Traceable Project Steps

Analyzing Customer Churn in the 'Bank Churners'

Methodological steps for data preparation, analysis, and visualization that were carried out to investigate customer churn based on two datasets.

1. Brief Presentation of the Problem Area / Introduction to the Topic

1.1 Content Overview. The core objective of this project, **"Bank Customer Churn Analysis,"** is to analyze and visualize customer data in order to gain insights into customer behavior and their tendencies to leave (churn). Two real-world datasets from the financial sector are explored and compared:

#BankChurners.csv
#Churn_Modelling.csv

The analysis is structured around the following key steps: **Data Analysis:** The analysis covers key customer information such as income categories, credit limits, usage behavior, and geographic origin. The goal is to identify which features influence customer churn. **Visualization:** Various visualization techniques are applied — including pie charts, bar charts, box plots, donut charts, and heatmaps — to make patterns and trends in the data visible. **Comparison & Segmentation:** Both datasets are segmented and compared based on defined criteria (e.g., income, geography, churn rate) to identify similarities and differences. **Deriving Insights:** Based on the analyses, conclusions are drawn about which customer groups are more likely to churn or which features support strong customer loyalty. **Communication of Results:** The findings are documented both visually (e.g., dashboards, charts) and in writing to make the analysis understandable and traceable.

Objectives of the Study. The main goal of this study is to generate data-driven insights into customer behavior, identify patterns and relationships, and derive actions to reduce customer churn. A special focus is placed on comparing two different banking datasets**. Data source**: Kaggle.com

1.2 Justification of the Topic. The analysis of customer churn is of high practical and economic relevance — especially in the banking and financial sector. The chosen topic “Bank Customer Churn Analysis” is especially significant for the following reasons: **Importance of Customer Retention:** In a highly competitive financial market, it is crucial for banks to retain their existing customers over the long term. Every lost customer results in financial loss. **Data-Driven Decision-Making:** The analysis of large datasets allows banks to develop targeted retention measures and detect churn risks early on. **Optimizing Service Quality and Customer Interaction:** By analyzing the data, banks can better understand which customer groups are at risk and where service offerings can be improved. **Competitive Advantage:** Companies that actively evaluate and use customer data are more competitive and capable of offering personalized services. **Development & Research:** The results of this project can serve as a foundation for further research — for example, in developing machine learning models for future churn prediction**.**

2. Traceable Steps

2.1 State of Research / Review of Existing Literature and Tutorials. The topic of customer churn in the banking sector has already been extensively covered in both academic research and practical case studies. It is evident that factors such as service quality, pricing, customer satisfaction, and digitalization are key drivers of churn [1]. Due to the large number of available providers and increasing market transparency, customers today switch service providers much more quickly. Therefore, churn is a central challenge for many banks [2]. The early detection of customers at risk of churn enables businesses to take targeted retention measures and minimize the costs of acquiring new customers [3, 4]. Customer data analysis and predictive churn detection models are now firmly integrated into data-driven banking strategies. Through the targeted analysis of relevant customer features (e.g., income, credit limit, usage behavior), high-risk customers can be identified and approached proactively [5, 6]. This project applies exactly these types of analytical and visualization methods to real banking customer data obtained from Kaggle datasets**.**

2.2 Research Question. The central research question of this study is: "How can customer churn in the banking industry be better understood, visualized, and reduced". The analysis is based on real-world customer data from two distinct banking datasets: #BankChurners.csv & #Churn_Modelling.csv

These datasets contain a wide range of customer attributes, including: •Credit Score •Age •Gender •Estimated Salary •Income Category •Geographic Origin (Geography) •Credit Limit •Behavioral Patterns (e.g., inactivity, contact frequency) The goal is to analyze and visually process this customer data in order to identify patterns, trends, and risk factors that contribute to customer churn.

