California Fiscal Health in 2019

Group 3:
Yechen Cao, Luning Ding, Xianya Fu, Yuexuan Wu,
Lingxi Zhang, Shiwei Chen

01
INTRODUCTION

02
REGION ANALYSIS

03
CITY ANALYSIS

04
LINEAR MODEL
&
CONCLUSION

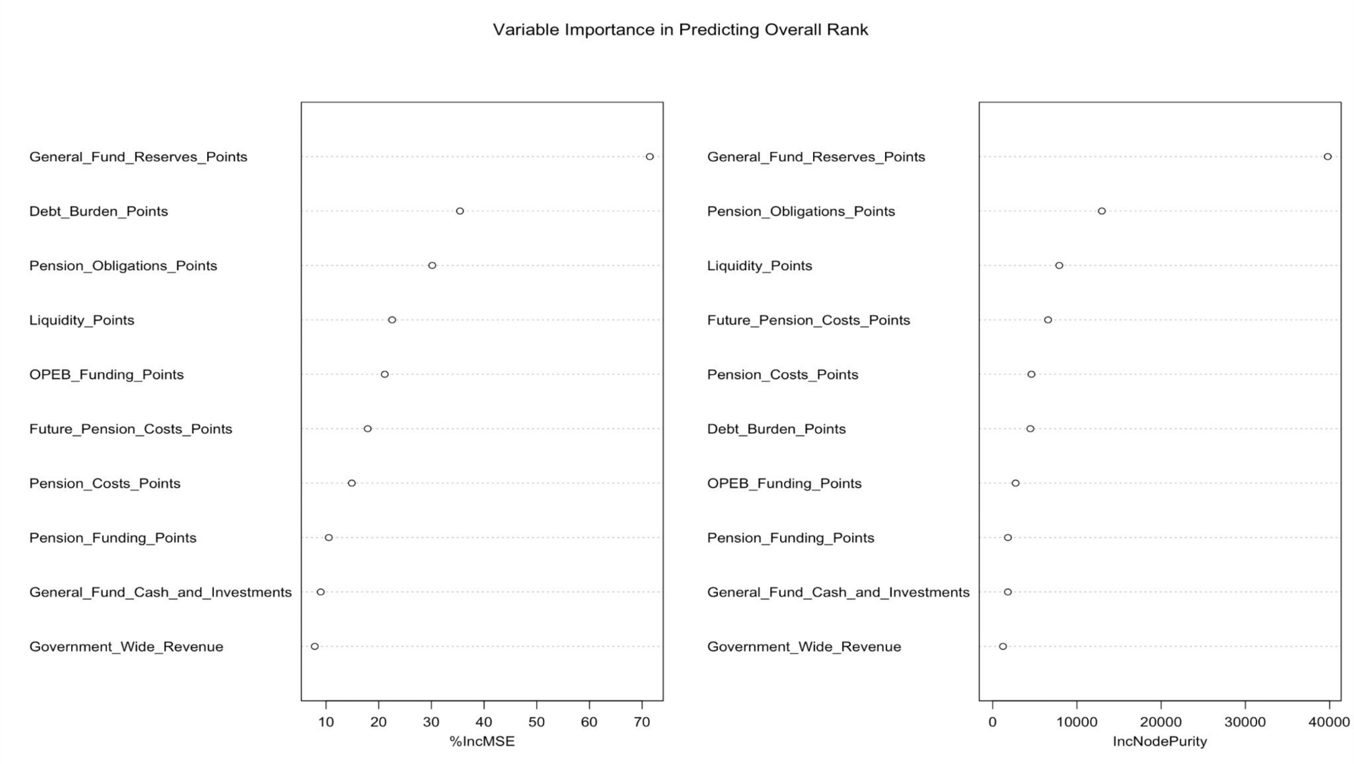
01.

INTRODUCTION

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.

IMPORTANCE



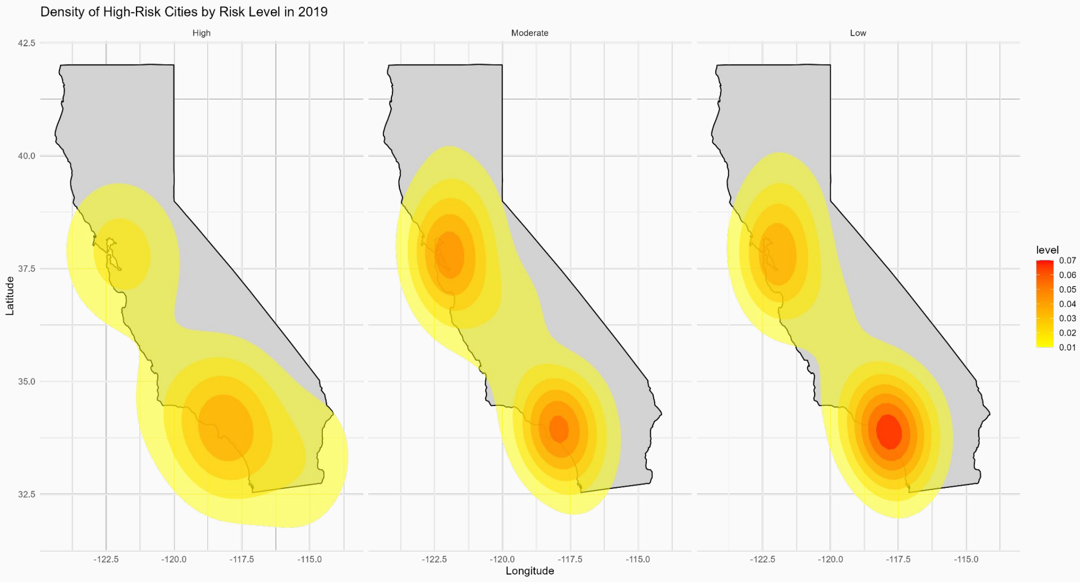
Variable Importance in Predicting Overall Rank
General_Fund_Reserves_Points
Debt_Burden_Points
Pension_Obligations_Points
Liquidity_Points
OPEB_Funding_Points
Future_Pension_Costs_Points
Pension_Costs_Points
Pension_Funding_Points
General_Fund_Cash_and_Investments
Government_Wide_Revenue
10 20 30 40 50 60 70
%IncMSE
General_Fund_Reserves_Points
Pension_Obligations_Points
Liquidity_Points
Future_Pension_Costs_Points
Pension_Costs_Points
Debt_Burden_Points
OPEB_Funding_Points
Pension_Funding_Points
General_Fund_Cash_and_Investments
Government_Wide_Revenue
0 10000 20000 30000 40000
IncNodePurity

