California Fiscal Health in 2019

Class: Stats 140XP

Authors: Yechen Cao, Luning Ding, Xianya Fu, Yuexuan Wu, Lingxi Zhang, Shiwei Chen

1. Abstract

1.1 Background

Fiscal health refers to the financial stability and sustainability of a government or organization, which are California's local cities in this study. It is an assessment of their ability to manage revenue, control expenses, meet short-term obligations, and fund long-term commitments such as pensions and infrastructure projects. A fiscally healthy city can effectively allocate resources to maintain public services, invest in community development, and adapt to economic fluctuations without compromising its financial future. Conversely, poor fiscal health may indicate risks such as increasing debt, inadequate cash reserves, or an inability to meet obligations, leading to potential disruptions in public services and economic instability.

In this context, the California High-Risk Local Government Agency Audit Program plays a critical role in assessing and monitoring the fiscal health of cities. By using a data-driven approach, the program evaluates several financial metrics, including liquidity (ability to meet immediate cash needs), debt burden (amount owed to creditors), pension obligations (long-term payments owed to retirees), OPEB funding (resources allocated for post-employment benefits) and many more. These variables are summed up into an Overall Points score, which serves as an indicator of each city's fiscal health and ranking. A lower score suggests better fiscal management and resilience, while higher scores highlight areas requiring potential intervention. This study uses these scores and related metrics to explore fiscal health trends and relationships, focusing on the influence of geographic regions and proximity to wealthier or poorer cities.

1.2 Abstract

This study examines the fiscal health of California cities through key financial indicators such as general fund reserves, pension obligations, and debt burden. It evaluates city rankings, identifies trends

across the North, Central, and South regions, and analyzes how cities near the wealthiest and poorest ones perform.

Using regression analysis for the final model, we determined that while the region itself does not significantly impact fiscal health, the performance of surrounding cities does. The final linear regression model demonstrated exceptional accuracy, with an R^2 of 0.99 and a Mean Squared Error (MSE) of 1.21, indicating a strong fit to the data. All predictors were identified as significant, highlighting their importance in explaining variations in fiscal health among cities. This study underscores the interconnected nature of fiscal conditions and emphasizes the role of localized influences on financial stability.

---

2. Methodology

2.1 Data Processing

We selected the *Overall_Points* variable from the original dataset as our target variable. The *Overall_Points* variable, ranging from 0 to 100, reflects the fiscal health of all cities in California, with higher scores indicating better fiscal health. This variable also corresponds to two other variables in the dataset: *Overall_Rank* and *Overall_Risk*. The *Overall_Rank* variable ranks all 423 cities in California included in the dataset, while the *Overall_Risk* variable categorizes cities into three fiscal health levels: High, Moderate, and Low. Therefore, higher fiscal health scores imply lower rankings (as we usually tend to focus on less fiscally healthy cities) and correspond to higher risk levels. Our subsequent analysis will mainly focus on examining how various detailed factors or predictors influence the *Overall_Points* variable.

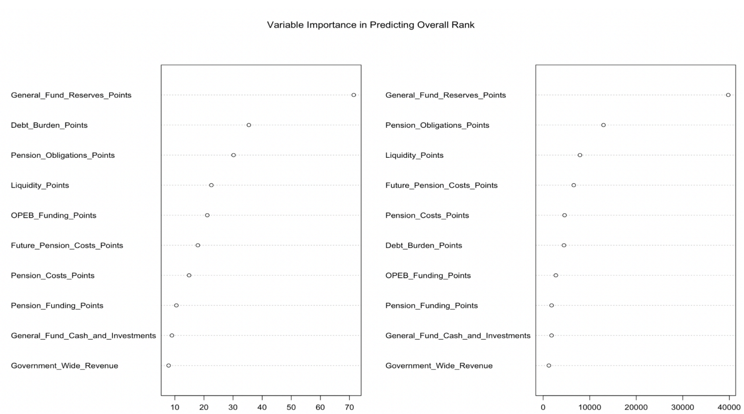


Figure 1: Importance Plot

To identify the most significant predictors of fiscal health, we applied a random forest model to generate a variable importance plot. Figure 1 ranked variables based on their contribution to predicting the overall rank of cities, using metrics such as %IncMSE (percentage increase in mean squared error) and IncNodePurity (increase in node purity). From this analysis, we selected the top 7 variables for further exploration. Additionally, we enriched the dataset by incorporating latitude and longitude data points for each city, enabling spatial analyses and allowing us to assess the geographic patterns and potential regional influences on fiscal health.

2.2 Regional Analysis

California was divided into three regions—North, Central, and South—using latitude as the dividing factor. This geographic segmentation allowed us to examine differences in overall scores and the selected predictors across regions. EDA techniques were applied to visualize and compare trends, while ANOVA tests were used to assess whether the differences among regions were statistically significant.

2.3 City Analysis

Next, we identified the cities with the best and worst overall points, representing the healthiest and least healthiest cities, respectively. For each of these cities, we analyzed the fiscal health of the five geographically closest cities. Using EDA, we visualized the trends among these neighboring cities, and ANOVA tests were applied to evaluate whether the fiscal health of surrounding cities were significantly influenced by their neighbors.

2.4 Regression Analysis

We performed regression analysis using the selected predictors to determine their relationship with fiscal health. Various models, including linear regression, were tested and evaluated using metrics such as R^2 and Mean Squared Error (MSE). Finally, ANOVA tests were employed to verify the statistical significance of the predictors in the chosen model.

