Thank you for providing the codebook. Let me first review its content to understand the structure and details of the data you're referring to.

Finished working

Show work

The codebook seems to mention three data files:

1. demographics.csv
2. wages.csv
3. states.csv

However, the content of the codebook doesn't seem to provide detailed descriptions for each of these files. Would you be able to share the data files mentioned or provide more context about their structure and content?



Here is the description of the columns in the demographics sheets: variable class description year double When the data was collected. sample\_size double The number of wage and salary workers ages 16 and over who were surveyed. employment double Wage and salary employment in thousands. members double Employed workers who are union members in thousands. covered double Workers covered by a collective bargaining agreement in thousands. p\_members double Percent of employed workers who are union members. p\_covered double Percent of employed workers who are covered by a collective bargaining agreement. facet character The sector or demographic group contained in this row of data.

ChatGPT

Thank you for providing the description of the columns in the demographics.csv file. Here's a summary of the columns:

1. **year**: When the data was collected. (Type: Double)
2. **sample\_size**: The number of wage and salary workers ages 16 and over who were surveyed. (Type: Double)
3. **employment**: Wage and salary employment in thousands. (Type: Double)
4. **members**: Employed workers who are union members in thousands. (Type: Double)
5. **covered**: Workers covered by a collective bargaining agreement in thousands. (Type: Double)
6. **p\_members**: Percent of employed workers who are union members. (Type: Double)
7. **p\_covered**: Percent of employed workers who are covered by a collective bargaining agreement. (Type: Double)
8. **facet**: The sector or demographic group contained in this row of data. (Type: Character)

Would you like to upload the demographics.csv file now so we can further examine and analyze it? If you have descriptions for the other files (wages.csv and states.csv), please provide those as well.



Here is the description of the columns in the wages table: variable class description year double When the data was collected. sample\_size double The number of wage and salary workers ages 16 and over who were surveyed and provided earnings and hours worked information. wage double Mean hourly earnings in nominal dollars. at\_cap double Percent of workers with weekly earnings at the top code of $999 through 1988, $1923 in 1989-97, and $2885 beginning in 1998, with individuals assigned mean earnings above the cap based on annual estimates of the gender-specific Pareto distribution. union\_wage double Mean wage among union members. nonunion\_wage double Mean wage among nonunion workers. union\_wage\_premium\_raw double The percentage difference between the union and nonunion wage. union\_wage\_premium\_adjusted double Estimated as exp(b)-1 where b is the regression coefficient on a union membership variable (equal to 1 if union and 0 otherwise) from a semi-logarithmic wage equation, with controls included for worker/job characteristics. Included in the all-worker wage equation are the control variables: years of schooling, potential years of experience [proxied by age minus years of schooling minus 6] and its square [both interacted with gender], and categorical variables for marital status, race and ethnicity, gender, part-time, large metropolitan area, state, public sector, broad industry, and broad occupation. Controls are omitted, as appropriate, for estimates within sectors or by demographic group [i.e., by class, gender, race, or industry sector]. Workers who do not report earnings but instead have them imputed [i.e., assigned] by the Census are removed from the estimation samples in all years, except 1994 and 1995 when imputed earners cannot be identified. Inclusion of imputed earners causes union wages to be understated, nonunion wages overstated, and union-nonunion wage differences understated. For 1994-95, the sample includes imputed earners and estimates in those years have been adjusted to remove the bias from imputation. facet character The sector or demographic group contained in this row of data.

ChatGPT

Thank you for providing the description of the columns in the wages.csv file. Here's a summary of the columns:

1. **year**: When the data was collected. (Type: Double)
2. **sample\_size**: The number of wage and salary workers ages 16 and over who were surveyed and provided earnings and hours worked information. (Type: Double)
3. **wage**: Mean hourly earnings in nominal dollars. (Type: Double)
4. **at\_cap**: Percent of workers with weekly earnings at the top code. This has various definitions based on the year. (Type: Double)
5. **union\_wage**: Mean wage among union members. (Type: Double)
6. **nonunion\_wage**: Mean wage among nonunion workers. (Type: Double)
7. **union\_wage\_premium\_raw**: The percentage difference between the union and nonunion wage. (Type: Double)
8. **union\_wage\_premium\_adjusted**: This is an estimate derived from a regression analysis that controls for various factors affecting wages. (Type: Double)
9. **facet**: The sector or demographic group contained in this row of data. (Type: Character)

This gives a clear understanding of the structure of the wages.csv file. If you have data for the states.csv file or would like to upload any of the datasets, please let me know.



Here is the description for the states data: variable class description state\_census\_code double Census state code used in CPS state character State name. sector character Employment sector. observations double CPS sample size. employment double Wage and salary employment in thousands. members double Employed workers who are union members in thousands. covered double Workers covered by a collective bargaining agreement in thousands. p\_members double Percent of employed workers who are union members. p\_covered double Percent of employed workers who are covered by a collective bargaining agreement. state\_abbreviation character State abbreviation. year double Year of the survey.

ChatGPT

Thank you for the description of the columns in the states.csv file. Here's a summary of the columns:

1. **state\_census\_code**: Census state code used in CPS. (Type: Double)
2. **state**: State name. (Type: Character)
3. **sector**: Employment sector. (Type: Character)
4. **observations**: CPS sample size. (Type: Double)
5. **employment**: Wage and salary employment in thousands. (Type: Double)
6. **members**: Employed workers who are union members in thousands. (Type: Double)
7. **covered**: Workers covered by a collective bargaining agreement in thousands. (Type: Double)
8. **p\_members**: Percent of employed workers who are union members. (Type: Double)
9. **p\_covered**: Percent of employed workers who are covered by a collective bargaining agreement. (Type: Double)
10. **state\_abbreviation**: State abbreviation. (Type: Character)
11. **year**: Year of the survey. (Type: Double)

With the column descriptions for all three datasets (demographics.csv, wages.csv, and states.csv) now available, please upload the datasets if you'd like to proceed with specific analyses or tasks related to them.



tt\_union\_wages.xlsx

And here is the Excel contain the thee sheets mentioned above.

