Enterprise Cloud Computing and Big Data (BUDT737)

Project Title: Analyzing and Predicting Wages in California

Team Members:

Kaveri Ramyakoumudi
Anusha Aligeti
Jayanth Sankar

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	Signature
	Kaveri Ramyakoumudi	**Ramyakoumudi Kaveri**
	Anusha Aligeti	**Anusha Aligeti**
	Jayanth Sankar	**Jayanth Sankar**

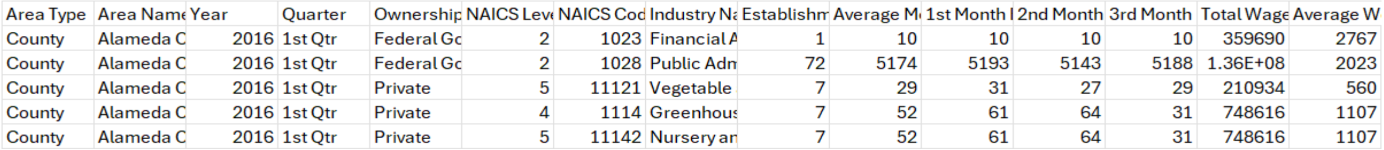
Executive Summary

Our research focused on analyzing the average weekly wages in California, aiming to uncover patterns, trends, and factors influencing wage variations. The significance of this study lies in providing valuable insights for businesses, policymakers, and stakeholders in understanding the dynamics of the state's labor market. We delved into the seasonal and quarterly patterns of wages, discovering that the 4th quarter consistently had the highest average weekly wages, while the 2nd quarter marked the lowest.

One key finding was the pronounced impact of ownership structures on wages, with private ownership consistently offering higher wages compared to various government levels. This highlights the influential role of the private sector in shaping both wage levels and employment stability. Additionally, our analysis unveiled significant regional disparities in wages across different areas in California. Our machine learning predictions identified Linear Regression as the most effective model for forecasting future weekly wages, providing a practical tool for anticipating wage trends. The clustering analysis revealed three distinct wage clusters, offering a nuanced understanding of wage distribution in the state. Overall, this research contributes valuable insights for informed decision-making, helping businesses and policymakers navigate the complexities of California's labor market.

Data Description

- Data source - DATA.GOV (Quarterly Census of Employment and Wages (QCEW))
- Link:https://catalog.data.gov/dataset/quarterly-census-of-employment-and-wages-qcew-a6fea/resource/515ceff0-009e-443d-8a29-976fc4d58c4a?inner_span=True
- By each variable we are measuring weekly wages of a individual row which measures in units US Dollars, it is treated as numerical variable.
- Sample size (n) is 1009955 and number of variables(k) are 18
- A small sample of observations from the data (a few observations to demonstrate the above points).



Area Type	Area Name	Year	Quarter	Ownership	NAICS Leve	NAICS Cod	Industry Na	Establishm	Average M	1st Month I	2nd Month	3rd Month	Total Wage	Average W
County	Alameda C	2016	1st Qtr	Federal Gc	2	1023	Financial A	1	10	10	10	10	359690	2767
County	Alameda C	2016	1st Qtr	Federal Gc	2	1028	Public Adn	72	5174	5193	5143	5188	1.36E+08	2023
County	Alameda C	2016	1st Qtr	Private	5	11121	Vegetable	7	29	31	27	29	210934	560
County	Alameda C	2016	1st Qtr	Private	4	1114	Greenhous	7	52	61	64	31	748616	1107
County	Alameda C	2016	1st Qtr	Private	5	11142	Nursery ar	7	52	61	64	31	748616	1107

The dataset is of interest cause:

- Looking at past wage information helps predict what the job market might be like in the future. This helps businesses, government officials, and investors make smart choices based on data. By knowing what wages have been like, they can guess what might happen next, find new chances to succeed, or spot possible problems ahead of time. This helps them use their money and resources wisely.
- Information about wages helps government leaders decide on things like minimum wage rules, how to regulate the job market, taxes, and programs to help people in need. With this info, policymakers can create rules to ensure fair pay, lower poverty rates, and make sure everyone has a chance to succeed in the economy.