DESCRIPTIVE STATISTICS

Variables	Mean	Medium	Lower bond	Upper bond	Variable meaning
Overall points	71.38969	72.26	60.7625	83.8525	Dependent variable
General Fund Reserves Points	18.91024	18.71500	11.75	30	A measure of the financial resources available
Debt burden	11.89289	13.24000	10.25	15	Amount of money owes to others
Liquidity	9.340284	10	10	10	The ability to meet cash obligations when due
Future Pension Cost	3.04981	3	2.63	3.38	Amount of future payments to the employee in retirement
Pension Obligations	6.873934	7.52	4.72	9.6	Amount of payments to the employee in retirement (long-term)
OPEB Funding	2.061137	1.6	0	4.625	Financial strategies and resources allocated by an organization
Pension Costs	3.247583	3.33	2.5	4.17	Expenses incurred to provide retirement benefits to employees

02.
REGION ANALYSIS

EDA

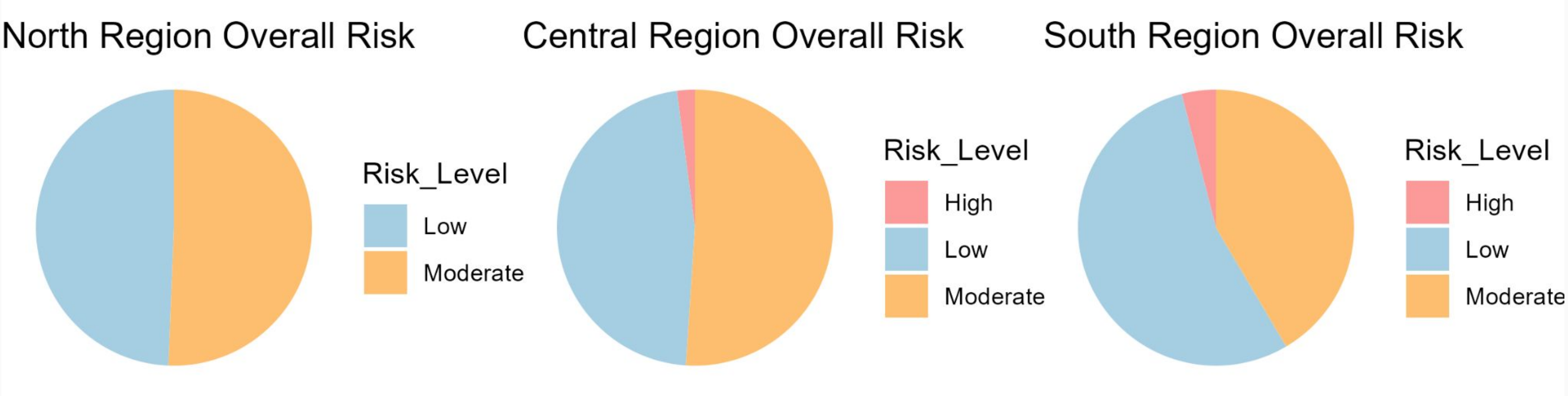


We use coordinates to classify all cities in California as northern, central, and southern. This is the density plots of cities categorized into three risk levels: High, Moderate, and Low according with the overall points.

Key observation:

1. High-risk areas are more tightly distributed and widespread in sorthern California.
2. Most low-risk cities are also distributed in sorthern california.

EDA



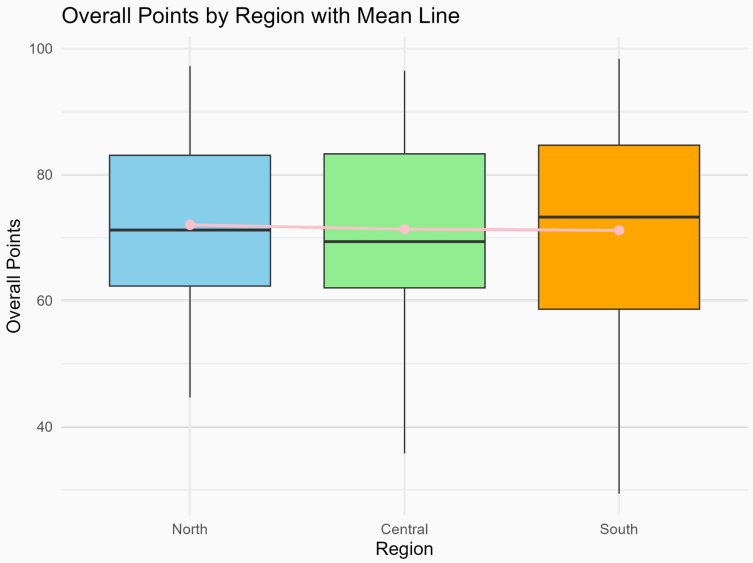
North Region: Risk is evenly split between low and moderate levels, with no high-risk cases, indicating a stable profile.

Central Region: Most risk is moderate, followed by low risk, and a small percentage is high risk, suggesting some areas of concern.

South Region: Moderate risk is dominant, with low risk second and slightly more high-risk cases than the Central Region, indicating specific vulnerabilities.

Overall, moderate risk is the most common, while high risk is minimal but present with more proportion in the South regions.

