Week 4 Project Report

(This week, our primary focus is still data processing specificity merging the data. The steps are added in the data section. The model part is also updated after our discussion. Our EDA centers on the key variables: price, quantity, and quality and examines their relationships through correlation analysis.

Specifically, this section will evaluate pairwise correlations among these variables to identify potential multicollinearity issues before proceeding to formal modeling. This diagnostic step ensures that the model specification is statistically sound and that the explanatory variables do not exhibit excessive collinearity.

The code is documented in the notebook folders, and the merged data is stored in the `data` directory under `merged_data`.)

Yinyi Xue, Jiayi Hou, Nikolas A Papa

Problem Statement

How do price, service quality, and regional market (circle) differences influence wireless subscriber demand in India?

India is one of the world's most competitive telecom markets, with thin margins, rapid technology upgrades, and strong regional heterogeneity. Carriers face a critical strategic tradeoff:

- Lower prices to protect volume, or
- Invest in quality and premium plans to drive growth and loyalty

However, these decisions are often made without a clear, data-driven understanding of how price, service quality, brand, and region interact to influence subscriber demand over time.

The conceptual model can be expressed as follow:

$$\text{Subscribers}_{i,j,t} = \alpha + \beta_1\,\text{Price}_{i,j,t} + \beta_2\,\text{Quality}_{i,j,t} + \epsilon_{i,j,t}$$

Where i,j, and t mean the company, circle, and time factors.

Articulation of value

India's wireless telecom market is one of the largest globally, with over 1.17 billion subscribers, according to the Telecom Regulatory Authority of India (TRAI, 2025). Wireless services account for more than 96% of all connections. The market is highly competitive and dominated by Reliance Jio, Bharti Airtel, and Vodafone Idea. Operators are shifting away from pure price competition. Premium plans and network quality now matter more. Understanding how price, quality, brand, and regional differences affect subscriber growth is therefore critical for pricing, investment, and growth decisions.

Mock Calculation: Potential Economic Value

Below is an illustrative calculation of the potential economic value:

Assumptions:

- Carrier has 100 million subscribers
- Average monthly ARPU = ₹200
- Monthly churn rate = 2%
- Insights from this analysis reduce churn by 0.1 percentage points (2.0% → 1.9%)

Step 1: Retained subscriber per month

- Baseline churn = 100M × 2.0% = 2.0M
- Improved churn = 100M × 1.9% = 1.9M
- Subscribers retained per month = 100,000

Step 2: Revenue retained

- Monthly revenue retained = 100,000 × ₹200 = ₹20M
- Annual revenue retained = ₹20M × 12 = ₹240M per year

This research project is valuable because a very small improvement in churn driven by better pricing and quality decisions can generate hundreds of millions of rupees in annual revenue. The Indian wireless market operates at a massive scale, even minor improvements identified through data analysis create meaningful economic value.

Project plan

Week 1 – Problem Definition

- Define the modeling problem and business value. Group Discussion.

  Deliverables: Written problem statement, business case, GitHub repo, JupyterHub setup

Week 2 – Data Ingestion

- Load data and explore variables

  Deliverables: Jupyter notebook, initial EDA, updated GitHub

Week 3 – Exploratory Data Analysis

- Split data. Perform EDA. Identify data issues.

  Deliverables: EDA notebook, written summary, code output

Week 4 – Exploratory Data Analysis,

- Clean and preprocess datasets.

  Deliverables: Preprocessed dataset, notebook, code output

Week 5 – Data Preprocessing, Feature Engineering

- Create and finalize features.

Deliverables: Feature engineering notebook, updated dataset

Week 6 – Baseline Modeling

- Build and evaluate simple models.

Deliverables: Baseline model notebook, evaluation metrics

Week 7 – More Complex Modeling

- Build and tune more complex models.

Deliverables: Model comparison notebook

Week 8 – Advanced Modeling

- Build and evaluate final model candidates.

Deliverables: Tuned models, evaluation results

Week 9 – Model Selection

- Select best model using test data.

Deliverables: Final model, test performance report

Week 10 – Data-Centric Improvement

- Improve model via data and features.

Deliverables: Refined dataset, improved results

Week 11 – Ethics & Explainability

- Analyze bias, risk, and feature importance.

Deliverables: Explainability analysis, ethics section

Week 12 – Deployment & Monitoring

- Package model and define monitoring plan.

Deliverables: Saved model, monitoring plan

Week 13 – Final Integration

- Assemble finished project.

Deliverables: Final report, final notebook, merged GitHub repo

Week 14 – Peer Review

- Review peer projects.

Deliverables: Written peer feedback

Dataset

The datasets were provided and compiled by Jiayi, covering Q1 2021 and Q1 2023. All datasets are structured at the company–plan–circle–time level, allowing them to be merged into a unified panel for regression analysis.