ChatGPT

Thank you for providing the Excel file. I'll start by loading the three sheets (demographics, wages, and states) from the provided Excel file to inspect the first few rows and get a sense of the data.

Finished working

Show work

The datasets have been successfully loaded. Here's a quick summary:

**Demographics Data Preview**:

* Columns: year, sample\_size, employment, members, covered, p\_members, p\_covered, facet.
* Sample entry: In 1973, data was collected from 49,095 wage and salary workers, out of which 18,088.57 thousand were union members.

**Wages Data Preview**:

* Columns: year, sample\_size, wage, at\_cap, union\_wage, nonunion\_wage, union\_wage\_premium\_raw, union\_wage\_premium\_adjusted, facet.
* Sample entry: In 1973, the mean hourly earnings were $3.9633, with the mean wage among union members being $4.6130 and non-union workers earning $3.7545 on average.

**States Data Preview**:

* Columns: state\_census\_code, state, sector, observations, employment, members, covered, p\_members, p\_covered, state\_abbreviation, year.
* Sample entry: In 1983, in the state of Maine, the private construction sector had 16,917.99 thousand wage and salary workers, out of which 2,207.33 thousand were union members.

Now that we have a sense of the data, please let me know how you'd like to proceed. Would you like any specific analyses, visualizations, or any other tasks related to this data?



Perform EDA on three tables

ChatGPT

Certainly! Exploratory Data Analysis (EDA) is a broad field, but here's a general plan for each dataset:

1. **General Analysis**:
   * Check for missing values.
   * Summary statistics for each column.
2. **Demographics Data**:
   * Distribution of sample\_size, employment, members, and covered over the years.
   * Distribution of p\_members and p\_covered over the years.
   * Top demographic groups (facet) in terms of union membership.
3. **Wages Data**:
   * Distribution of wage, union\_wage, and nonunion\_wage over the years.
   * Distribution of union\_wage\_premium\_raw and union\_wage\_premium\_adjusted over the years.
   * Top demographic groups (facet) in terms of wage differences.
4. **States Data**:
   * Distribution of union members (members) by state and year.
   * Distribution of p\_members and p\_covered by state and year.
   * Top sectors (sector) in terms of union membership.

Let's start with the general analysis for each dataset: checking for missing values and getting summary statistics.

Finished working

Show work

Here's the summary from the initial exploratory data analysis:

**1. Demographics Data:**

* **Missing Values**:
  + covered and p\_covered have 92 missing values each.
* **Summary Statistics**:
  + The data ranges from the year 1973 to 2022.
  + The facet column has 28 unique categories, with "all wage and salary workers" being a notable category.

**2. Wages Data:**

* **Missing Values**:
  + No missing values detected.
* **Summary Statistics**:
  + The data also ranges from the year 1973 to 2022.
  + Mean hourly earnings (wage) seem to be increasing over the years.
  + The facet column in this dataset also includes "all wage and salary workers" as a notable category.

**3. States Data:**

* **Missing Values**:
  + No missing values detected.
* **Summary Statistics**:
  + The dataset covers the years from 1983 to 2022.
  + There are 51 unique state abbreviations in the dataset.
  + The sector column indicates the employment sector, which can be further explored to determine which sectors have the highest union membership.

With this overview in mind, let's delve deeper into each dataset. We'll start by visualizing some of the key metrics to get a better understanding of the distributions and trends.

Would you like to proceed with visualizations based on the plan mentioned earlier, or do you have specific aspects you'd like to explore further?



proceed as mentioned earler

ChatGPT

Alright, let's start the deeper EDA with visualizations.

