.
 ↪02)
    ax.set_xlabel('Percentage')
    ax.set_ylabel('')
    ax.legend(title=target, loc='center left', bbox_to_anchor=(1.0, 0.5))
    plt.tight_layout()
    plt.show()

    # Tabel gabungan (Count dan Persentase) dengan reset index
    table = counts.copy()
    for status in actual_order:
        table[f'{status} (%)'] = percentages[status].apply(lambda x: f'{x:.
 ↪1f}%')

    table = table.reset_index()
```

```
    display(table)
```

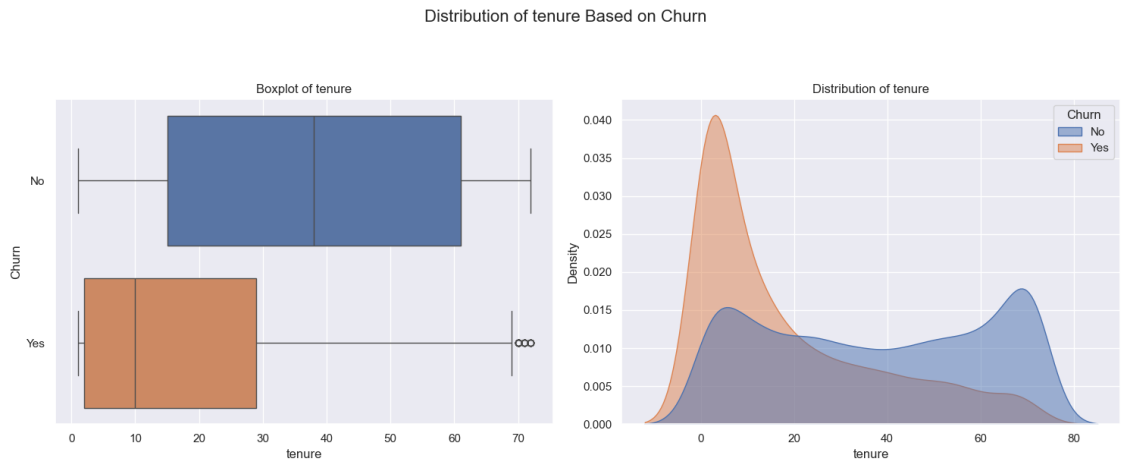
Fungsi `plot_stacked_barh_churn` digunakan untuk memvisualisasikan distribusi variabel kategorikal berdasarkan target seperti 'Churn' dalam bentuk stacked bar chart horizontal. Fungsi ini menghitung jumlah dan persentase tiap kategori target dalam setiap kelas variabel, lalu mengurutkan data berdasarkan persentase kategori 'No' secara menaik. Visualisasi menampilkan batang tersegmentasi untuk masing-masing kategori target, disertai anotasi berupa persentase dan jumlah absolut untuk mempermudah interpretasi. Fungsi ini sangat berguna untuk membandingkan proporsi churn antar kategori dan mengidentifikasi kelompok dengan tingkat churn tinggi.

[33]:
```
def mannwhitney_test(df, col, target='Churn'):
    # Pastikan hanya ada dua kategori pada target
    categories = df[target].dropna().unique()
    if len(categories) != 2:
        raise ValueError(f"Kolom target '{target}' harus memiliki tepat 2␣
↪kategori (misal: 'Yes' dan 'No').")

    # Ambil data dari masing-masing grup
    group1 = df[df[target] == categories[0]][col].dropna()
    group2 = df[df[target] == categories[1]][col].dropna()

    # Uji Mann-Whitney U dua arah
    stat, p_value = mannwhitneyu(group1, group2, alternative='two-sided')

    # Tampilkan hasil
    print(f"P-value untuk uji Mann-Whitney U antara '{categories[0]}' dan␣
↪'{categories[1]}' pada kolom '{col}': {p_value:.4f}")
    if p_value < 0.05:
        print(f"Tolak Ho, terima Ha. Terdapat perbedaan signifikan antara grup␣
↪'{categories[0]}' dan '{categories[1]}' untuk '{col}'.")
    else:
        print(f"Gagal tolak Ho. Tidak terdapat perbedaan signifikan antara grup␣
↪'{categories[0]}' dan '{categories[1]}' untuk '{col}'.")
```

Fungsi ini bertujuan untuk menguji Bagaimana perbedaan yang signifikan secara statistik antara dua kelompok pada kolom numerik tertentu berdasarkan kategori target seperti 'Churn'. Ia menggunakan uji non-parametrik Mann-Whitney U, yang cocok digunakan saat data tidak diasumsikan berdistribusi normal. Fungsi ini membandingkan distribusi nilai antara dua grup (misalnya 'Yes' dan 'No'), menghasilkan nilai p-value, dan memberikan interpretasi Bagaimanakah hipotesis nol dapat ditolak atau tidak. - Ho: Tidak ada perbedaan signifikan antara kedua grup. - Ha: Terdapat perbedaan signifikan antara kedua grup.

2.6.1 Apakah terdapat perbedaan lama pelanggan berlangganan antara yang churn dan tidak churn?

[34]:
```
plot_box_and_kde_churn(df, 'tenure')
mannwhitney_test(df, 'tenure')
```



Churn	count	mean	std	min	25%	50%	75%	max	IQR
No	5163.0	37.650010	24.076940	1.0	15.0	38.0	61.0	72.0	46.0
Yes	1869.0	17.979133	19.531123	1.0	2.0	10.0	29.0	72.0	27.0

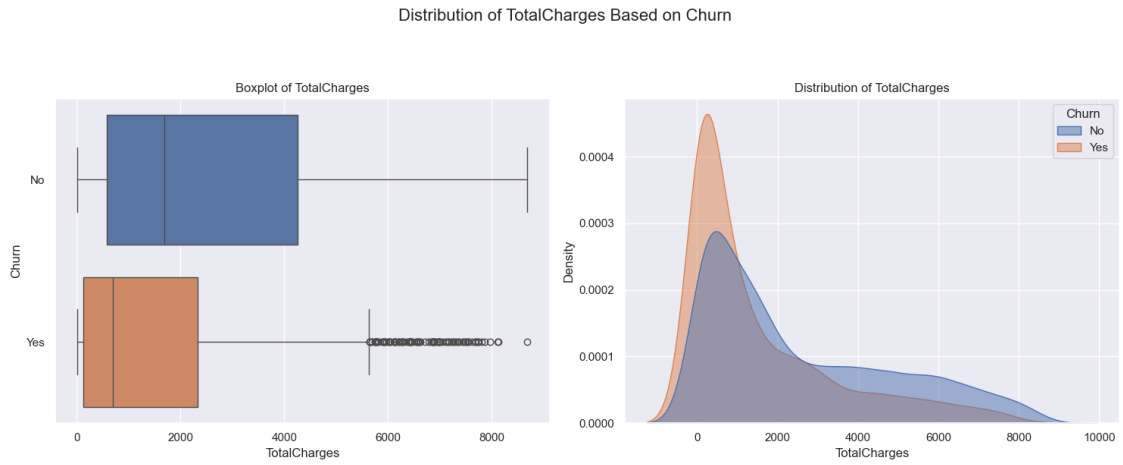
```
P-value untuk uji Mann-Whitney U antara 'No' dan 'Yes' pada kolom 'tenure':
0.0000
Tolak Ho, terima Ha. Terdapat perbedaan signifikan antara grup 'No' dan 'Yes'
untuk 'tenure'.
