Telecommunications Customer Churn Analysis

Part 1: Data Loading & Initial Inspection (with Code)

This section documents the first stage of the project: loading the datasets and performing an initial inspection to understand structure, size, and data quality. Python (pandas) was used for all operations.

1. Import Required Libraries

```python
import pandas as pd import numpy as np
```
Output: Libraries imported successfully.

2. Load the Datasets

```python
churn_df = pd.read_csv('telecom_customer_churn.csv') zipcode_df = pd.read_csv('telecom_zipcode_population.csv')
```
Output: Two DataFrames loaded without errors.

3. Inspect First Rows

```python
churn_df.head()
```
Output (sample): Displays customer demographics, services, contract details, charges, and customer status for the first five records.

4. Dataset Dimensions

```python
churn_df.shape
```
Output: (7043, ~50 columns) indicating 7,043 customers.

5. Data Types & Schema

```python
churn_df.info()
```
Output (summary): Mix of numerical and categorical variables; several columns contain missing values requiring further inspection.

6. Statistical Summary

```python
churn_df.describe()
```
Output (summary): Shows distributions of tenure, monthly charges, total charges, and age; no extreme anomalies detected at this stage.

Key Observations from Initial Inspection

1. Dataset contains 7,043 customer records from California.
2. Customer status includes Churned, Joined, and Stayed.
3. Missing values appear concentrated in service-related columns.
4. Data cleaning and feature engineering are required before analysis.

Part 2: Data Cleaning & Feature Engineering (with Code)

This section documents the data cleaning and preparation steps applied to the telecom churn dataset. The focus is on handling missing values correctly, fixing data types, and creating derived features to support downstream analysis.

1. Create a Working Copy

```python
df = churn_df.copy()
```

Output: Working copy of the dataset created.

2. Convert Incorrect Data Types

```python
df['Total Charges'] = pd.to_numeric(df['Total Charges'], errors='coerce')
```

Output: Total Charges converted to numeric; invalid values set to NaN.

3. Handle Structural Missing Values

```python
df['Churn Reason'] = df['Churn Reason'].fillna('Not Churned') df['Churn
Category'] = df['Churn Category'].fillna('Not Churned') df['Offer'] =
df['Offer'].fillna('No Offer')
```

Output: Churn-related fields populated with meaningful business labels.

4. Internet & Phone Service Imputation

```python
internet_cols = ['Internet Type','Unlimited Data','Streaming
Music','Streaming Movies', 'Streaming TV','Premium Tech Support','Device Protection
Plan', 'Online Security','Online Backup'] for col in internet_cols: df[col] =
df[col].fillna('No Internet') df['Avg Monthly GB Download'] = df['Avg Monthly GB
Download'].fillna(0) df['Multiple Lines'] = df['Multiple Lines'].fillna('No Phone
Service') df['Avg Monthly Long Distance Charges'] = df['Avg Monthly Long Distance
Charges'].fillna(0)
```

Output: Service-related missing values resolved based on service availability.

5. Feature Engineering

```python
df['Senior'] = (df['Age'] >= 65).astype(int) df['Age Group'] = pd.cut(
df['Age'], bins=[18, 30, 45, 60, 75, 100],
labels=['18–30','31–45','46–60','61–75','76+'] ) df['Customer Value'] = pd.qcut(
df['Monthly Charge'], q=3, labels=['Low','Medium','High'] )
```

Output: New analytical features created successfully.

6. Final Validation

```python
df.isnull().sum().sum()
```

Output: 0 (no remaining missing values).

Key Takeaways from Data Cleaning

1. Missing values were structural, not random.
2. Business-meaningful imputations preserved data integrity.
3. Derived features enabled robust churn segmentation.
4. Dataset is now analysis-ready.

Part 3: Churn Rate, Customer Status & Join Analysis (with Code)

This section establishes the key baseline metrics for the analysis, including overall churn rate, customer status distribution, and recent customer acquisition trends.

1. Customer Status Distribution

```python
df['Customer Status'].value_counts()
```

Output (summary): Customers are classified into Churned, Stayed, and Joined categories.