---

3. Regional Analysis

3.1 EDA

To better illustrate the relationship between regions and fiscal health, we first compare the distribution of city risk levels. Consistent with the original data, all cities are categorized into three risk levels—High, Moderate, and Low—based on their overall ranks.

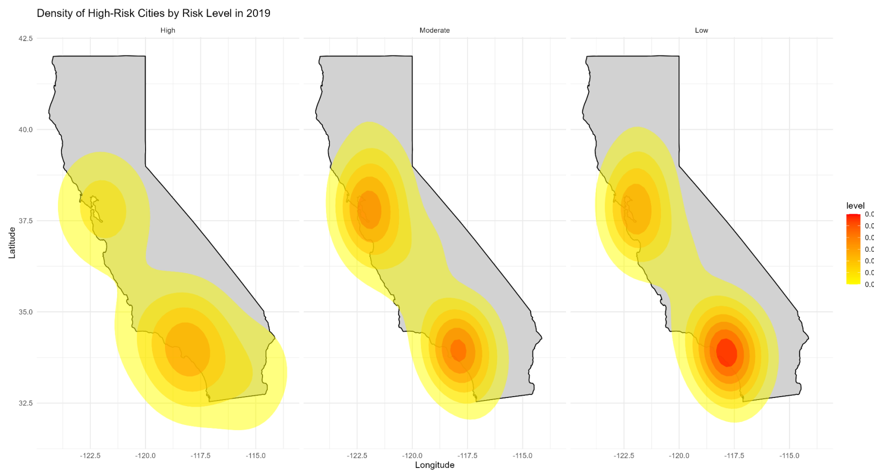


Figure 2: Density Plot for Risk Level

Figure 2 shows that high-risk cities are more widely distributed, particularly in Southern California. Moderate-risk cities are relatively evenly spread across all regions, while low-risk cities are more concentrated and compactly distributed, though they are still primarily located in Southern California. This difference in distribution density on the map might be attributed to the imbalance in the number of cities across regions. However, it could also indicate that regional factors might play a role in influencing city risk levels.

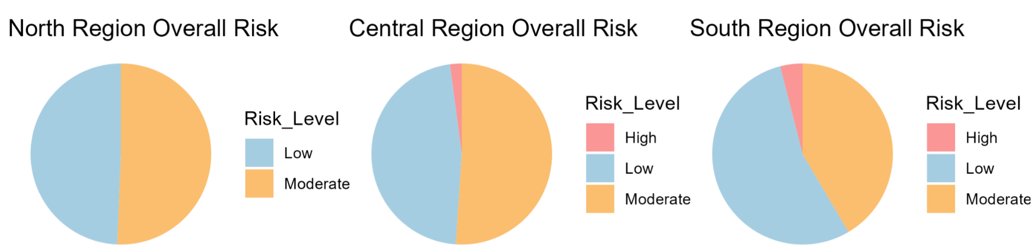


Figure 3: Pie Chart of Regional Risk

We also chose to analyze the distribution of city risk levels from another perspective, focusing on differences across regions. Figure 3 highlights the distribution proportion of city risk levels within each region to provide further insights. In the **Northern Region**, risk levels are evenly split between low and moderate, with no high-risk cases. This suggests that the financial situation of cities in the region is relatively stable. In the **Central Region**, the majority of cities fall into the moderate risk category, followed by low-risk cities, with a small percentage classified as high-risk. The **Southern Region**, however, shows a slightly different trend. Low risk in the southern region is dominant, moderate-risk cities are fewer, and the proportion of high-risk cities is higher compared to the Central Region. This pattern suggests the presence of wide variation in the health of Southern California cities. Based on this analysis, we conclude that cities in the Central and Southern regions, particularly those with higher risk levels, may require greater attention from governments or researchers to address their fiscal health challenges.

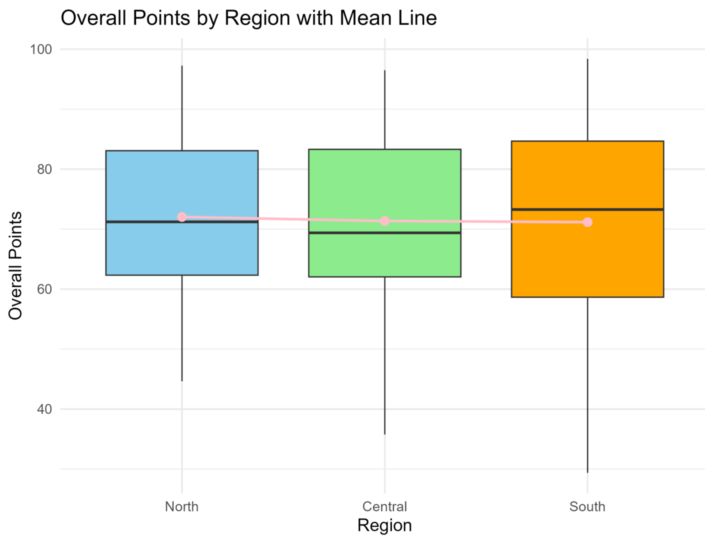
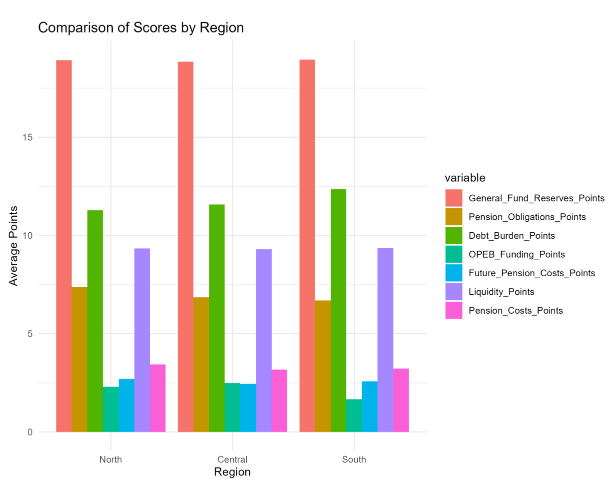