Research Questions

Questions that were investigated using this data

1. How do average weekly wages vary across different industries? Which industries offer higher wages?
2. Do average weekly wages fluctuate seasonally or based on quarterly trends?
3. Are there particular quarters or times of the year when wages tend to be higher or lower?
4. How do wages differ between different types of ownership (e.g., private, public, non-profit)?
5. Are there noticeable differences in wage levels and employment stability based on ownership structure?
6. Are there significant variations in wages between different regions or areas?
7. How do different features such as industries, ownership types, regions, and quarters contribute to the variation in average weekly wages?
8. Can we build a reliable predictive model to estimate average weekly wages based on these features?
9. Which machine learning algorithms perform best for this task, and how can we enhance the accuracy of the best-performing model?
10. How do industries cluster based on average weekly wages, establishments, and employment?
11. Can we identify meaningful segments or clusters of industries based on these variables?
12. How can we utilize graph-based algorithms or network analysis techniques to uncover patterns or associations within the data?
13. What is the importance of different industries, ownership types, or regions in the context of average weekly wages?
14. How can we quantify the influence or centrality of various nodes (representing industries, ownership types, or regions) using PageRank?
15. What does the triangle count represent in terms of connections between industries, ownership types, or regions?

Methodology

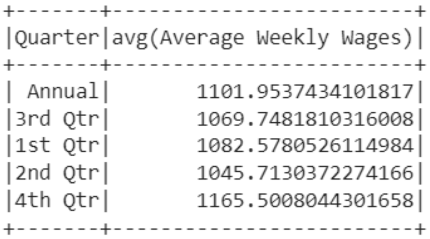
What techniques did you use and why?

- The first step involved EDA, descriptive analysis, grouping, and aggregation to understand the wages data structure. Visualizations were used for trend interpretation, while statistical methods could enhance analysis depth.
- The methodology for enhancing the dataset involves leveraging user-defined functions (UDFs) to compute additional columns such as employment size and ownership category. Initially, UDFs are created to calculate the employment size and categorize ownership based on predefined criteria. These functions are then applied to the dataset, resulting in the creation of new columns that capture employment size and ownership information. Furthermore, another UDF is developed to compute the increase in employment between consecutive months. This function is applied to the dataset, generating a new column representing the change in employment over time. This approach enriches the dataset with valuable information that can provide deeper insights into employment dynamics, without explicitly involving regression models.

- The methodology comprises several stages for constructing regression models, employing techniques such as string indexing, one-hot encoding, and vector assembler. Initially, the dataset undergoes preprocessing, addressing missing values and converting categorical variables into numerical representations using string indexing. These numerical representations are further transformed into binary vectors through one-hot encoding. Subsequently, the features, encompassing both encoded categorical variables and numerical attributes, are consolidated into a unified vector using vector assembler. The processed data is then partitioned into training and testing sets. Regression models, including linear regression, random forest regression, and decision tree regression, are trained on the designated training data. After training, the models undergo evaluation using regression metrics such as R2 and Mean Squared Error (RMSE). The model exhibiting superior performance, as determined by evaluation metrics, is chosen for deployment or additional analysis. This structured methodology ensures an organized approach to constructing and selecting regression models, utilizing techniques like string indexing, one-hot encoding, and vector assembler to streamline feature engineering.
- Clustering involved KMeans, vector assembler, and pipelines. It was created based on the Average Weekly Wages - High, Medium and Low Wage Industries.
- The methodology involves creating graphs with nodes representing industry names, area types, and ownership categories. This is achieved by utilizing a subset of the dataset and defining relations between these nodes. Each node represents a unique entity within its respective category.