EDA



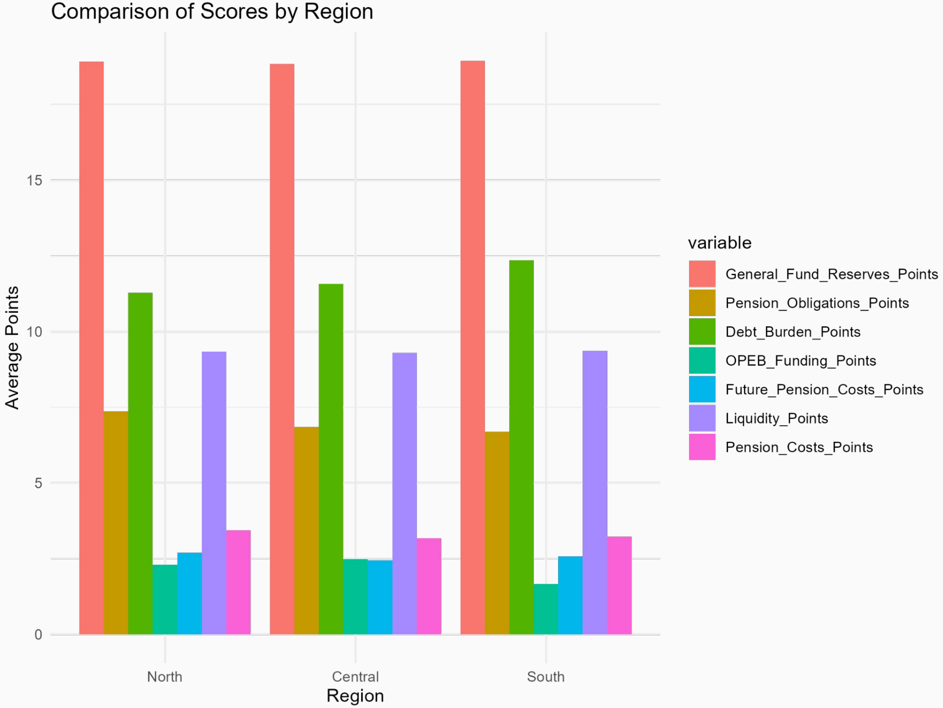
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Region	2	42	21.15	0.097	0.908
Residuals	419	91527	218.44		

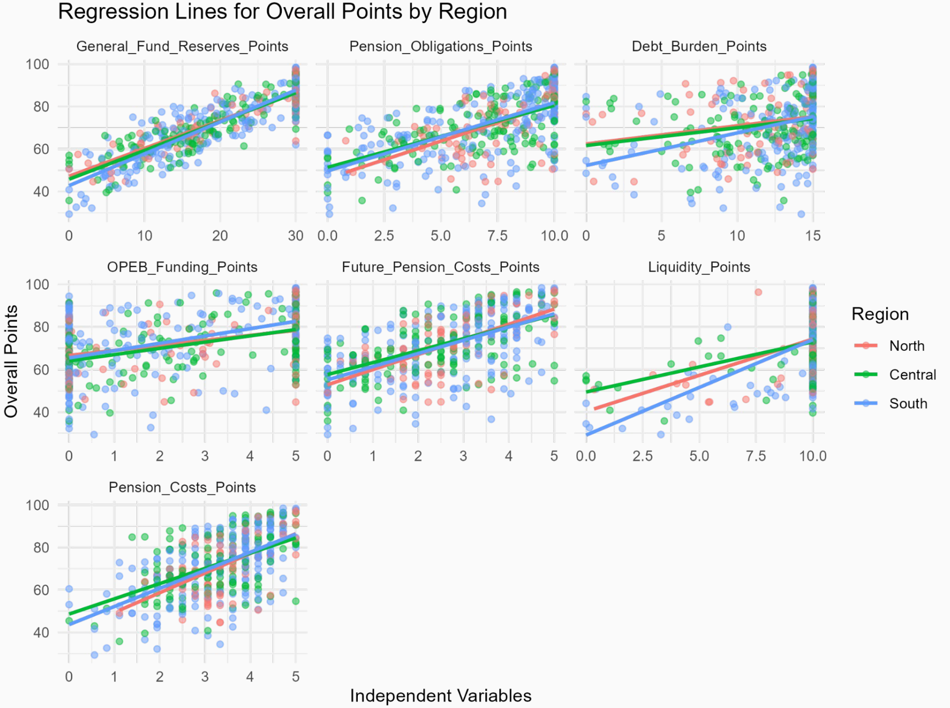
This boxplot shows the distribution of overall points for the North, Central, and South regions, with a pink line marking the average points. All regions have similar average points, but the spread of scores varies slightly. The North and South regions have a wider range of points, while the Central region is more consistent.

From the previous pie chart, we can see that regions with high risk also have more low-risk cases, which balances the overall scores and makes the means similar, but the medians differ.

According to the anova table, we failed to reject the null hypothesis that the region has no effect on the overall pointswith a high p-value.

EDA





The bar chart compares average scores across regions, showing consistent patterns for most variables, with General Fund Reserves dominating in all regions.

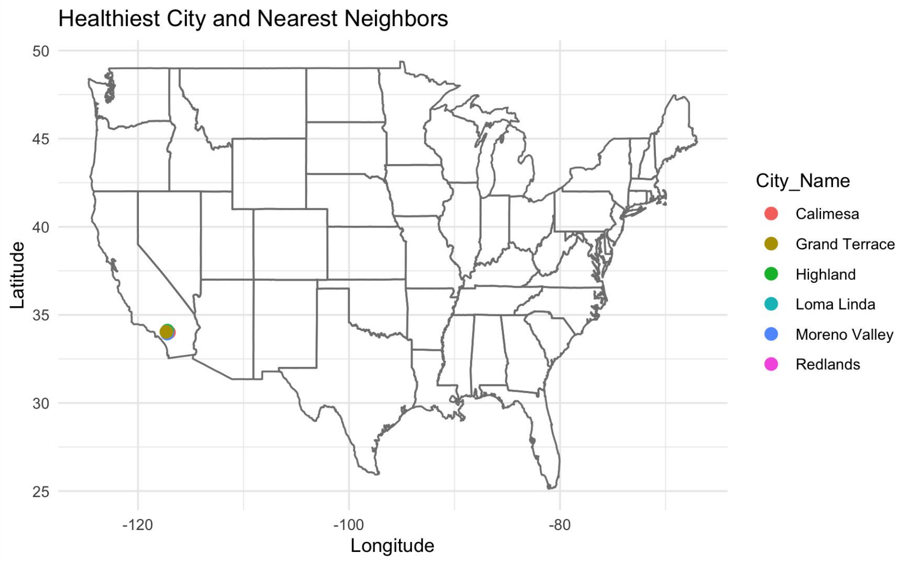
The regression plots show positive relationships between all variables and overall points, with slight differences in slopes across regions.

03.