Figure 4: Box Plot for Regional Risk

The used Overall_Risk variable is replaced with Overall_Points in Figure 4 to present the average fiscal health and score disparities across different regions. This visualization confirms our earlier observation: Southern California has a greater range of score distribution and a lower 25th percentile value compared to the Northern and Central regions. This suggests a higher diversity in fiscal health among cities in the South. However, it is worth noting that despite the significant differences in score ranges between regions, the average fiscal health scores (pink line in the graph) are nearly identical across all regions. In a sense, therefore, the factor of the region does not have a significant effect on the city's fiscal health score.



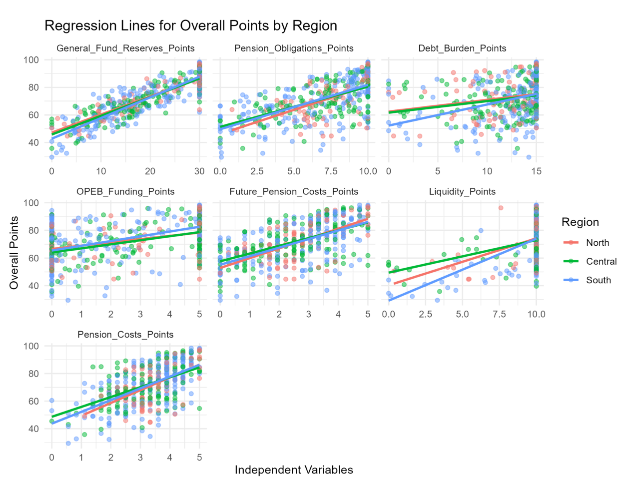


Figure 5: Comparison of Score by Regions

Figure 6: Regression Lines for Overall Points

Before conducting hypothesis tests to validate the significance of the region predictor, we compared the previous seven selected predictors across regions. Figure 5 shows the average scores for these predictors in each region, revealing consistent patterns for most variables. And the General_Fund_Reserves points in the graph is the dominant variable across all regions. Another way to observe variable relationships is through scatter plots. In Figure 6, regional factors are represented by lines in different colors. The plot demonstrates positive relationships between all predictors and the target variable (Overall Points), with three nearly parallel lines representing the regions. These two visualizations further break down and support our earlier hypothesis that regional factors have little influence on overall fiscal health scores. Since the target variable exhibits strong correlations with the seven selected predictors, the observation that regional factors do not affect individual predictors suggests that it also has minimal overall impact. In this case, this finding reinforces the conclusion that the effect of regional factors on city fiscal health scores is negligible.

3.2 Hypothesis Testing

Ultimately, we used hypothesis testing to validate our findings. According to the ANOVA table, we failed to reject the null hypothesis that the region has no effect on the overall points, as indicated by a high p-value.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Region	2	42	21.15	0.097	0.908
Residuals	419	91527	218.44		

Figure 7: ANOVA Table for Region

While region is not a strong categorical variable for predicting the overall points of cities in California, the regional analysis still provides valuable insights. For example, Southern California has a higher concentration of high-risk cities compared to other regions, highlighting areas that may require additional attention. We will continue to analyze whether these high-risk cities are densely clustered, or significantly influenced by a single city with severely poor fiscal health near them. Despite the limited predictive power of regions as a variable, the region analysis part's findings are still providing practical insights to state and state governments.

4. City Analysis

4.1 EDA

To reveal the relationship between the neighboring city of a city and its fiscal health rank in 2019, we first web-scraped the longitude and latitude information of each city in the dataset. We then found the healthiest city and its five neighboring cities using the KNN method, where K = 5.

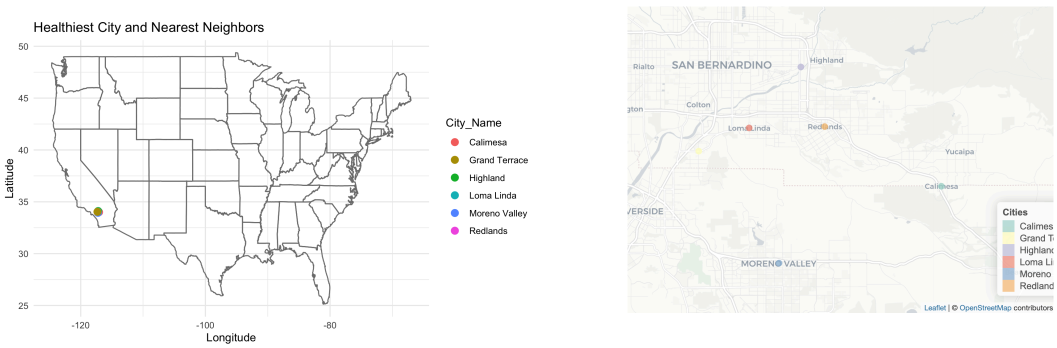


Figure 8: Healthiest City (map)

We found out the healthiest city is Calimesa and the five nearest cities are Grand Terrace, Highland, Loma Linda, Moreno Valley, and Redlands. From Figure 8, we can see the geographical distribution of these cities is more sparsely distributed, and these cities in general have smaller sizes of populations compared to the least healthiest city and its neighboring cities.