```

Business Insights

- Pelanggan yang tidak churn biasanya sudah lama berlangganan (median 38 bulan), sedangkan yang churn cenderung berhenti dalam waktu singkat (median 10 bulan). Artinya, risiko churn paling besar terjadi di awal masa langganan, terutama dalam 1 tahun pertama. Menjaga pelanggan di fase awal sangat penting untuk mengurangi churn.

2.6.2 Apakah terdapat perbedaan total pengeluaran pelanggan antara yang churn dan tidak churn?

[35]:
```
plot_box_and_kde_churn(df, 'TotalCharges')
mannwhitney_test(df, 'TotalCharges')
```



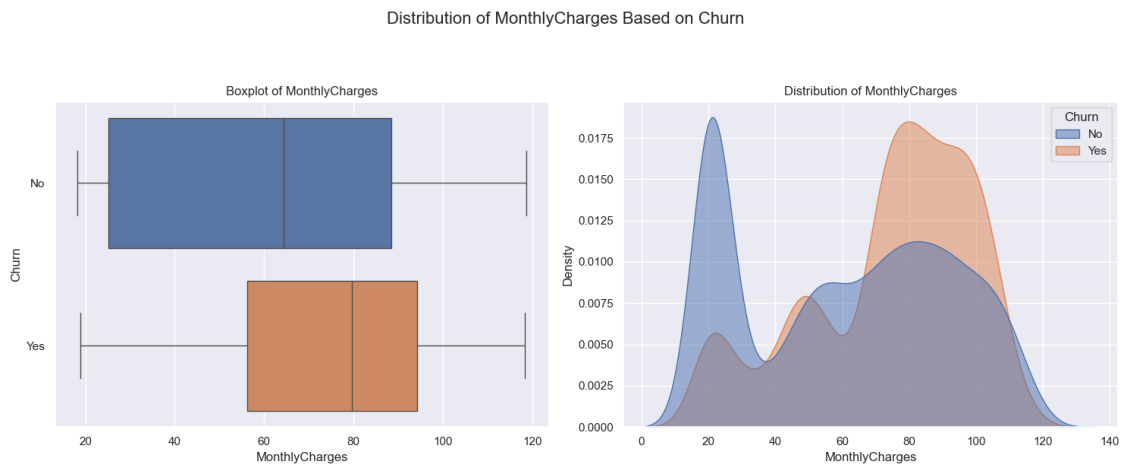

```
         count         mean          std    min      25%      50%       75%  \
Churn
No      5163.0  2555.344141  2329.456984  18.80  577.825  1683.60  4264.125
Yes     1869.0  1531.796094  1890.822994  18.85  134.500   703.55  2331.300

            max     IQR
Churn
No      8672.45  3686.3
Yes     8684.80  2196.8

P-value untuk uji Mann-Whitney U antara 'No' dan 'Yes' pada kolom
'TotalCharges': 0.0000
Tolak Ho, terima Ha. Terdapat perbedaan signifikan antara grup 'No' dan 'Yes'
untuk 'TotalCharges'.
```

Business Insights

- Pelanggan yang tidak churn memiliki total pengeluaran jauh lebih tinggi (median $1.683) dibanding pelanggan yang churn (median $704). Ini menunjukkan bahwa pelanggan churn sering kali berhenti sebelum memberikan nilai ekonomi maksimal. Banyak churn terjadi saat nilai pelanggan masih rendah.

2.6.3 Bagaimana perbedaan biaya bulanan pelanggan antara yang churn dan tidak churn?

[36]:
```
plot_box_and_kde_churn(df, 'MonthlyCharges')
mannwhitney_test(df, 'MonthlyCharges')
```



	count	mean	std	min	25%	50%	75%	max	\
Churn									
No	5163.0	61.307408	31.094557	18.25	25.10	64.45	88.475	118.75	
Yes	1869.0	74.441332	24.666053	18.85	56.15	79.65	94.200	118.35	

	IQR
Churn	
No	63.375
Yes	38.050

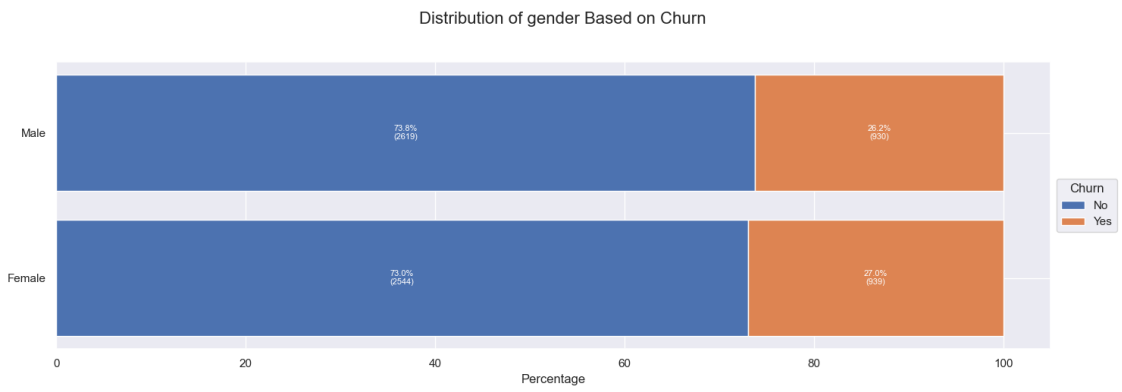
```
P-value untuk uji Mann-Whitney U antara 'No' dan 'Yes' pada kolom
'MonthlyCharges': 0.0000
Tolak Ho, terima Ha. Terdapat perbedaan signifikan antara grup 'No' dan 'Yes'
untuk 'MonthlyCharges'.
```

Business Insights

- Pelanggan yang churn cenderung memiliki tagihan bulanan lebih tinggi (median $79.65) dibanding yang tidak churn (median $64.45). Artinya, semakin mahal biaya bulanan, semakin besar kemungkinan pelanggan untuk berhenti.