CITY ANALYSIS

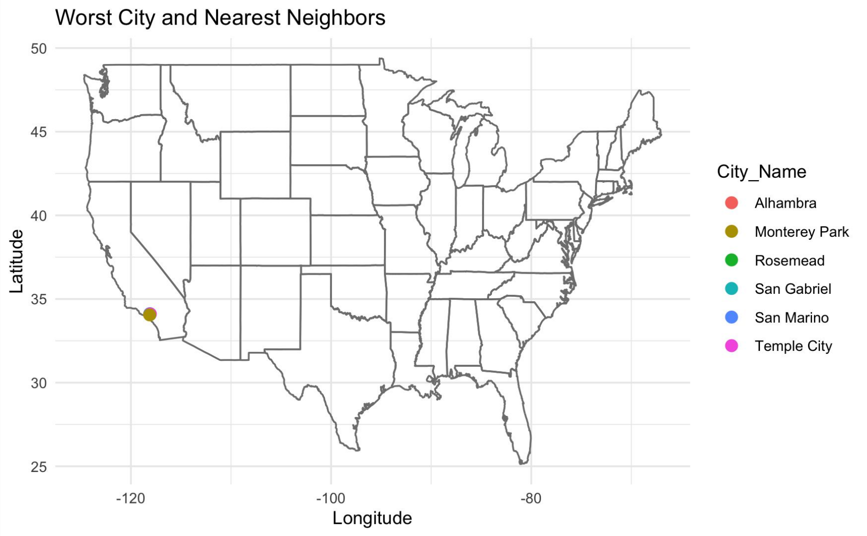
WITH 5 NEAREST NEIGHBOURS

HEALTHIEST CITY ANALYSIS IN 2019



- Longitude and latitude information about California cities are web-scrapped from online sources
- Healthiest City is **Calimesa** (green dot on the right graph, selected by filtering Overal Rank = 423)
- Nearest Neighors are selected using KNN, where k=5
- Neighbors near the healthiest city are more distributed.

LEAST HEALTHIEST CITY ANALYSIS IN 2019



- Least healthiest City is **San Gabriel** (selected by filtering Overal Rank = 1 aftering removing data with NAs)
- Nearest Neighors are selected using KNN, where k=5
- Neighbors near the least healthiest city are generally close to each other. (Near Los Angeles & us!)
- * Los Angeles only has a Overall_Rank of 27.

LINEAR HYPOTHESIS ON THE EFFECT OF CITY IN 2019

- **Null Hypothesis**: The coefficient for City_Type is equal to 0. This means that the city type (whether a neighbor of the healthiest city or not) does not have an effect on the fiscal health rank (Overall_Points) of the neighboring cities.

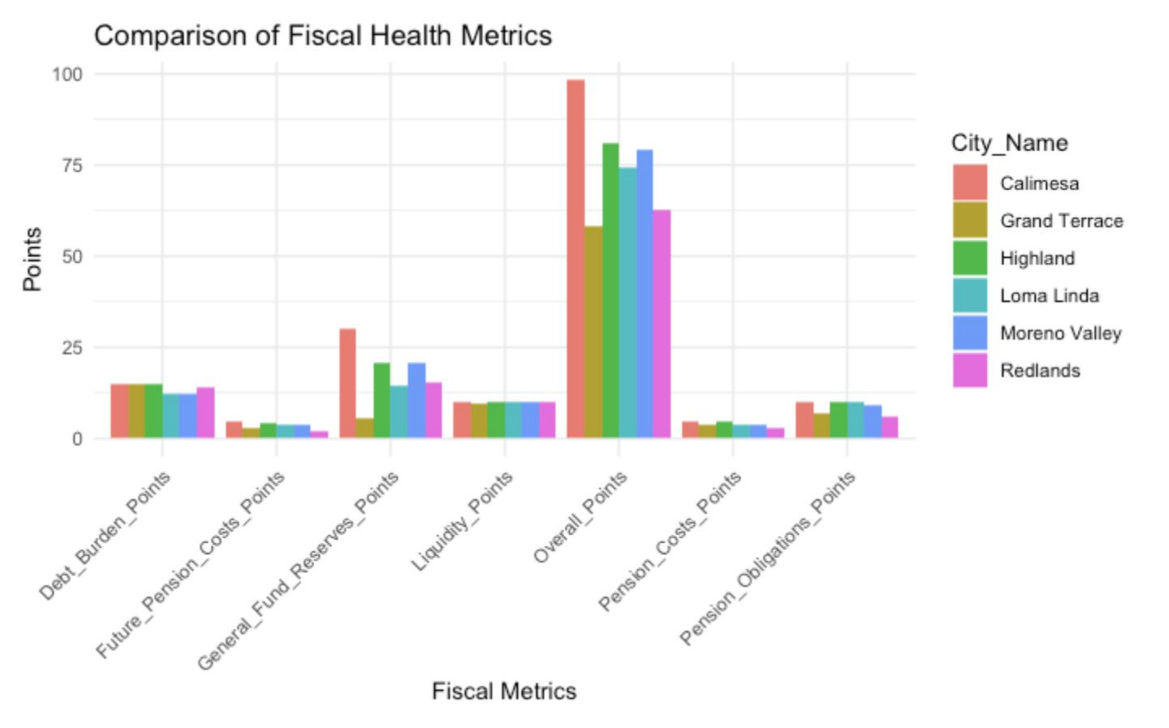
- **Alternative Hypothesis**: The coefficient for City_Type is not equal to 0. This implies that the city type has a significant effect on fiscal health rank.

```
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
```

