📊 IBM Telco Customer Churn Prediction

Exploratory Data Analysis (EDA) & Preprocessing Report

📌 Project Overview

Customer churn (i.e., when customers stop using a company's services) is a major concern for telecom companies. The objective of this project is to explore the IBM Telco Customer Churn dataset, perform exploratory data analysis (EDA), preprocessing, and identify important factors influencing churn.

This report covers:

- Dataset description
- Tools & libraries used
- Exploratory Data Analysis (EDA)
- Preprocessing steps
- Statistical tests applied
- Key findings & insights

📁 Dataset Description

- **Dataset:** IBM Telco Customer Churn
- **Rows:** ~7,000+
- **Columns:** 50 (customer demographics, service details, account info, churn-related attributes)

🔑 Key Columns

- **Customer Info:** Customer ID, Gender, Age, Married, Dependents, Number of Dependents
- **Geography:** Country, State, City, Zip Code, Latitude, Longitude, Population
- **Service Details:** Phone Service, Internet Service, Multiple Lines, Streaming TV/Movies/Music, Online Security, Device Protection Plan, Tech Support
- **Financials:** Monthly Charge, Total Charges, Total Refunds, Total Extra Data Charges, Total Revenue
- **Churn Indicators:** Customer Status, Churn Label, Churn Score, CLTV, Churn Category, Churn Reason

---

🛠️ Tools & Libraries Used

- **Python**
- **Pandas, NumPy** → Data cleaning & manipulation

- **Seaborn, Matplotlib** → Data visualization
- **Statistics** → Hypothesis testing (t-test, chi-square test)
- **Scikit-learn** (planned for modeling phase)

🔎 Exploratory Data Analysis (EDA)

1. Univariate Analysis

- **Numerical features:** Histograms & boxplots revealed distributions of Age, Tenure in Months, Monthly Charges, Total Charges.
- **Categorical features:** Countplots showed customer distribution across gender, contract types, payment methods, and churn labels.

2. Bivariate Analysis

- **Correlation heatmap** highlighted strong relationships between Monthly Charge, Total Charges, and Revenue.
- **Scatterplots** between Tenure in Months vs. Monthly Charge showed clustering patterns.
- **Countplots** showed higher churn in month-to-month contracts compared to long-term contracts.

3. Statistical Testing

- **t-test:** Verified significant differences in Monthly Charges between churned vs. non-churned customers.
- **Chi-square test:** Found associations between categorical features (Contract, Payment Method, Internet Service) and churn.

⚙️ Preprocessing Steps

- **Handling Missing Values:** Imputed missing data in Total Charges, Dependents, and other columns.
- **Encoding Categorical Variables:** Converted categorical variables (Gender, Contract, Payment Method) into numeric form using label encoding / one-hot encoding.
- **Feature Scaling:** Normalized continuous variables (Monthly Charge, Total Charges) for future model use.
- **Outlier Detection:** Used boxplots & IQR method to identify extreme values in Monthly Charges & Total Charges.
- **Feature Selection:** Dropped irrelevant columns like Customer ID, Latitude, Longitude, Zip Code.

📈 Key Findings

1. **Contract Type:** Customers on **month-to-month contracts** churn more often compared to yearly contracts.
2. **Payment Method:** Electronic check users show higher churn rates.
3. **Age Group:** Younger customers (<30 years) churn less compared to middle-aged groups.
4. **Monthly Charges:** Customers with **higher monthly charges** are more likely to churn.
5. **Services:** Lack of value-added services (e.g., no online security, no tech support) increases churn probability.
6. **Statistical Tests:** Confirmed churn is significantly associated with Contract, Payment Method, and Internet Service.

🚀 Next Steps (Future Work)

- Feature engineering (derive new insights from tenure, revenue, etc.)
- Build ML models (Logistic Regression, Random Forest, XGBoost)
- Evaluate models using accuracy, precision, recall, F1-score
- Build churn prediction dashboard (Power BI / Tableau)
- Deployment via Flask/Django + Docker (MLOps pipeline)

🏆 Conclusion

This project provided insights into **factors influencing customer churn** through EDA and preprocessing. The findings highlight the importance of **contract type, payment method, and additional services** in predicting churn. These insights will guide the next phase of **machine learning model building and deployment**.