2.6.4 Bagaimana perbedaan tingkat churn antara pelanggan laki-laki dan perempuan?

[37]:
```
plot_stacked_barh_churn(df, 'gender')
```



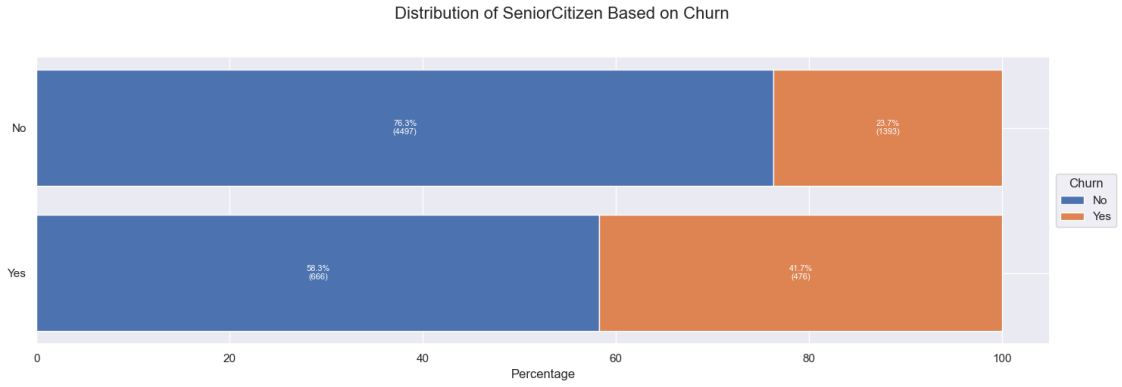
Churn	gender	No	Yes	No (%)	Yes (%)
0	Female	2544	939	73.0%	27.0%
1	Male	2619	930	73.8%	26.2%

Business Insights

- Tingkat churn antara pelanggan laki-laki (26.2%) dan perempuan (27.0%) hampir sama. Artinya, gender bukan faktor penentu utama dalam keputusan pelanggan untuk berhenti berlangganan.

2.6.5 Bagaimana perbedaan tingkat churn antara pelanggan lansia dan non-lansia?

[38]:
```
plot_stacked_barh_churn(df, 'SeniorCitizen')
```



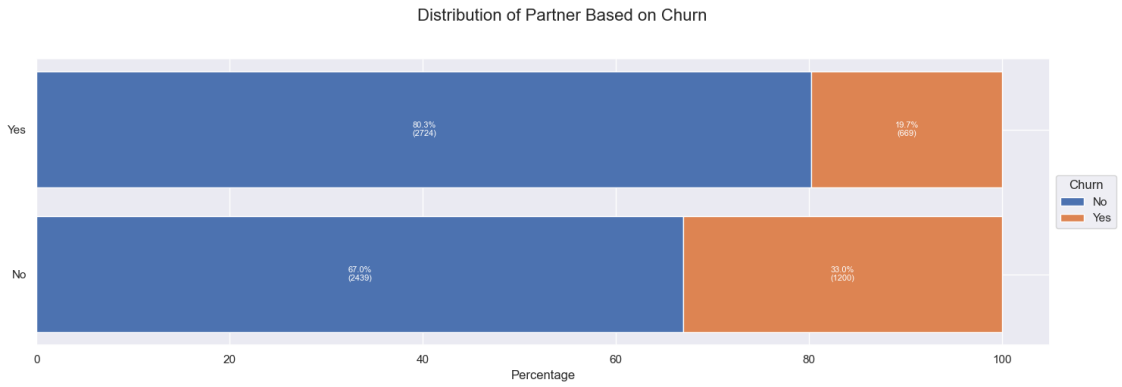
Churn	SeniorCitizen	No	Yes	No (%)	Yes (%)
0	Yes	666	476	58.3%	41.7%
1	No	4497	1393	76.3%	23.7%

Business Insights

- Pelanggan lansia (SeniorCitizen = Yes) memiliki tingkat churn yang jauh lebih tinggi (41.7%) dibandingkan pelanggan non-lansia (23.7%). Ini menunjukkan bahwa pelanggan lansia cenderung lebih rentan untuk berhenti berlangganan.

2.6.6 Bagaimana perbedaan tingkat churn antara pelanggan yang memiliki pasangan dan yang tidak?

[39]:
```
plot_stacked_barh_churn(df, 'Partner')
```



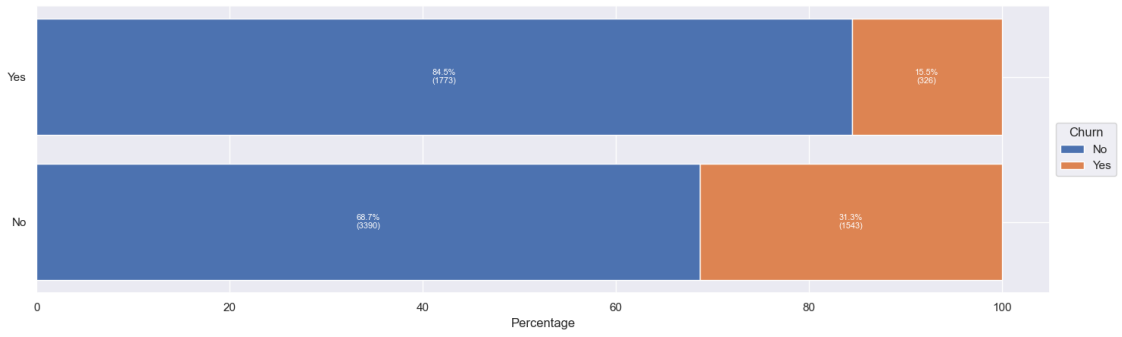
Churn	Partner	No	Yes	No (%)	Yes (%)
0	No	2439	1200	67.0%	33.0%
1	Yes	2724	669	80.3%	19.7%

Business Insights

- Pelanggan tanpa pasangan (Partner = No) memiliki tingkat churn yang jauh lebih tinggi (33.0%) dibandingkan pelanggan yang memiliki pasangan (19.7%). Ini menunjukkan bahwa pelanggan yang hidup sendiri cenderung lebih mudah meninggalkan layanan.