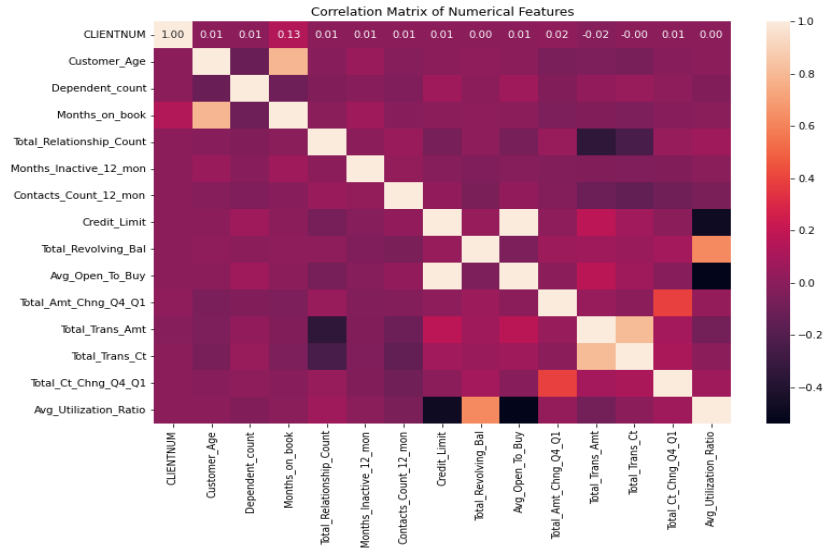
2.3 Methodological Approach / Implementation Steps. The following methodological approach was applied:
Step 1: Data Analysis & Preparation: • Import of required libraries •Data loading with error handling •Data cleaning (removal of duplicates, filling missing values) • Data type conversion for improved analysis

```
%%
Enhanced data loading with error handling ('BankChurners.csv')
Laden des Datensatzes mit Fehlerbehandlung
def load_data(file_path):
    try:
        if os.path.exists(file_path):
            df = pd.read_csv(file_path)
            print("Data loaded successfully.")
            print(df.head())
            return df
        else:
            print(f"File not found: {file_path}")
            return None
    except Exception as e:
        print(f"An error occurred: {e}")
        return None

Clean & preprocess data # Datenbereinigung und Vorverarbeitung
Removes duplicates, fills missing values, converts object columns to categor
def preprocess_data(df):
    df = df.drop_duplicates()
    df = df.ffill()
    for col in df.select_dtypes(include=['object']).columns:
        df[col] = df[col].astype('category')
    return df
```

Step 2: Exploratory Data Analysis (EDA): •Creation of visualizations for initial pattern recognition and segmentation •Identification of correlations and notable trends in customer behavior •Reporting functions including Excel report generation:

```
# Perform extended data analysis # Erweiterte Datenanalyse
# Generates a correlation matrix heatmap of numerical features
def analyze_data(df):
    numeric_df = df.select_dtypes(include=[np.number])
    correlation_matrix = numeric_df.corr()
    plt.figure(figsize=(12, 8))
    sns.heatmap(correlation_matrix, annot=True, fmt=".2f")
    plt.title('Correlation Matrix of Numerical Features')
    plt.show()
    print("Data analysis completed.")
```

```
# Generate Excel report from dataset # Reporting-Funktionen
# Saves raw data & descriptive stats to an Excel file
def generate_report(df):
    with pd.ExcelWriter('data_analysis_report.xlsx') as write
        df.to_excel(writer, sheet_name='Raw Data')
        df.describe().to_excel(writer, sheet_name='Descriptiv
        print("Report generated successfully!")
```



Visualizations & Analysis Functions

VisualizationType	Objective	Function	Analysis Content
Pie Charts	Distribution & frequency	plotting_pie() shows top 10 categories per column	Income_Category, Education_Level, Card_Category, etc.
Bar Charts	Distribution of numerical values	plotting_bar() uses histplot (for numerical), countplot (for categorical)	Credit_Limit, Balance, Age, Education_Level, Marital_Status
Swarmplot / Stripplot	Relationships between variables	plotting_swarm() shows individual data points per category; stripplot also for Age vs. Churn	Income_Category vs. Credit_Limit, Age vs. Churn
Boxplot	Spread and outlier detection	plot_boxplot_credit() reveals risk: customers with low scores	CreditScore or Balance vs. Churn

Step 3: Comparative Analysis. Comparison of *BankChurners* and *Churn_Modelling* datasets, including churn rate analysis by: • Geography (country) • Income Category • Credit Limit • Balance

Step 4: Dashboard & Reporting: Creation of a final dashboard including visual summaries and key recommendations

3. Data Compilation and Preparation

3.1 Preparation Phase: Importing Libraries and Loading Data. In the first step of the project, the necessary Python libraries were imported and the dataset was loade

```
%%
 Importing necessary libraries # Importieren der
mport os
mport numpy as np
mport pandas as pd
mport matplotlib.pyplot as plt
mport seaborn as sns
rom tqdm import tqdm
```

•OS for used to check whether the file exists and to manage file paths when loading CSV files (via os.path.exists())
•NumPy for numerical operations
•Pandas for processing, structuring, and analyzing tabular
data
•Matplotlib and Seaborn for creating professional charts and visualizations
•TQDM for displaying progress bars during iterative processes

Data Import and Initial Processing
•The dataset BankChurners.csv was imported using pd.read_csv(). •A copy (df_copy) of the original DataFrame was created to allow data manipulation without affecting the original data. •The first five rows of the data were displayed using df.head() to gain an overview of the dataset's structure.