Results and Findings

The analysis indicates that there are indeed particular quarters or times of the year when wages tend to be higher or lower. Specifically, the 4th quarter stands out as having the highest average weekly wages, with an average of 1165.50. Conversely, the 2nd quarter has the lowest average weekly wages, with an average of $1045.71. This suggests a seasonal pattern where wages peak towards the end of the year and dip in the middle of the year. Further exploration into the factors driving these seasonal variations could provide valuable insights for understanding wage dynamics and informing workforce management strategies.



```
+-------+-------------------------+
|Quarter|avg(Average Weekly Wages)|
+-------+-------------------------+
| Annual|       1101.9537434101817|
|3rd Qtr|       1069.7481810316008|
|1st Qtr|       1082.5780526114984|
|2nd Qtr|       1045.7130372274166|
|4th Qtr|       1165.5008044301658|
+-------+-------------------------+
```

The analysis reveals notable differences in wages between different types of ownership. Here's a breakdown of the average total wages across various ownership categories:

Total U.I. Covered: 3.18 trillion

Total Government: 278.46 billion

Total Covered: 69.69 billion

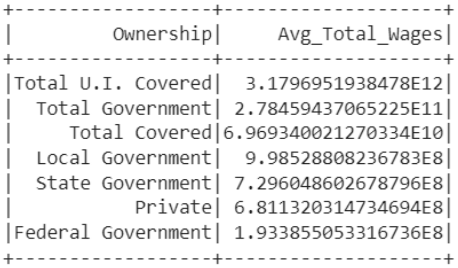
Local Government: 998.53 million

State Government: 729.60 million

Private: 681.13 million

Federal Government: 193.39 million

These figures indicate significant variations in average total wages across different ownership categories. Particularly, the private sector appears to have the highest average total wages, followed by various levels of government (local, state, and federal). Understanding these wage disparities can provide insights into the economic landscape, labor market dynamics, and the distribution of resources across different sectors of the economy. Further analysis may be necessary to explore the factors contributing to these differences and their implications for various stakeholders.



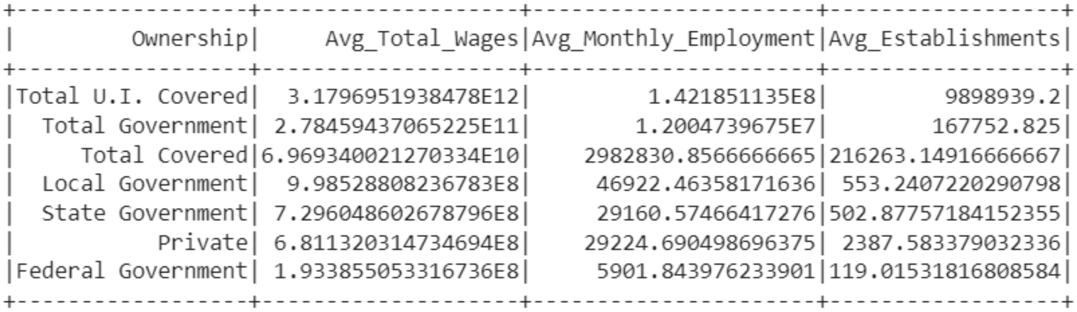
```
+------------------+--------------------+
|         Ownership|     Avg_Total_Wages|
+------------------+--------------------+
|Total U.I. Covered|  3.1796951938478E12|
|  Total Government| 2.78459437065225E11|
|     Total Covered|6.969340021270334E10|
|  Local Government|  9.98528808236783E8|
|  State Government| 7.296048602678796E8|
|           Private| 6.811320314734694E8|
|Federal Government| 1.933855053316736E8|
+------------------+--------------------+
```

Wage Levels:

Total U.I. Covered has the highest average total wages, followed by Total Government and Total Covered. Private ownership also shows relatively high average total wages compared to Federal Government, State Government, and Local Government.

Employment Stability:

Total U.I. Covered has the highest average monthly employment, indicating potentially more stable employment within this ownership category. Private ownership also shows a relatively high average monthly employment level compared to Federal Government, State Government, and Local Government. Overall, ownership structure appears to influence both wage levels and employment stability, with certain categories such as Total U.I. Covered showing higher values in both metrics compared to others. Further analysis may be required to understand the underlying factors contributing to these differences.



```
+------------------+--------------------+----------------------+------------------+
|         Ownership|     Avg_Total_Wages|Avg_Monthly_Employment|Avg_Establishments|
+------------------+--------------------+----------------------+------------------+
|Total U.I. Covered|  3.1796951938478E12|         1.421851135E8|         9898939.2|
|  Total Government| 2.78459437065225E11|        1.2004739675E7|        167752.825|
|     Total Covered|6.969340021270334E10|    2982830.8566666665|216263.14916666667|
|  Local Government|  9.98528808236783E8|     46922.46358171636| 553.2407220290798|
|  State Government| 7.296048602678796E8|     29160.57466417276|502.87757184152355|
|           Private| 6.811320314734694E8|    29224.690498696375| 2387.583379032336|
|Federal Government| 1.933855053316736E8|     5901.843976233901|119.01531816808584|
+------------------+--------------------+----------------------+------------------+
```

