A case study of economic dynamics in the USA from the 20th and 21st centuries; median compensation, productivity and poverty level trends

This case study explores the economic dynamics of the USA by examining trends in median compensation, productivity, and poverty-level wages over the course of the 20th and 21st centuries. The study primarily investigates how changes in productivity and median compensation have influenced annual poverty level wages during two distinct time periods: the 1980-1990s and the 2010-2020s. Additionally, the study delves into patterns based on gender, seeking to understand how these economic factors have differently impacted men and women over time. Through this analysis, the study aims to uncover significant trends and insights into the evolving economic landscape of the USA.

1. Question

The main question this study aims to answer is: ***How has productivity and median compensation affected the annual poverty level wages in the USA from two distinct time periods?***

The study also seeks further insights which can be broken down into the following questions.

- What are some major patterns from the 2 time periods (1980-1990 vs 2010-2020)?
- What are some major patterns based on gender for these parameters?

2. Data sources

	Data source 1	Data source 2
Name	Productivity and Hourly Compensation (1948-2021)	Poverty-Level Wages in the USA Dataset (1973-2022)
Source	Kaggle	Kaggle
License	Both datasets are licensed under CC0: Public domain. It explicitly waives copyright and does not require attribution. The data has been sourced from Economic Policy Institute's State of Working America Data Library	
Data URL	https://www.kaggle.com/datasets/asaniczka/productivity-and-hourly-compensation-1948-2021	https://www.kaggle.com/datasets/asaniczka/poverty-level-wages-in-the-usa-dataset-1973-2022
Data format and structure	CSV Tabular data	CSV Tabular data
Data Quality	Data is quite well-structured, consistent and complete. Descriptive statistics have been applied; measure of central tendency (median).	Data is quite well-structured, consistent and complete. There are no missing values in the dataset.
Content Overview	It contains information about the productivity and hourly compensation trends in the USA from 1948 to 2021. The hourly compensation is represented in 2021 US dollars. The compensation for men and women is represented	It provides information on poverty-level wages in the USA from 1973 to 2022. It includes data on both annual and hourly poverty-level wages, as well as wage shares for different income brackets. Wages mentioned are in nominal dollars. The Census Bureau's annual poverty-level

	separately. Moreover, it contains information on both total compensation and compensation specifically for production and non-supervisory employees.	wage is used for 2022 and deflated using the CPI-U-RS (Consumer Price Index for All Urban Consumers – Research Series). Annual wages are converted to hourly wages for full-time, year-round workers.
Reason behind selection	Information related to productivity and hourly compensation could be extracted for the years required. Moreover, the data is consistent for different times.	The data provides information about the share of men and women earning below and above the poverty level wage (wide range) along with annual poverty level wage. The greater range might lead to better results in analyzing trends related to economic gender disparity.

3. Data Pipeline

The automated data pipeline has been built using Python in VS Code and follows the ETL structure (Extract, Transform and Load). It extracts data from two sources via Kaggle API, automatically unzips the data folders, reads the .csv files, performs relevant transformation on the data and saves it into an SQLite database. Some modules and libraries used are pandas, os, sys, sqlite3.

Transformation and Cleaning steps	Error handling
• Only the columns relevant to the study and from our target years (1980-1990 and 2010-2020) are kept. For example, 'net_productivity_per_hour_worked', 'median_compensation'. • The datasets are merged based on the column 'year' as data mutual to these years is needed. • Although the datasets utilized do not have any missing values, to keep the pipeline more structured, a separate function has been created to handle this case and input a mean value for it.	• For connectivity, invalid URL, missing key errors etc., try-except blocks have been utilized within the functions. • For the problem statement, since we are interested only in data from 2 decades, a check has been kept in place that the resultant data should only be from 20 years. • As part of transformation, since the required columns are specified, error in case of invalid or missing columns has been handled. • For datasets merge, error in case of invalid key column is handled.

Problem encountered	Solution
Personally, the major hurdle for me was not related to data transformation but it was to select the right parameters (columns) to best answer the question statement. For example, datasource2 has many columns i.e. 33, a lot of them representing share (different %) of men/women below/above the poverty level which seemed a bit unnecessary.	After careful consideration, to keep it relatively simple and highlight outliers (if any), I selected the extremes to keep in the data output: 'wo(men)_share_below_poverty_wages' 'wo(men)_300%+_of_poverty_wages'.

4. Result and limitations

Result	The output merged data from two original datasets is well-structured, compact and narrowed down to perform comprehensive data analysis based on the question statement. Columns related to net productivity, men/women compensation, annual poverty level wages, share of men/women above/below the poverty levels from 1980-1990 and 2010-2020 have been kept.
Why SQLite is selected as an output data format?	The pipeline output is in SQLite format since it easily integrates with python, befits medium sized dataset and is easy to share.
Limitations	• The annual poverty-level wage in 2022 may not match the Census Bureau's published wage due to rounding hourly wages to the nearest cent. • The data for the later years (2019-2020) in the second time frame (2010-2020) may misrepresent the results due to the corona virus pandemic. It caused global economic slowdowns, widespread unemployment, and disruptions in industries, which is why, these years may introduce outliers due to the pandemic's economic impact.
Critical Analysis	The patterns from this data can be used to explore differences in economic productivity from two centuries for both genders. The results can be utilized to highlight wage inequalities, wage stagnation, inequitable income distribution (if any), potentially leading to promoted economic growth and better decision making by identifying the underlying causes.