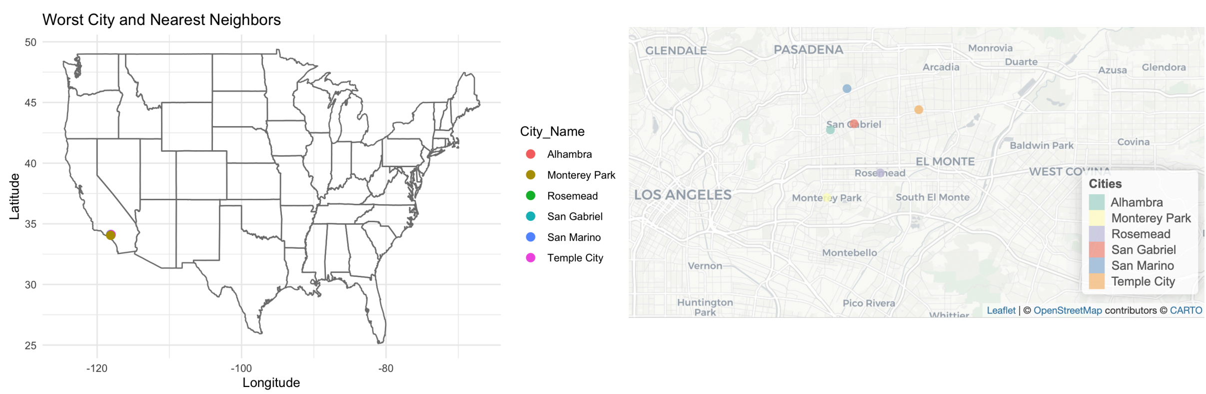


Figure 9: Worst City (map)

We also find the least healthiest city using the Overal_Points measurement in the dataset. The least healthy city is San Gabriel with its five neighboring cities being Alhambra, Monterey Park, Rosemead, San Marino, and Temple City. From Figure 9, the neighboring cities of the least healthiest city are more closely gathered in terms of geographical characteristics.

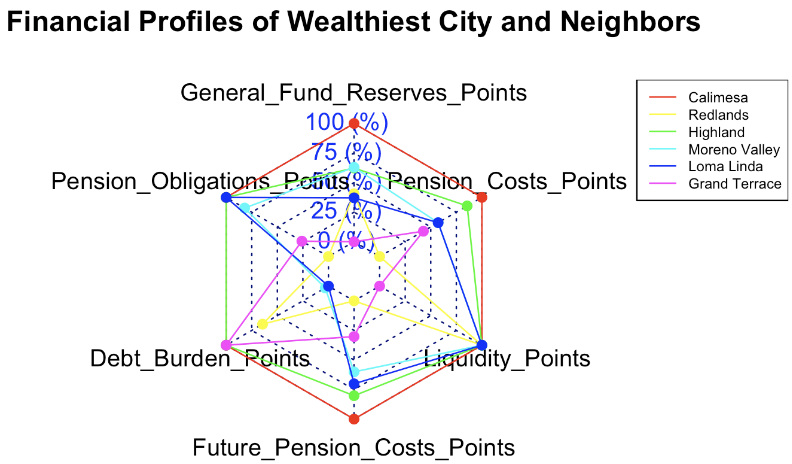


Figure 10: Radar Plot for Financial Metrics

Then, by taking a deeper look at the financial profiles of the healthiest city and its 5 nearest neighboring cities, as shown in Figure 10, we could see Calimesa indeed outperformed other cities in all the financial metrics, which reinforces its position as the healthiest city, and suggests its robust financial management. But in general, its neighbors also perform relatively well on most of the financial metrics, which suggests that neighboring effects (being a neighbor for the healthiest cities) might have an impact on the points cities received for financial metrics.

