Introduction

In the digital age, data analysis has become an essential tool for uncovering insights and making informed decisions. This project delves into the relationship between racial demographics and voting patterns in the United States. Specifically, we aim to understand how the majority race in a county influences the likelihood of that county voting Republican (REP) or Democrat (DEM). By leveraging data from the 2020 U.S. elections and the U.S. Census, we seek to uncover statistically significant trends that can shed light on the socio-political landscape of the nation.

Data Introduction

The primary dataset used in this analysis is the 2020 U.S. Election dataset, sourced from Kaggle. This dataset contains detailed information about voting outcomes at the county level across the United States. Each record in the dataset represents a county and includes variables such as the total number of votes for each party, the winning party, and other relevant electoral data.

To enhance our analysis, we integrated this election data with demographic data from the U.S. Census. The census data provides additional context, particularly the racial composition of each county. This allows us to analyse how the racial majority within a county correlates with the county's voting behaviour.

High-Level Explanation

The core of our analysis revolves around two key variables:

- **Majority Race (majority race)**: The predominant racial group in each county, as determined by the census data.

- **Voting Outcome (winning party)**: Whether a county voted Republican or Democrat in the 2020 election.

The primary goal was to investigate whether the racial composition of a county significantly influences its likelihood of voting Republican. We applied a series of statistical tests and visualizations to explore this relationship.

Explanation of Calculations and Data Processing

1. Data Cleaning and Preparation

- **Objective**: Combine U.S. election data with census demographic data to analyse voting patterns based on the majority race.
- **Steps**:
    - **Data Merging**: Aligned election results with demographic data using state and county as keys.
    - **Data Cleaning**: Removed incomplete records, resulting in a 29% data loss, primarily from small, remote counties (e.g., in Alaska or deserts), which are mostly white. Despite this, the remaining data is robust for analysis.

2. Statistical Analysis

- **Objective**: Determine the likelihood of voting Republican or Democrat based on race.
- **Steps**:
    - **Proportion Test (Z-Test)**: Showed White-majority counties are significantly more likely to vote Republican (Z=65.739, p=0.0).
    - **Chi-Square Test**: Found a strong association between race and voting patterns (Chi2=690.51, p=3.96e-148).
    - **T-Test**: Confirmed White-majority counties have a higher likelihood of voting Republican compared to non-white-majority counties (T=18.561, p=1.62e-51).

3. Probability Calculation

- **Objective**: Calculate probabilities of each racial group voting Republican or Democrat.
- **Steps**:
    - **Data Grouping**: Grouped by majority race and voting outcome.
    - **Probability Calculation**: Calculated the likelihood for each race and saved results to the output directory.

Results and Detailed Explanations

1. Proportion Test (Z-Test): Likelihood of Voting Republican in White-Majority Counties

- **Z-statistic**: 65.739
- **P-value**: 0.0

What Does This Mean?

- **Z-Statistic (65.739)**: The Z-statistic is a measure of how far your sample proportion is from the null hypothesis proportion (in this case, 50%), in units of the standard error. A Z-statistic of 65.739 is extraordinarily high, indicating that the proportion of White-majority counties voting Republican is much higher than 50%.
    - **Interpretation**: This extremely high Z-score suggests that the observed proportion is not just due to random chance. In fact, the chance that this result is due to randomness is so low that we can confidently say White-majority counties are significantly more likely to vote Republican.
- **P-Value (0.0)**: The p-value tells us the probability of observing a test statistic as extreme as the Z-statistic (or more extreme) if the null hypothesis were true (i.e., if 50% of White-majority counties are expected to vote Republican).
    - **Interpretation**: A p-value of 0.0 (or remarkably close to zero) indicates that there is almost no chance that the observed voting pattern happened by random chance. Thus, we reject the null hypothesis and conclude that White-majority counties are highly likely to vote Republican.

2. Chi-Square Test: Association Between Race and Voting Patterns

- **Chi-Square Statistic**: 690.51
- **P-value**: 3.96e-148

What Does This Mean?

- **Chi-Square Statistic (690.51)**: The Chi-square statistic measures how much the observed counts (votes for REP or DEM) differ from the expected counts if there were no association between race and voting patterns. A higher Chi-square statistic indicates a greater difference between observed and expected counts.

    - **Interpretation**: A Chi-square statistic of 690.51 is extremely high, suggesting that the observed voting patterns differ significantly from what would be expected if race had no effect on voting behaviour.

- **P-Value (3.96e-148)**: The p-value here is effectively zero, indicating that the probability of observing such a strong association between race and voting patterns by chance is almost non-existent.

    - **Interpretation**: The extremely low p-value confirms that there is a strong and statistically significant association between the majority race in a county and the likelihood of voting Republican or Democrat.

3. T-Test: Comparing Likelihood of Voting Republican Between White-Majority and Non-White-Majority Counties

- **T-statistic**: 18.561

- **P-value**: 1.62e-51

What Does This Mean?

- **T-Statistic (18.561)**: The T-statistic measures the difference between the means of two groups (in this case, White-majority vs. non-white-majority counties) relative to the variation in the data. A T-statistic of 18.561 is extremely high, indicating that the difference in voting patterns between these two groups is large and unlikely to be due to chance.

    - **Interpretation**: The high T-statistic suggests that White-majority counties have a significantly different voting pattern compared to non-white-majority counties, with a strong tendency to vote Republican.

- **P-Value (1.62e-51)**: The p-value here is also remarkably close to zero, indicating that the likelihood of observing such a major difference in voting patterns purely by chance is virtually impossible.

    - **Interpretation**: The extremely low p-value means we can confidently conclude that White-majority counties are statistically more likely to vote Republican compared to non-white-majority counties.

Summary of Findings

- **Proportion Test (Z-Test)**: The significantly high Z-score (65.739) and the p-value of 0.0 suggest that White-majority counties overwhelmingly tend to vote Republican.

- **Chi-Square Test**: The large Chi-square statistic (690.51) and the p-value of nearly zero indicate a strong association between the majority race in a county and its voting outcome, proving that race plays a crucial role in voting behaviour.
- **T-Test**: The T-statistic (18.561) and its near-zero p-value show a substantial difference in voting patterns between White-majority and non-white-majority counties, with White-majority counties being much more likely to vote Republican.

These results provide compelling evidence that racial demographics are a significant factor in predicting whether a county will vote Republican or Democrat. Each statistical test reinforces the conclusion that race is a key determinant in voting behaviour across U.S. counties.