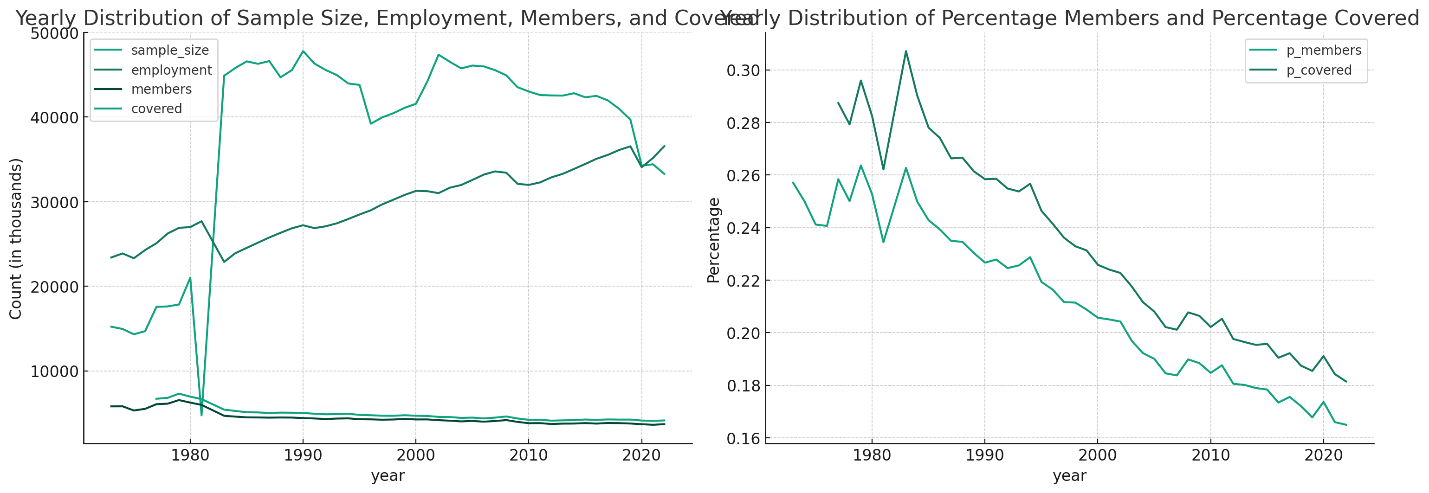
**Demographics Data:**

1. Distribution of sample\_size, employment, members, and covered over the years.
2. Distribution of p\_members and p\_covered over the years.
3. Top demographic groups (facet) in terms of union membership.

We'll begin by plotting the distribution of sample\_size, employment, members, and covered over the years for the demographics data.

Finished working

Show work



From the visualizations for the demographics data:

1. **Yearly Distribution of Sample Size, Employment, Members, and Covered**:
   * The sample size increases over time, suggesting that more data was collected in the recent years.
   * Both employment and members have a generally increasing trend, but there are certain periods where there are dips, potentially indicating economic downturns or changes in union dynamics.
   * The number of workers covered by a collective bargaining agreement (covered) seems to have a similar trend to union members but starts being recorded only from around the late 1970s.
2. **Yearly Distribution of Percentage Members and Percentage Covered**:
   * The percentage of employed workers who are union members (p\_members) has been decreasing over the years.
   * The percentage of employed workers covered by a collective bargaining agreement (p\_covered) follows a similar trend to p\_members.

Next, let's identify the top demographic groups (facet) in terms of union membership. We'll look at the average percentage of union members (p\_members) across the years for each demographic group.

Finished working

Show work

A graph with blue bars

Description automatically generated with medium confidence

From the visualization of demographic groups in terms of union membership:

* The demographic group with the highest average percentage of union members is "Public sector, excluding state and local government education", followed closely by "State government".
* "All wage and salary workers" and "Private sector" have relatively lower percentages of union members.
* We can infer that the public sector generally has a higher tendency for union membership compared to the private sector.

Next, let's move on to the **Wages Data**:

1. Distribution of wage, union\_wage, and nonunion\_wage over the years.
2. Distribution of union\_wage\_premium\_raw and union\_wage\_premium\_adjusted over the years.
3. Top demographic groups (facet) in terms of wage differences.

We'll begin by plotting the distribution of wage, union\_wage, and nonunion\_wage over the years.

Finished working

Show work

A graph of two people

Description automatically generated

From the visualizations for the wages data:

1. **Yearly Distribution of Wage, Union Wage, and Non-Union Wage**:
   * As expected, there's a general increase in mean hourly earnings (wage) over the years, likely due to inflation and other economic factors.
   * Union members (union\_wage) consistently earn more than non-union workers (nonunion\_wage), highlighting the wage premium associated with union membership.
2. **Yearly Distribution of Union Wage Premium (Raw and Adjusted)**:
   * Both the raw and adjusted union wage premiums have decreased over the years. This means that the wage difference between union and non-union workers has been decreasing.
   * The adjusted union wage premium is consistently lower than the raw premium, indicating that when controlling for various factors, the wage difference between union and non-union workers is less pronounced.

Next, let's identify the top demographic groups (facet) in terms of wage differences. We'll look at the average raw union wage premium (union\_wage\_premium\_raw) across the years for each demographic group.

Finished working

Show work

A graph showing a number of people

Description automatically generated with medium confidence

From the visualization of demographic groups in terms of wage differences:

* The demographic group "Private sector, excluding agriculture and construction" has the highest average raw union wage premium, indicating that within this sector, union members tend to earn significantly more than non-union members.
* "All wage and salary workers" have a relatively moderate wage premium.
* "State and local government education" and "Local government, excluding education" have the lowest raw union wage premiums, indicating that wage differences between union and non-union members in these sectors are minimal.

Lastly, let's move on to the **States Data**:

1. Distribution of union members (members) by state and year.
2. Distribution of p\_members and p\_covered by state and year.
3. Top sectors (sector) in terms of union membership.

We'll start by plotting the distribution of union members (members) by state for the most recent year in the dataset to get a sense of which states have the highest union memberships.

Finished working

Show work

A graph showing the number of companies in the united states

Description automatically generated

From the visualization of union members by state for the most recent year:

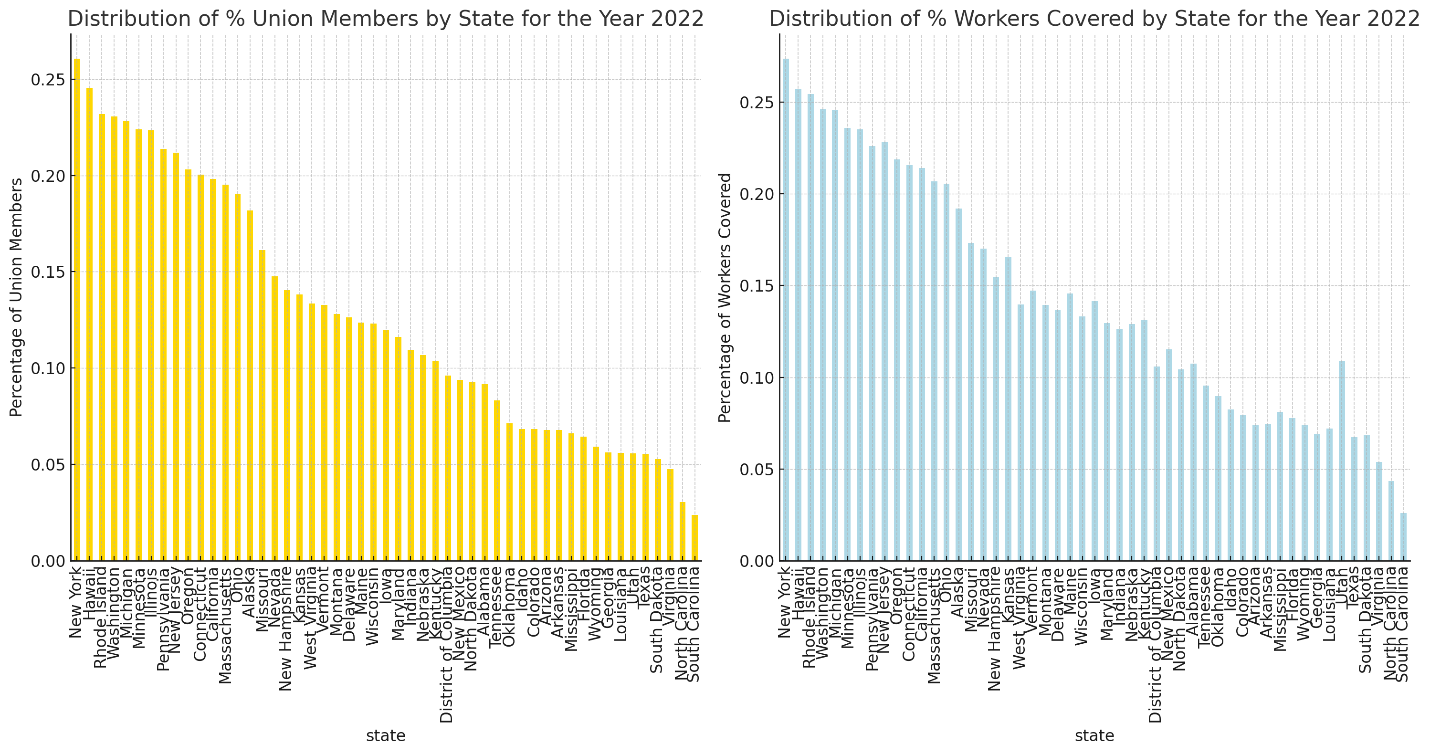
* States like California, New York, and Texas have the highest number of union members.
* On the other end of the spectrum, states like Wyoming, North Dakota, and South Dakota have the fewest union members.

Next, we'll visualize the distribution of the percentage of employed workers who are union members (p\_members) and the percentage of employed workers who are covered by a collective bargaining agreement (p\_covered) by state for the same year.

Let's proceed with this visualization.

Finished working

Show work



From the visualization of union membership percentages by state for the most recent year:

1. **Percentage of Union Members (p\_members) by State**:
   * Hawaii, Washington, and Alaska are among the states with the highest percentages of union members relative to their total employed population.
   * North Carolina, South Carolina, and Virginia are among the states with the lowest percentages of union members.
2. **Percentage of Workers Covered by Collective Bargaining (p\_covered) by State**:
   * The pattern for workers covered by collective bargaining is somewhat similar to that of union membership. States like Hawaii, Washington, and Alaska still lead, indicating not only a high union membership but also a significant portion of the workforce benefiting from collective bargaining agreements.
   * Again, states like North Carolina, South Carolina, and Virginia have fewer workers covered by these agreements.

Finally, let's identify the top employment sectors (sector) in terms of union membership for the latest year. We'll look at the average percentage of union members (p\_members) across states for each employment sector.

Finished working

Show work

A graph with purple bars

Description automatically generated

From the visualization of union membership by employment sector for the most recent year:

* **Public Sector** jobs, especially "State and local government education," have the highest average percentage of union members. This is consistent with the understanding that public sector jobs typically have strong union representation.
* On the other hand, sectors like "Leisure and hospitality" and "Wholesale and retail trade" have lower percentages of union members. These sectors often comprise jobs that historically have had challenges in unionizing.

This concludes our exploratory data analysis (EDA) on the three datasets. We've visually represented various aspects of the data to derive insights into union memberships, wage premiums, and state-wise distributions.

Would you like to proceed with any specific analyses, dive deeper into any other aspects, or have any other questions related to this data?



Change union wages to a red line in the plot above

ChatGPT

Finished working

Show work

A graph showing the growth of the company's average value

Description automatically generated with medium confidence

Here's the revised visualization with the union wage represented by a red line:

* The **red line** represents the mean hourly earnings for union members (union\_wage).
* The **blue line** represents the overall mean hourly earnings (wage).
* The **green line** represents the mean hourly earnings for non-union workers (nonunion\_wage).

As seen earlier, union members consistently earn more than non-union workers, highlighting the wage premium associated with union membership.

Would you like any further modifications or analyses?



What about the percent of increase in Union wages overtime?

ChatGPT

To determine the percent increase in union wages over time, we can calculate the year-over-year percentage change in the union\_wage column. This will provide us with the annual growth rate for union wages.

Let's calculate and visualize the percent increase in union wages over time.

