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

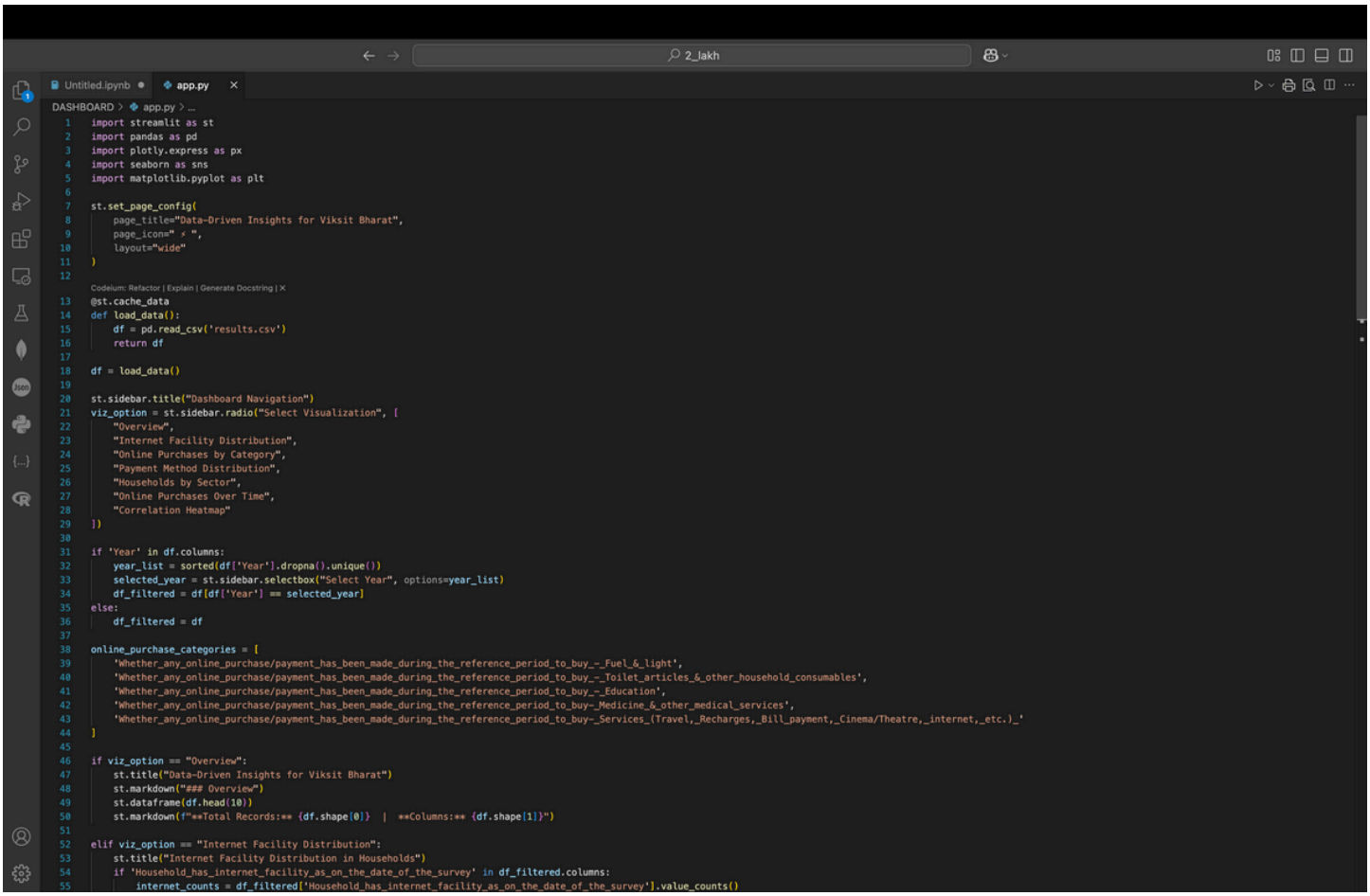
This analysis underscores the potential of HCES 2022-23 data to inform equitable policy-making. While data sparsity limited granular insights, the findings highlight critical gaps in

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*.