This phase provides the foundation for all subsequent analysis and visualization steps.

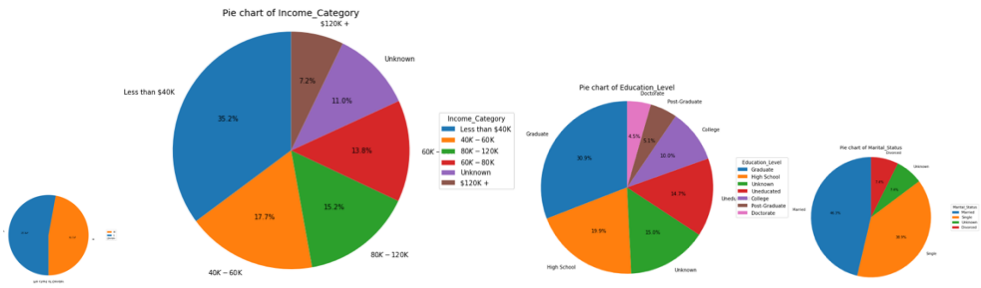
3.2 Presentation of High-Level Variables.

```
# Pie charts visualization # Visualisierungsfunktionen
def plotting_pie(df):
    exclude_columns = ['Card_Category']  # Optional
    categorical_columns = df.select_dtypes(include=['object', 'category']).columns

    for column in tqdm(categorical_columns):
        if column in exclude_columns:
            continue

        counts = df[column].value_counts().iloc[:10]

        plt.figure(figsize=(6, 6))
        wedges, texts, autotexts = plt.pie(
            counts,
            labels=counts.index,
            autopct='%1.1f%%',
            startangle=90,
            textprops={'fontsize': 10, 'color': 'black'}
        )

        plt.title(f'Pie chart of {column}', fontsize=14)
        plt.legend(wedges, counts.index, title=column, loc='center left', bbox_to_anchor=(1, 0.5))
        plt.axis('equal')  # Force circle shape # Kreisform erzwingen
        plt.tight_layout()
        plt.savefig(f'pie_{column}.png', bbox_inches='tight')
        plt.show()

# Bar Plot and Histogram visualization depending on feature type
def plotting_bar(df):
    exclude_columns = ['Gender', 'Education_Level', 'Marital_Status']  # Exclude these columns # Die

    for column in tqdm(df.columns):
        if column in exclude_columns:
            continue  # Skip excluded columns # Überspringe ausgeschlossene Spalten

        plt.figure(figsize=(10, 5))
        if pd.api.types.is_numeric_dtype(df[column]):
            sns.histplot(df[column], kde=True, bins=30) # Reduce bin count for better performance
        else:
            counts = df[column].value_counts().nlargest(10)  # Top 10 only
            sns.barplot(x=counts.index, y=counts.values)
            plt.xticks(rotation=45)
        plt.title(f'Bar chart of {column}')
        plt.savefig(f'bar_{column}.png', bbox_inches='tight')
        plt.show()

# Swarm Plot visualization
def plotting_swarm(df, x_col, y_col):
    sample_size = min(1000, len(df))  # Limit to 1000 samples for speed
    df_sample = df.sample(sample_size, random_state=42)  # Downsampling

    plt.figure(figsize=(10, 6))
    sns.swarmplot(x=df_sample[x_col], y=df_sample[y_col], size=2)  # Using smaller subset
    plt.title(f'Swarm plot of {y_col} by {x_col}')
    plt.savefig(f'swarm_{y_col}_by_{x_col}.png', bbox_inches='tight')
    plt.show()
```

3.2.1 Pie Charts. Overview of Most Frequent Categories. Using the plotting_pie() function, pie charts were created for selected categorical columns.

