Final Project Report: Exploratory Data Analysis for Telco Customer Churn

1. Data Summary

The dataset used for this project is sourced from a telecommunications company ("Telco"). The primary objective is to identify the profile of customers who are likely to discontinue their service.

- **Dataset Size**: The collection consists of **7,043 rows** (individual customers) and **21 columns** (features).
- **Target Variable**: Churn (categorical: Yes/No), indicating whether the customer left within the last month.

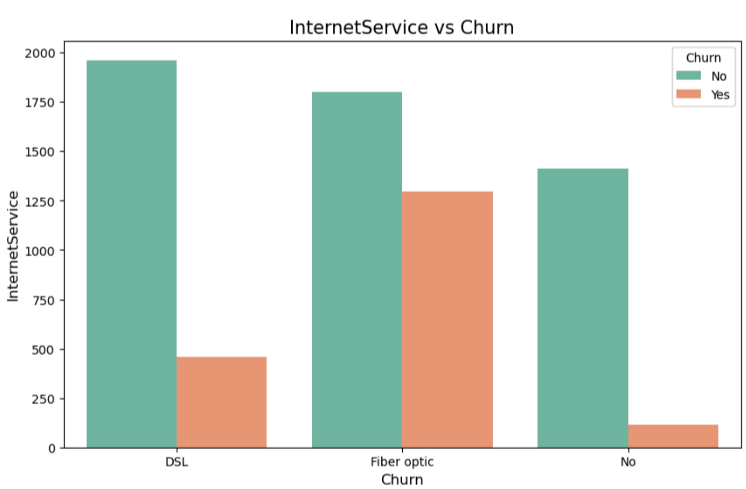
2. Data Exploration Plan

The analysis follows a logical four-phase vision to ensure insightful results:

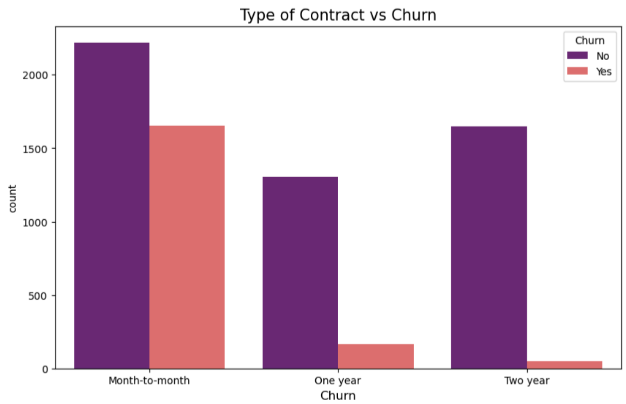
1. **Initial Inspection**: Identify data types, unique values, and anomalies (specifically investigating the TotalCharges column).
2. **Cleaning and Transformation**: Fix data type inconsistencies and handle missing values to ensure mathematical integrity.
3. **Univariate and Bivariate Analysis**: Visualize individual feature distributions and their direct correlation with the target variable (Churn).

3. Exploratory Data Analysis (EDA) Results

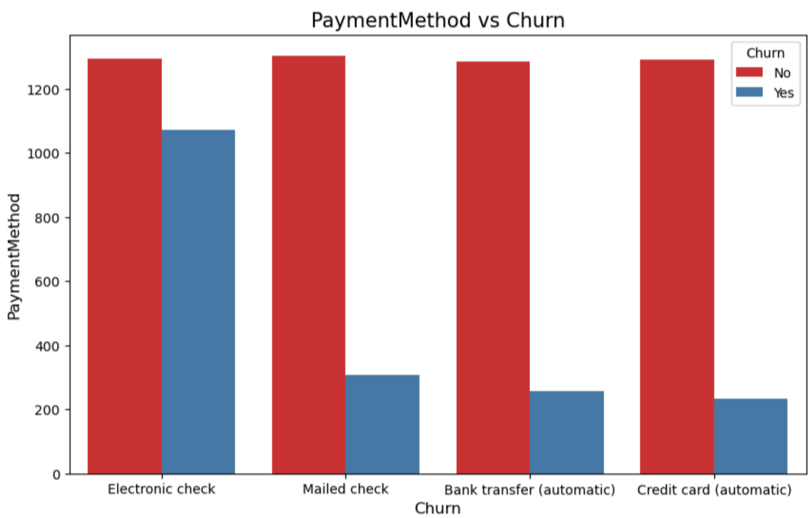
- **Churn Distribution**: Approximately 26.5% of the customer base has churned.
- **Internet Service Impact**: Customers using **Fiber Optic** internet show a significantly higher churn rate compared to DSL users. This is an insightful finding as Fiber Optic is typically the more expensive service.



- **Contract Vulnerability**: There is a critical churn risk associated with **"Month-to-month"** contracts. The vast majority of customers who leave belong to this group.



- **Payment Methods**: Customers using **"Electronic check"** are much more likely to churn compared to those using automated bank transfers or credit cards.