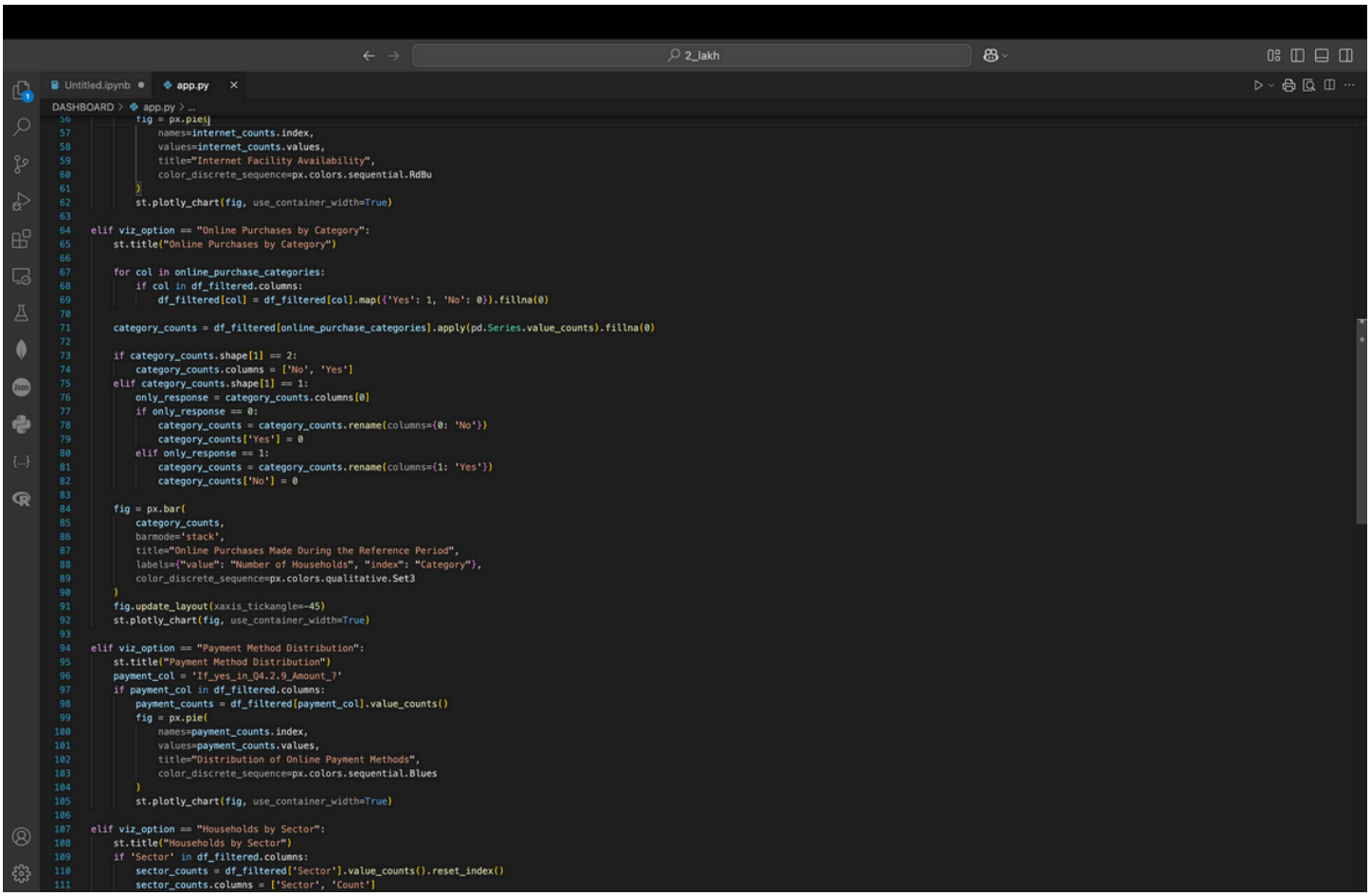
SOURCE CODE (ALSO AVAILABLE ON GITHUB ON:

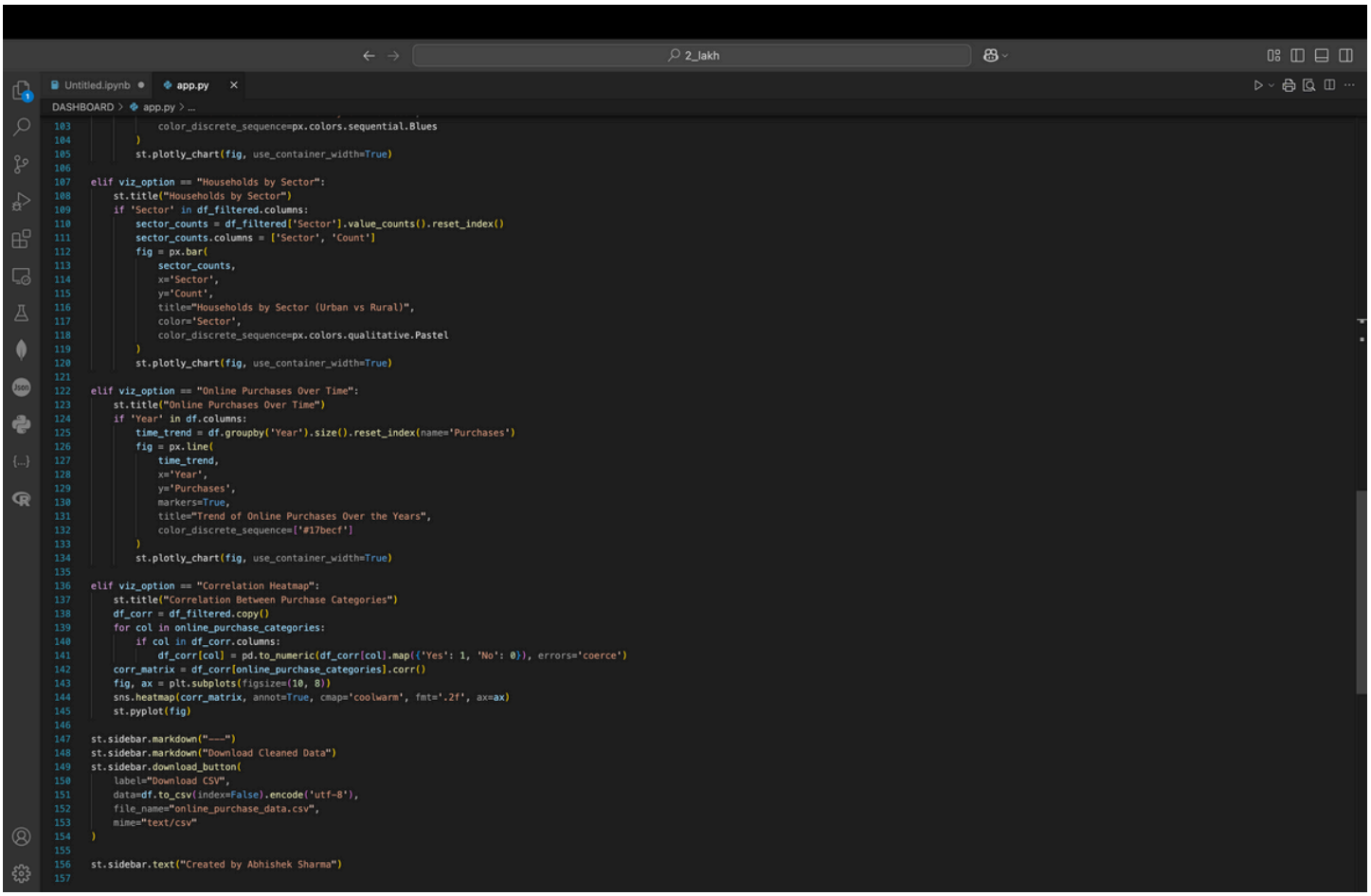
abhi-1408-shek/Innovate_with_GoIStats

STREALIT DASHBOARD ----> app.py

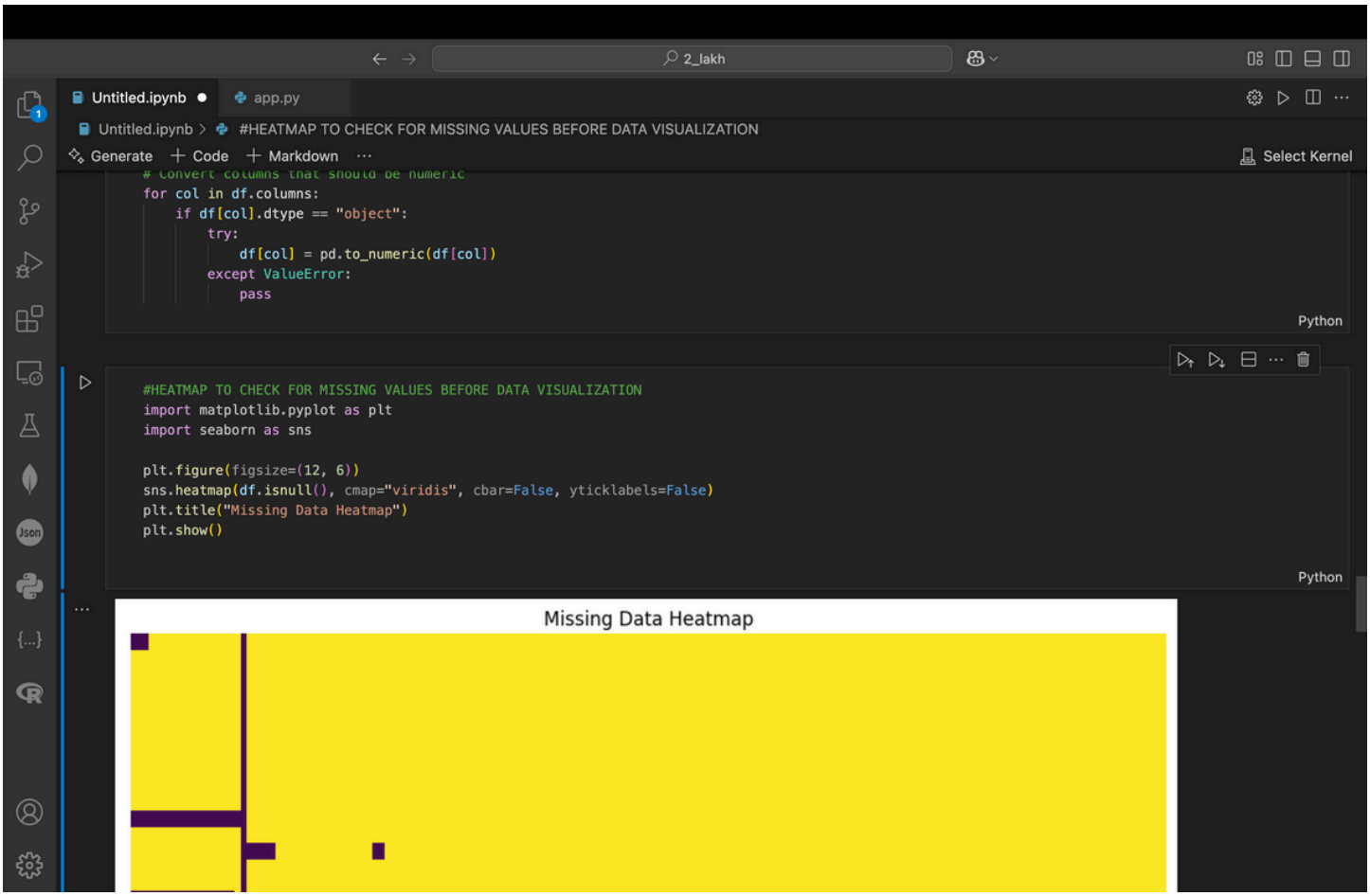


```python
import streamlit as st
import pandas as pd
import plotly.express as px
import seaborn as sns
import matplotlib.pyplot as plt

st.set_page_config(
    page_title="Data-Driven Insights for Viksit Bharat",
    page_icon=" ⚡ ",
    layout="wide"
)

Codeium: Refactor | Explain | Generate Docstring | ×
@st.cache_data
def load_data():
    df = pd.read_csv('results.csv')
    return df

df = load_data()

st.sidebar.title("Dashboard Navigation")
viz_option = st.sidebar.radio("Select Visualization", [
    "Overview",
    "Internet Facility Distribution",
    "Online Purchases by Category",
    "Payment Method Distribution",
    "Households by Sector",
    "Online Purchases Over Time",
    "Correlation Heatmap"
])

if 'Year' in df.columns:
    year_list = sorted(df['Year'].dropna().unique())
    selected_year = st.sidebar.selectbox("Select Year", options=year_list)
    df_filtered = df[df['Year'] == selected_year]
else:
    df_filtered = df

online_purchase_categories = [