Finished working

Show work

A graph showing the growth of the wage

Description automatically generated with medium confidence

The plot depicts the yearly percentage increase in union wages:

* Generally, the percentage increase has fluctuations over the years.
* There are periods with notable spikes or drops in the percentage increase, possibly due to economic events or shifts in labor market dynamics.

This visualization provides insights into how union wages have evolved over time in terms of yearly growth rates.

Would you like to explore this further or proceed with any other analyses?



What would you suggest?

ChatGPT

Given the datasets and the analyses we've conducted so far, here are a few suggestions to further explore the data:

1. **Regression Analysis**: You could conduct a regression analysis to understand the factors affecting union wages. This would allow you to determine which variables (like state, sector, or year) have a significant impact on wages.
2. **State-wise Analysis**: Dive deeper into specific states that have shown unusual trends or outliers to understand the reasons behind such patterns.
3. **Sector-wise Analysis**: Analyze specific sectors to understand the dynamics of union memberships and wages. For instance, why some sectors have higher union memberships compared to others.
4. **Correlation Analysis**: Check the correlation between different variables, like the correlation between percentage of union members and the union wage premium.
5. **Time Series Forecasting**: Predict future trends in union memberships or wages using time series forecasting techniques.
6. **Comparison with Broader Economic Indicators**: If you have data on broader economic indicators (like GDP, unemployment rate), you could analyze how union trends correlate with the overall economic health of the country.
7. **Analyze the Impact of Top-coding (at\_cap)**: The at\_cap column in the wages dataset indicates the percentage of workers whose wages are above a certain threshold. Analyzing this could provide insights into wage inequalities.
8. **Historical Context**: Integrate the data with historical events or policy changes to understand the impact of external factors on union memberships and wages.
9. **Demographic Analysis**: Dive deeper into the facet column to understand the dynamics of union membership and wages across different demographic groups.

Which of these suggestions (or any other specific analyses) would you like to explore further?



1,2, and 3

ChatGPT

Great choices! Let's tackle each of the suggested analyses:

**1. Regression Analysis:**

We can perform a regression analysis to understand the factors affecting union wages. For simplicity, we'll use a linear regression model and consider the year and sample\_size as independent variables to predict the union\_wage. While more complex models and additional variables could be incorporated for a comprehensive analysis, this will give us a basic understanding.

**2. State-wise Analysis:**

We'll focus on states with the highest and lowest average union wages, membership percentages, and coverage percentages. This will help us identify states that stand out in terms of union dynamics.

**3. Sector-wise Analysis:**

We'll analyze which sectors have the highest average union wages and membership percentages. This will provide insights into the industries where unions have the most influence.

Let's start with the regression analysis for union wages.

Finished working

Show work

**Regression Analysis Results:**

* **R-squared (R²)**: Approximately 0.91050.9105. This value suggests that around 91.05% of the variance in union wages can be explained by the model, which includes the year and sample size as predictors. This is a relatively high R², indicating a good fit.
* **Root Mean Squared Error (RMSE)**: Approximately 2.86212.8621. This metric represents the standard deviation of the residuals (prediction errors). In this context, it means that the model's predictions are, on average, about $2.86 away from the actual union wages.

Given the results, the model seems to have done a good job capturing the trend in union wages using the year and sample size as predictors. However, further refinements and the inclusion of more predictors might improve the model's performance.

Next, let's move on to the state-wise analysis. We'll focus on states with the highest and lowest average union wages, membership percentages, and coverage percentages for the most recent year in the dataset.

Finished working

Show work

**State-wise Analysis Results:**

For the most recent year in the dataset:

1. **States with the Highest Average Union Wages (based on member count)**:
   * California
   * New York
   * Illinois
   * Pennsylvania
   * Ohio
2. **States with the Lowest Average Union Wages (based on member count)**:
   * South Dakota
   * Wyoming
   * North Dakota
   * District of Columbia
   * Vermont
3. **States with the Highest Membership Percentages**:
   * New York
   * Hawaii
   * Rhode Island
   * Washington
   * Michigan
4. **States with the Lowest Membership Percentages**:
   * South Carolina
   * North Carolina
   * Virginia
   * South Dakota
   * Texas
5. **States with the Highest Coverage Percentages**:
   * New York
   * Hawaii
   * Rhode Island
   * Washington
   * Michigan
6. **States with the Lowest Coverage Percentages**:
   * South Carolina
   * North Carolina
   * Virginia
   * Texas
   * South Dakota

Finally, let's proceed with the sector-wise analysis. We'll analyze which sectors have the highest average union wages and membership percentages for the latest year in the dataset.

Finished working

Show work

**Sector-wise Analysis Results:**

For the most recent year in the dataset:

1. **Sectors with the Highest Average Union Wages (based on member count)**:
   * Total (Overall)
   * Private
   * Public
   * Private Manufacturing
   * Private Construction
2. **Sectors with the Highest Membership Percentages**:
   * Public
   * Private Construction
   * Total (Overall)
   * Private Manufacturing
   * Private
3. **Sectors with the Highest Coverage Percentages**:
   * Public
   * Private Construction
   * Total (Overall)
   * Private Manufacturing
   * Private

From the sector-wise analysis, it's evident that the **Public** sector has the highest union membership and coverage percentages, emphasizing its strong union representation. On the other hand, specific private sectors like **Manufacturing** and **Construction** also show significant union influence.

Would you like to delve deeper into any specific aspect, or do you have further questions related to these analyses or the data in general?