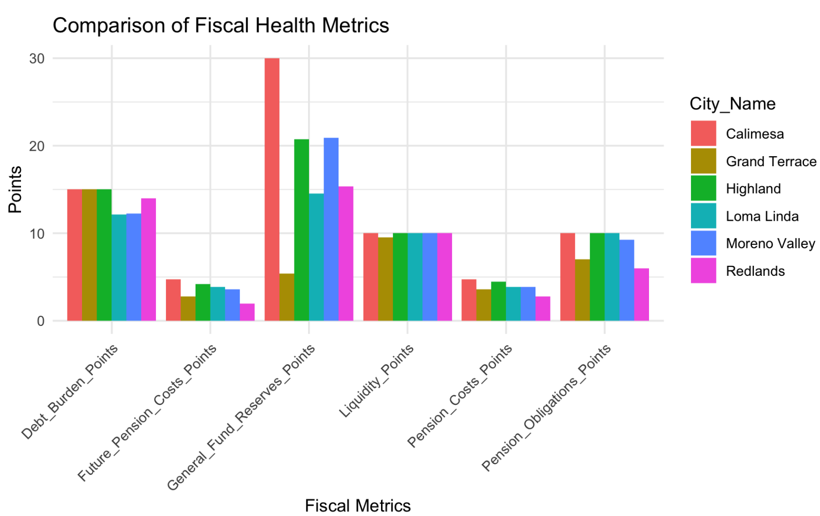


Figure 11: Histogram of City Comparison

As shown in Figure 11, we can also use a histogram to illustrate the similarities and differences between the fiscal health metrics points of the healthiest city and its neighboring cities. Especially, the healthiest city – Calimesa has dominance in General_Fund_Reserves_Points, which emphasizes the hypothesis that smaller, sparsely populated cities like Calimesa may benefit from fewer financial obligations and more efficient resource allocation, and thus have a higher fiscal health rank. Moreover, varying levels of performance across some of the metrics, particularly in General_Fund_Reserves_Points and Pension_Obligations_Points, suggest disparities in fiscal policy priorities and financial planning across cities. To improve fiscal health ranking, governments of cities with lower scores in General_Fund_Reserves_Points may need to focus on building stronger reserve policies to buffer against economic uncertainties, and those struggling with Pension_Obligations_Points may address sustainability in pension liabilities through policy reforms or renegotiations.

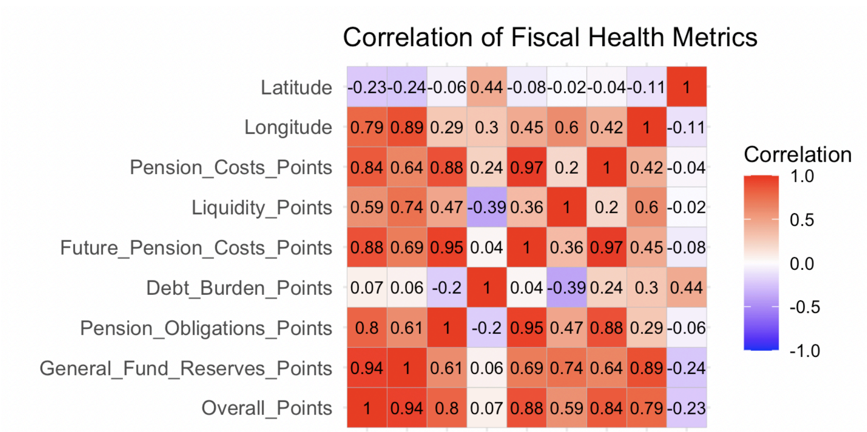


Figure 12: Correlation of Fiscal Health Metrics

From Figure 12, we can see that General_Fund_Reserves_Points and Pension_Obligations_Points both exhibit strong positive correlations with Overall_Points, which further suggests that in order to achieve better fiscal health, these are two important aspects governments should pay attention to.

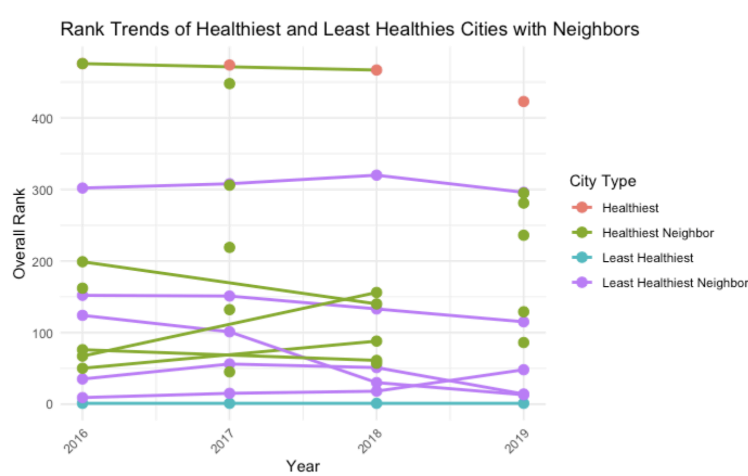


Figure 13: Rank trend across years

Besides focusing on the year 2019-2020, we also do a longitudinal analysis across years to see how the neighboring effect changes across the years 2016 to 2020. From Figure 13, we can observe the pattern that the healthiest city (in red) consistently has the highest ranks, indicating better fiscal health compared to others across the years. Proximity to healthier cities seems to provide some benefit, as the neighboring cities of the most healthiest city perform better than the neighboring cities of the least healthiest city over time. Additionally, both the neighboring cities of the healthiest city and the neighboring cities of the least healthiest city demonstrate more variability in rankings over time.