    'Whether_any_online_purchase/payment_has_been_made_during_the_reference_period_to_buy_-_Fuel_&_light',
    'Whether_any_online_purchase/payment_has_been_made_during_the_reference_period_to_buy_-_Toilet_articles_&_other_household_consumables',
    'Whether_any_online_purchase/payment_has_been_made_during_the_reference_period_to_buy_-_Education',
    'Whether_any_online_purchase/payment_has_been_made_during_the_reference_period_to_buy-_Medicine_&_other_medical_services',
    'Whether_any_online_purchase/payment_has_been_made_during_the_reference_period_to_buy-_Services_(Travel,_Recharges,_Bill_payment,_Cinema/Theatre,_internet,_etc.)_'
]

if viz_option == "Overview":
    st.title("Data-Driven Insights for Viksit Bharat")
    st.markdown("### Overview")
    st.dataframe(df.head(10))
    st.markdown(f"**Total Records:** {df.shape[0]}  |  **Columns:** {df.shape[1]}")

elif viz_option == "Internet Facility Distribution":
    st.title("Internet Facility Distribution in Households")
    if 'Household_has_internet_facility_as_on_the_date_of_the_survey' in df_filtered.columns:
        internet_counts = df_filtered['Household_has_internet_facility_as_on_the_date_of_the_survey'].value_counts()
```