2.6.7 Bagaimana perbedaan tingkat churn antara pelanggan yang memiliki tanggungan dan yang tidak?

[40]:
```
plot_stacked_barh_churn(df, 'Dependents')
```



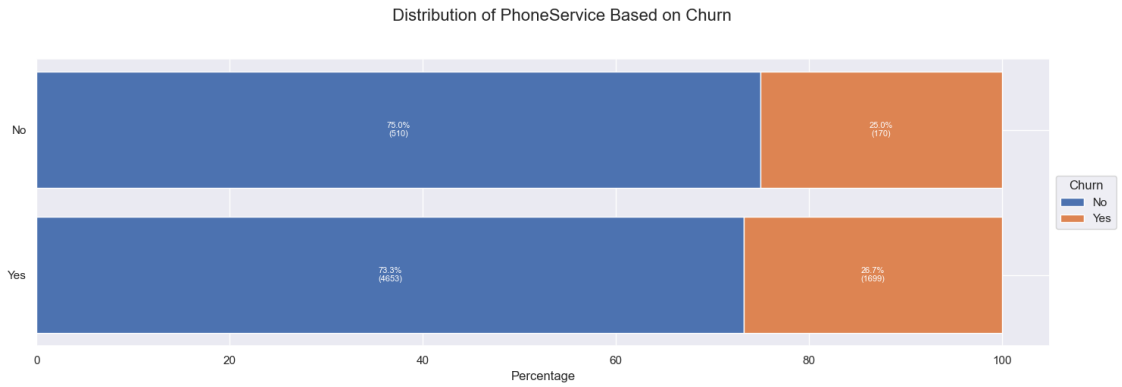
Churn	Dependents	No	Yes	No (%)	Yes (%)
0	No	3390	1543	68.7%	31.3%
1	Yes	1773	326	84.5%	15.5%

Business Insights

- Pelanggan tanpa tanggungan (Dependents = No) memiliki tingkat churn yang jauh lebih tinggi (31.3%) dibandingkan pelanggan dengan tanggungan (15.5%). Ini menunjukkan bahwa pelanggan dengan tanggungan cenderung lebih stabil dan loyal terhadap layanan.

2.6.8 Bagaimanakah tingkat churn berbeda antara pelanggan yang menggunakan layanan telepon dan yang tidak?

[41]:
```
plot_stacked_barh_churn(df, 'PhoneService')
```



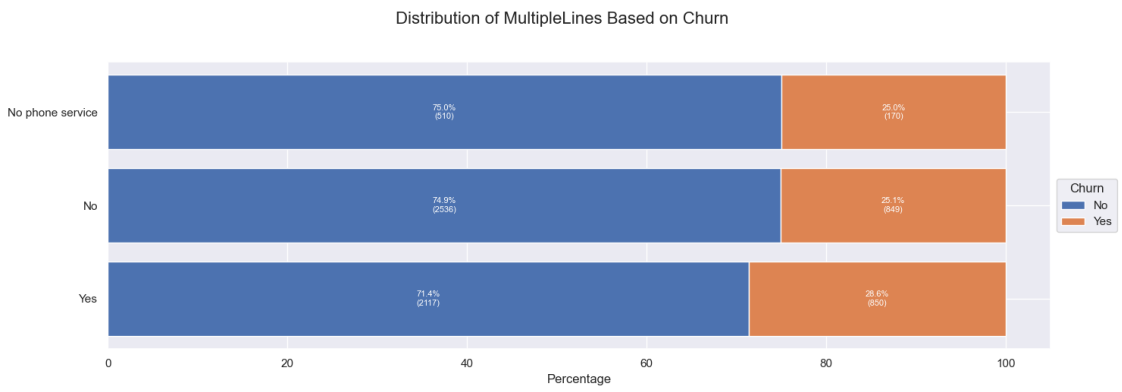
Churn	PhoneService	No	Yes	No (%)	Yes (%)
0	Yes	4653	1699	73.3%	26.7%
1	No	510	170	75.0%	25.0%

Business Insights

- Tingkat churn antara pelanggan yang memiliki layanan telepon (26.7%) dan yang tidak memiliki (25.0%) relatif serupa. Ini menunjukkan bahwa keberadaan layanan telepon tidak memiliki pengaruh signifikan terhadap churn.

2.6.9 Bagaimanakah tingkat churn berbeda antara pelanggan yang memiliki beberapa saluran telepon, satu saluran, atau tidak menggunakan layanan telepon?

[42]:
```
plot_stacked_barh_churn(df, 'MultipleLines')
```



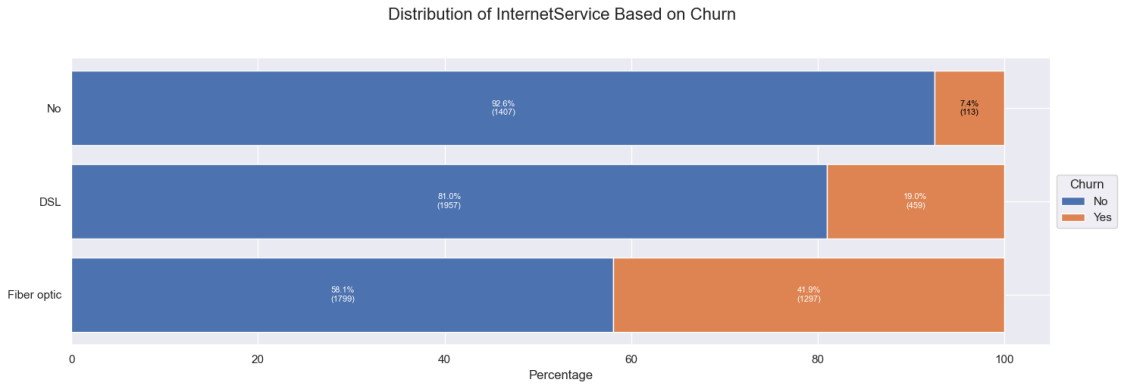
Churn	MultipleLines	No	Yes	No (%)	Yes (%)
0	Yes	2117	850	71.4%	28.6%
1	No	2536	849	74.9%	25.1%
2	No phone service	510	170	75.0%	25.0%

Business Insights

- **Pelanggan dengan layanan *Multiple Lines* memiliki tingkat *churn* yang sedikit lebih tinggi (sekitar 29%) dibandingkan pelanggan dengan satu jalur telepon (sekitar 25%).** Hal ini menunjukkan bahwa layanan tambahan ini, meskipun meningkatkan pendapatan, juga berkorelasi dengan risiko *churn* yang sedikit lebih besar.

2.6.10 Bagaimanakah tingkat churn berbeda antara pelanggan yang menggunakan DSL, fiber optic, atau tidak menggunakan layanan internet?

