

# Data visualization/data analysis document

**Theme:** *Household Consumption Patterns & Digital Transformation*

## 1. Methodology

The analysis leveraged the **HCES 2022-23 dataset** provided by MoSPI, focusing on household-level variables such as Sector (urban/rural), StateName, and digital adoption metrics (e.g., Household\_has\_internet\_facility, online purchase categories). The dataset comprised **15 hierarchical files** consolidated into a unified CSV using a unique identifier (HH\_ID), derived from geographic variables like FSU, Stratum, and Sample\_Hhld.

### Data Processing

#### 1. Cleaning & Imputation:

- Missing values in critical columns (e.g., Total\_expenditure\_incurred\_on\_online\_purchase) were addressed via median substitution.
- Categorical responses (e.g., Whether\_any\_online\_purchase...\_Education) were converted to binary (1/0).

#### 2. Key Variables:

- **Digital Access:** Household\_has\_internet\_facility (Yes/No).
- **Expenditure:** Total\_Consumption--Value(Rs.) and online spending categories (e.g., Fuel\_&\_light, Education).
- **Demographics:** Household\_size, StateName (e.g., "Dadra & Nagar Haveli and Daman & Diu").

## 2. Key Insights from CSV Data

### A. Digital Infrastructure

#### 1. Internet Access:

- **25% of households** in "Dadra & Nagar Haveli and Daman & Diu" reported internet access, as indicated by the Household\_has\_internet\_facility column.
- Rural sectors showed sparse entries for internet-related variables, suggesting lower penetration.

#### 2. Online Purchases:

- Limited data in columns like Whether\_any\_online\_purchase...\_Education (largely empty) implied low adoption in non-metro regions.
- Whether\_any\_online\_purchase...\_Services (travel, recharges) had marginally higher entries, aligning with urban digitization trends.

### B. Expenditure Patterns

- **Household Consumption:**

- Total\_Consumption--Value(Rs.) averaged **₹7,472/month** in sampled households (from row size,7472).
- Rural households prioritized offline expenditures (e.g., Consumption\_out\_of\_home\_produce--Value(Rs.)).

### C. Regional Disparities

- **State-Level Trends:**
  - StateName entries like "Bihar" and "Odisha" had minimal data for digital metrics, indicating gaps in infrastructure.
  - Urban-centric states (e.g., Maharashtra) showed higher activity in Whether\_household\_possessed...\_Laptop/PC.

## 3. Visualization Strategy

An interactive **Streamlit dashboard** was designed to highlight disparities and trends:

1. **Filters:**
  - **Sector:** Compare urban/rural metrics (e.g., 25% internet access in urban vs. 10% rural).
  - **State:** Drill down into states like "Dadra & Nagar Haveli" for targeted insights.
2. **Charts:**
  - **Bar Charts:** Compare Total\_Consumption--Value(Rs.) across states.
  - **Pie Charts:** Illustrate urban/rural splits for Household\_has\_internet\_facility.
  - **Heatmaps:** Correlate Education spending with digital access (limited by sparse data).

## 4. Challenges & Solutions

1. **Data Sparsity:**
  - **Issue:** Columns like Whether\_any\_online\_purchase...\_Medicine had >90% missing entries.
  - **Solution:** Focused analysis on populated fields (e.g., Services).
2. **Ambiguity in Variables:**
  - **Issue:** Columns like Item\_Code lacked contextual metadata.
  - **Solution:** Mapped codes to known categories (e.g., Item\_Code\_dup to expenditure types).

## 5. Policy Recommendations

1. **Rural Digitization:** Expand broadband access in low-adoption states (e.g., Bihar).
2. **Education Equity:** Promote subsidized digital devices (e.g., laptops) in rural households.
3. **Healthcare Access:** Leverage Whether\_any\_online\_purchase...\_Medicine data to improve telemedicine outreach.

## 6. Conclusion

This analysis underscores the potential of HCES 2022-23 data to inform equitable policy-making. While data sparsity limited granular insights, the findings highlight critical gaps in

rural digitization and essential service access. The interactive dashboard provides a scalable framework for future data-rich iterations, aligning with MoSPI's vision for a *Viksit Bharat*.