•The function loops through all columns of the DataFrame •The ten most frequent values per column are extracted and visualized as proportions •The charts are circular and display percentage values •Each chart is given a descriptive title and saved for reuse

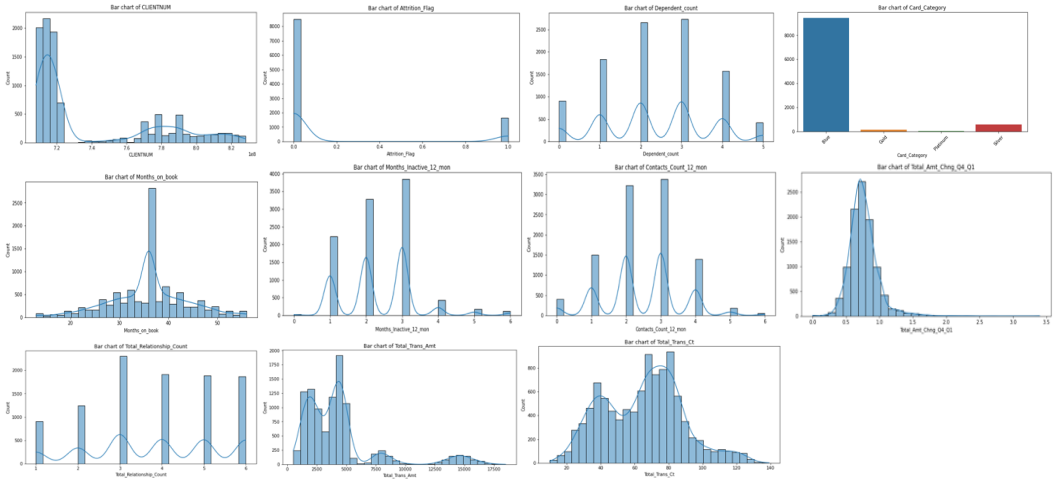
This type of visualization is ideal for identifying dominant categories - for example, in Education Level, Marital Status, or Gender.
Figure: Pie Chart



3.2.2 Bar and Histogram Charts. The plotting_bar() function creates distributions for numerical and categorical data:

•Numerical columns are visualized as histograms with density curves (KDE) •Categorical columns are shown as bar charts (Top 10 categories) •Some irrelevant or overly dominant columns (e.g., Gender**,** Marital_Status**)** are excluded

These diagrams make it possible to identify central tendencies and distributions**,** such as typical salary ranges **or** frequent credit limits**.** See: Illustration (Bar Chart / Histogram)

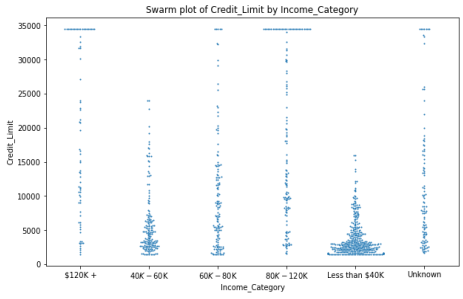


3.2.3 Swarm Plot. Relationship between Income and Credit Limit. An important visualization is the relationship between income category and credit limit, implemented using a swarm plot (s. below)**:**

python
KopierenBearbeiten
sns.swarmplot(x='Income_Category', y='Credit_Limit', data=df, size=2)

- The **X-axis** displays the income categories, and the **Y-axis** shows the credit limits
- Each point represents an individual customer
- The distribution provides insights into whether customers with higher incomes tend to have higher credit limits

```
# Interaction main interface
def main():
    file_path = input("Please enter the path to the dataset: ")  # Updated input prompt for clarity
    df = load_data(file_path)
    if df is not None:
        df = preprocess_data(df)
        analyze_data(df)

        while True:
            plot_choice = input("Enter 'pie', 'bar', 'swarm', or 'exit': ").lower()
            if plot_choice in ['pie', 'bar', 'swarm']:
                if plot_choice == 'swarm':
                    x_col = input("Enter X column name for swarm plot: ") #Income_Category
                    y_col = input("Enter Y column name for swarm plot: ") #Credit_Limit
                    if x_col in df.columns and y_col in df.columns:
                        plotting_swarm(df, x_col, y_col)
                    else:
                        print("Invalid column names. Try again.")
                elif plot_choice == 'pie':
                    plotting_pie(df)
                elif plot_choice == 'bar':
                    plotting_bar(df)
            elif plot_choice == 'exit':
                print("Exiting program.")
                break

# Main program start
if __name__ == "__main__":
    main()
```



4. Data Analysis and Visualization

4.1 Analysis Objective. The main objective of this analysis is to gain insights into customer behavior and churn based on real banking data. The purpose of the Exploratory Data Analysis (EDA) is to identify patterns, correlations, and influencing factors that affect customer retention.