2. Overall Churn Rate Calculation

```python
churn_rate = (df['Customer Status'] == 'Churned').mean() * 100 churn_rate
```

Output: Displays overall churn rate as a percentage of total customers.

3. Total Customers Joined

```python
total_joined = (df['Customer Status'] == 'Joined').sum() total_joined
```

Output: Total number of newly joined customers.

4. Customers Joined in the Last Quarter

```python
last_quarter_joins = df[(df['Customer Status'] == 'Joined') & (df['Tenure in Months'] <= 3)] last_quarter_joins.shape[0]
```

Output: Number of customers who joined within the last three months.

Key Insights from Churn & Join Analysis

1. Overall churn rate provides a baseline measure of customer attrition.
2. A meaningful number of customers have joined recently, indicating active acquisition.
3. New customers are concentrated in low-tenure segments.
4. Early customer lifecycle is critical for retention.

Part 4: Customer Profiles – Churned vs Joined vs Stayed (with Code)

This section compares customer profiles across churned, joined, and stayed customers to understand whether specific demographic and service-related characteristics are associated with churn behavior.

1. Age Profile by Customer Status

```python df.groupby('Customer Status')['Age'].mean() ```

Output (summary): Displays average customer age for churned, joined, and stayed segments.

2. Age Group Distribution

```python df.groupby('Customer Status')['Age Group'].value_counts(normalize=True) * 100 ```

Output (summary): Age distribution percentages across customer statuses.

3. Contract Type Profile

```python df.groupby('Customer Status')['Contract'].value_counts(normalize=True) * 100 ```

Output (summary): Month-to-month contracts dominate churned customers, while long-term contracts are common among retained customers.

4. Tenure Comparison

```python df.groupby('Customer Status')['Tenure in Months'].mean() ```

Output: Churned customers have significantly lower average tenure.

5. Monthly Charges Comparison

```python df.groupby('Customer Status')['Monthly Charge'].mean() ```

Output: Churned customers tend to have higher monthly charges.

Key Profile Insights

1. Age differences exist but are not the primary churn driver.
2. Contract type strongly differentiates churned and retained customers.
3. Low tenure is strongly associated with churn.
4. Higher monthly charges increase churn risk.
```

Part 5: Key Drivers of Customer Churn (with Code)

This section identifies and validates the primary drivers of customer churn by analyzing contractual, financial, and service-related factors. The objective is to isolate variables that show strong associations with churn behavior.

1. Churn by Contract Type

```
```python df.groupby('Contract')['Customer Status'].value_counts(normalize=True) * 100 ```
```

Output (summary): Month-to-month contracts exhibit substantially higher churn rates compared to one-year and two-year contracts.

2. Churn by Internet Type

```
```python df.groupby('Internet Type')['Customer Status'].value_counts(normalize=True) * 100 ```
```

Output: Fiber internet customers show higher churn relative to DSL and non-internet customers.

3. Add-on Services & Churn

```
```python addon_cols = ['Online Security','Online Backup','Premium Tech Support','Device Protection Plan'] for col in addon_cols: print(df.groupby(col)['Customer Status'].value_counts(normalize=True) * 100) print() ```
```

Output (summary): Customers without add-on services are significantly more likely to churn, indicating that bundled services increase customer retention.

4. Customer Value Segments & Churn

```
```python pd.crosstab(df['Customer Value'], df['Customer Status'], normalize='index') * 100 ```
```

Output: High-value customers demonstrate a disproportionately higher churn rate, posing a revenue risk to the company.

Key Churn Drivers Identified

1. Month-to-month contracts are the strongest churn driver.
2. High monthly charges increase churn probability.
3. Fiber internet customers churn more frequently.
4. Lack of add-on services reduces customer stickiness.
5. High-value customers represent a critical retention risk.

Part 6: Geographic Analysis – Zipcode Population & Churn (with Code)

This section incorporates zipcode-level population data to explore whether geographic population density has any relationship with customer churn. The goal is to add spatial context without overcomplicating the analysis.

1. Standardize Zip Code Data Types

```
```python df['Zip Code'] = df['Zip Code'].astype(str) zipcode_df['Zip Code'] = zipcode_df['Zip Code'].astype(str) ```
```

Output: Zip Code columns standardized for accurate merging.

2. Merge Zipcode Population Data

```
```python df = df.merge(zipcode_df, on='Zip Code', how='left') ```
```

Output: Population data successfully merged into main dataset.