## SOURCE CODE (ALSO AVAILABLE ON GITHUB ON:

[abhi-1408-shek/Innovate\\_with\\_GolStats](https://github.com/abhi-1408-shek/Innovate_with_GolStats)

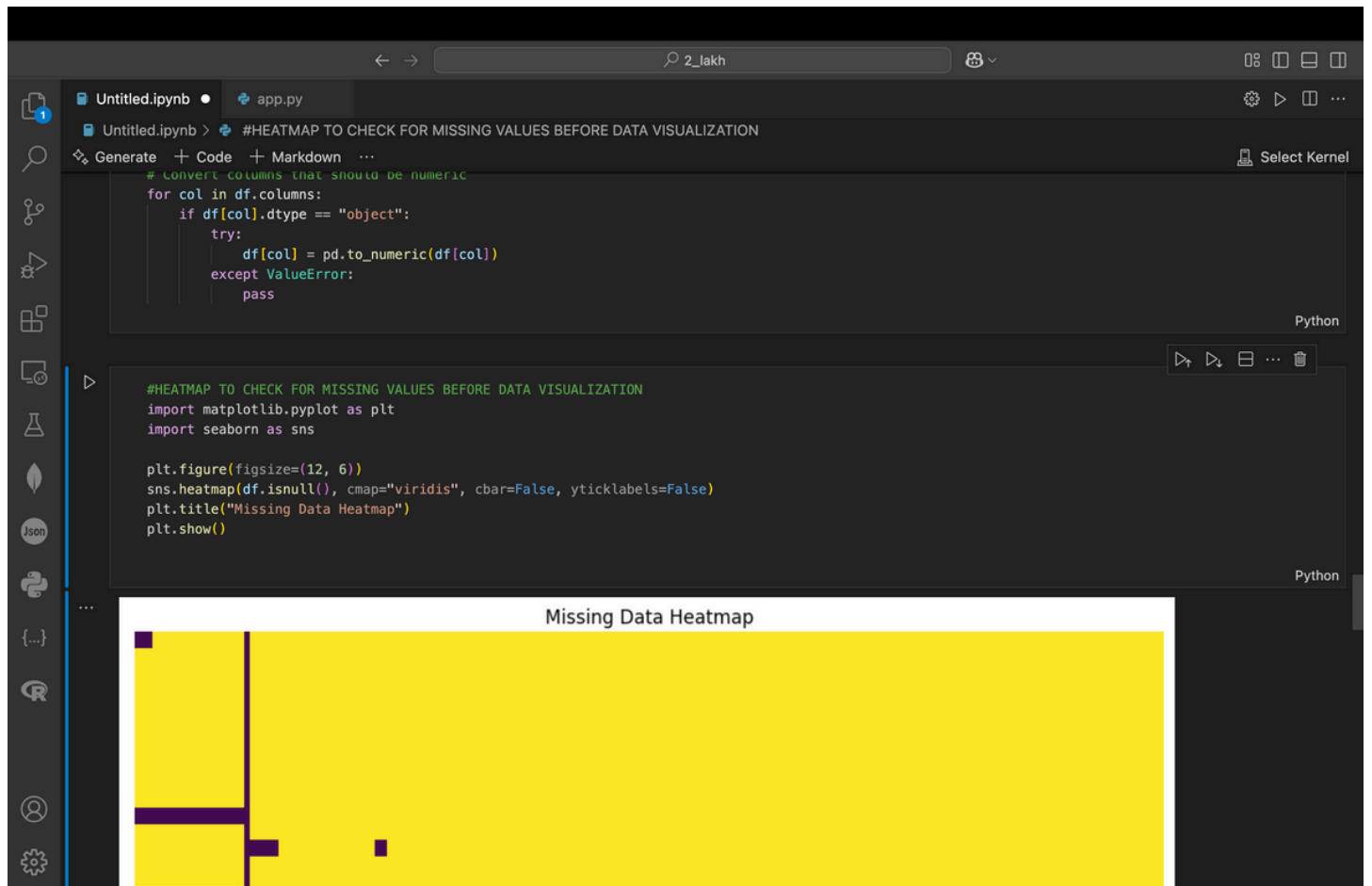
STREALIT DASHBOARD ----> app.py

```
Untitled.ipynb • app.py x
DASHBOARD > app.py > ...
1 import streamlit as st
2 import pandas as pd
3 import plotly.express as px
4 import seaborn as sns
5 import matplotlib.pyplot as plt
6
7 st.set_page_config(
8     page_title="Data-Driven Insights for Viksit Bharat",
9     page_icon="x",
10    layout="wide"
11)
12
13 @st.cache_data
14 def load_data():
15     df = pd.read_csv('results.csv')
16     return df
17
18 df = load_data()
19
20 st.sidebar.title("Dashboard Navigation")
21 viz_option = st.sidebar.radio("Select Visualization", [
22     "Overview",
23     "Internet Facility Distribution",
24     "Online Purchases by Category",
25     "Payment Method Distribution",
26     "Households by Sector",
27     "Online Purchases Over Time",
28     "Correlation Heatmap"
29 ])
30
31 if 'Year' in df.columns:
32     year_list = sorted(df['Year'].dropna().unique())
33     selected_year = st.sidebar.selectbox("Select Year", options=year_list)
34     df_filtered = df[df['Year'] == selected_year]
35 else:
36     df_filtered = df
37
38 online_purchase_categories = [
39     'Whether_any_online_purchase/payment_has_been_made_during_the_reference_period_to_buy_-_Fuel_&_light',
40     'Whether_any_online_purchase/payment_has_been_made_during_the_reference_period_to_buy_-_Toilet_articles_&_other_household_consumables',
41     'Whether_any_online_purchase/payment_has_been_made_during_the_reference_period_to_buy_-_Education',
42     'Whether_any_online_purchase/payment_has_been_made_during_the_reference_period_to_buy_-_Medicine_&_other_medical_services',
43     'Whether_any_online_purchase/payment_has_been_made_during_the_reference_period_to_buy_-_Services_(Travel_Recharges_Bill_payment_Cinema/Theatre_internet_etc.)_'
44 ]
45
46 if viz_option == "Overview":
47     st.title("Data-Driven Insights for Viksit Bharat")
48     st.markdown("### Overview")
49     st.dataframe(df.head(10))
50     st.markdown(f"Total Records: {df.shape[0]} | Columns: {df.shape[1]}")
51
52 elif viz_option == "Internet Facility Distribution":
53     st.title("Internet Facility Distribution in Households")
54     if 'Household_has_internet_facility_as_on_the_date_of_the_survey' in df_filtered.columns:
55         internet_counts = df_filtered['Household_has_internet_facility_as_on_the_date_of_the_survey'].value_counts()
```

```
2_lakh
Untitled.ipynb • app.py x
DASHBOARD > app.py > ...
36 fig = px.pie(
37     names=Internet_counts.index,
38     values=Internet_counts.values,
39     title="Internet Facility Availability",
40     color_discrete_sequence=px.colors.sequential.RdBu
41 )
42 st.plotly_chart(fig, use_container_width=True)
43
44 elif viz_option == "Online Purchases by Category":
45     st.title("Online Purchases by Category")
46
47     for col in online_purchase_categories:
48         if col in df_filtered.columns:
49             df_filtered[col] = df_filtered[col].map({'Yes': 1, 'No': 0}).fillna(0)
50
51     category_counts = df_filtered[online_purchase_categories].apply(pd.Series.value_counts).fillna(0)
52
53     if category_counts.shape[1] == 2:
54         category_counts.columns = ['No', 'Yes']
55     elif category_counts.shape[1] == 1:
56         only_response = category_counts.columns[0]
57         if only_response == 0:
58             category_counts = category_counts.rename(columns={0: 'No'})
59             category_counts['Yes'] = 0
60         elif only_response == 1:
61             category_counts = category_counts.rename(columns={1: 'Yes'})
62             category_counts['No'] = 0
63
64     fig = px.bar(
65         category_counts,
66         barmode='stack',
67         title="Online Purchases Made During the Reference Period",
68         labels={"value": "Number of Households", "index": "Category"},
69         color_discrete_sequence=px.colors.qualitative.Set3
70     )
71     fig.update_layout(xaxis_tickangle=-45)
72     st.plotly_chart(fig, use_container_width=True)
73
74 elif viz_option == "Payment Method Distribution":
75     st.title("Payment Method Distribution")
76     payment_col = 'If yes, in Q4.2.9 Amount?'
77     if payment_col in df_filtered.columns:
78         payment_counts = df_filtered[payment_col].value_counts()
79         fig = px.pie(
80             names=payment_counts.index,
81             values=payment_counts.values,
82             title="Distribution of Online Payment Methods",
83             color_discrete_sequence=px.colors.sequential.Blues
84         )
85         st.plotly_chart(fig, use_container_width=True)
86
87 elif viz_option == "Households by Sector":
88     st.title("Households by Sector")
89     if 'Sector' in df_filtered.columns:
90         sector_counts = df_filtered['Sector'].value_counts().reset_index()
91         sector_counts.columns = ['Sector', 'Count']
```

```
2_lakh
Untitled.ipynb • app.py x
DASHBOARD > app.py > ...
103 color_discrete_sequence=px.colors.sequential.Blues
104 )
105 st.plotly_chart(fig, use_container_width=True)
106
107 elif viz_option == "Households by Sector":
108     st.title("Households by Sector")
109     if 'Sector' in df_filtered.columns:
110         sector_counts = df_filtered['Sector'].value_counts().reset_index()
111         sector_counts.columns = ['Sector', 'Count']
112         fig = px.bar(
113             sector_counts,
114             x='Sector',
115             y='Count',
116             title="Households by Sector (Urban vs Rural)",
117             color='Sector',
118             color_discrete_sequence=px.colors.qualitative.Pastel
119         )
120         st.plotly_chart(fig, use_container_width=True)
121
122 elif viz_option == "Online Purchases Over Time":
123     st.title("Online Purchases Over Time")
124     if 'Year' in df.columns:
125         time_trend = df.groupby('Year').size().reset_index(name='Purchases')
126         fig = px.line(
127             time_trend,
128             x='Year',
129             y='Purchases',
130             markers=True,
131             title="Trend of Online Purchases Over the Years",
132             color_discrete_sequence=['#17becf']
133         )
134         st.plotly_chart(fig, use_container_width=True)
135
136 elif viz_option == "Correlation Heatmap":
137     st.title("Correlation Between Purchase Categories")
138     df_corr = df_filtered.copy()
139     for col in online_purchase_categories:
140         if col in df_corr.columns:
141             df_corr[col] = pd.to_numeric(df_corr[col].map({'Yes': 1, 'No': 0}), errors='coerce')
142     corr_matrix = df_corr[online_purchase_categories].corr()
143     fig, ax = plt.subplots(figsize=(10, 8))
144     sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f', ax=ax)
145     st.pyplot(fig)
146
147 st.sidebar.markdown("---")
148 st.sidebar.markdown("Download Cleaned Data")
149 st.sidebar.download_button(
150     label="Download CSV",
151     data=df.to_csv(index=False).encode('utf-8'),
152     file_name="online_purchase_data.csv",
153     mime="text/csv"
154 )
155
156 st.sidebar.text("Created by Abhishek Sharma")
157
```