[43]:
```
plot_stacked_barh_churn(df, 'InternetService')
```



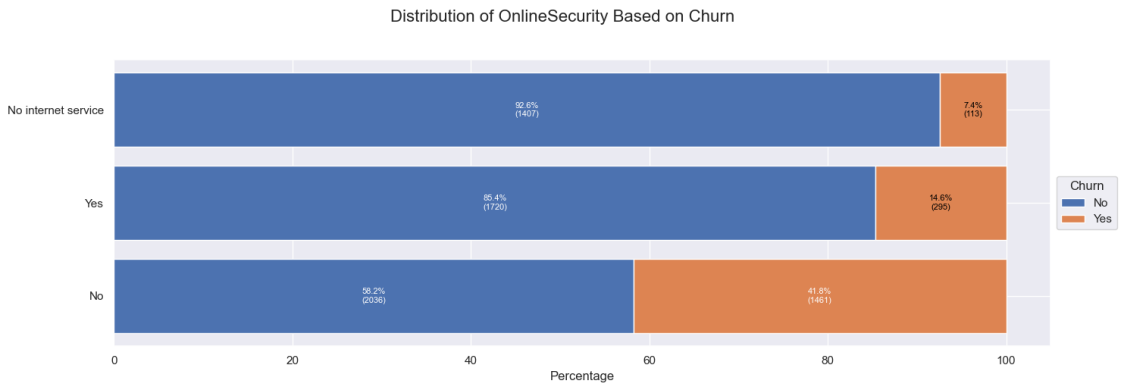
Churn	InternetService	No	Yes	No (%)	Yes (%)
0	Fiber optic	1799	1297	58.1%	41.9%
1	DSL	1957	459	81.0%	19.0%
2	No	1407	113	92.6%	7.4%

Business Insights

- Pelanggan dengan lebih dari satu saluran telepon (`MultipleLines = Yes`) memiliki tingkat churn sedikit lebih tinggi (28.6%) dibanding yang hanya punya satu saluran (25.1%). Meskipun selisihnya tidak besar, hal ini bisa mencerminkan bahwa pelanggan dengan banyak saluran mungkin lebih demanding dan sensitif terhadap layanan.

2.6.11 Bagaimanakah tingkat churn berbeda antara pelanggan yang menggunakan perlindungan keamanan online, tidak menggunakan, atau tidak memiliki layanan internet?

[44]:
```
plot_stacked_barh_churn(df, 'OnlineSecurity')
```



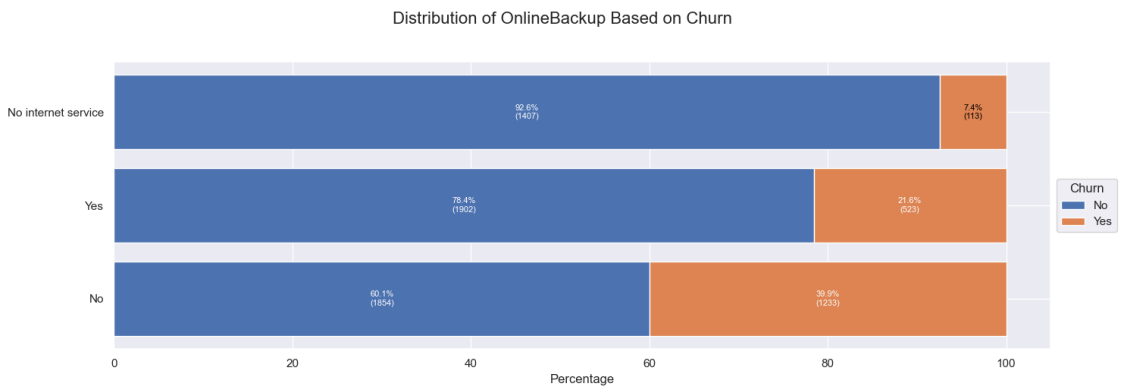
Churn	OnlineSecurity	No	Yes	No (%)	Yes (%)
0	No	2036	1461	58.2%	41.8%
1	Yes	1720	295	85.4%	14.6%
2	No internet service	1407	113	92.6%	7.4%

Business Insights

- Tingkat churn jauh lebih rendah pada pelanggan yang menggunakan layanan `OnlineSecurity` (14.6%) dibanding yang tidak menggunakannya (41.8%). Bahkan, pelanggan tanpa internet pun menunjukkan churn sangat rendah (7.4%), yang menegaskan pentingnya layanan keamanan digital sebagai faktor retensi utama.

2.6.12 Bagaimanakah tingkat churn berbeda antara pelanggan yang menggunakan cadangan data online, tidak menggunakan, atau tidak memiliki layanan internet?

[45]:
```
plot_stacked_barh_churn(df, 'OnlineBackup')
```



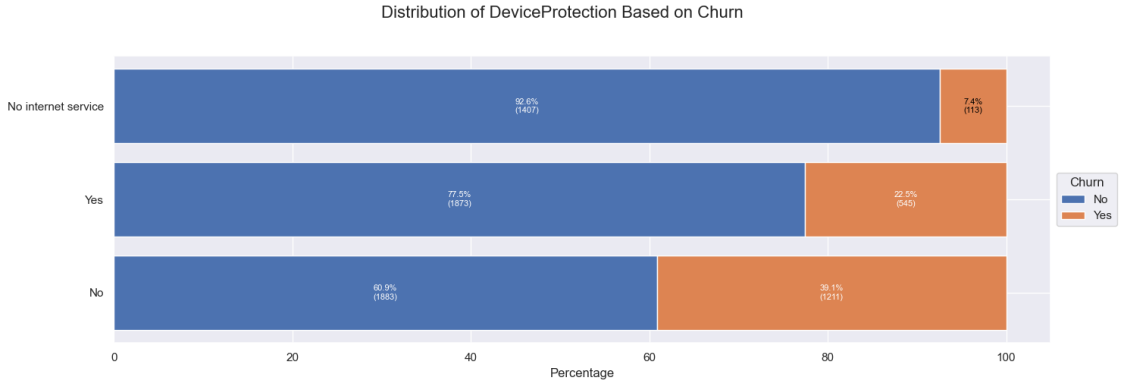
Churn	OnlineBackup	No	Yes	No (%)	Yes (%)
0	No	1854	1233	60.1%	39.9%
1	Yes	1902	523	78.4%	21.6%
2	No internet service	1407	113	92.6%	7.4%

Business Insights

- Pelanggan yang menggunakan layanan `OnlineBackup` memiliki tingkat churn lebih rendah (21.6%) dibanding yang tidak menggunakannya (39.9%). Ini menunjukkan bahwa layanan pendukung seperti backup data dapat meningkatkan loyalitas. Bahkan pelanggan tanpa layanan internet menunjukkan churn sangat rendah (7.4%).

2.6.13 Bagaimanakah tingkat churn berbeda antara pelanggan yang menggunakan perlindungan perangkat, tidak menggunakan, atau tidak memiliki layanan internet?

