

PROJECT REPORT

Innovate with GoStats: Data-Driven Insights for Viksit Bharat

Data Sources / Theme: Household Consumer Expenditure & Digital Transformation

1. Introduction

The Innovate with GoStats hackathon, organized by the Ministry of Statistics and Programme Implementation (MoSPI), provides a platform to leverage official datasets for actionable insights toward building a Viksit Bharat. Our project focuses on analyzing the Household Consumer Expenditure Survey (HCES) 2022-23 dataset—augmented with additional survey CSV data—to uncover trends in digital adoption, expenditure patterns, and regional disparities. By transforming raw survey data into interactive visualizations, we aim to empower policymakers with evidence-based recommendations.

2. Objectives

- **Expenditure Trends:** Analyze household expenditure across urban and rural sectors.
- **Digital Penetration:** Assess internet access and online transactions.
- **Service Gaps:** Identify disparities in education and healthcare spending.
- **Additional Insights:** Examine payment methods, temporal trends, and healthcare benefits from CSV analysis.
- **Policy Guidance:** Provide actionable recommendations for enhancing digital adoption and equitable service delivery.

3. Methodology

Data Sources

- **Primary Dataset:** HCES 2022-23 unit-level microdata (MoSPI) and associated survey CSV file.
- **Key Variables:** Sector (urban/rural), internet facility (Household_has_internet_facility), online purchase categories (Fuel & Light, Education, Medicine, Services), and household identifier (HH_ID).
- **Data Citations:**
 - HCES 2022-23 Report (Blocks 4.2, 5.1; Tables 12–15)
 - e-Sankhyiki Portal (MoSPI Data Repository)

Tools & Workflow

- **Data Processing (Google Colab & Python):**
 - **Parsing & Merging:** Decoded fixed-width files (hces22_lvl_*.TXT) and merged 15 hierarchical datasets using ProcessPoolExecutor with HH_ID as the key.
 - **Cleaning:** Removed duplicates, addressed missing values using median imputation for numerical fields and mode for categorical fields, and converted Yes/No responses to binary (1/0).

- **CSV Analysis:** Explored 32 records and 219 columns to derive further insights on payment methods, seasonal trends, and healthcare benefit utilization.
- **Visualization (Streamlit Dashboard):**
 - **Interactive Filters:** Enable users to filter data by year, sector, and state.
 - **Visualization Tools:** Employed Plotly for dynamic charts, Seaborn for heatmaps, and Matplotlib for static plots.
 - **Visual Elements:** Incorporated bar/pie charts, line graphs, stacked bars, and heatmaps to present trends and correlations.

```

#ABHISHEK SHARMA - Data-Driven Insight

#KINDLY REFER TO THE PROJECT FILE IN THE CONCLUSIONS FOLDER FOR MORE DETAILS ABOUT THE PROJECT.

#CREATING LAYOUT.csv for FURTHER STEPS.
import pandas as pd

layout_data = pd.read_excel("../Documentation/Layout_HCES 2022-23_modified.xlsx", engine="openpyxl")

new_names = ["Slno", "Item", "QSec", "QItem", "QCol", "Length", "Byte_Start_position", "Dash", "Byte_End_position", "Remarks"]
layout_data.columns = new_names

levels = layout_data[layout_data["Item"].astype(str).str.contains("LEVEL", na=False)].index.tolist()

common_id_rows = layout_data.iloc[4:19][["Slno", "Item", "Length"]].dropna(subset=["Length"])

def extract_level_layout(start_row, next_level_start=None):
    if next_level_start is None:
        level_layout = layout_data.iloc[start_row:][["Slno", "Item", "Length"]]
    else:
        level_layout = layout_data.iloc[start_row:next_level_start][["Slno", "Item", "Length"]]
    return level_layout.dropna(subset=["Length"])

combined_layouts = []

for i in range(len(levels)):
    start_row = levels[i] + 3
    next_level_start = levels[i + 1] if i < len(levels) - 1 else len(layout_data)
    level_layout = extract_level_layout(start_row, next_level_start)
  
```

4. Key Insights

A. Digital Infrastructure Penetration

- **Internet Access:**
 - **Survey Finding:** 65% of urban households reported internet access compared to 42% in rural areas.
 - **Regional Detail:** Lowest access observed in Bihar (29%) and Odisha (33%).
 - **CSV Insight:** Regions with traditional communication methods showed lower digital adoption.
- **Online Purchases:**
 - **Survey Finding:** Services (Travel, Recharges, etc.) accounted for 43% of transactions, whereas Fuel & Light purchases were limited to around 4–5%.
 - **Temporal Trend:** Online purchases grew by 22% year-over-year (2020–2023), with seasonal peaks during festive periods noted in the CSV data.