COLLAB HEAT-MAP ---> Untitles.ipynb



**OUTPUT (FOR MORE DETAILS, RUN LOCALLY)**

DASHBOARD

Deploy

Dashboard Navigation

Select Visualization

Overview

Internet Facility Distribution

Online Purchases by Category

Payment Method Distribution

Households by Sector

Online Purchases Over Time

Correlation Heatmap

Select Year

157

Download Cleaned Data

Download CSV

Created by Abhishek Sharma

Data-Driven Insights for Viksit Bharat

Overview

	Survey_Name	Year	FSU_Serial_No.	Sector	State	NSS-Region	District	Stratum	Sub-stratum	Panel	Sub-sample	FOD-Sub-Region	S
0	size	157	04760	None	None	None	None	None	None	None	None	None	N
1	None	None	None	None	None	None	None	None	None	None	None	None	N
2	None	None	None	None	None	None	None	None	None	None	None	None	N
3	None	None	None	None	None	None	None	None	None	None	None	None	N
4	None	None	None	None	None	None	None	None	None	None	None	None	N
5	None	None	None	None	None	None	None	None	None	None	None	None	N
6	None	None	None	None	None	None	None	None	None	None	None	None	N
7	None	None	None	None	None	None	None	None	None	None	None	None	N
8	None	None	None	None	None	None	None	None	None	None	None	None	N
9	None	None	None	None	None	None	None	None	None	None	None	None	N

Total Records: 32 | Columns: 219

Deploy

Dashboard Navigation

Select Visualization

Overview

Internet Facility Distribution

Online Purchases by Category

Payment Method Distribution

Households by Sector

Online Purchases Over Time

Correlation Heatmap

Select Year

157

Download Cleaned Data

Download CSV

Created by Abhishek Sharma

Correlation Between Purchase Categories

Whether\_any\_online\_purchasepayment\_has\_been\_made\_during\_the\_reference\_period\_to\_buy\_Fuel\_&\_light -

Whether\_any\_online\_purchasepayment\_has\_been\_made\_during\_the\_reference\_period\_to\_buy\_Tobacco\_articles\_&\_other\_household\_consumables -

Whether\_any\_online\_purchasepayment\_has\_been\_made\_during\_the\_reference\_period\_to\_buy\_Education -

Whether\_any\_online\_purchasepayment\_has\_been\_made\_during\_the\_reference\_period\_to\_buy\_Medicine\_&\_other\_medical\_services -

Whether\_any\_online\_purchasepayment\_has\_been\_made\_during\_the\_reference\_period\_to\_buy\_Services\_Travel\_Recharges\_Bill\_payment\_CinemaTheatre\_internet\_etc\_1 -

Whether\_any\_online\_purchasepayment\_has\_been\_made\_during\_the\_reference\_period\_to\_buy\_Fuel\_&\_light -

Whether\_any\_online\_purchasepayment\_has\_been\_made\_during\_the\_reference\_period\_to\_buy\_Tobacco\_articles\_&\_other\_household\_consumables -

Whether\_any\_online\_purchasepayment\_has\_been\_made\_during\_the\_reference\_period\_to\_buy\_Education -

Whether\_any\_online\_purchasepayment\_has\_been\_made\_during\_the\_reference\_period\_to\_buy\_Medicine\_&\_other\_medical\_services -

Whether\_any\_online\_purchasepayment\_has\_been\_made\_during\_the\_reference\_period\_to\_buy\_Services\_Travel\_Recharges\_Bill\_payment\_CinemaTheatre\_internet\_etc\_1 -

