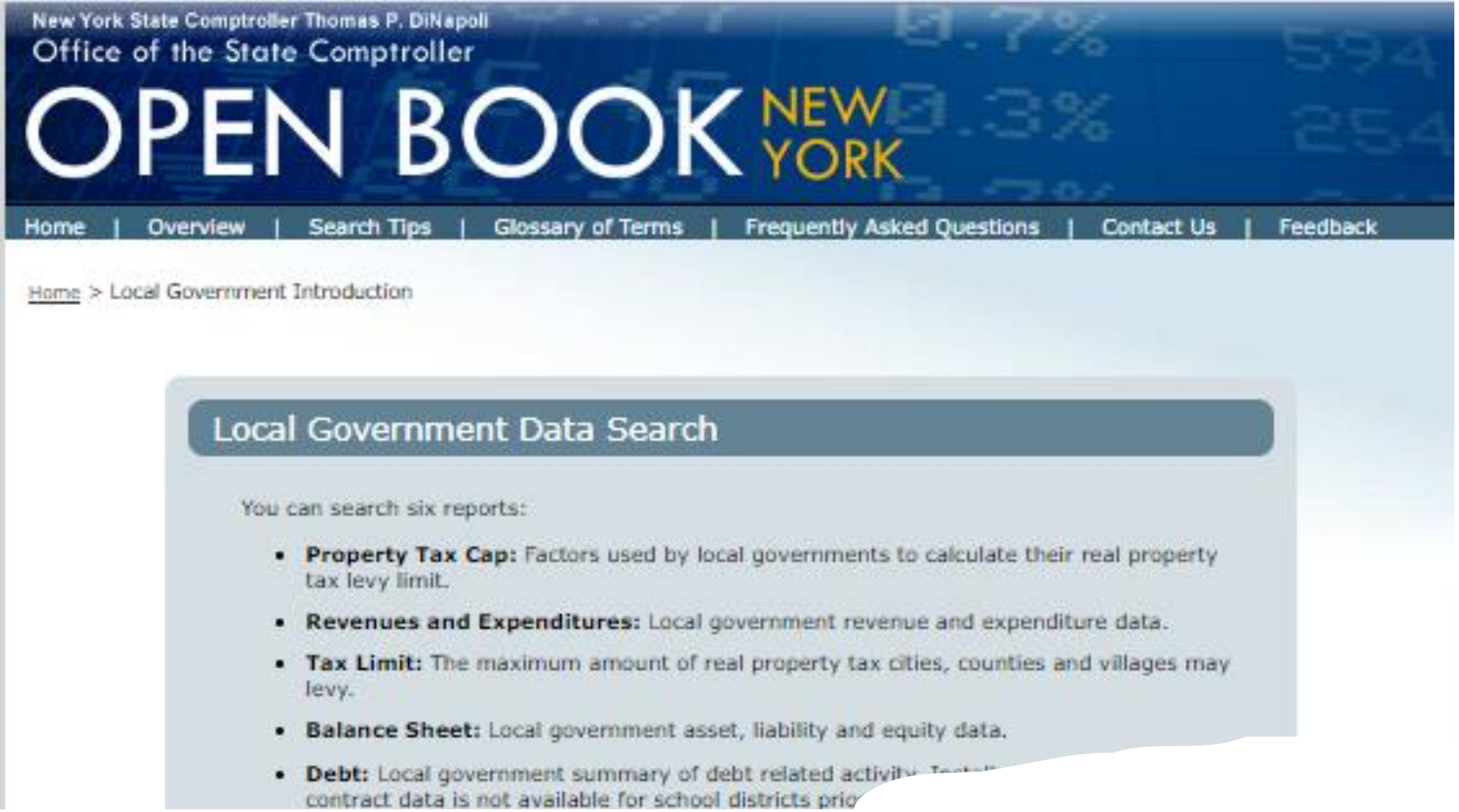


"Analyzing the association between Socioeconomic Profiles and Fiscal Health of New York State Municipalities"

THE WALL STREET JOURNAL.