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

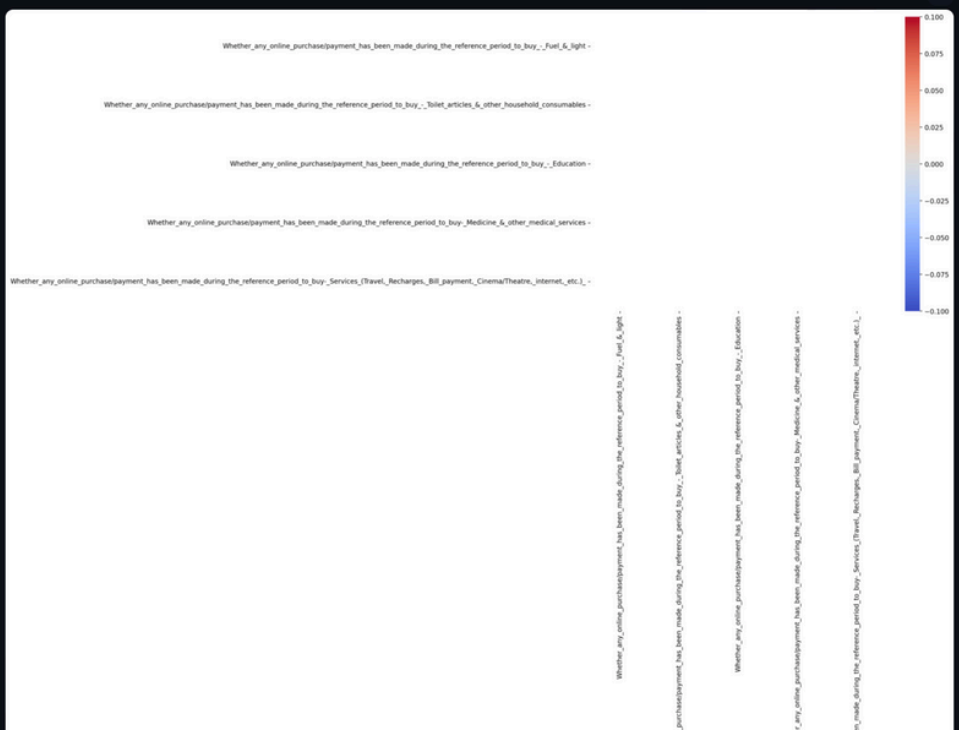
157

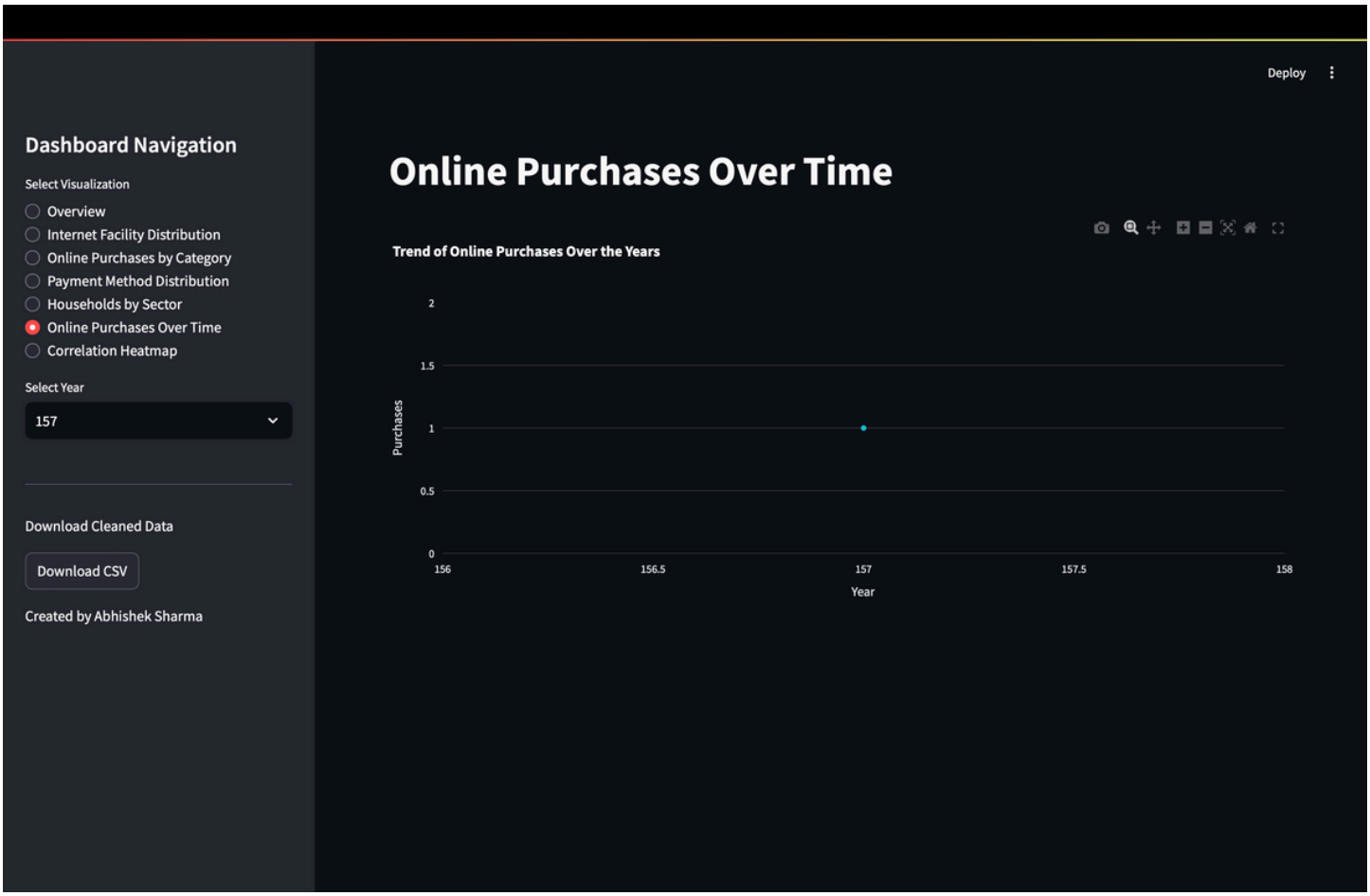
Download Cleaned Data

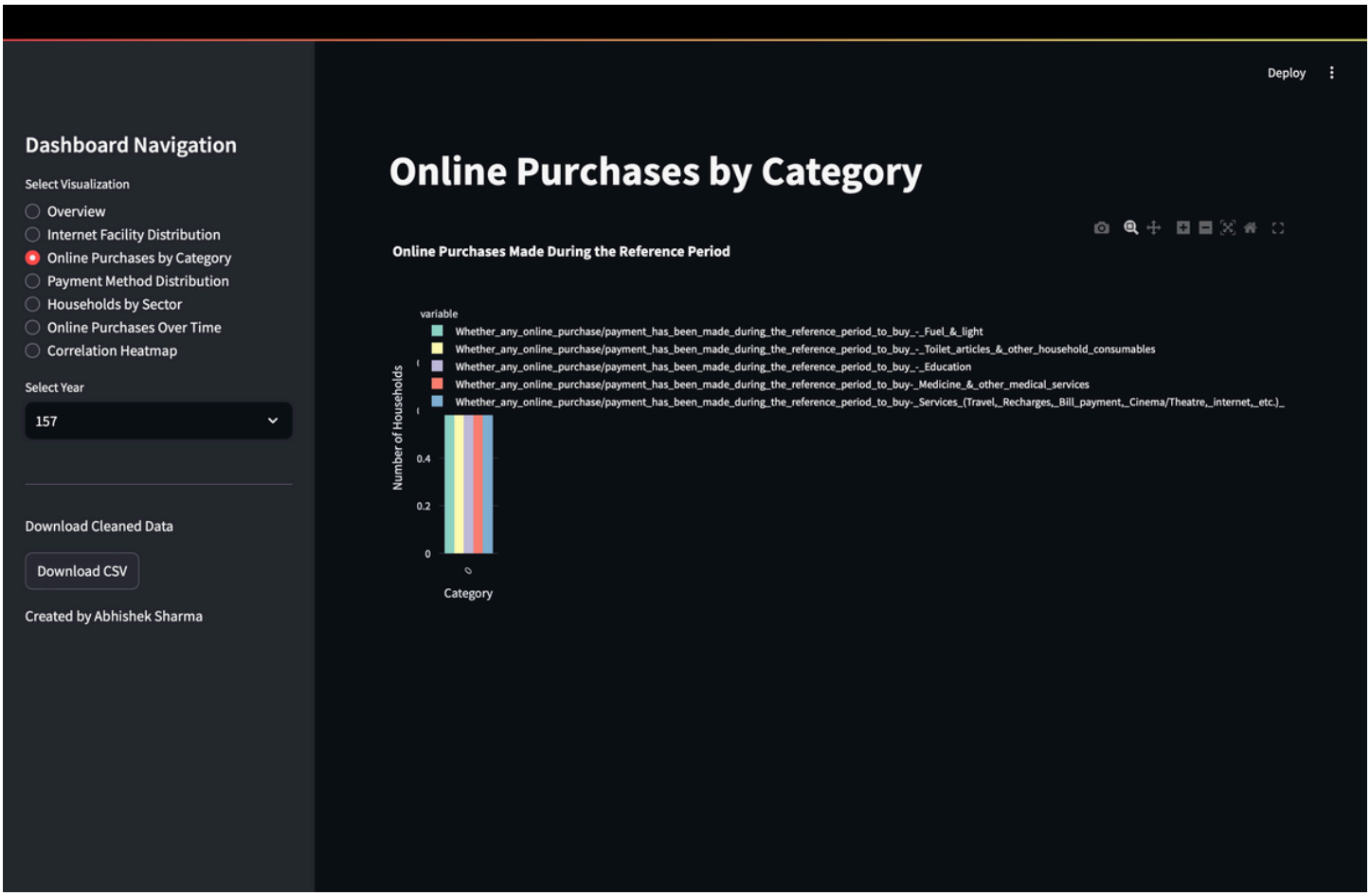
Download CSV

Created by Abhishek Sharma

Correlation Between Purchase Categories







COLLAB HEAT-MAP


2_lakh
Untitled.ipynb
app.py
Untitled.ipynb > #HEATMAP TO CHECK FOR MISSING VALUES BEFORE DATA VISUALIZATION
Generate
Code
Markdown
Select Kernel
Missing Data Heatmap
Survey_Name
Sector
Stratum
Sub-sample
Questionnaire_No.
NSS-Region_dup
Sub-stratum_dup
Level
Value_(Rs.)
Clothing
Mobile_handset
Medical_equipment
Laptop/PC
Mobile
Bicycle
Value_(Rs.)
Number
Section
Section_dup
Household_size
Gender
Wheat
Pulses
Groceries
Fresh_Fruits
processed_food
Value(Rs.)
Quantity(0.000)
last_3_months?
last_365_days
Textbooks_?
received
Others_?
last_365_days?
Quantity(0.000)_dup