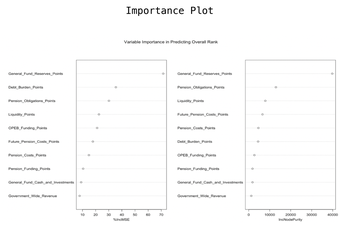
Analysis for California Fiscal Health in 2019

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

Department of Statistics and Data Science

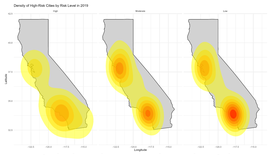
Introduction

This analysis examines the fiscal health of cities across California, focusing on key economic indicators such as general fund reserves, pension obligations, and debt burden. In addition to evaluating overall city rankings and regional patterns across the North, Central, and South regions, we analyzed how cities surrounding the richest and poorest cities perform, exploring potential influences or trends. Using regression analysis for the final model, we identified significant predictors of fiscal health and their relationships.

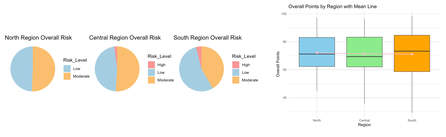


This graph, generated using random forest importance, highlights the most influential variables in predicting overall rank. General Fund Reserves Points stand out as the most significant predictor, with Pension Obligations and Liquidity Points also contributing strongly, indicating their critical role in fiscal health assessment.

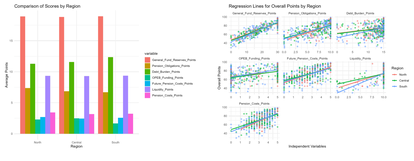
Region Analysis



1. High-risk areas are tightly clustered and widespread in northern CA
2. Most low-risk cities are concentrated in northern CA



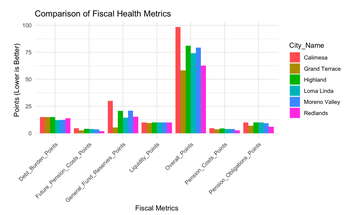
Moderate risk dominates all regions, with minimal high risk in the Central and South, while the Central region shows a more consistent distribution than the North and South.



The bar chart highlights consistent average scores across regions, with General Fund Reserves leading in all. Regression plots reveal positive relationships between variables and overall points, with slight slope variations by region.

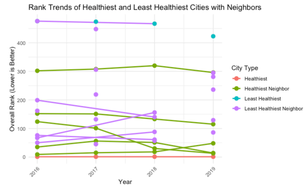
City Analysis

Healthiest City is Calimesa (indicated by the green dot on the *left* map, selected by filtering Overal Rank = 423)
Least healthiest City is San Gabriel (pink dot on the *right* graph, selected by filtering Overal Rank = 1)
Nearest Neighers are selected using KNN, where k=5.
Less healthy cities are generally close to each other and are in more populated areas. (Near Los Angeles & us!)



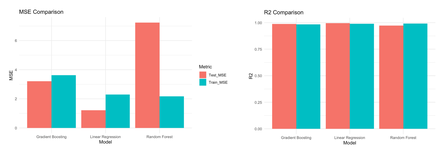
Comparison between healthiest neighbors:
Calimesa: Outperforms other cities across most metrics, particularly in overall points, general fund reserves, and liquidity.
Moreno Valley and Redlands: Show lower overall points.
Grand Terrace, Highland, and Loma Linda: Perform moderately well across all metrics, maintaining stable fiscal health.

Financial Metrics could also be visualized in radar plot.
We can see many of them are very correlated.



Key Observations:
The healthiest cities maintain a consistently low rank (better performance).
Proximity to healthier cities seems to provide some benefit, as the healthiest neighbors perform better than the least healthiest neighbors.
The healthiest cities and their neighbors demonstrate more stable rankings over time while Least healthiest cities and their neighbors show more variability.

Linear Model



Comparison of the three models with 10-fold cross-validation:
1.**Gradient Boosting** 2. **Linear Regression** 3. **Random Forest**
Linear Regression is the best-performing model based on its lowest MSE for both train and test datasets, indicating highly accurate predictions close to the actual values.
Additionally, its highest R^2 on the test dataset suggests that the model effectively captures a large proportion of the variance in the target variable, demonstrating a strong fit to the data.

Model	Train_MSE	Test_MSE	Train_R2	Test_R2	Train_MAE	Test_MAE
Linear Regression	2.295323	1.210444	0.9898451	0.9959610	0.9534597	0.8415929
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

Lower MSE, higher MAE, higher R^2 indicate better model performance. Therefore, we choose linear regression as our model.

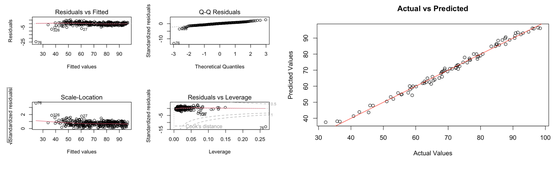
```
Call:
lm(formula = .outcome ~ ., data = dat)

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
```

p-values of all the predictors are less than 0.05, which indicates that all predictors are significant.



The residuals are evenly distributed around the red line, with consistent variance across fitted values, indicating a well-fitted model, although points like 76 and 327 may warrant further investigation due to high leverage or influence.
Most points align closely with the diagonal line (y=x), suggesting highly accurate predictions for the majority of observations.

Model Equation:
Overall Points = 13.086 + 0.989 General Fund Reserves Points + 1.074 Pension Obligations Points + 1.024 Debt Burden Points + 1.266 Future Pension Costs Points + 1.152 Liquidity Points + 1.120 Pension Costs Points + 1.113 OPEB funding Points

Hypothesis Testing

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

```
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
```

According to the *left* ANOVA table, we failed to reject the null hypothesis that the region has no effect on the overall points with a high p-value.
According to the *right* ANOVA table, since the p-value is 0.034 (< 0.05), we reject the null hypothesis, indicating that city type significantly affects the fiscal health rank of neighboring cities.

```
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
```

The ANOVA results ($p < 2.2e{-}16$) indicate that the full model, including predictors like General Fund Reserves Points and Pension Obligations Points, significantly improves the explanation of Overall_Points, showing that these metrics are highly relevant in explaining its variability.

Conclusion

The density and proportion plots reveal that while low- and moderate-risk cities dominate across regions, the southern region has a higher proportion of high-risk cities. Despite these differences, the average overall points are similar across regions, and hypothesis testing shows that region is not a statistically significant factor influencing overall points.

The city analysis highlighted that being a neighbor to a particular city can significantly affect the fiscal health rank of that city. These findings provided a strong foundation for the subsequent statistical modeling.

In the modeling phase, Linear Regression emerged as the most robust model, with the lowest Mean Squared Error (MSE) and highest R^2 on both training and test datasets, effectively capturing variance in the target variable. All predictors were statistically significant at a 0.05 confidence level, with forward and backward selection validating their inclusion.

The modeling results highlight the importance of all selected predictors in determining fiscal health, with Linear Regression offering a stable and interpretable approach supported by validation metrics. This analysis provides actionable insights for policymakers to address regional disparities and improve the fiscal health of California cities. Future research could incorporate temporal data or additional predictors to deepen understanding and explore trends over time.