There are significant variations in wages between different regions or areas:

Highest Average Wages:

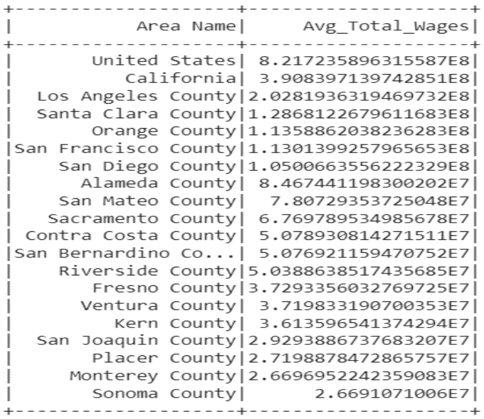
The United States has the highest average total wages, followed by California. Los Angeles County, Santa Clara County, and Orange County also have relatively high average total wages.

Variations Across Counties:

There are noticeable variations in average wages across different counties within California. For example, San Francisco County, San Diego County, and Alameda County have relatively high average total wages, while other counties like Fresno County and Kern County have lower average total wages.

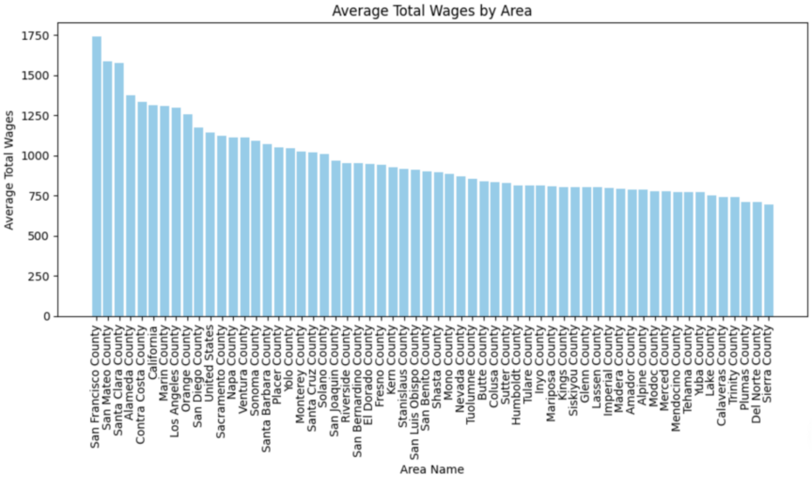
Regional Disparities:

There are disparities in wages between regions, with some areas experiencing higher average wages compared to others. This indicates that wage levels may vary significantly depending on the geographical location or area. Overall, the analysis highlights significant variations in wages between different regions or areas, suggesting the presence of regional economic disparities. Further investigation may be required to understand the underlying factors contributing to these variations.

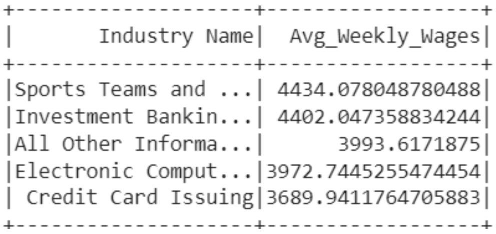


```
+--------------------+--------------------+
|           Area Name|     Avg_Total_Wages|
+--------------------+--------------------+
|       United States| 8.217235896315587E8|
|          California| 3.908397139742851E8|
|  Los Angeles County|2.0281936319469732E8|
|  Santa Clara County|1.2868122679611683E8|
|       Orange County|1.1358862038236283E8|
|San Francisco County|1.1301399257965653E8|
|    San Diego County|1.0500663556222329E8|
|      Alameda County| 8.467441198300202E7|
|    San Mateo County|  7.80729353725048E7|
|   Sacramento County| 6.769789534985678E7|
| Contra Costa County| 5.078930814271511E7|
|San Bernardino Co...| 5.076921159470752E7|
|    Riverside County|5.0388638517435685E7|
|       Fresno County|3.7293356032769725E7|
|      Ventura County| 3.719833190700353E7|
|         Kern County| 3.613596541374294E7|
|  San Joaquin County|2.9293886737683207E7|
|       Placer County|2.7198878472865757E7|
|     Monterey County|2.6696952242359083E7|
|       Sonoma County|       2.6691071006E7|
+--------------------+--------------------+
```