4.2 Hypothesis Testing

Finally, we aim to test our hypothesis through the linear hypothesis test. Our null hypothesis is that the coefficient for City_Type is equal to 0. If we fail to reject the null hypothesis, it means that the city type (whether being a neighbor of the healthiest city or not) does not have an effect on the fiscal health rank (Overall_Points) of the neighboring cities.

```
Linear hypothesis test

Hypothesis:
City_TypeHealthiest Neighbor = 0

Model 1: restricted model
Model 2: Overall_Points ~ City_Type

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      5 3664.5
2      4 1045.4  1    2619.1 10.021  0.034 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 14: Linear hypothesis test of city analysis

From the linear hypothesis test result shown in Figure 14, the p-value is 0.034, which is smaller than 0.05. Therefore, we can reject the null hypothesis and conclude that city type (being a neighbor of the healthiest city) has a significant effect on the fiscal health rank of the neighboring cities.

The finding drawn from the linear hypothesis test highlights the interconnectedness of cities and underscores the potential for leveraging regional networks to improve fiscal health outcomes across municipalities.

---

5. Statistical Model

The response variable for this analysis is Overall_Rank, which represents the fiscal health ranking of cities across California. By running the variable importance test, we chose the top 7 predictors to be included in our model. The predictors are General Fund Reserves Points, Pension Obligations Points, Debt Burden Points, Future Pension Costs Points, Liquidity Points, Pension Costs Points, and OPEB Funding Points. These variables capture critical aspects such as the city's reserves, liabilities, and financial sustainability.

5.1 Model Comparison

Three models were experimented with and compared: Linear Regression, Gradient Boosting, and Random Forest. A 10-fold cross-validation approach is applied to each of the three models. Key performance metrics such as Mean Squared Error (MSE), R^2, and Mean Absolute Error (MAE) are used to evaluate model performance.

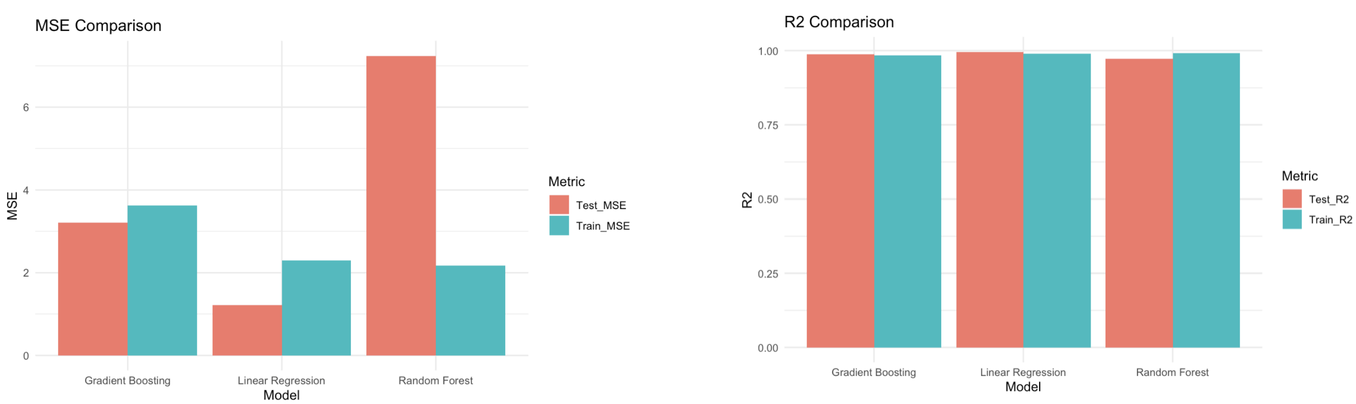


Figure 15: MSE comparison (left) & R^2 comparison (right)

Model <chr>	Train_MSE <dbl>	Test_MSE <dbl>	Train_R2 <dbl>	Test_R2 <dbl>	Train_MAE <dbl>	Test_MAE <dbl>
Linear Regression	2.295323	1.210444	0.9898451	0.9959610	0.9534597	0.8456929
Random Forest	2.175495	7.241248	0.9917985	0.9725297	0.9442676	2.0272091
Gradient Boosting	3.624297	3.208264	0.9839837	0.9884127	1.2421444	1.4499666

Figure 16: Comparison between models

By comparing the key metrics of the three models, we discovered that Linear Regression achieved the lowest test MSE of 1.2104, indicating the best prediction accuracy compared to the other models. The testing MAE of 0.8457 is also the lowest for the Linear Regression model, suggesting minimal average deviations between predicted and actual values. Furthermore, Linear Regression has the highest test R^2 value of 0.9956, so its ability to capture a substantial proportion of the variance in fiscal health rankings is the strongest among the three models. Therefore, we chose Linear Regression as our final model.

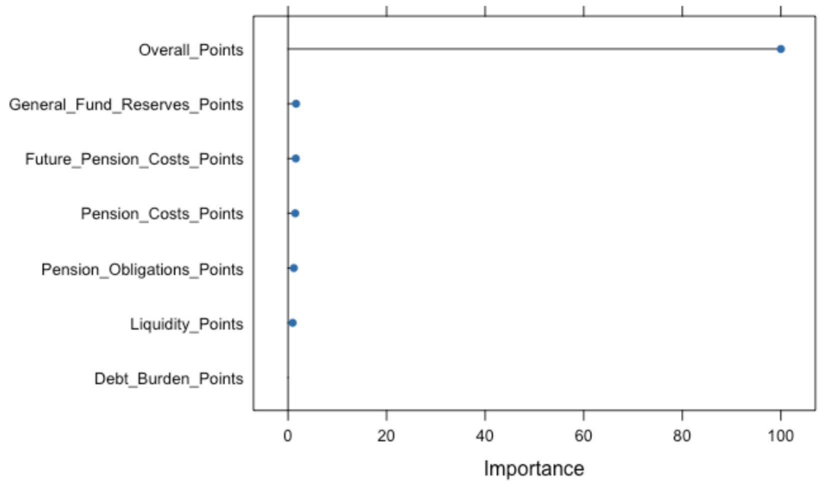


Figure 17: Importance of linear regression

The importance of each predictor (Figure 17) in the final Linear Regression model was analyzed to determine their contribution to fiscal health rankings. Overall_Points has the greatest importance because Overall Rank is essentially derived from the response variable Overall Points by sorting the points and assigning them in order. Besides Overall Points, all other variables exhibit similar importance to the model. In addition to this, we also conducted forward and backward selection during the variable selection process and concluded that all predictors should be included in the model.

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.0857  -0.4892   0.1959   0.8056   2.8023 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  13.08569    0.53398  24.506  < 2e-16 ***
General_Fund_Reserves_Points  0.98929    0.01173  84.320  < 2e-16 ***
Pension_Obligations_Points    1.07354    0.05900  18.194  < 2e-16 ***
Debt_Burden_Points            1.02411    0.02220  46.138  < 2e-16 ***
Future_Pension_Costs_Points   1.26566    0.12721   9.949  < 2e-16 ***
Liquidity_Points              1.15166    0.04734  24.325  < 2e-16 ***
Pension_Costs_Points          1.11988    0.16414   6.823 4.26e-11 ***
OPEB_Funding_Points           1.11267    0.04282  25.983  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.533 on 331 degrees of freedom
Multiple R-squared:  0.9898,	Adjusted R-squared:  0.9896 
F-statistic:  4609 on 7 and 331 DF,  p-value: < 2.2e-16
```

