**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***"Understanding Voting Patterns Across Counties and Demographics in California"***

Submitted by

**Ruban Ciby**

Registration No-12309016

Programme and Section- BTECH(CSE) & K23GD

Course Code -INT375

Under the Guidance of

**Mrs. Baljinder Kaur**

**Discipline of CSE/IT**

****

**CERTIFICATE**

This is to certify that Ruban Ciby bearing Registration no. 12309872 has completed INT375 project titled, *"Understanding Voting Patterns Across Counties and Demographics in California"*under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 13/04/2024

**DECLARATION**

I, Ruban Ciby , student of BTECH (CSE) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 13/04/2025

Registration No. 12309016

**ACKNOWLEDGEMENT**

First and foremost, I would like to express my sincere gratitude to my project guide, Mrs. BALJINDER KAUR, for his valuable guidance, continuous support, and encouragement throughout the course of this project. His mentorship has helped me grow academically and personally.

I also extend my thanks to the faculty members and peers at Lovely Professional University for providing an environment that encourages learning and innovation.

**TABLE OF CONTENTS**

1. Introduction
2. Source of Dataset
3. EDA Process
4. Analysis on Dataset
5. Conclusion
6. Future Scope
7. References

**1. INTRODUCTION**

This project presents an exploratory data analysis (EDA) of voter registration trends across California. The dataset contains detailed records of registered voters spanning multiple years, counties, and racial/ethnic demographics. Each entry includes essential metrics such as the number of registered voters (numerator), voting age population (vap), registration rates, and geographic identifiers like county and region codes.

The primary objective of this analysis is to uncover insights into how voter registration patterns vary by year, demographic group, and geographic region. By visualizing these patterns, the project aims to identify disparities, trends, and potential outliers in voter participation. The study further explores correlations between variables and applies clustering techniques to group counties with similar registration behaviour.

**2. SOURCE OF DATASET**

The dataset used in this project—eda.csv—appears to be derived from **California voter registration statistics** compiled across various years and demographic segments. It includes detailed information such as:

* Number of registered voters (numerator)
* Voting age population (vap)
* Demographic identifiers (e.g., race/ethnicity, region, county)
* Year-wise trends (reportyear)
* Statistical estimates (e.g., confidence intervals, standard error)

**3. EDA PROCESS**

The EDA was conducted using Python and libraries such as Pandas, Matplotlib, Seaborn, and NumPy. The overall EDA workflow includes:

* Data Cleaning: Checking for null values, inconsistent entries, and duplicates.
* Data Transformation: Converting data types, filtering relevant columns.
* Data Aggregation: Grouping data to derive macro-level insights.
* Visualization: Creating various plots to understand distributions, trends, and comparisons.

Key Tools Used:

* Pandas for data manipulation
* Seaborn and Matplotlib for visualization
* NumPy for numerical operations

**4. ANALYSIS ON DATASET**

**1. Understanding the Dataset Structure**

* **Objective:** Load the dataset and understand the number of rows, columns, data types, and summary statistics.
* **Why:** Gives a snapshot of what’s available — such as numerator, vap, report year, and race \_eth\_ name — and helps identify null values, data ranges, or categorical vs numeric columns.

**2. Cleaning the Dataset**

* **Objective:** Handle missing values, drop irrelevant or duplicate rows, and fix type mismatches (e.g., converting report year to integer).
* **Why:** Clean, consistent data ensures trustworthy insights and allows smooth plotting and transformations.

**3. Filtering for Total Race/Ethnicity (All Groups Combined)**

* **Objective:** Filter rows where race \_eth \_name == 'Total' to focus on population-wide metrics without demographic splits.
* **Why:** Allows for unbiased analysis of overall registration trends across counties and time without group-specific influence.

**4. Analyzing Voter Registration Over Time**

* **Objective:** Visualize total registered voters (numerator) across different years.
* **Why:** Helps track civic participation trends over time and correlates with election cycles or policy changes.