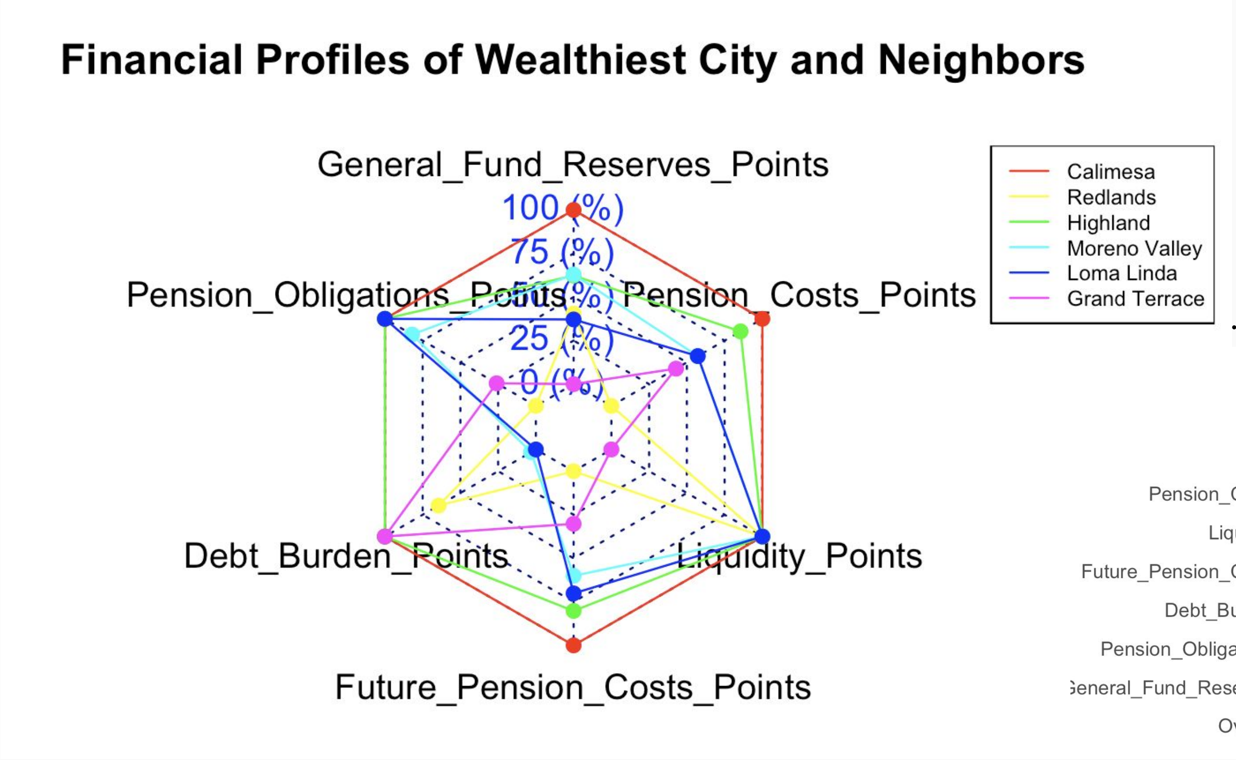
Since the p-value is 0.034 (< 0.05), we reject the null hypothesis.

This indicates that the city type (being a neighbor of the healthiest city) has a significant effect on the fiscal health rank of the neighboring cities.

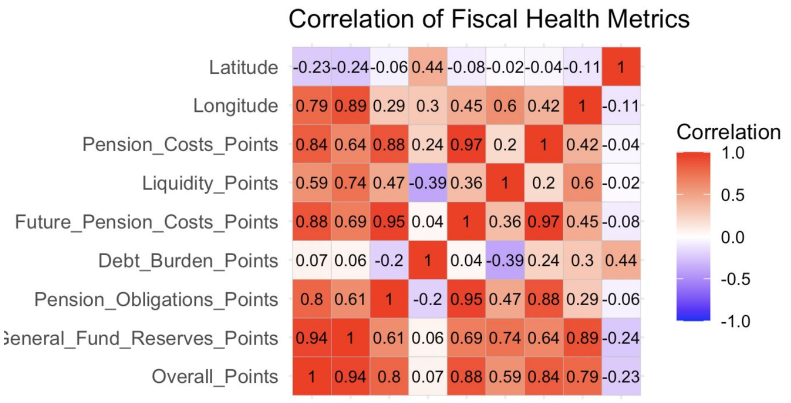
COMPARISON BETWEEN HEALTHIEST NEIGHBORS



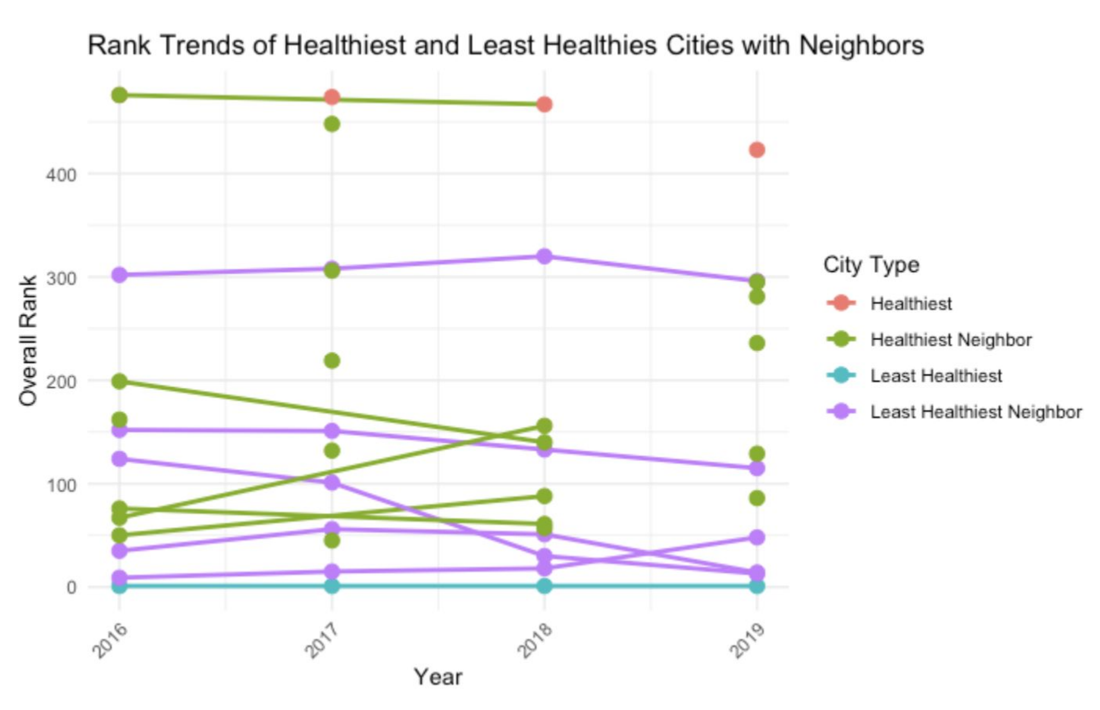
- Calimesa: Outperforms other cities across most metrics, particularly in overall points, and general fund reserves points.
- Both the healthiest city and its neighbors have similar debt burden points, liquidity points, pension obligation points, and low future pension costs points



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



ANALYSIS ACROSS YEARS



Key Observations:

- The healthiest cities consistently have the highest ranks, indicating better fiscal health compared to others.