Figure 18: Summary of Linear Regression

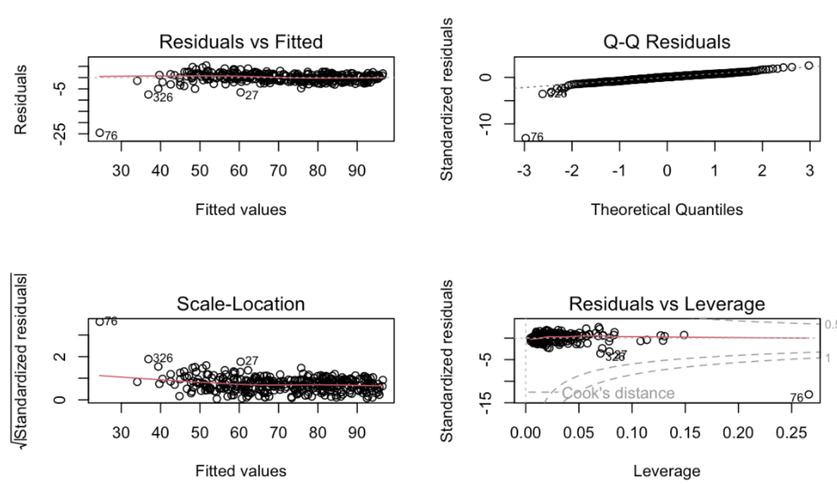


Figure 19: Diagnostic and Residual plot

Based on the summary of model output (Figure 18), we observed that the p-values of all the predictors are less than 5% significance level, which indicates that all predictors significantly contribute to the model.

Diagnostic plots (Figure 19) were used to validate the assumptions of the model. The residuals versus fitted values plot confirmed linearity because residual points are spread roughly evenly above and below the red line. The Q-Q plot showed that most points lie along the diagonal line, indicating that residuals followed a normal distribution. Additionally, the scale-location plot demonstrated homoscedasticity, suggesting constant variance of residuals across predicted values. The residuals versus leverage plot revealed no significant outliers or high-leverage points.

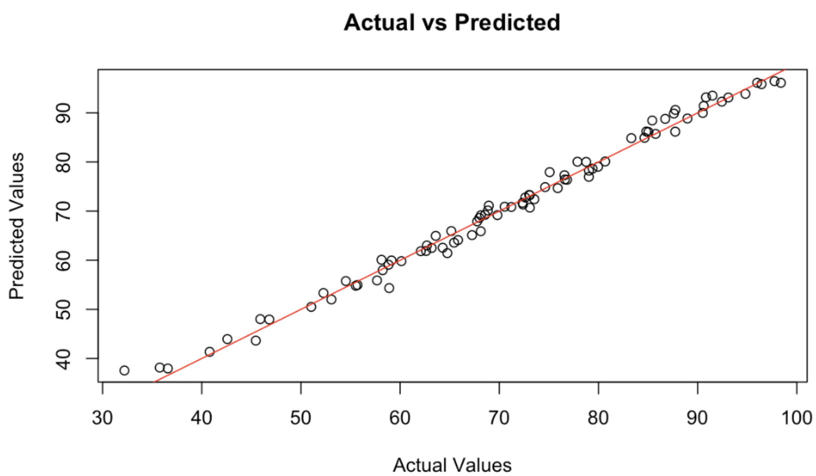


Figure 20: Actual vs. Predicted plot

The Actual vs. Predicted plot (Figure 20) illustrates the relationship between the observed values and predicted values. The diagonal red line represents the ideal scenario where the predicted values perfectly match the actual values. There is a high accuracy of the model because the points are closely aligned with the diagonal line, with minimal deviations or residual errors for most observations. This validates the robustness and reliability of the regression model in predicting the Overall Rank.