[46]:
```
plot_stacked_barh_churn(df, 'DeviceProtection')
```



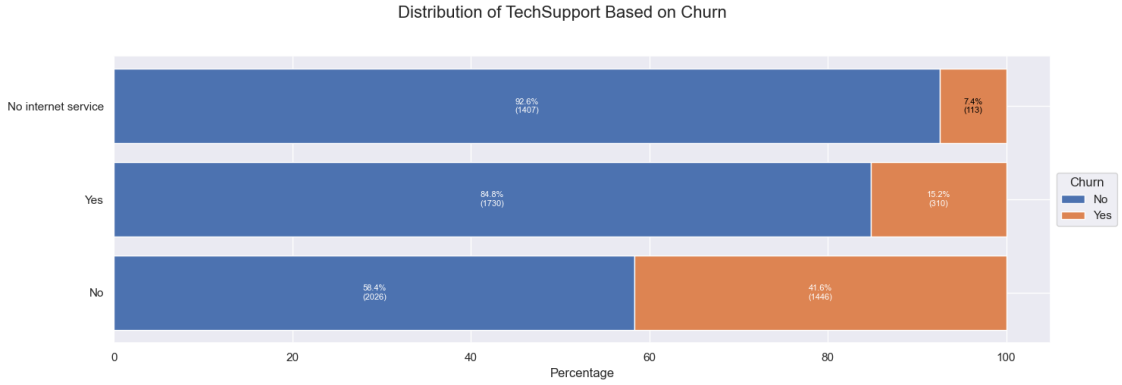
Churn	DeviceProtection	No	Yes	No (%)	Yes (%)
0	No	1883	1211	60.9%	39.1%
1	Yes	1873	545	77.5%	22.5%
2	No internet service	1407	113	92.6%	7.4%

Business Insights

- Pelanggan yang menggunakan layanan `DeviceProtection` memiliki tingkat churn lebih rendah (22.5%) dibanding yang tidak menggunakannya (39.1%). Hal ini mengindikasikan bahwa layanan tambahan yang memberikan rasa aman terhadap perangkat berkontribusi pada retensi pelanggan. Pelanggan tanpa internet menunjukkan churn sangat rendah (7.4%), menandakan keterkaitan kuat antara layanan internet dan kebutuhan akan proteksi perangkat.

2.6.14 Bagaimanakah tingkat churn berbeda antara pelanggan yang menggunakan dukungan teknis, tidak menggunakan, atau tidak memiliki layanan internet?

[47]:
```
plot_stacked_barh_churn(df, 'TechSupport')
```



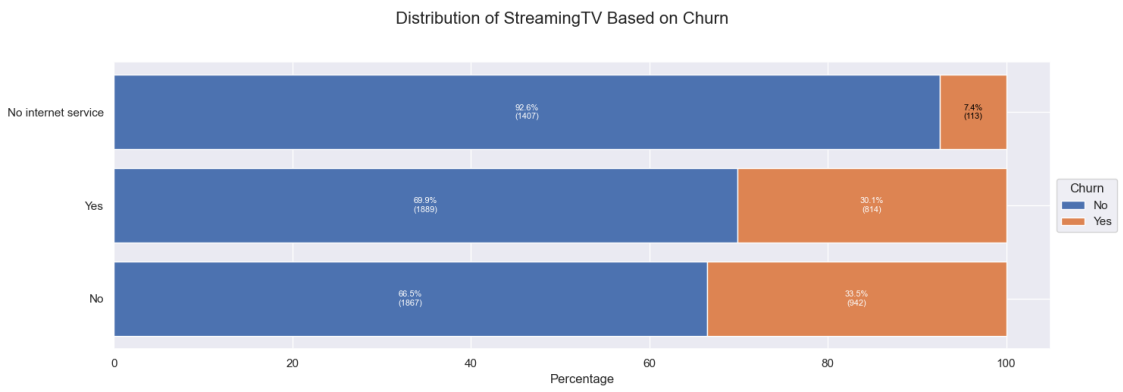
Churn	TechSupport	No	Yes	No (%)	Yes (%)
0	No	2026	1446	58.4%	41.6%
1	Yes	1730	310	84.8%	15.2%
2	No internet service	1407	113	92.6%	7.4%

Business Insights

- **Pelanggan internet yang tidak memiliki *Tech Support* memiliki tingkat *churn* yang sangat tinggi, yaitu sekitar 42%.** Ini menjadikan mereka sebagai salah satu segmen pelanggan dengan risiko tertinggi untuk berhenti berlangganan.
- **Sebaliknya, pelanggan yang memiliki *Tech Support* tingkat *churn*-nya sangat rendah, yaitu sekitar 15%.** Hal ini membuktikan bahwa layanan *Tech Support* adalah fitur yang sangat vital untuk mempertahankan pelanggan.
- **Secara keseluruhan, ketiadaan layanan *Tech Support* adalah salah satu prediktor *churn* terkuat.** Pelanggan internet tanpa dukungan teknis ini hampir tiga kali lebih mungkin untuk berhenti berlangganan.

2.6.15 Bagaimanakah tingkat churn berbeda antara pelanggan yang menggunakan layanan streaming TV, tidak menggunakan, atau tidak memiliki layanan internet?

[48]:
```
plot_stacked_barh_churn(df, 'StreamingTV')
```



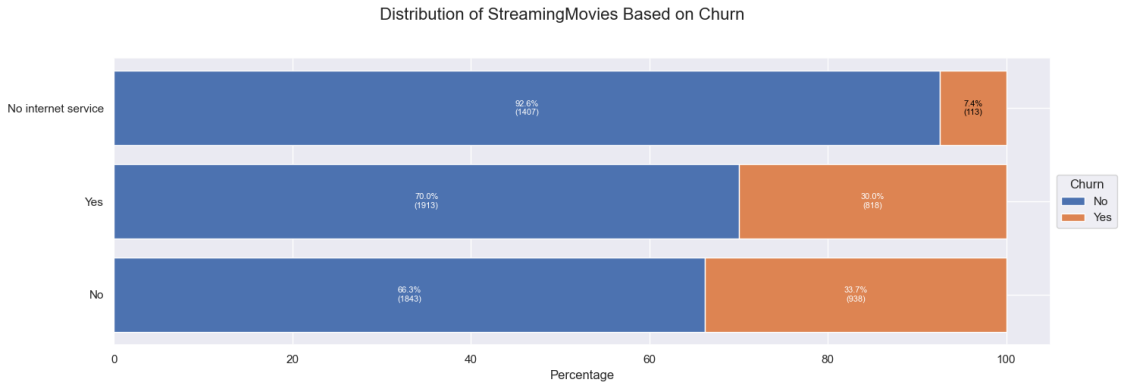
Churn	StreamingTV	No	Yes	No (%)	Yes (%)
0	No	1867	942	66.5%	33.5%
1	Yes	1889	814	69.9%	30.1%
2	No internet service	1407	113	92.6%	7.4%

Business Insights

- Pelanggan yang menggunakan layanan `TechSupport` memiliki tingkat churn jauh lebih rendah (15.2%) dibandingkan yang tidak menggunakannya (41.6%). Ini menunjukkan bahwa dukungan teknis merupakan layanan bernilai yang meningkatkan kepuasan dan loyalitas pelanggan. Sebaliknya, pelanggan tanpa akses internet juga menunjukkan churn sangat rendah (7.4%).