- Proximity to healthier cities seems to provide some benefit, as the healthiest neighbors perform better than the least healthiest neighbors over time.

- Both healthiest neighbors and least healthiest neighbors demonstrate more variability in rankings over time

04. STATISTICAL MODELING

Response Variable:
Overall_Rank

Predictors:
General Fund Reserves Points
Pension Obligations Points
Debt Burden Points
Future Pension Costs Points
Liquidity Points
Pension Costs Points
OPEB Funding Point

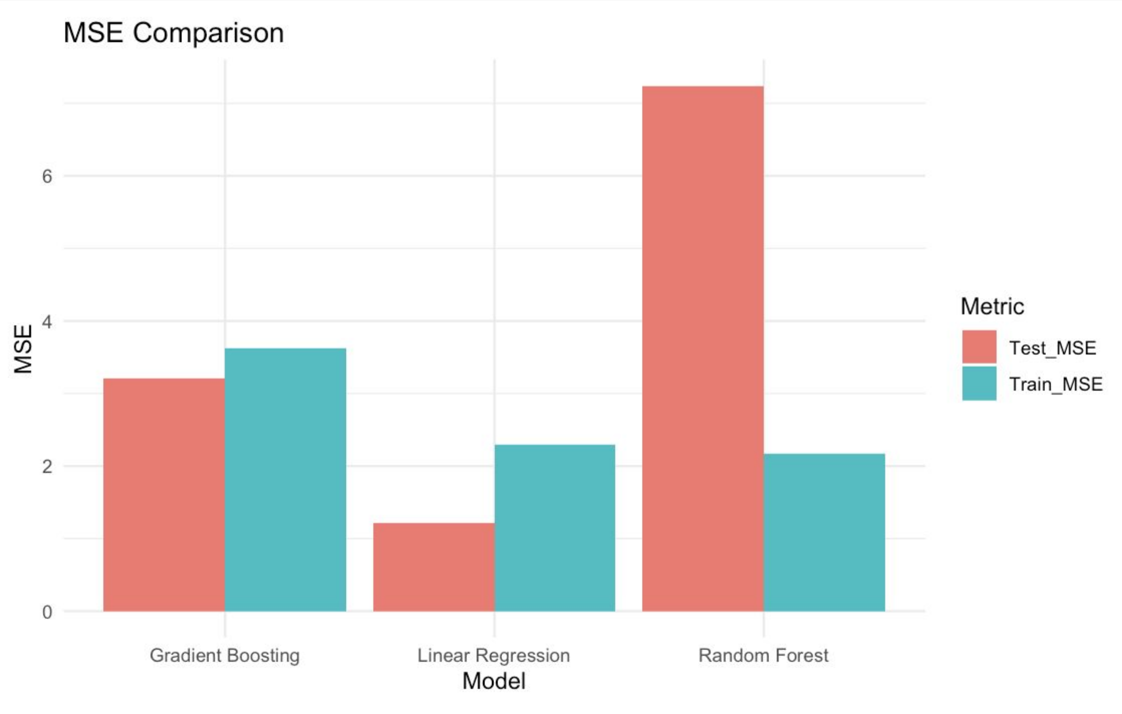
MODEL SELECTION

We performed three models (Linear Regession, Random Forest, Gradient Boosting) using 10-fold validation.

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

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

MSE COMPARISON

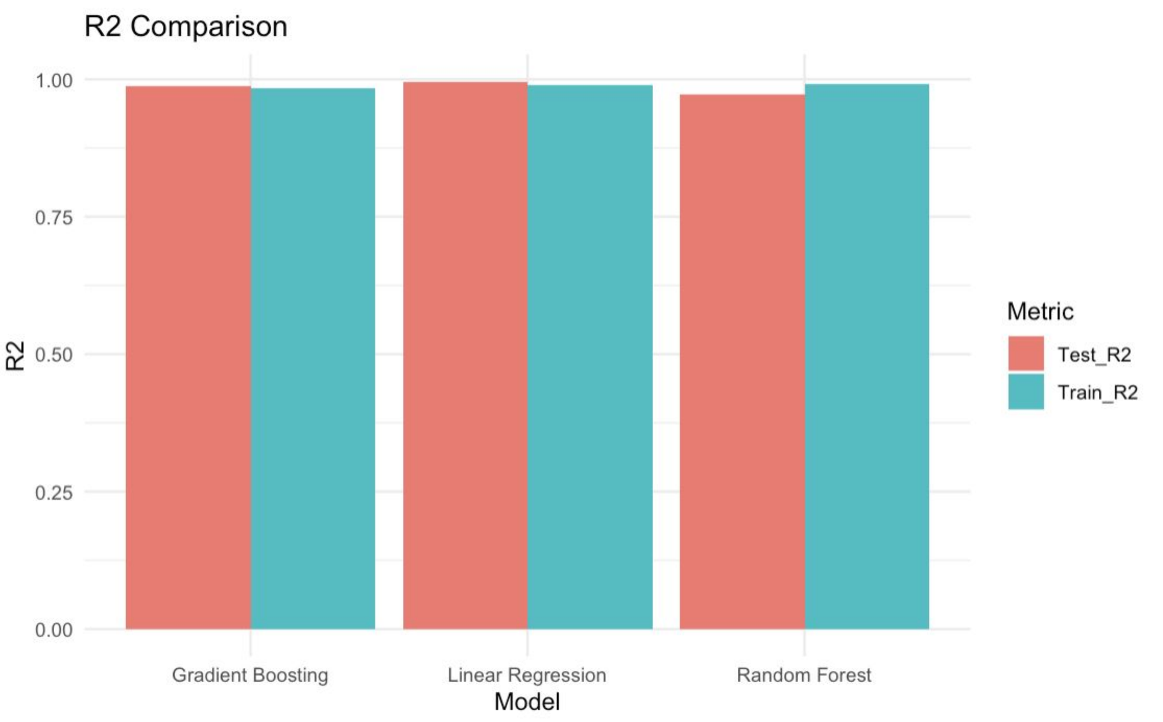


MSE Comparison of the three models with 10-fold cross-validation:

1. Gradient Boosting
2. Linear Regression
3. Random Forest

According to the MSE of the train and test dataset, Linear Regression appears to the best because of its lowest MSE, which indicates that the model's predictions are very close to the actual values, meaning the model has high accuracy in its predictions.