The table shows the average weekly wages by area, with San Francisco County having the highest average weekly wages at approximately $1740.88. Following closely are San Mateo County and Santa Clara County with average weekly wages of $1588.34 and $1577.82, respectively. Alameda County and Contra Costa County also have relatively high average weekly wages at $1376.00 and $1333.89, respectively.



The average weekly wages for each Industry ordered descending:



```
+--------------------+------------------+
|       Industry Name|  Avg_Weekly_Wages|
+--------------------+------------------+
|Sports Teams and ...| 4434.078048780488|
|Investment Bankin...| 4402.047358834244|
|All Other Informa...|      3993.6171875|
|Electronic Comput...|3972.7445255474454|
| Credit Card Issuing|3689.9411764705883|
+--------------------+------------------+
```

Three models were used on the data to end up with the best model to predict future weekly wages - Linear Regression, Decision Trees Regression, and Random Forest Regression.

Machine Learning:

Feature engineering was performed before ML models, EDA was done and new features were created from existing columns.

The input columns used for the ML models are - "Area Type","Area Name","Year","Quarter", "Industry Name", "NAICS Level","NAICS Code","Ownership", "Employment_Size", "Total Wages (All Workers)","Average Monthly Employment","Establishments" and the variable of interest is “Average Weekly Wages”.

The categorical variables were - "Area Type","Area Name","Year","Quarter", "Industry Name", "NAICS Level","NAICS Code","Ownership", "Employment_Size". String Indexing, OneHotEncoding, and Vector Assembler were used prior to the pipeline where models were implemented.

8. The performance metrics we used for the selection of best model among the three models are R^2 value and RMSE value.

For Logistic Regression:

```
R-squared: 0.5312309442025673

rmse: 502.16601954940745
```

For Decision Tree Regression:

```
RMSE:655.1991419414471
R- Squared:0.20198522511352224
```

For Random Forest Regression:

```
RMSE:641.892902637405
R- Squared:0.23406936621093455
```

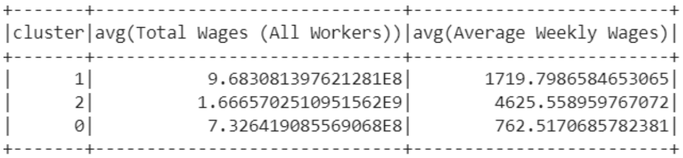
Based on the above performance metrics, linear regression was chosen as the best model and made predictions on the test data.

```
+------------------+-------------------+

|   predictions    |      2016-2019     |

+------------------+-------------------+
|1471.3863452215053|             1494.0|

|1828.6681476789154|             2696.0|

|1104.4341837962265|                0.0|

|1509.2919953036153|             2576.0|

|1812.9632145247847|             1629.0|

+------------------+-------------------+
```

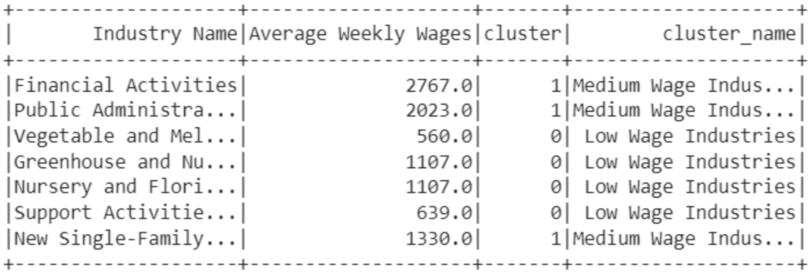
The average predicted change in wages is: -1.19%