Latest World Business U.S. Politics Economy Tech Finance Opinion Arts & Culture Lifestyle Real Estat

Understanding China
News & analysis

Great Power Conflict Tensions With U.S. Taiwan Relations China Endgame Controlling Companies

WORLD | ASIA | CHINA

China's Cities Struggle Under Trillions of Dollars of Debt

Financial obligations built up during pandemic weigh on growth as China's legislature meets to address economic needs

By Stella Yifan Xie Follow , *Yoko Kubota* Follow and *Cao Li* Follow
March 6, 2023 10:00 am ET

Michigan Advance

On this day in 2013: The city of Detroit files for bankruptcy • Michigan Advance

On July 18, 2013, the city of Detroit filed for Chapter 9 bankruptcy. It was the largest municipal bankruptcy filing in U.S. history.

introduction:

After the pandemic, many local governments in China find themselves on the brink of a financial crisis that is disturbingly similar to the collapse of Detroit. This precarious situation highlights the urgent need for forecasting strategies that can preemptively signal financial distress.

Study Area:

Why do I choose town and city instead of county?

Local government finance: the government of COUNTY does not control specific local affairs. (County data not available)

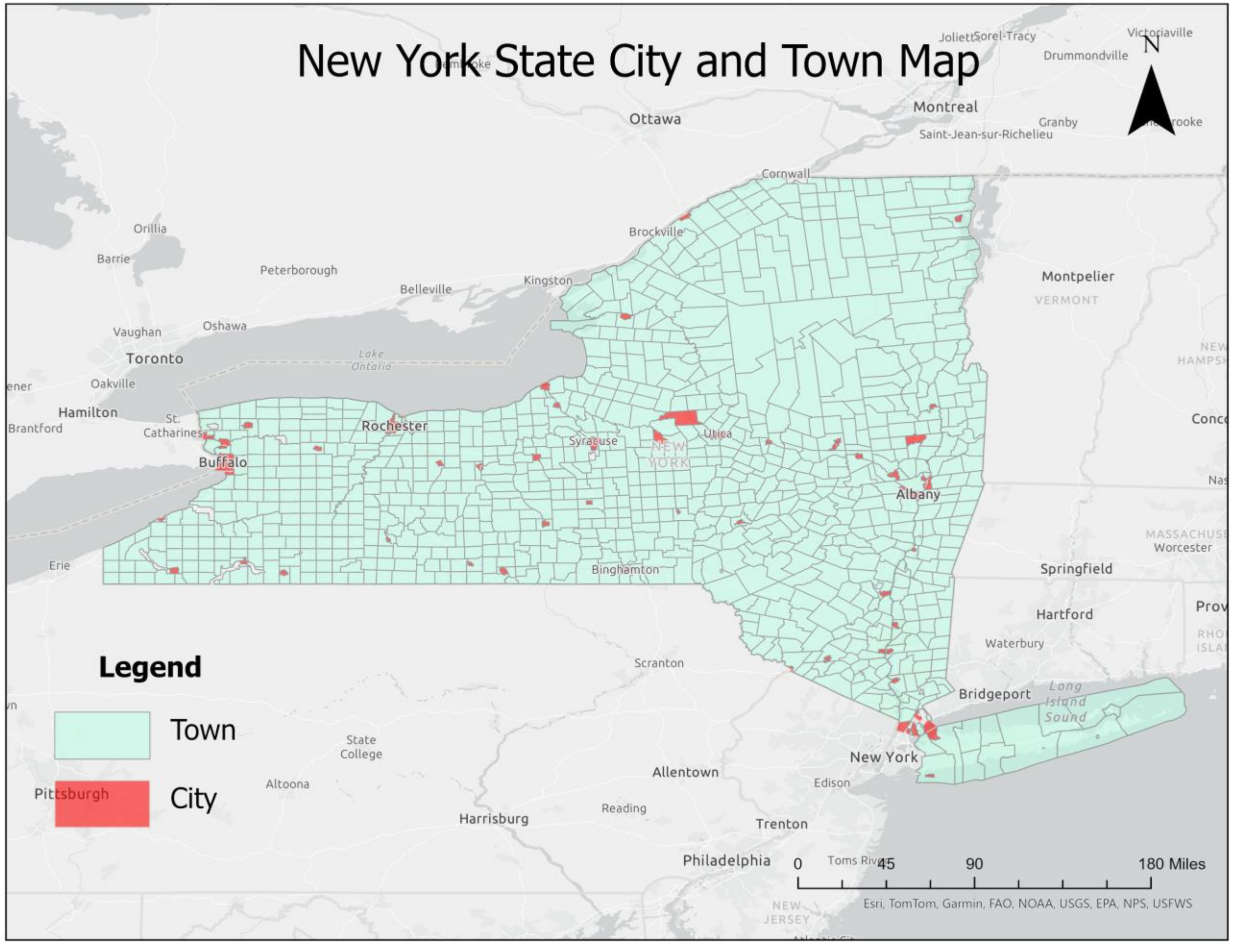
For city data on government revenues and expenditures: **19998**rows