3. Create Population Segments

```
```python df['Population Segment'] = pd.qcut( df['Population'], q=3, labels=['Low Population','Medium Population','High Population'] ) ```
```

Output: Customers grouped into population-based segments.

4. Churn Analysis by Population Segment

```
```python pd.crosstab( df['Population Segment'], df['Customer Status'], normalize='index' ) * 100 ```
```

Output (summary): Slightly higher churn observed in high-population zip codes.

Key Geographic Insights

1. Customers in densely populated areas show marginally higher churn.
2. Competitive pressure may be greater in urban regions.
3. Geography has a secondary, not primary, impact on churn.

Part 7: Data Visualization – Communicating Insights (with Code)

This section presents key visualizations used to communicate churn insights effectively. The focus is on clarity and business relevance rather than excessive charting.

1. Customer Status Distribution

```
```python df['Customer Status'].value_counts().plot(kind='bar') plt.title('Customer Status Distribution') plt.ylabel('Number of Customers') plt.show() ```
```

Output (description): Bar chart showing the relative size of churned, stayed, and joined customers.

2. Churn Rate by Contract Type

```
```python contract_churn = df.groupby('Contract')['Customer Status'] \ .value_counts(normalize=True).rename('Percentage').mul(100).reset_index() contract_churn[contract_churn['Customer Status']=='Churned'] \ .plot(x='Contract', y='Percentage', kind='bar') plt.title('Churn Rate by Contract Type') plt.ylabel('Churn Percentage') plt.show() ```
```

Output: Month-to-month contracts show significantly higher churn than long-term contracts.

3. Monthly Charges vs Customer Status

```
```python df.boxplot(column='Monthly Charge', by='Customer Status') plt.title('Monthly Charges by Customer Status') plt.suptitle('') plt.show() ```
```

Output: Churned customers exhibit higher median monthly charges.

4. Customer Value Distribution by Status

```
```python pd.crosstab(df['Customer Status'], df['Customer Value'], normalize='index') \ .plot(kind='bar', stacked=True) plt.title('Customer Value Distribution by Status') plt.ylabel('Proportion') plt.show() ```
```

Output: High-value customers form a substantial portion of churned customers.

Visualization Takeaways

1 Visuals clearly reinforce contract type as the strongest churn driver.
2 Pricing pressure is visually evident among churned customers.
3 High-value customer loss is clearly communicated through stacked charts.
4 Simple visuals effectively support business decisions.

Part 8: Executive Summary & Business Recommendations

This final section consolidates insights from the entire analysis into a concise executive summary. It highlights the scale of customer churn, identifies the most important drivers, and proposes actionable recommendations to improve customer retention and revenue stability.

Executive Summary

1. The overall customer churn rate is significant, indicating meaningful revenue risk.
2. Churn is most prevalent among customers on month-to-month contracts.
3. Customers with shorter tenure are far more likely to churn than long-tenured customers.
4. High monthly charges and fiber internet subscriptions are associated with increased churn.
5. High-value customers account for a disproportionate share of churned customers.
6. Geographic population density shows a weak but observable relationship with churn.

Is the Company Losing High-Value Customers?

Yes. Analysis of customer value segments reveals that high-value customers exhibit a higher-than-average churn rate. This presents a direct financial risk, as the loss of these customers disproportionately impacts revenue compared to low-value churn.

Strategic Business Recommendations

1. Incentivize contract upgrades by offering discounts or benefits to convert month-to-month customers to one-year or two-year contracts.
2. Implement early-tenure engagement programs focused on the first 90 days to reduce initial churn risk.
3. Develop targeted retention strategies for high-value customers, including loyalty rewards and proactive customer outreach.
4. Bundle add-on services such as security and premium support to increase customer stickiness.
5. Review fiber internet pricing and value positioning to remain competitive in high-density markets.

This project demonstrates a full end-to-end data analytics workflow, from raw data ingestion to actionable business insights. The findings provide a data-driven foundation for reducing customer churn and improving long-term customer value.