**5. County-wise Registration Trends**

* **Objective:** Identify counties with the highest and lowest voter registrations.
* **Why:** Highlights areas of high civic engagement or neglect, useful for targeted outreach or support.

**6. Analyzing Registration by Race/Ethnicity**

* **Objective:** Group data by race\_eth\_name and reportyear to assess registration disparities.
* **Why:** Crucial for understanding equity in access and representation in the voting system.

**7. Calculating and Analyzing Registration Rate**

* **Objective:** Compute registration rate as numerator / vap and study it over time and by county.
* **Why:** Registration numbers are only meaningful in context — this rate shows what portion of eligible citizens are registered.

**8. Identifying Counties with Extreme Registration Rates**

* **Objective:** Use Z-score to find counties with significantly high or low registration rates.
* **Why:** Outlier counties might signal overperformance (effective campaigns) or underperformance (barriers to access).

**9. Clustering Counties Based on Registration Behavior**

* **Objective:** Apply KMeans clustering on features like average vap and numerator per county.
* **Why:** Groups similar counties together to spot regional behavioral patterns in voter registration.

**10. Visualizations and Styling**

* **Objective:** Use well-styled plots (with titles, axis labels, legends, consistent colors).
* **Why:** Makes data stories easy to understand, presentable, and visually compelling for reports or presentations.

**5. CONCLUSION**

This analysis provided valuable insights into voter registration patterns across California's counties, years, and demographic groups. By cleaning and exploring the dataset thoroughly, we discovered key trends and disparities in civic engagement. Notable findings include:

* **Steady Growth in Voter Registration:** There has been a general increase in the number of registered voters over the years, reflecting either population growth, heightened political awareness, or both.
* **Disparities Among Counties:** Certain counties consistently report higher registration numbers, while others lag behind. These variations may be influenced by factors such as population size, socio-economic status, or outreach effectiveness.
* **Race and Ethnicity Trends:** Registration patterns varied significantly among racial and ethnic groups, indicating potential gaps in access or engagement that may warrant further investigation or policy intervention.
* **Registration Rate as a Key Metric:** The registration rate (registered voters divided by the voting age population) offered a more normalized and insightful view of participation, helping to identify underperforming regions more accurately.
* **Clustering Revealed Regional Behavior:** KMeans clustering grouped counties into distinct behavior patterns, suggesting regional similarities that could guide future voter engagement strategies.
* **Outliers Highlight Potential Issues:** The Z-score analysis surfaced counties with unusually high or low registration rates, which could point to anomalies worth further exploration — such as administrative inefficiencies, campaign success, or data inconsistencies.

**6. FUTURE SCOPE**

**1. Time-Series Forecasting**

* Goal: Predict future voter registration trends using models like ARIMA or LSTM.
* Why: Anticipating voter turnout can help government agencies and political groups prepare campaigns and allocate resources.

**2. Advanced Demographic Analysis**

* **Goal:** Dive deeper into age, gender, income level, and education if available.
* **Why:** More granular segmentation can reveal barriers or motivators to voter participation within specific population segments.

**3. Machine Learning for Prediction**

* **Goal:** Predict low-registration areas using classification or regression models.
* **Why:** Helps target intervention efforts, such as civic engagement programs or mobile registration drives.

**4. Policy Impact Studies**

* **Goal:** Analyze how laws (e.g., automatic registration, vote-by-mail) affected registration rates over time.
* **Why:** Supports evidence-based decision making on electoral reforms.

**5. Real-Time Dashboards**

* **Goal:** Create an interactive dashboard using tools like Dash, Streamlit, or Power BI.
* **Why:** Enables continuous monitoring and easy communication of findings to stakeholders.

**7. REFERENCES**  
[1] [https://seaborn.pydata.org](https://seaborn.pydata.org/)  
[2] [https://matplotlib.org](https://matplotlib.org/)

[3] [https://pandas.pydata.org](https://pandas.pydata.org/)