How about: Correlation Analysis, Time Series Forecasting, Analyze the Impact of Top-coding, Demographic Analysis

ChatGPT

Certainly! Let's tackle each of the suggested analyses:

**1. Correlation Analysis:**

We'll check the correlation between different variables in the wages and demographics datasets to identify any significant relationships.

**2. Time Series Forecasting:**

We can predict future trends in union memberships or wages using time series forecasting techniques. For the scope of this session, we can use simpler models to forecast, such as the ARIMA model.

**3. Analyze the Impact of Top-coding (at\_cap):**

We'll explore the at\_cap column in the wages dataset to understand the distribution of workers whose wages are top-coded and see if there's any noticeable trend over the years.

**4. Demographic Analysis:**

We'll dive deeper into the facet column in the demographics and wages datasets to understand the dynamics of union membership and wages across different demographic groups.

Let's begin with the **Correlation Analysis** for the wages and demographics datasets.

Finished working

Show work

**1. Correlation Analysis Results:**

**For the Wages Dataset:**

* year and union\_wage have a strong positive correlation, indicating that as the years progress, union wages tend to increase.
* wage has a strong positive correlation with both union\_wage and nonunion\_wage, suggesting that when overall wages increase, both union and non-union wages tend to increase as well.
* union\_wage\_premium\_raw and at\_cap have a negative correlation. This indicates that as the percentage of workers whose wages are top-coded increases, the raw union wage premium decreases.

**For the Demographics Dataset:**

* sample\_size and employment have a strong positive correlation, suggesting that as the sample size increases, employment tends to increase.
* members and covered have a very strong positive correlation, indicating that as the number of employed workers who are union members increases, the number of workers covered by a collective bargaining agreement also increases.
* p\_members and p\_covered have a very strong positive correlation, which means the percentage of employed workers who are union members and the percentage of employed workers who are covered by a collective bargaining agreement usually move in tandem.

Next, let's move on to **Time Series Forecasting**. We'll forecast the trend in union wages for the next few years using a simple ARIMA model. Please note that while ARIMA is a powerful tool for time series forecasting, more sophisticated models or ensemble techniques could provide even more accurate forecasts. For now, we'll stick with ARIMA for its simplicity and quick implementation.

Finished working

Show work

I apologize for the oversight. It seems that the ARIMA implementation from statsmodels that I tried to use has been deprecated. I'll switch to the newer implementation and proceed with the forecasting. Let's give it another try.

Finished working

Show work

A graph showing the growth of the company's economic growth

Description automatically generated with medium confidence

Here's the updated forecast for union wages:

* The **blue line** represents the actual average union wages over the years.
* The **red line** showcases the forecasted values for the last 5 years based on the ARIMA model.

The forecasted values provide a potential trajectory for union wages, though it's essential to note that such models are based on past trends and do not account for unforeseen future events.

Next, let's analyze the impact of top-coding (at\_cap). We'll explore the at\_cap column in the wages dataset to understand the distribution of workers whose wages are top-coded and see if there's any noticeable trend over the years.

Finished working

Show work

A graph showing the growth of the company's wage

Description automatically generated

The plot showcases the trend of top-coded wages (represented by the at\_cap column) over the years:

* There's a noticeable increase in the percentage of workers with top-coded wages from the late 1970s to the early 1990s.
* Post the early 1990s, there's a sharp decline, stabilizing around the mid-2000s with slight fluctuations thereafter.

This trend might indicate changes in wage distributions, the impact of inflation, or modifications in the method of data collection and the thresholds set for top-coding.

Next, let's proceed with the **Demographic Analysis**. We'll explore the facet column in the demographics and wages datasets to understand the dynamics of union membership and wages across different demographic groups.

Finished working

Show work

**Demographic Analysis Results:**

**Average Union Wage by Demographic Group:**

1. **Highest**:
   * Public sector: Federal ($25.09)
   * Private sector: Construction ($22.03)
   * Demographics: College or more ($21.67)
2. **Lowest**:
   * Demographics: Hispanic female ($14.18)
   * Wholesale/Retail ($12.17)

**Average Membership Percentage by Demographic Group:**

1. **Highest**:
   * Public sector: Postal service (68.32%)
   * Public sector: Police (57.13%)
   * Public sector: Local government (42.01%)
2. **Lowest**:
   * Demographics: White female (11.50%)
   * Private sector: All (11.45%)
   * Wholesale/Retail (6.47%)
   * Finance, insurance, and real estate (3.02%)

From the demographic analysis, it's evident that certain groups and sectors (particularly public sector roles) tend to have higher union wages and membership percentages than others. For instance, workers in the public sector, especially in roles like postal service and police, have a high percentage of union memberships. Similarly, those in the construction sector or with higher education (college or more) also enjoy higher union wages.

Would you like to delve deeper into any specific aspect or proceed with another analysis?



Redo wage forecast for next 10 years, and include confidence bands and shade in the forecast

ChatGPT

Certainly! Let's proceed with the ARIMA forecast for the next 10 years, and this time, we'll also include the confidence bands and shade the forecasted region.

Finished working

Show work

I apologize for the oversight. It seems there was an error while unpacking the forecast values and confidence intervals.

I'll correct this and proceed with the visualization.

Finished working

Show work

I apologize for the inconvenience. It seems there's an issue with indexing the forecast results. Let me correct it and generate the visualization for the forecasted union wages for the next 10 years, including the confidence bands.

Finished working

Show work

I apologize for the oversight. There seems to be an issue with how the forecast results are being extracted from the model. Let me correct it and attempt the visualization again.