4.2 Procedure and Methods. *The analysis was carried out in several steps:*

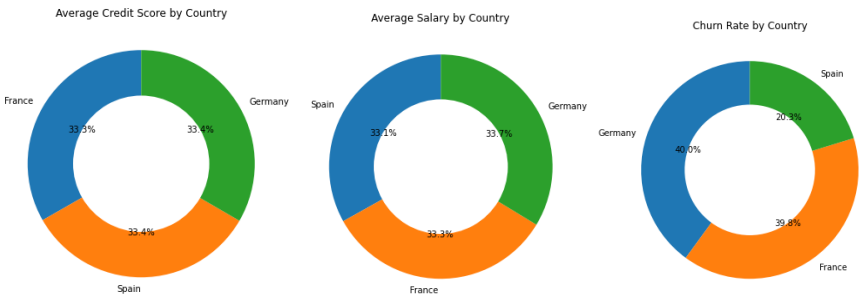
1. **Numerical Correlation Matrix**:
   A heatmap was used to visualize relationships between numerical variables (e.g., credit limit, inactivity, credit utilization).
   → **Goal**: Identify potential influencing factors on churn.
2. **Categorical Visualizations (Pie & Bar Plots)**:
   The data was segmented by features such as *Income_Category*, *Card_Category*, *Marital_Status*.
   → **Goal**: Display distribution and dominant groups.
3. **Relationship Analysis using Swarm/Box/Strip Plots**:
   - **Swarm plot**: Relationship between *Income_Category* and *Credit_Limit*
   - **Strip plot**: Age distribution based on churn status
   - **Box plot**: Analysis of *CreditScore* based on churn

4. **Country Comparison (Geography)**:
   One of the most informative insights came from analyzing data by country
   → Identified country-specific churn tendencies.
5. **Comparative Visualizations Between Datasets**:
   The two datasets (*BankChurners.csv* & *Churn_Modelling.csv*) were compared to highlight differences in:
   - Churn rate
   - Customer demographics (age, salary, score)
   - Segment behavior
6. **Dashboard Creation & Churn Prediction Simulation**:
   Based on behavioral factors (e.g., inactivity, contact frequency, credit utilization), a churn prediction was simulated and visually compared (predicted vs. actual).
   → **Goal**: Apply data-driven risk assessment.

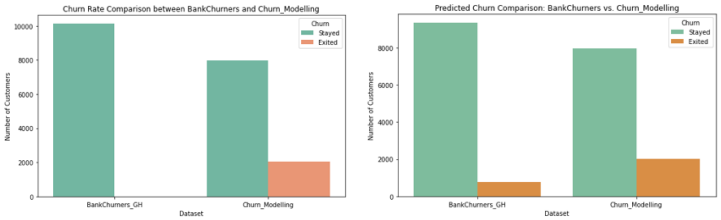
4.3 Visualization Results

Visualization Type	**Purpose & Findings**
Heatmap (Correlations)	Highlights relationships between numerical variables (e.g., *Balance* & *Credit_Limit*)
Donut Chart (Churn by Country)	Emphasizes countries with the highest churn rates
Swarmplot	Shows differences in credit limits across income groups
Histogram (Balance, Salary)	Visualizes distributions across the customer base
Dashboard Comparison	Final comparison of predicted vs. actual churn (e.g., 17% vs. 20%)

Visualization 1: *Credit Score & Salary by Country*: Average values varied significantly by country. **Visualization 2**: *Estimated Salaries*: Average customer salaries were similarly distributed across all three countries, indicating comparable income levels. **Visualization 3**: *Customer Churn by Country*: Certain countries exhibited significantly higher churn rates.