Clustering:

The data has been clustered based on the Average Weekly Wages and separated into 3 clusters(k=3) - High Paying, Medium Paying and Low Paying. The average weekly wages and total wages in the respective clusters are:



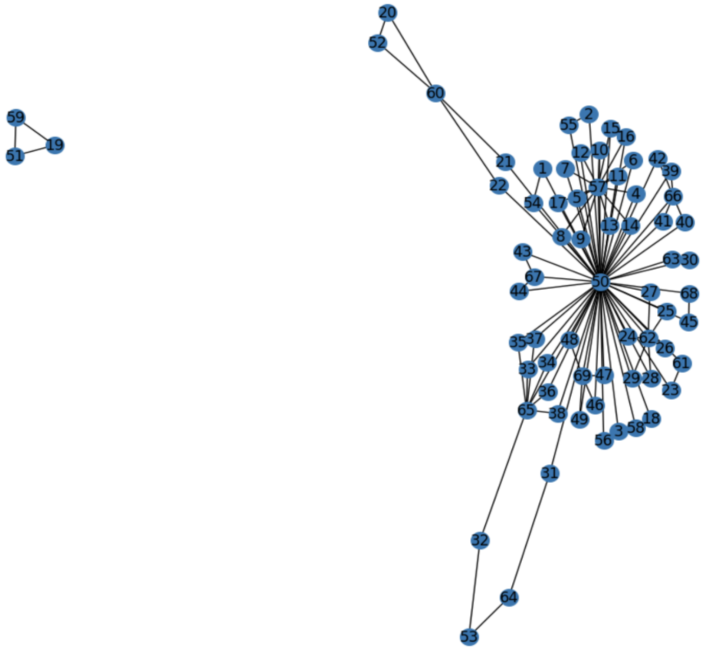
```
+-------+-----------------------------+------------------------+
|cluster|avg(Total Wages (All Workers))|avg(Average Weekly Wages)|
+-------+-----------------------------+------------------------+
|      1|          9.683081397621281E8|      1719.7986584653065|
|      2|         1.6665702510951562E9|       4625.558959767072|
|      0|          7.326419085569068E8|       762.5170685782381|
+-------+-----------------------------+------------------------+
```



```
+--------------------+--------------------+-------+--------------------+
|       Industry Name|Average Weekly Wages|cluster|        cluster_name|
+--------------------+--------------------+-------+--------------------+
|Financial Activities|              2767.0|      1|Medium Wage Indus...|
|Public Administra...|              2023.0|      1|Medium Wage Indus...|
|Vegetable and Mel...|               560.0|      0| Low Wage Industries|
|Greenhouse and Nu...|              1107.0|      0| Low Wage Industries|
|Nursery and Flori...|              1107.0|      0| Low Wage Industries|
|Support Activitie...|               639.0|      0| Low Wage Industries|
|New Single-Family...|              1330.0|      1|Medium Wage Indus...|
+--------------------+--------------------+-------+--------------------+
```

Graph DataFrame:

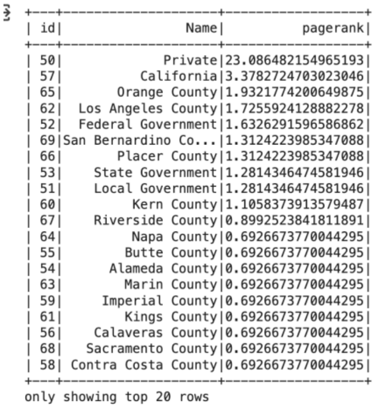
A subset of the main data was taken as there were over a million entries and graphs consume a lot of memory. The subset was taken in proportion to the Industry Names category.



The goal of the graph was to connect the Counties, Industry Names, and Ownership and analyze the relationships between these.