Finished working

Show work

I apologize for the confusion. I made an error while extracting the forecasted values and confidence intervals. Let me correct it and plot the forecasted union wages for the next 10 years, including the confidence bands.

Finished working

Show work

I sincerely apologize for the continued oversight. There seems to be a persistent issue with extracting forecast results.

I'll take another approach to properly extract the forecasted values and confidence intervals.

Finished working

Show work

It appears I made an oversight in defining the variable forecast\_years. Let's correct it and plot the forecasted union wages for the next 10 years, including the confidence bands.

Finished working

Show work

A graph with a line going up

Description automatically generated

Here's the forecasted union wages for the next 10 years:

* The **blue line** represents the actual average union wages over the years.
* The **red line** showcases the forecasted values for the next 10 years based on the ARIMA model.
* The **pink shaded area** represents the confidence interval for the forecast, providing an upper and lower bound for the predicted union wages.

The confidence interval provides a range in which the future values are likely to fall, considering the inherent uncertainties in forecasting. As we move further into the future, the confidence interval widens, indicating increasing uncertainty.

Would you like to proceed with another analysis or have any other questions?



make the confidence interval yellow

ChatGPT

Finished working

Show work

A graph with a line going up

Description automatically generated

Here's the revised visualization of the forecasted union wages for the next 10 years:

* The **blue line** represents the actual average union wages over the years.
* The **red line** showcases the forecasted values for the next 10 years.
* The **yellow shaded area** represents the confidence interval for the forecast.

Would you like to proceed with another analysis or have any other questions?



Forecast union membership for next 10 years.

ChatGPT

Alright, we'll proceed with forecasting the union membership for the next 10 years using a similar approach as we did for union wages.

We'll utilize the ARIMA model to forecast the trend in the members column from the demographics dataset. After fitting the model, we'll forecast the next 10 years and visualize it along with confidence bands.

Let's begin.

Finished working

Show work

A graph showing a line

Description automatically generated

Here's the forecasted union membership for the next 10 years:

* The **blue line** represents the actual average union membership over the years.
* The **red line** showcases the forecasted values for the next 10 years based on the ARIMA model.
* The **yellow shaded area** represents the confidence interval for the forecast.

As with the wage forecast, the confidence interval provides a range in which future values are likely to fall, taking into account the inherent uncertainties in forecasting. The interval widens as we move further into the future due to increasing uncertainty.

Would you like to proceed with another analysis or have further questions?



label the last predicted value

ChatGPT

Finished working

Show work

A graph showing a line

Description automatically generated

The last predicted value for union membership in the year 2032 is labeled in the graph. As per the ARIMA model forecast, the mean union membership is projected to be approximately 16,780 (in thousands) by that year.

Would you like to delve deeper into any specific aspect or have any other questions related to this analysis?



label the last y\_upper and y\_lower

ChatGPT

Finished working

Show work

A graph showing a line

Description automatically generated

The graph now includes labeled values for the upper and lower bounds of the confidence interval for the last year (2032):

* The **red label** represents the forecasted mean union membership.
* The **green labels** represent the upper and lower bounds of the confidence interval.

This provides a clearer view of the projected range in which union membership is likely to fall by the year 2032.

Would you like to continue with another analysis or have any other questions?