R SQARED COMPARISON

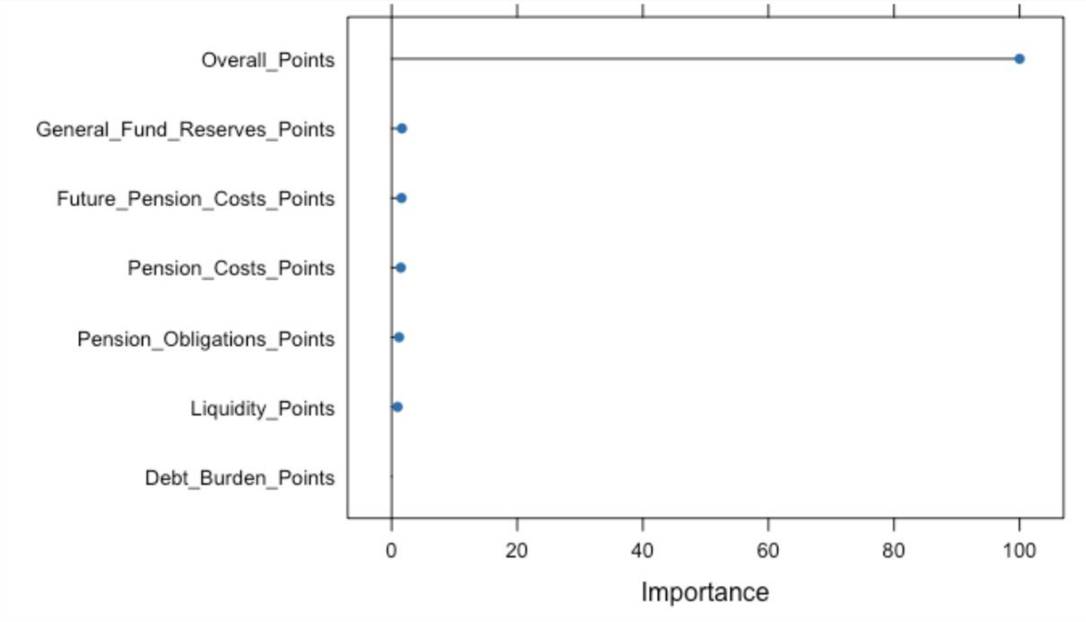


R^2 Comparison of the three models with 10-fold cross-validation:

1. Gradient Boosting
2. Linear Regression
3. Random Forest

According to the R^2 of the train and test dataset, Linear Regression appears to the best because of its highest R^2 for test dataset, which suggests that the model captures a large proportion of the variance in the target variable and that the model fits the data well.

MODEL SELECTION



The output will show the relative importance of each predictor in the model. Variables with higher importance scores contribute more to predicting Overall_Rank.

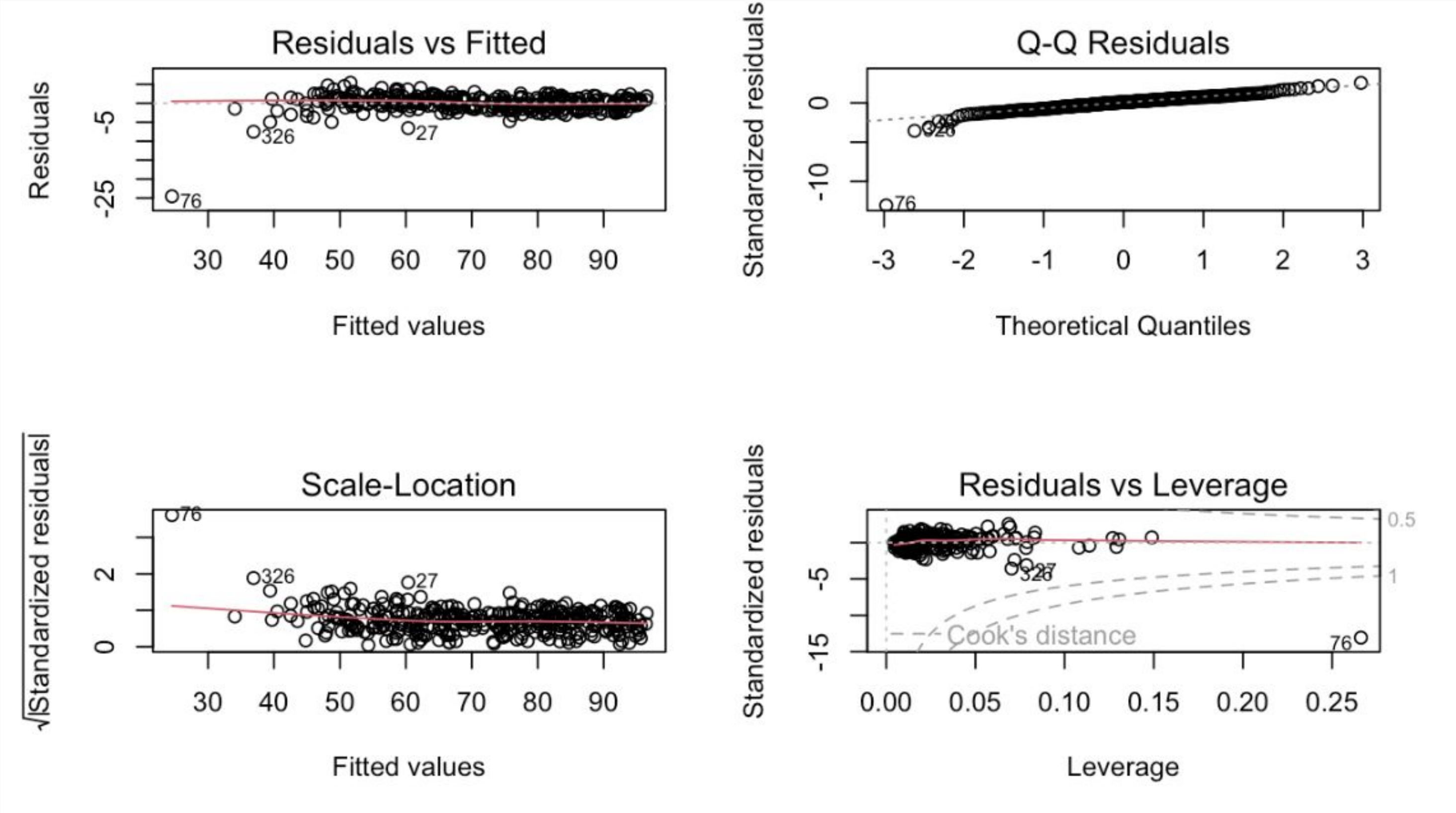
Both **backward** and **forward selection** shows that all predictors should be included in the model.