Government revenue and expenditure data for towns: **187332**rows

After clean:
There are a total of **53** Cities
And total of **848** towns

CITY and TOWN can cover most of New York State (excluding New York City)

Research Area



Fiscal Stress Monitoring System 2022 List and Details

Name	Class	County	Municode	Region	Fiscal Year End	Type of Stress	Fiscal Sco	Environmental Rating	Environmental Score	Snapshot Date
Amsterdam	City	Montgomery	270202000000	Mohawk Valley	06/30	No Designation	1.7	No Designation	26.7	2/28/2023
Auburn	City	Cayuga	050203000000	Central New York	06/30	No Designation	23.8	No Designation	23.3	2/28/2023
Batavia	City	Genesee	180204000000	Finger Lakes	03/31	No Designation	3.3	No Designation	0.0	2/28/2023
Buffalo	City	Erie	140207000000	Western New York	06/30	No Designation	6.3	Moderate Environmental Stress	43.3	2/28/2023
Corning	City	Steuben	460210000000	Southern Tier	06/30	No Designation	0.0	No Designation	3.3	2/28/2023
Hornell	City	Steuben	460219000000	Southern Tier	03/31	No Designation	0.0	Moderate Environmental Stress	43.3	2/28/2023
Lackawanna	City	Erie	140225000000	Western New York	07/31	***Not filed***				2/28/2023
Long Beach	City	Nassau	280228000000	Long Island	06/30	No Designation	20.8	No Designation	0.0	2/28/2023
Olean	City	Cattaraugus	040239000000	Western New York	05/31	***Not filed***				2/28/2023
Rensselaer	City	Rensselaer	380247000000	Capital District	07/31	***Not filed***				2/28/2023
Rochester	City	Monroe	260248000000	Finger Lakes	06/30	No Designation	0.0	Susceptible Environmental Stress	33.3	2/28/2023
Salamanca	City	Cattaraugus	040251000000	Western New York	03/31	***Not filed***				2/28/2023
Syracuse	City	Onondaga	310255000000	Central New York	06/30	***Not filed***				2/28/2023
Utica	City	Oneida	300258000000	Mohawk Valley	03/31	No Designation	0.0	Moderate Environmental Stress	43.3	2/28/2023
Watertown	City	Jefferson	220259000000	North Country	06/30	No Designation	0.0	Susceptible Environmental Stress	33.3	2/28/2023
White Plains	City	Westchester	550261000000	Mid-Hudson	06/30	No Designation	0.0	No Designation	6.7	2/28/2023
Yonkers	City	Westchester	550262000000	Mid-Hudson	06/30	No Designation	5.0	No Designation	13.3	2/28/2023
Adams	Village	Jefferson	220400100010	North Country	05/31	No Designation	3.3	Moderate Environmental Stress	40.0	2/28/2023
Addison	Village	Steuben	460400200020	Southern Tier	05/31	***Not filed***				2/28/2023
Afton	Village	Chenango	080400300030	Southern Tier	05/31	No Designation	0.0	Moderate Environmental Stress	40.0	2/28/2023

Summary | Financial Scoring | Indicator 1 | Indicator 2 | Indicator 3 | Indicator 4 | Indicator 5 | Indicator 6 | Indicator 7 | Indicator 8

Type of Stress
No Designation
No Designation
No Designation
No Designation
No Designation
No Designation
Not filed
No Designation
Not filed
Not filed
No Designation
Not filed
Not filed
No Designation
No Designation
No Designation
No Designation
No Designation
Not filed
No Designation