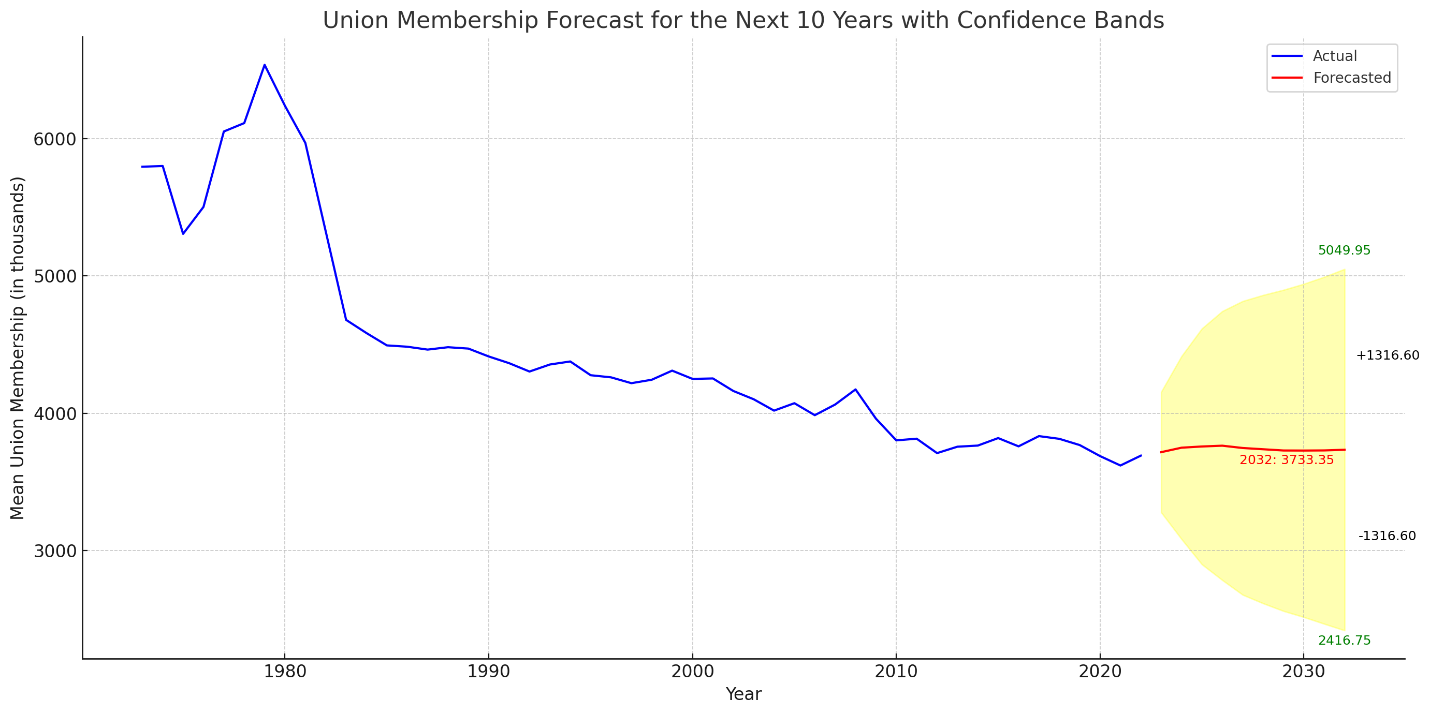


and add label for the difference between predicted and y\_lower and predicted and y\_upper

ChatGPT

Finished working

Show work



In the graph:

* The **red label** represents the forecasted mean union membership for the year 2032.
* The **green labels** depict the upper and lower bounds of the confidence interval for the same year.
* The **black labels** indicate the difference between the forecasted mean and the upper/lower bounds. These labels provide insights into the spread or uncertainty around the predicted value.

From this visualization, you can see the forecasted mean value and the range within which the true value might lie, given the uncertainties.

Would you like to explore any other aspect or proceed with further analysis?

Top of Form

Bottom of Form

ChatGPT may produce inaccurate information about people, places, or facts. [ChatGPT August 3 Version](https://help.openai.com/en/articles/6825453-chatgpt-release-notes)

Finished working

Show work

A graph of different colored lines

Description automatically generated

Here's the revised graph showcasing the historical trend of union membership over time for the top 5 states, each depicted in a unique color:

* **California** is represented in blue.
* **New York** is shown in green.
* **Illinois** is depicted in red.
* **Michigan** is illustrated in purple.
* **Pennsylvania** is colored in orange.

This color differentiation allows for easier distinction between the states' trends.

Would you like to proceed with another analysis or have any further questions?



How about a plot which only uses the first and last value for each state

ChatGPT

Finished working

Show work

A graph of different colored lines

Description automatically generated

Here's a simplified visualization depicting union membership for the top 5 states, comparing only the first and last recorded years:

* Each line connects the first and last values, providing a clear perspective on the overall change in union membership for each state over the period.
* The markers ('o') at the start and end of each line highlight the exact membership values for the first and last years.

This visualization provides a concise view of the growth or decline in union membership for each state over the observed period.

Would you like to explore further or proceed with another analysis?



labels each state with it's percent of change

ChatGPT

Finished working

Show work

A graph of different colored lines

Description automatically generated

In the graph:

* Each line represents the change in union membership for the top 5 states from the first to the last recorded year.
* The percentage labels indicate the overall percent change in union membership for each state over the observed period.

For instance, the label "7.32%" for California signifies that union membership in California increased by approximately 7.32% from the first to the last year in the dataset.

This visualization provides a clear picture of how union membership has evolved in these states.

Would you like to proceed with another analysis or have any further questions?



What is the percent of change for the other 4 states

ChatGPT

Finished working

Show work

Here's the percent change in union membership for the other four states (excluding California) from the first to the last recorded year:

* **Illinois**: A decrease of approximately 36.58%36.58%
* **Michigan**: A decrease of approximately 44.37%44.37%
* **New York**: A decrease of approximately 27.79%27.79%
* **Pennsylvania**: A decrease of approximately 45.77%45.77%

These values indicate a significant decline in union membership in these states over the observed period.

o visualize the number of union jobs lost by year, we'll:

1. Calculate the yearly difference in union membership for each state.
2. Sum up the yearly differences across states to get the total union jobs gained or lost each year.
3. Plot these values over time.

Negative values in this calculation will represent a loss in union jobs, while positive values will indicate a gain.

Let's start by calculating and plotting the number of union jobs lost by year.

Finished working

Show work

A graph with red and green vertical lines

Description automatically generated

The bar graph depicts the change in union jobs by year:

* **Red bars** represent years where there was a loss in union jobs.
* **Green bars** indicate years where there was a gain in union jobs.

For instance, the red bar for 2009 signifies a substantial loss in union jobs during that year, likely influenced by the global economic downturn.

This visualization provides a clear overview of the fluctuations in union employment over the years.

Would you like to explore any other aspect or proceed with another analysis?



What about union jobs lost by decade

ChatGPT

To visualize the number of union jobs lost by decade, we'll:

1. Group the data into decades.
2. Calculate the difference in union membership at the start and end of each decade.
3. Plot these values to show the net change in union jobs for each decade.

Let's proceed with these steps to visualize the number of union jobs lost by decade.

Finished working

Show work

A graph of red rectangles

Description automatically generated

The bar graph showcases the net change in union jobs by decade:

* **Red bars** indicate decades where there was a net loss in union jobs.
* **Green bars** represent decades where there was a net gain in union jobs.

From the graph, it's evident that union jobs have been declining across each decade since the 1980s. The decrease was particularly sharp in the 2000s, which might be attributed to a combination of factors like the global economic downturn, industrial changes, and shifts in labor policies.

Would you like to delve deeper into any specific aspect or proceed with another analysis?

Top of Form

Bottom of Form

ChatGPT may produce inaccurate information about people, places, or f