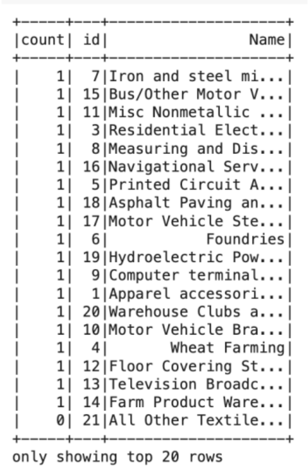
The result was that most of the industries were owned by Private players in California which was confirmed using the Page Rank method.

Page Rank:



```
+---+--------------------+------------------+
| id|                Name|          pagerank|
+---+--------------------+------------------+
| 50|             Private|23.086482154965193|
| 57|          California|3.3782724703023046|
| 65|       Orange County|1.9321774200649875|
| 62|  Los Angeles County|1.7255924128882278|
| 52|  Federal Government|1.6326291596586862|
| 69|San Bernardino Co...|1.3124223985347088|
| 66|       Placer County|1.3124223985347088|
| 53|    State Government|1.2814346474581946|
| 51|    Local Government|1.2814346474581946|
| 60|         Kern County|1.1058373913579487|
| 67|    Riverside County|0.8992523841811891|
| 64|         Napa County|0.6926673770044295|
| 55|        Butte County|0.6926673770044295|
| 54|      Alameda County|0.6926673770044295|
| 63|        Marin County|0.6926673770044295|
| 59|     Imperial County|0.6926673770044295|
| 61|        Kings County|0.6926673770044295|
| 56|    Calaveras County|0.6926673770044295|
| 68|   Sacramento County|0.6926673770044295|
| 58| Contra Costa County|0.6926673770044295|
+---+--------------------+------------------+
only showing top 20 rows
```

Triangle Count:



```
+-----+---+--------------------+
|count| id|                Name|
+-----+---+--------------------+
|    1|  7|Iron and steel mi...|
|    1| 15|Bus/Other Motor V...|
|    1| 11|Misc Nonmetallic ...|
|    1|  3|Residential Elect...|
|    1|  8|Measuring and Dis...|
|    1| 16|Navigational Serv...|
|    1|  5|Printed Circuit A...|
|    1| 18|Asphalt Paving an...|
|    1| 17|Motor Vehicle Ste...|
|    1|  6|           Foundries|
|    1| 19|Hydroelectric Pow...|
|    1|  9|Computer terminal...|
|    1|  1|Apparel accessori...|
|    1| 20|Warehouse Clubs a...|
|    1| 10|Motor Vehicle Bra...|
|    1|  4|       Wheat Farming|
|    1| 12|Floor Covering St...|
|    1| 13|Television Broadc...|
|    1| 14|Farm Product Ware...|
|    0| 21|All Other Textile...|
+-----+---+--------------------+
only showing top 20 rows
```

Conclusion

The financial analysis of average weekly wages in California yielded several key findings, shedding light on critical aspects of the state's labor market. Seasonal trends emerged, with the 4th quarter showcasing the highest average weekly wages and the 2nd quarter experiencing the lowest, offering valuable insights for strategic decision-making among businesses and policymakers. The influence of ownership structures on wages was pronounced, highlighting that private ownership consistently provides higher average total wages compared to various government levels. Notably, the private sector played a pivotal role in shaping both wage levels and employment stability.

Regional disparities in wages were evident across diverse regions and counties within California, with areas such as San Francisco, San Mateo, and Santa Clara exhibiting notably higher average weekly wages. Machine learning predictions underscored the effectiveness of Linear Regression as the best-performing model for forecasting future weekly wages, emphasizing the importance of feature engineering and categorical encoding in enhancing model performance. Clustering analysis revealed three distinct clusters based on average weekly wages—High Paying, Medium Paying, and Low Paying—offering a nuanced understanding of wage distribution in the state.

Graph-based insights further illuminated the economic landscape, showcasing the prevalence of private ownership across various industries in California. PageRank analysis confirmed the central role of private players in driving the state's economic dynamics. As we move forward, continuous monitoring of wage trends and further exploration into the factors influencing regional and ownership-based variations are recommended. The valuable insights derived from this study can inform policies related to minimum wage regulations, economic development initiatives, and strategic workforce planning, contributing to a more equitable and informed decision-making process in California.