4. Data Cleaning and Feature Engineering

Several steps were taken to prepare the data for analysis and modeling:

- **Handling Missing/Erroneous Values**: TotalCharges was initially detected as an "object" type due to empty strings. These corresponded to new customers with a tenure of 0. I forced the conversion to numeric and imputed the missing values with 0.
- **Feature Engineering (Binning)**: I created a new feature called tenure_group to segment customers by years of service.
- **Cleaning Results**: By applying .apply() and .to_numeric(), the data was successfully cleaned. Visual verification using .dtype confirmed that TotalCharges is now a float.

It seems the column "TotalCharges" is numerical but treated as an object in this dataset. Why?

```
[88]: data["TotalCharges"].nunique()
```

```
[88]: 6531
```

```
Only 6531 rows are unique, but we have 7043 non null rows, what does this mean?
```

```
[86]: data["TotalCharges"]= pd.to_numeric(data["TotalCharges"],errors='coerce' )
      data['TotalCharges'] = data['TotalCharges'].fillna(0)
```

```
[92]: print(data["TotalCharges"].dtype)
```

```
float64
```

Now every row in Total Charges has a numeric type!

5. Summary of Key Insights

Synthesizing the EDA results, it is clear that churn is not random. The highest-risk customers share a specific profile:

1. They are in their **first year** of service.
2. They are on **month-to-month** contracts.
3. They utilize **non-automated** payment methods (Electronic checks).

6. Hypotheses

Based on the data exploration, the following three hypotheses were formulated:

1. **Hypothesis 1**: Customers with higher monthly charges are significantly more likely to churn than those with lower charges.
2. **Hypothesis 2**: Customers with a contract of one year or longer have a churn rate below 10%.
3. **Hypothesis 3**: Subscribing to "Tech Support" services reduces the probability of churn for Fiber Optic users.

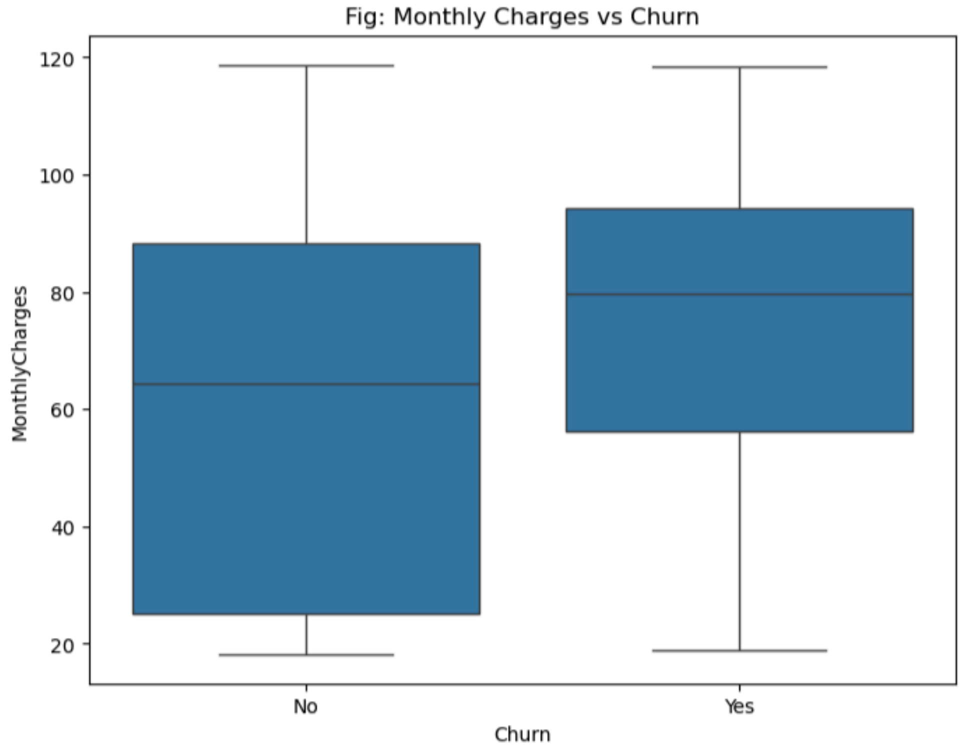
7. Significance Test Analysis

I selected **Hypothesis 1** for a thorough significance test.

- **Null Hypothesis (H_0)**: The monthly charges for departing customers are less than or equal to those for customers who stay. ($u_1 <= u_2$).
- **Alternative Hypothesis (H_1)**: The monthly charges for departing customers are significantly higher than those for customers who stay. ($u_1 > u_2$).
- **Methodology**: An **Independent Samples T-test** was performed with a significance level of alpha = 0.05.
- **Results**:
  - **T-statistic**: 16.53

- **P-value**: 1.35e-60 (zero).

- **Interpretation**: Since the p-value is extremely lower than 0.05, **the null hypothesis is rejected**. There is overwhelming statistical evidence that churning customers pay higher monthly fees. This confirms that price sensitivity is a major driver of customer loss.



8. Final Conclusions and Next Steps

Conclusions: The churn profile is heavily linked to high short-term costs and a lack of long-term contractual commitment. Cleaning the TotalCharges column was vital, as it revealed that "missing" data actually represented new customers, preventing biased averages.

Next Steps:

1. **Predictive Modeling**: Develop a classification model (e.g., Logistic Regression or Random Forest) to proactively identify at-risk customers.