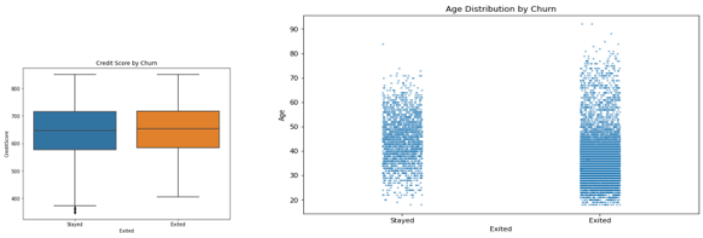


Visualization 4: *Comparison of Churn Rates*: A side-by-side comparison of the actual and predicted churn rates from both datasets:



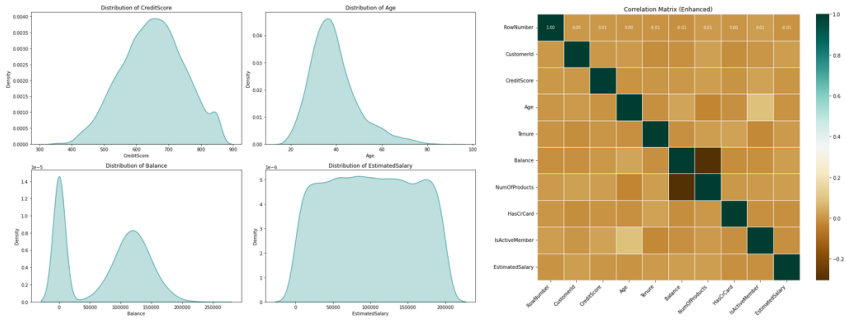
Visualization 5: *Dashboard Overview*: A combined presentation of bar charts, boxplots, pie charts, correlation matrix, and comparison plots for a holistic overview of the results.

1. **Boxplot** (see below): Credit Score vs. Churn Status – customers with similar credit scores are found in both groups (Stayed and Exited), indicating no clear correlation.
2. **Plot**: Customers who churned ("Exited") are, on average, older than those who stayed ("Stayed").



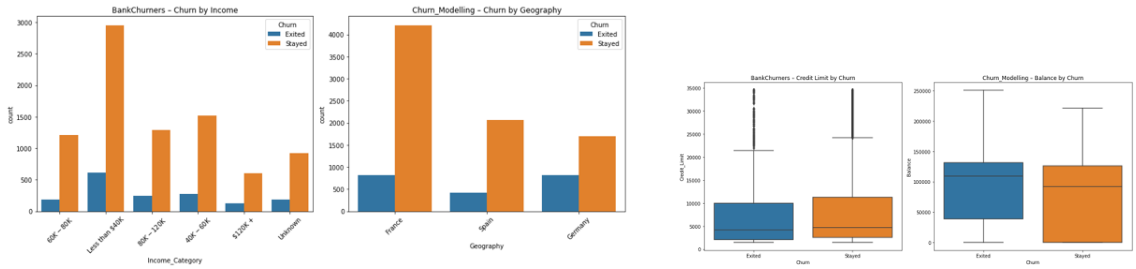
3. KDE Plots (see below): Kernel Density Estimates for four numerical variables in the dataset: •**CreditScore**: Slightly right-skewed distribution, concentrated between 550 and 750 points •**Age**: Right-skewed distribution with a peak around 35 years. •**Balance**: Bimodal distribution – some customers have very low balances, others much higher amounts. •**EstimatedSalary**: Almost uniformly distributed across the value range. → This visualization helps identify central tendencies and outliers for key numeric features in the customer base.

4. Correlation Matrix (Enhanced): Correlations among the dataset's numeric variables: •Overall, only weak correlations between most variables are observed •Slight negative correlation (~ –0.2**)** found between *Balance* and *NumOfProducts* •Other features like *Age*, *CreditScore*, or *EstimatedSalary* appear largely independent. → The correlation matrix helps assess which features may influence churn - or which might be redundant.

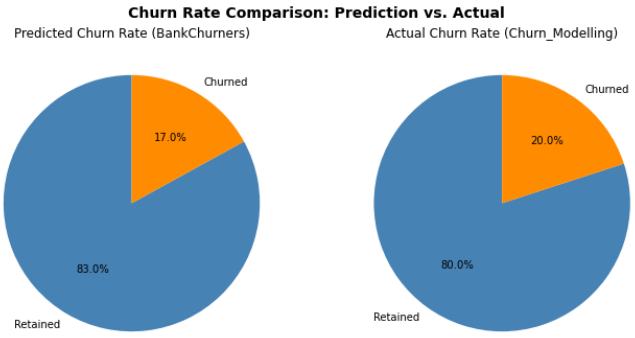


5. Churn by Income and Country (see below)**: Left**: In lower income brackets (e.g., *"Less than 40K"*), the share of churned customers (*Exited*) is noticeably higher. **Right**: France shows a significantly **higher churn rate** compared to Spain and Germany in the Churn_Modelling dataset.