- **Subscriber Quantity (Demand)**
  This dataset reports wireless subscriber counts by company, service plan (generation), regional market (circle), and time period. It includes both total and active subscribers and represents the demand-side outcome variable. The data are sourced from official subscription reports published by the Telecom Regulatory Authority of India (TRAI).
- **Price**
  The price dataset contains tariff and plan information for each company, segmented by plan type (e.g., 2G, 3G, 4G), circle, and time period. It captures plan-level prices offered to consumers, enabling comparison across carriers and regions.
- **Service Quality Metrics**
  The service quality dataset includes quantitative measures of network performance at the company, plan, and circle level. Key metrics include service activation success rates, data transmission success rates, latency, drop rates, and average data throughput. These variables capture the quality dimension of wireless service provision.

Data Sources and Preprocessing

For each dataset and each time period, the data sources and preprocessing procedures are outlined below. All datasets were then merged using company, circle, plan, and time period as unique keys, yielding a comprehensive panel suitable for regression analysis.

Quantity (Subscriber Count)

- **Source:**
  Data are obtained from the official subscription reports published by the Telecom Regulatory Authority of India (TRAI), specifically from the monthly reports available at TRAI Telecom Subscriptions Reports, covering March 2021 and March 2023.

- **Main variables:**
  - Company name
  - Circle (regional market)
  - Time period (March 2021, March 2023)
  - Total number of wireless subscribers

    - Number of active subscribers (VLR-based)
    - Urban/rural subscriber segmentation
    - Market share by company
- **Preprocessing steps:**
    - Extract tabular data from published PDF reports for the relevant periods
    - Standardize company and circle names for consistency
    - Align circle definitions across periods
    - Merge urban and rural data if required for aggregate analysis
    - Handle missing or inconsistent entries through manual review

Price

- **Source:**
  Price data were collected using a custom Python automation script, combined with the Wayback Machine to access historical snapshots of Economic Times Telecom Recharge Plans for March 2021 and March 2023. Data was scraped for all circles and companies.
- **Main variables:**
    - Company name
    - Service plan/generation (2G, 3G, 4G, plan ID)
    - Plan price (INR)
    - Validity (number of days)
    - Data allowance (GB)
    - Circle (region)
    - Time period (March 2021, March 2023)
    - Additional benefits (if applicable)
- **Preprocessing steps:**
    - Automated extraction and parsing of all plan information per company and circle
    - Unification of price formats and units
    - Identification and removal of duplicate or discontinued plans
    - Standardization of plan naming conventions
    - Validation against official sources when possible

Service Quality Metrics

- **Source:**
  Quality metrics are sourced from the quarterly Performance Monitoring Reports (PMR) on

Wireless Data Services published by TRAI, specifically from the reports for March 2021 and March 2023, available at TRAI Wireless Data Reports.

- **Main variables:**
    - Company name
    - Circle (regional market)
    - Plan/generation (2G, 3G, 4G)
    - Time period (March 2021, March 2023)
    - Service activation/provisioning success rate (%)
    - Data transmission success rate (download/upload, %)
    - Latency (ms)
    - PDP context activation success rate (%)
    - Drop rate (%)
    - Average throughput (Kbps)
    - Minimum download speed (plan level)
- **Preprocessing steps:**
    - Extraction of tabular quality indicators from PDF reports
    - Mapping of quality metrics to the correct company, circle, and plan
    - Standardization of metric units and definitions
    - Handling missing or outlier values per TRAI reporting standards
    - Aggregation at desired level (e.g., by company-circle-plan-time)

Model

This project uses a supervised learning approach. The primary modeling task is a regression problem, as the outcome variable: wireless subscriber count is continuous. Regression-based models will be used to estimate how subscriber counts relate to price, service quality, company, and regional market differences. Model complexity will increase gradually, allowing assessment of the individual and combined contributions of each factor.

As mentioned previously the general model is:

$$\text{Subscribers}_{i,j,t} = \alpha + \beta_1\,\text{Price}_{i,j,t} + \beta_2\,\text{Quality}_{i,j,t} + \epsilon_{i,j,t}$$

To rigorously evaluate how price, company, regional market (circle), and service quality affect wireless subscriber counts, a series of linear regression models were developed and estimated using the cleaned and integrated datasets.

- Baseline Model:
    - The initial model regresses subscriber count solely on average plan price. This specification provides a benchmark for assessing the isolated impact of pricing on subscription levels.

Appendix: Data Source

Telecom Regulatory Authority of India. (2025). Telecom subscription reports (Q1 2021, Q1 2023). https://www.trai.gov.in/release-publication/reports/telecom-subscriptions-reports