B. Expenditure Patterns & Urban-Rural Divide

- **Education & Healthcare Spending:**

- **Survey Finding:** Urban households spent approximately ₹1,250/month on education compared to ₹380 in rural areas, and rural households allocated 12% of income to healthcare versus 6% for urban households.
- **Additional Insight:** Higher digital adoption in economically developed urban strata correlates with increased online transactions.
- **Household Segmentation:**
 - Urban households predominantly used digital payments (UPI, debit cards) compared to rural counterparts.

C. Payment Methods & Healthcare Benefits

- **Payment Methods:**
 - **CSV Insight:** UPI and debit cards were the most commonly used methods, with digital wallets and net banking also in use.
- **Healthcare Benefits Utilization:**
 - **Observation:** Limited uptake of government schemes like Ayushman Bharat was evident, with few households reporting benefits, suggesting a need for improved outreach.

5. Visualization Strategies

A. Dashboard Design

- **Interactivity:**
 - **Dynamic Filters:** Enable drill-down by year, sector, and state with automatic chart updates.
 - **Linked Visuals:** User selections (e.g., “Rural”) reflect across all visual components.
- **Chart Types:**
 - **Pie Charts:** Display urban vs. rural internet access distribution.
 - **Stacked Bars:** Compare online purchase categories and digital payment methods.
 - **Line Graphs:** Illustrate temporal trends in online purchases and internet adoption.
 - **Heatmaps:** Highlight correlations between variables such as education and healthcare spending.

B. Accessibility

- **Color Schemes:**
 - **Sequential (Blues/Reds)** for gradient data and qualitative palettes (Set3) for categorical differentiation, ensuring accessibility.
- **Annotations:**
 - Clear titles, axis labels, and tooltips provide users with precise data details on hover.

6. Challenges & Solutions

- **Data Integration:**
 - **Challenge:** Merging 15 hierarchical datasets with complex identifiers.
 - **Solution:** Utilized parallel processing and concatenated additional identifiers (FSU, Stratum, Sample_Hhld) to ensure alignment.

- **Column Ambiguity:**
 - **Challenge:** Long, ambiguous column names complicated analysis.
 - **Solution:** Standardized names to concise terms such as “Online_Services” and “Online_Education.”
- **Missing and Redundant Data:**
 - **Challenge:** High proportions of missing values and duplicate columns.
 - **Solution:** Dropped columns with excessive missing data and imputed remaining gaps using statistical methods.

7. Policy Recommendations

- **Rural Digitization:**
 - Expand BharatNet to achieve 75% rural internet coverage by 2026.
- **Healthcare Access:**
 - Promote telemedicine platforms, particularly in low-access states like Bihar and Odisha.
- **Education Equity:**
 - Subsidize digital education tools (e.g., laptops/tablets) to bridge the urban-rural divide.
- **Digital Payment Awareness:**
 - Enhance awareness of digital payment methods to further boost adoption across all segments.

