



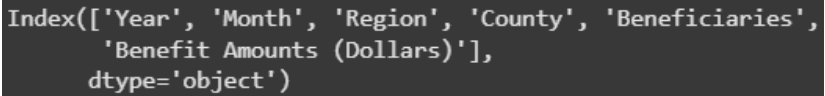
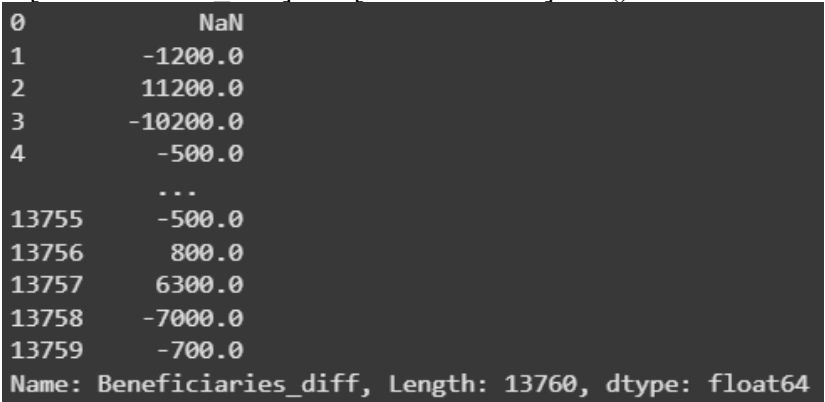
Data Collection and Preprocessing Phase

| | |
|---------------|--|
| Date | 18 June 2025 |
| Team ID | AS PS VS VV |
| Project Title | Unemployed Insurance Beneficiary Forecasting |
| Maximum Marks | 6 Marks |

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

| Section | Description |
|------------------------|--|
| Data Overview | Calculated dataset shape with <code>df.shape</code> and checked data types and structure using <code>df.info()</code> . Dataset contains monthly records of beneficiaries, benefit amounts, regions, and counties. |
| Univariate Analysis | Used <code>df.describe()</code> and line plots to explore distributions and summary statistics (mean, median, min, max) for variables like 'Beneficiaries' and 'Benefit Amounts (Dollars)'. |
| Bivariate Analysis | Examined relationships between pairs of variables, such as plotting 'Beneficiaries' over time for different counties and comparing beneficiary counts across regions using bar plots. |
| Multivariate Analysis | Generated boxplots for multiple numeric columns to identify patterns, distributions, and outliers across several variables simultaneously. |
| Outliers and Anomalies | Detected outliers using boxplots and summary statistics; treated anomalies by reviewing data points and deciding whether to keep, transform, or remove them as appropriate. |

| Data Preprocessing Code Screenshots | |
|-------------------------------------|---|
| Loading Data | <pre>uploaded = files.upload() df = pd.read_csv('unemployment-insurance-beneficiaries-and-benefit-amounts-paid-beginning-2001-1 (1).csv')</pre>  |
| Handling Missing Data | <pre>print(df.isna().sum())</pre>  |
| Data Transformation | <pre>df.columns = df.columns.str.strip() Index(['Year', 'Month', 'Region', 'County', 'Beneficiaries', 'Benefit Amounts (Dollars)'], dtype='object')</pre>  |
| Feature Engineering | <pre>df['Beneficiaries_diff'] = df['Beneficiaries'].diff()</pre>  |
| Save Processed Data | <pre>df.to_csv('processed_unemployment_data.csv', index=False)</pre> |