6. Churn and Credit Behavior: **Left**: In the BankChurners dataset, customers with **lower credit limits** are more likely to churn. **Right**: In the Churn_Modelling dataset, churned customers tend to have a **higher account balance (Balance)** on average.



7. Comparison of Predicted vs. Actual Churn Rate: The left graph shows a **predicted churn rate of 17%** in the *BankChurners* dataset, while the actual churn rate **in** the *Churn_Modelling* dataset is **20%** - indicating **similar churn patterns** across both data sources.



In both datasets, customer churn was significantly influenced by: • Low customer activity • Infrequent contact by the bank • High credit utilization ratio**.**

Summary & Recommendations. The applied methodology provided valuable insights into the behavior of customers at risk of churn. **Recommendations:** • Inactive customers should be proactively approached with personalized offers • Customer retention programs are essential to reduce churn rates. **Future Potential:** •Implementation of **ML**/ machine learning models for churn prediction •Deeper segmentation and profiling of customers based on actual behavior

5. Final Results and Key Insights. The data analysis and visualizations carried out in the “Bank Customer Churn Analysis” project provided valuable insights into customer behavior and churn rates in the banking sector. **Key findings from both datasets:**

1. **Churn Rate Comparison: BankChurners vs. Churn_Modelling**
   • BankChurners.csv: Approx. 17% churn rate
   • Churn_Modelling.csv: Approx. 20% churn rate
   → Conclusion: Churn rates are at a similar level in both banks
2. **Customer Behavior by Income and Credit Limit**. In both datasets, higher-income customers had significantly higher credit limits. → Recommendation: Introduce premium customer loyalty programs tailored to high-value clients.
3. **Impact of Card Category on Churn**. Especially in BankChurners: Customers with "Blue" cards had a much higher churn rate compared to those with "Silver," "Gold," or "Platinum" cards.→ These customers tend to be more price-sensitive and switch banks more easily.
4. **Effect of Inactivity and Contact Frequency**. Customers with long periods of inactivity and low contact from the bank had a higher churn risk. → Recommendation: Engage inactive customers with personalized offers or consultations.
5. **Credit Behavior as a Churn Indicator**. Customers with high revolving balances and high utilization ratios were more likely to churn. → Suggests financial instability and the need for targeted support strategies.
6. **Correlation Between Credit Limit and Utilization**. Customers with lower credit limits showed higher utilization ratios. → Recommendation: Review whether increasing credit limits could improve retention.

Overall Conclusion. The conducted analysis provides a solid foundation for making data-driven decisions to reduce customer churn. → Visual analytics gives banks powerful tools to: •Reveal hidden behavioral patterns •Identify risk groups early •Deliver targeted, individualized customer support

6. Outlook. Strategic Recommendations and Future Potential. Based on the data analysis, the following strategic actions and future measures are recommended: **1. Strengthen Customer Retention Strategies**. Focus on at-risk segments, especially customers with low-tier cards or low activity. Measures: •Personalized consultations •Loyalty and bonus programs •Proactive communication •Individualized support for at-risk customers **2. Review and Optimize Credit Policy**. The correlations identified between credit limits, utilization, and churn behavior point to room for improvement: •Increase limits for stable customers •Provide credit usage training programs •Offer flexible credit options to enhance satisfaction **3. Expand Data Analytics and Leverage AI**. For long-term retention, banks should increasingly use Data Analytics and Machine Learning: •Develop churn prediction models •Segment risk groups based on behavior •Detect at-risk customers early and act proactively •Automate churn identification **4. Improve Customer Service and Communication**. Inactive customers are contacted less frequently - this needs to be addressed: •Use digital channels (apps, chatbots, newsletters) •Send automated reminders or offers to inactive customers •Introduce CRM systems for systematic customer engagement **5. Continuous Monitoring and Success Measurement**. Establish professional monitoring systems to track and evaluate retention strategies: •Use KPIs such as churn rate by customer segment •Visualize performance in dashboards •Conduct regular reporting cycles

Summary of Outlook. This data analysis and visualization project provides a strong basis for data-driven decision-making in the banking industry. With targeted retention efforts, modern analytical tools, and improved customer service, a bank can: •Increase competitiveness •Reduce churn sustainably •Improve customer satisfaction and long-term loyalty.