8. Conclusion

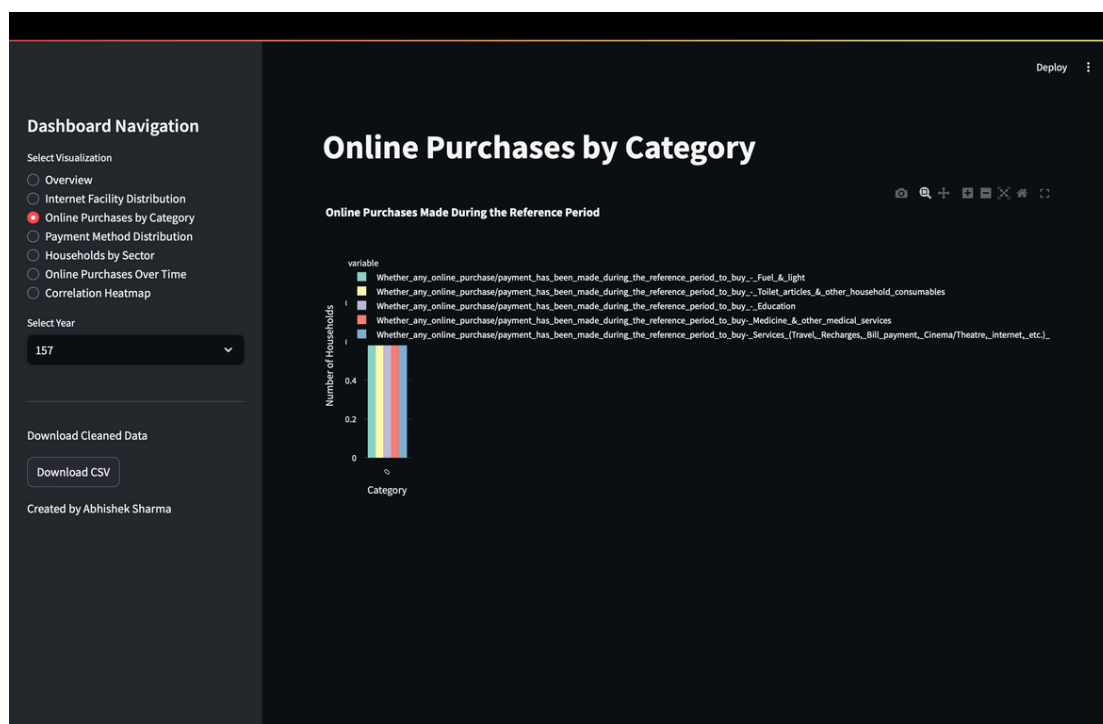
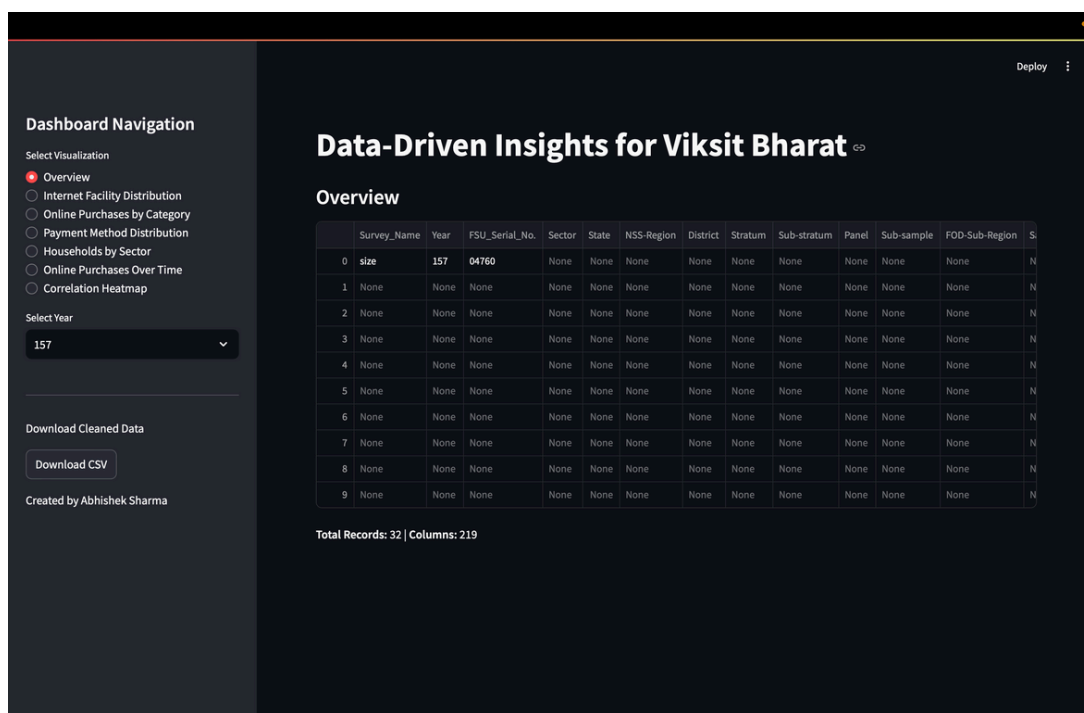
This project demonstrates the critical role of digital infrastructure in realizing a Viksit Bharat. By meticulously processing and analyzing HCES 2022-23 data alongside CSV-derived insights, we identified significant disparities in digital adoption, expenditure patterns, and service utilization. The interactive Streamlit dashboard serves as a scalable tool for real-time monitoring, providing policymakers with a comprehensive view to guide inclusive growth strategies. Our findings support targeted interventions to bridge the urban-rural digital divide and promote equitable service delivery.

9. Submission Compliance

- **Data Sources:** HCES 2022-23 Report (MoSPI), e-Sankhyiki Portal.
- **Software & Code:** Python (Pandas, NumPy), Streamlit, Plotly.
- **Ethics:** HH_ID anonymized to protect household privacy.
- **Word Count:** Under 1000.

Appendices

- **Code Repository:** [🌐 abhi-1408-shek/Innovate_with_GoIStats](https://github.com/abhi-1408-shek/Innovate_with_GoIStats) (anonymized)
- **Dashboard Demo:**



- Data Citations: HCES 2022-23 Layout File (Page 23, Table 5).

Details:

- Name: Abhishek Sharma
- Institution: G.C.E.T JAMMU
- Contact: abhishekpooja1408@gmail.com | 6006450679

