Penthouse Pandas - US Immigrants Regression Analysis

- Dhrumil Buch
- Sharad Param
- Atharva Sankpal

The analysis in this project focuses on the "US Immigration Statistics" dataset available on Kaggle, spanning from 1980 to 2020, containing diverse immigration metrics. Initial scrutiny revealed inconsistent data types, particularly caused by commas in numeric columns leading to storage as objects. This hindered quantitative analyses and visualizations. To remedy this, a meticulous data cleaning process was executed, involving comma removal and conversion of affected columns to integers, rendering the dataset suitable for modeling and exploration.

Upon completion of data cleaning, the project progressed to visualization to extract crucial patterns and insights. Time series plots were generated using Python libraries such as Matplotlib and Seaborn to depict trends in each immigration metric over time. Bar graphs facilitated direct comparisons of metric magnitudes, elucidating the relative significance of different immigration indicators. These visuals served as a foundation for further investigations and provided a comprehensive overview of immigration trends. Additionally, key challenge addressed was incorporating presidential and party data to explore correlations with immigration metrics. This likely involved merging datasets and creating plots to visualize relationships between immigration indicators and presidential/party affiliations.

Regression analysis was conducted to delve deeper into the dataset and uncover potential correlations among variables. Specifically, linear and polynomial regression models were constructed to explore the relationship between the year and noncitizen removals. Model efficacy was evaluated by partitioning the dataset into training and testing subsets and utilizing assessment metrics like mean squared error (MSE) and R-squared. Moreover, regression analysis provided a deeper understanding of the dataset, uncovering potential correlations among variables. Ultimately, this project yielded valuable insights into immigration trends and patterns, demonstrating the utility of regression modeling in predicting and comprehending fluctuations in immigration-related factors over time.