2.6.16 Bagaimanakah tingkat churn berbeda antara pelanggan yang menggunakan layanan streaming film, tidak menggunakan, atau tidak memiliki layanan internet?

[49]:
```
plot_stacked_barh_churn(df, 'StreamingMovies')
```



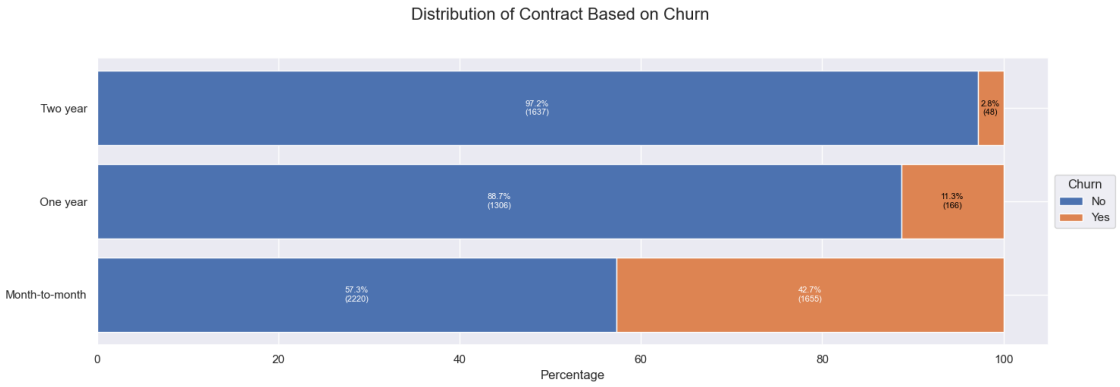
Churn	StreamingMovies	No	Yes	No (%)	Yes (%)
0	No	1843	938	66.3%	33.7%
1	Yes	1913	818	70.0%	30.0%
2	No internet service	1407	113	92.6%	7.4%

Business Insights

- Tingkat churn pelanggan yang menonton film (`Yes`) dan yang tidak (`No`) cukup mirip (30.0% vs 33.7%). Artinya, layanan `StreamingMovies` tidak terlalu berpengaruh dalam menekan churn. Sementara itu, pelanggan tanpa layanan internet kembali menunjukkan churn yang sangat rendah (7.4%).

2.6.17 Bagaimanakah tingkat churn berbeda berdasarkan jenis kontrak langganan yang dipilih pelanggan?

[50]:
```
plot_stacked_barh_churn(df, 'Contract')
```



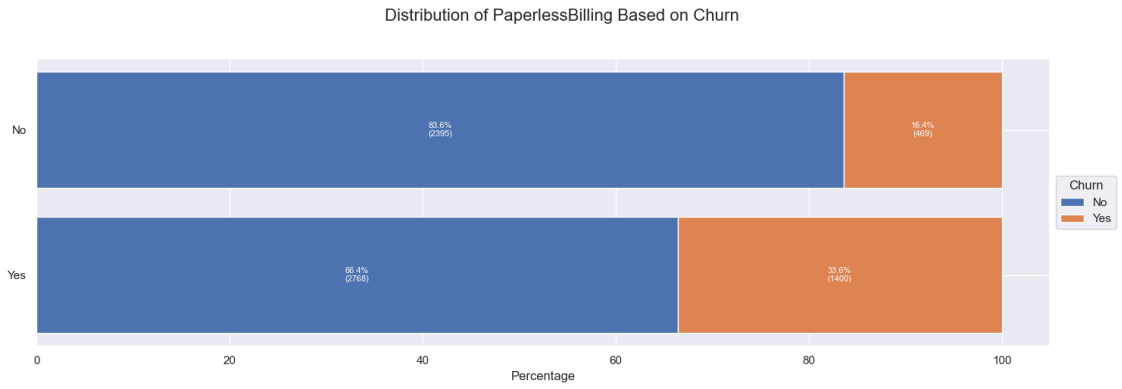
Churn	Contract	No	Yes	No (%)	Yes (%)
0	Month-to-month	2220	1655	57.3%	42.7%
1	One year	1306	166	88.7%	11.3%
2	Two year	1637	48	97.2%	2.8%

Business Insights

- Pelanggan dengan kontrak **month-to-month** memiliki tingkat churn tertinggi (42.7%), jauh lebih tinggi dibandingkan dengan kontrak **one year** (11.3%) dan **two year** (2.8%). Artinya, semakin lama durasi kontrak, semakin kecil kemungkinan pelanggan untuk churn.

2.6.18 Bagaimanakah tingkat churn berbeda antara pelanggan yang menggunakan tagihan tanpa kertas dan yang tidak?

[51]:
```
plot_stacked_barh_churn(df, 'PaperlessBilling')
```



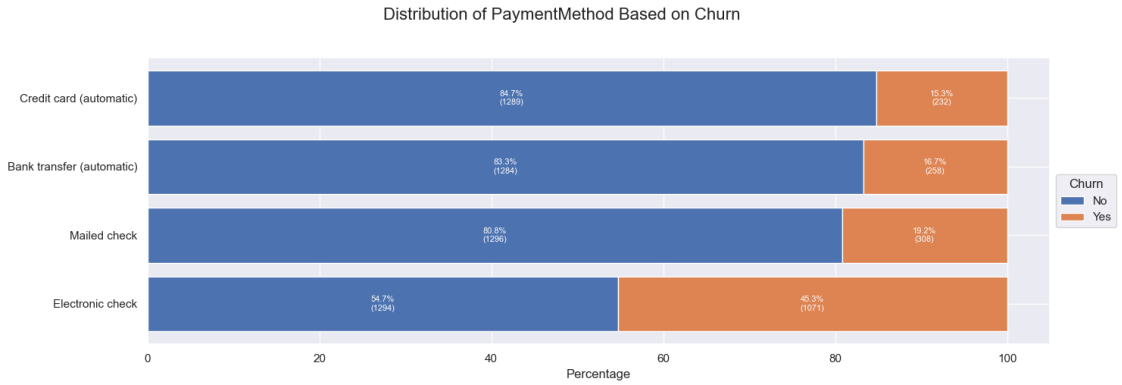
Churn	PaperlessBilling	No	Yes	No (%)	Yes (%)
0	Yes	2768	1400	66.4%	33.6%
1	No	2395	469	83.6%	16.4%

Business Insights

- Pelanggan yang menggunakan **paperless billing** memiliki tingkat churn lebih tinggi (33.6%) dibandingkan dengan yang tidak menggunakan (16.4%). Ini menunjukkan bahwa pelanggan digital cenderung lebih mudah berhenti berlangganan.