VARIABLE SELECTION & SIGNIFICANCE OF PREDICTORS

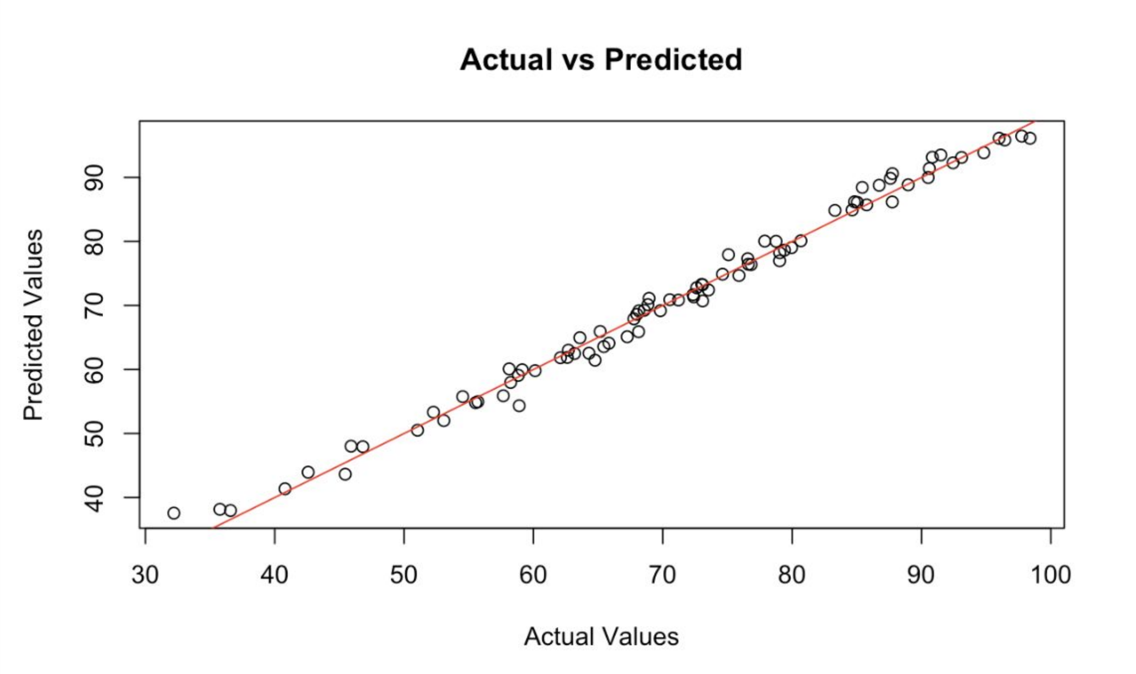
```
Call:
lm(formula = .outcome ~ ., data = dat)

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**.

DIAGNOSTIC AND RESIDUAL ANALYSIS



VARIABLE SELECTION & SIGNIFICANCE OF PREDICTORS



The points in the plot align closely with the diagonal red line (y=x), which indicates that the **model's predictions are highly accurate for most observations**.

The closer the points are to the red line, the smaller the errors between actual and predicted values.

ANOVA TABLE (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
```

Null Hypothesis: For all indicators, the regression coefficient (βi) is equal to zero, which means that our chosen metrics have no significant impact on the overall points

Alternative Hypothesis: Our chosen metrics have a significant impact on the overall points.

The ANOVA results show that the full model, including predictors such as General Fund Reserves Points and Pension Obligations Points, significantly improves the explanation of Overall_Points compared to the null model ($p<2.2e-16$). This indicates that the included predictors are highly relevant in explaining the variability in Overall Points.

MODEL EQUATION

$$
\begin{aligned}
\text{Predicted Value (Overall Points)} = 13.086 \\
&+ 0.989 \cdot \text{General Fund Reserves Points} \\
&+ 1.074 \cdot \text{Pension Obligations Points} \\
&+ 1.024 \cdot \text{Debt Burden Points} \\
&+ 1.266 \cdot \text{Future Pension Costs Points} \\
&+ 1.152 \cdot \text{Liquidity Points} \\
&+ 1.120 \cdot \text{Pension Costs Points} \\
&+ 1.113 \cdot \text{OPEB Funding Points}
\end{aligned}
$$

04.

CONCLUSION

RIGIONAL ANALYSIS

The density plots and proportion plots illustrate differences in the distribution of city risk levels across regions. While low- and moderate-risk cities predominate in each region, the southern region exhibits a notably higher proportion of high-risk cities compared to other regions. However, despite these differences, the average overall points are nearly identical across regions. Furthermore, hypothesis testing confirms that region is not a statistically significant variable influencing overall points.

CITY ANALYSIS

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.

CONCLUSION OF MODELING

In the modeling phase, we tested three models (Linear Regression, Random Forest, and Gradient Boosting) using 10-fold cross-validation. We figured out the Linear Regression model as the most robust model with lowest Mean Squared Error (MSE) and highest R^2 on both training and test datasets. The Linear Regression model accurately captures variance in the target variable and offers reliable predictions using the 7 metrics we selected. All predictors were confirmed to be statistically significant when confidence level = 0.05, and both forward and backward selection methods validated their inclusion.

VIEW OF THE RESULTS

The modeling results underscore the importance of all selected predictors in determining fiscal health, highlighting their collective impact on the overall points. The reliance on Linear Regression, supported by validation metrics, emphasizes the stability and interpretability of this approach for policy-oriented applications. This comprehensive analysis could help policymakers and stakeholders with actionable insights to address regional disparities and enhance the fiscal health of California cities. Future research can build on these findings by incorporating temporal data to examine trends over time or expanding the scope to include additional predictors for a more nuanced understanding of fiscal dynamics.


THANK YOU!!