STC Data Analyst Project – Week 2 Report: Data Cleaning

Name: Favour Ogundipe

Week: 2 – Data Cleaning and Preparation

Tool Used: Microsoft Excel

1. Tool Selection and Rationale

The entire data cleaning and preparation phase was performed using **Microsoft Excel - Power Query**

Excel was chosen as the primary tool for its strong capabilities in **visual data inspection** (using conditional formatting and filtering) and **structured data transformation**. It was the most efficient tool for:

- Quickly identifying visual patterns of anomalies, such as blanks, duplicates, and negative numbers.
- Applying logical, rule-based transformations using functions like IF, VLOOKUP, TRIM, and ISBLANK.
- Performing the required one-time **table merge** with the ZIP code population data.

2. Issues Identified in the Dataset

The initial profiling of the Telecom Customer Churn dataset (which originally contained **38 columns**) revealed several data quality issues requiring systematic attention:

Issue Type	**Specific Problem**	**Affected Columns**
Missing Values (Nulls)	Blank cells in critical numerical columns.	Avg. Monthly Long Distance Charges, Avg. Monthly GB Download
	Blank cells in key categorical service columns.	Multiple Lines, Internet Type, Online Security, Online Backup, Device Protection Plan, Premium Tech Support, Streaming TV, Streaming Music, Streaming Movies, Unlimited Data

	Blank cells in churn classification fields.	Churn Category, Churn Reason
Invalid Data	Values that are logically impossible for the metric.	**Negative values** found in the Monthly Charge column.
Integrity Check	Need to confirm the uniqueness of the primary identifier.	Customer ID
	Need to validate lookup key for merging.	Zip Code

3. Specific Cleaning Steps and Justifications

Each identified issue was systematically addressed using appropriate Excel functions to ensure data quality and integrity.

Issue	**Cleaning Action Taken**	**Reason / Justification**
Numerical Nulls	Replaced null or blank cells with $\mathbf{0}$ (zero).	These columns represent usage metrics. A null value implies no record of usage or zero usage for that period, maintaining data continuity without skewing usage averages.
Categorical Nulls (Service)	Replaced blanks in all service columns (e.g., Online Security) with **"UNKNOWN"**.	A blank could be a data entry error or genuine unavailability of information. Using **"UNKNOWN"** preserves the record and creates a distinct category, preventing the false assumption of "No" service.

Categorical Nulls (Churn)	Replaced blanks in Churn Reason with **"Don't know"** and Churn Category with **"Other"**.	These were existing categories in the dataset. Using them maintains categorical consistency and ensures no missing classifications in the final summaries.
Negative Monthly Charge	Negative entries were **flagged for review** and temporarily treated as 0 for non-revenue-based analysis.	A customer charge cannot logically be negative. These records were identified as potentially indicating a refund or severe input error and must be verified before applying the absolute value or excluding them entirely.
Duplicate Records	Used Excel's Remove Duplicates feature on the entire dataset, based on the Customer ID column.	**Verification confirmed no duplicate records.** This step guarantees a reliable, one-to-one relationship between the customer and their associated data.

4. Column Counts and Table Merging

A. Column Reduction

The original dataset began with **38 columns** and was retained

B. Table Merging with Demographic Data

The cleaned dataset was enriched by merging it with a **Zipcode Population dataset** using the **Zip Code** column as the common key.

- **Process in Excel:** The MERGE function was used to merge the **Population** and **ZIP CODE** columns from the ZIP dataset into the main customer data.
- **Purpose:** The merge adds crucial **demographic and geographic context**, enabling richer analysis like exploring churn trends segmented by state or population density.
- **Final Column Count:** After the successful merge, the final dataset contains **39 columns** (38 original relevant columns + 1 new columns: Population).

5. Data Integrity and Validation Checks

A series of final checks confirmed the dataset’s readiness for analysis:

- **Customer ID Uniqueness:** A COUNTIF check across the Customer ID column confirmed that the dataset contains **7,044 unique records**, validating the absence of any duplicate entries.
- **Data Type Assignment:** All columns were reviewed and set to the correct data type (e.g., numerical values for charges and usage, Date format for date fields, and Text/General for categorical fields) to ensure proper calculations and filtering during the analysis phase.
- **Numeric Consistency:** Confirmed all numerical columns were formatted correctly and only contained valid numeric entries (with the exception of the flagged negative Monthly Charge values).
- **Categorical Uniformity:** All categorical fields (e.g., Gender, Multiple Lines) were verified to have consistent, cleaned values (e.g., "Yes," "No," or "UNKNOWN").
- **Post-Merge Validation:** The record count was verified pre- and post-merge to ensure no data loss occurred during the VLOOKUP process.

6. Final Dataset Status

The dataset is now **Clean and Ready for Analysis**.

All missing, inconsistent, and invalid entries have been logically addressed. The enriched dataset, with its added geographic and demographic context, is now prepared for exploratory data analysis and modeling.

- **Assumption:** The primary assumption made was that missing values in the categorical service columns implied "Unknown" availability rather than explicitly "No" service, preventing an artificial bias toward "No."
- **Documentation:** All transformations and assumptions are fully documented to ensure the entire cleaning process is reproducible.