2.6.19 Bagaimanakah tingkat churn berbeda berdasarkan metode pembayaran yang digunakan pelanggan?

[52]:
```
plot_stacked_barh_churn(df, 'PaymentMethod')
```



Churn	PaymentMethod	No	Yes	No (%)	Yes (%)
0	Electronic check	1294	1071	54.7%	45.3%
1	Mailed check	1296	308	80.8%	19.2%
2	Bank transfer (automatic)	1284	258	83.3%	16.7%
3	Credit card (automatic)	1289	232	84.7%	15.3%

Business Insights

- Pelanggan yang menggunakan **Electronic check** memiliki tingkat churn paling tinggi (45.3%), jauh di atas metode pembayaran otomatis seperti **Credit card (15.3%)** atau **Bank transfer (16.7%)**.
- Metode manual seperti *mailed check* juga menunjukkan churn yang relatif rendah (19.2%).

1.1.7 2.7 Ringkasan EDA

- **Tenure (Lama Berlangganan)**
  Sebagian besar pelanggan berhenti dalam 9 bulan pertama, menunjukkan churn awal sangat tinggi. Pelanggan yang bertahan cenderung jauh lebih lama, dengan median 38 bulan dibandingkan 10 bulan untuk yang churn.
- **Total Charges (Total Pengeluaran)**
  Pelanggan churn biasanya belum memberi kontribusi besar secara ekonomi. Median pengeluaran mereka hanya $704, jauh lebih rendah dibandingkan $1.683 pada pelanggan non-churn.
- **Monthly Charges (Tagihan Bulanan)**
  Pelanggan dengan tagihan bulanan lebih tinggi lebih rentan churn. Median tagihan pelanggan churn mencapai $79.65, dibandingkan $64.45 pada pelanggan yang tetap.
- **Churn Rate (Tingkat Churn)**
  Tingkat churn dalam data ini cukup tinggi, yaitu 26,6%. Dataset juga tidak seimbang karena pelanggan non-churn hampir tiga kali lebih banyak.
- **Gender (Jenis Kelamin)**
  Distribusi gender seimbang antara pria dan wanita. Tidak ditemukan perbedaan signifikan dalam tingkat churn antar gender.
- **Senior Citizen (Lansia)**
  Pelanggan lansia memiliki tingkat churn jauh lebih tinggi (41,7%). Sementara itu, pelanggan non-lansia hanya 23,7%.
- **Partners (Status Pasangan)**
  Pelanggan yang tidak memiliki pasangan lebih sering churn, yaitu sebesar 33%. Sedangkan pelanggan yang memiliki pasangan hanya 19,7%.
- **Dependents (Tanggungan)**
  Mayoritas pelanggan tidak memiliki tanggungan, dan kelompok ini lebih rentan churn. Tingkat churn mereka adalah 31,3%, dibandingkan 15,5% pada yang memiliki tanggungan.
- **Internet Service**
  Layanan internet merupakan produk utama perusahaan, digunakan oleh 78,4% pelanggan. Ini menjadi layanan inti dalam portofolio.
- **Value-Added Services (Security, Backup, Tech Support)**
  Pelanggan yang menggunakan layanan tambahan seperti keamanan atau bantuan teknis cenderung lebih loyal. Layanan ini terbukti menurunkan churn secara signifikan.
- **Streaming Services**
  Penggunaan layanan streaming seperti TV dan film cukup seimbang. Namun, tidak terlihat pengaruh besar terhadap tingkat churn.

- **Contract Type (Jenis Kontrak)**
  Kontrak bulanan paling banyak dipilih namun paling berisiko, dengan churn 42,7%. Sebaliknya, kontrak tahunan atau dua tahun jauh lebih stabil.
- **Paperless Billing**
  Pelanggan yang menggunakan tagihan digital lebih sering churn (33,6%). Sedangkan yang masih memakai tagihan fisik memiliki churn lebih rendah (16,4%).
- **Payment Method (Metode Pembayaran)**
  Electronic check merupakan metode dengan churn tertinggi sebesar 45,3%. Metode otomatis seperti kartu kredit dan transfer bank jauh lebih stabil.

1.1.8 2.8 Rekomendasi Bisnis

- **Fokus pada Pelanggan Baru**
  Banyak pelanggan berhenti di 9 bulan pertama. Berikan edukasi, promo awal, dan dukungan aktif untuk meningkatkan retensi sejak awal langganan.
- **Bangun Program Loyalitas Bertahap**
  Pelanggan yang bertahan lebih dari 2 tahun memberi nilai lebih tinggi. Tawarkan program loyalitas seperti diskon, poin, atau layanan tambahan untuk mempertahankan mereka.
- **Tawarkan Paket Harga Menarik**
  Pelanggan dengan tagihan bulanan tinggi lebih mudah churn. Sediakan bundling atau paket harga menengah agar mereka merasa mendapatkan nilai lebih.
- **Segmentasi Pelanggan Rentan**
  Kelompok lansia, pelanggan tanpa pasangan, dan tanpa tanggungan lebih sering churn. Berikan pendekatan komunikasi dan penawaran khusus yang sesuai karakteristik mereka.
- **Dorong Penggunaan Layanan Tambahan**
  Layanan seperti Online Security dan Tech Support terbukti menurunkan churn. Berikan uji coba gratis atau promosi agar pelanggan tertarik menggunakannya.
- **Arahkan ke Kontrak Jangka Panjang**
  Pelanggan kontrak bulanan lebih sering churn. Beri insentif agar mereka pindah ke kontrak 1–2 tahun untuk meningkatkan retensi dan pendapatan tetap.
- **Perbaiki Metode Pembayaran**
  Metode *electronic check* memiliki churn tertinggi. Dorong pelanggan untuk pindah ke pembayaran otomatis seperti kartu kredit atau transfer bank.
- **Evaluasi Paperless Billing**
  Pelanggan paperless billing lebih sering churn. Tinjau kembali pengalaman digital untuk memastikan prosesnya mudah dan tidak membingungkan.
- **Lakukan Analisis Churn Lebih Dalam**
  Dengan churn cukup tinggi dan data tidak seimbang, perlu analisis lebih lanjut untuk memahami pola churn. Ini penting untuk membangun model prediksi yang lebih akurat dan adil.

1.1.9 2.9 Save Dataset

[53]:
```
df.to_csv('../data/telco_data_eda.csv', index=False)
```