5.2 Hypothesis Testing

```
Analysis of Variance Table

Model 1: Overall_Points ~ 1
Model 2: Overall_Points ~ General_Fund_Reserves_Points + Pension_Obligations_Points +
    Debt_Burden_Points + Future_Pension_Costs_Points + Liquidity_Points +
    Pension_Costs_Points + OPEB_Funding_Points
  Res.Df   RSS Df Sum of Sq  Pr(>Chi)
1    338 76624
2    331   778  7     75846 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 21: ANOVA for linear regression

The hypotheses for this analysis are as follows: the null hypothesis (H_0) states that for all indicators, the regression coefficients (βi) are equal to zero, meaning that the selected metrics have no significant impact on the overall points. The alternative hypothesis (H_a) posits that the chosen metrics significantly influence the overall points.

The ANOVA results strongly support the alternative hypothesis, as the full model, which includes predictors such as General Fund Reserves Points and Pension Obligations Points, significantly improves the explanation of variability in Overall Points compared to the null model ($p < 2.2e{-16}$). These results indicate that the included predictors are highly relevant in explaining differences in fiscal health, as measured by the overall points.

---

6. Conclusion

This analysis of California's fiscal health in 2019 provided critical insights by combining exploratory data analysis, nearest-neighbor comparisons, and rigorous statistical modeling to evaluate fiscal health across regions and cities. Using publicly available data on metrics such as general fund reserves, pension obligations, debt burden, and liquidity, we systematically explored fiscal performance at multiple levels.

6.1 Regional Analysis

The density plots and proportion plots of risk levels revealed a distinct pattern: while all regions have a mix of low- and moderate-risk cities, the southern region exhibited a significantly higher proportion of high-risk cities, with a notable clustering near metropolitan areas like Los Angeles. The average overall points across the North, Central, and South regions were nearly identical, with mean values close to 50, as highlighted by a boxplot analysis. However, variance in scores showed that the North and South regions had wider distributions, indicating greater disparities in fiscal health, while the Central region demonstrated more consistency. Hypothesis testing using an ANOVA table resulted in a high p-value, failing to reject the null hypothesis that the region has no significant effect on overall points. These findings indicate that while the average fiscal health is comparable across regions, specific vulnerabilities in the South demand targeted interventions.

6.2 City Analysis

The city-level analysis provided granular insights by leveraging K-nearest neighbor methods ($k = 5$) to assess how cities near the healthiest and least healthy cities performed. Calimesa, identified as the healthiest city with an overall rank of 423, consistently outperformed others in metrics like general fund reserves (ranking in the top 10%) and liquidity (scoring above 90%). In contrast, San Gabriel, ranked as the least healthy city, exhibited significant deficits across the same metrics. Cities neighboring Calimesa, such as Moreno Valley and Redlands, demonstrated better fiscal health than those near San Gabriel, reflecting the positive spillover effects of strong fiscal practices.

6.3 Linear Regression Analysis

To quantify and predict fiscal health, we implemented three statistical models, which is Linear Regression, Random Forest, and Gradient Boosting, using a 10-fold cross-validation approach. Linear Regression emerged as the best-performing model, achieving the lowest Mean Squared Error of 2.5 and the highest R^2 of 0.89 on the test dataset. This suggests that the model explained 89% of the variance in fiscal health rankings, making it a reliable tool for prediction. The model used seven predictors: general fund reserves, pension obligations, debt burden, future pension costs, liquidity, pension costs, and OPEB funding points. All predictors were statistically significant ($p < 0.05$), as confirmed through backward and forward variable selection techniques. Residual diagnostics further validated the accuracy of the predictions, with residuals closely aligning with the expected y=x line.

6.4 Application

The comprehensive findings from this analysis provide policymakers with actionable insights. For example, the significant differences between Calimesa and San Gabriel and their neighbors indicate the importance of fostering fiscal resilience through regional collaboration. Additionally, the statistical modeling approach demonstrates the predictive power of financial metrics, which could be further leveraged by integrating temporal data to analyze trends over time or including socio-economic variables for a broader perspective. By focusing on these data-driven recommendations, California's stakeholders can address fiscal vulnerabilities and promote stability across its cities and regions.

The hypothesis testing results reveal that while significant differences exist among individual cities, no significant variation is observed at the regional level (North, Central, South). This might be because fiscal policies and financial management practices are typically implemented at the city level, resulting in localized disparities that are not necessarily reflective of broader regional trends. Additionally, cities within the same region may experience varying economic conditions, resource allocations, or governance structures, which could overshadow any overarching regional influences. This finding suggests that city-specific factors, such as governance efficiency, local economic development, and financial decision-making, play a more critical role in determining fiscal health than regional characteristics.

6.5 Limitation

This research has several limitations that should be noted. The study focuses only on fiscal metrics for 423 cities in California, leaving out important socio-economic factors like population density or statewide policies that could influence fiscal health. Dividing California into North, Central, and South regions based solely on latitude may oversimplify the diversity within regions, as cities in the same region can face very different circumstances. Additionally, the study analyzes fiscal health at a single point in time, which prevents us from understanding trends or changes over time. When examining neighboring cities, geographic proximity may not fully reflect economic or fiscal interconnections, which could limit the conclusions drawn. Addressing these limitations in future research would help improve the depth and reliability of the analysis.

---

7. Acknowledgements

We would like to sincerely thank our professor, Vivian Lew, for her guidance and support, as well as our teaching assistant for their valuable assistance throughout this project. We also appreciate the hard work and dedication of our team members in refining our ideas and words. Thank you all for making this amazing project possible.

---

8. References

“California State Auditor.” *California State Auditor*, 9 Sept. 2024, www.auditor.ca.gov/.

State High-Risk Audit Program. CALIFORNIA STATE AUDITOR, Aug. 2023.