New York State Comptroller
THOMAS P. DiNAPOLI

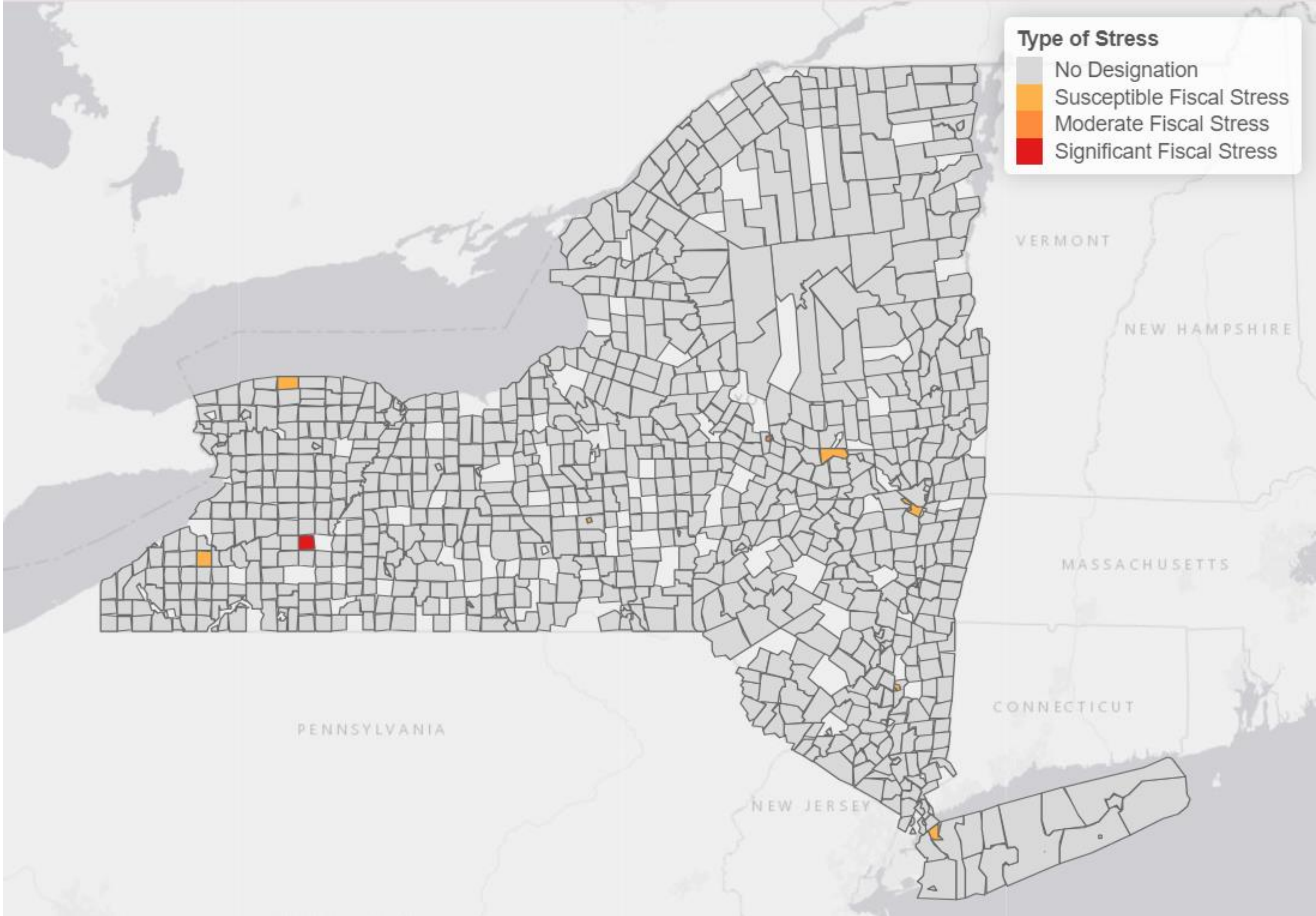
Fiscal Stress Monitoring System®

Municipalities in Stress Fiscal Years Ending 2022

Type of Stress:

- Susceptible Fiscal Stress
- No Designation
- Moderate Fiscal Stress
- Significant Fiscal Stress

That's what I'm trying to predict.



We can see the data imbalance in this chart, but it will be fixed afterward

My "X" data (data used for forecasting)

Socio-economic Data part 1

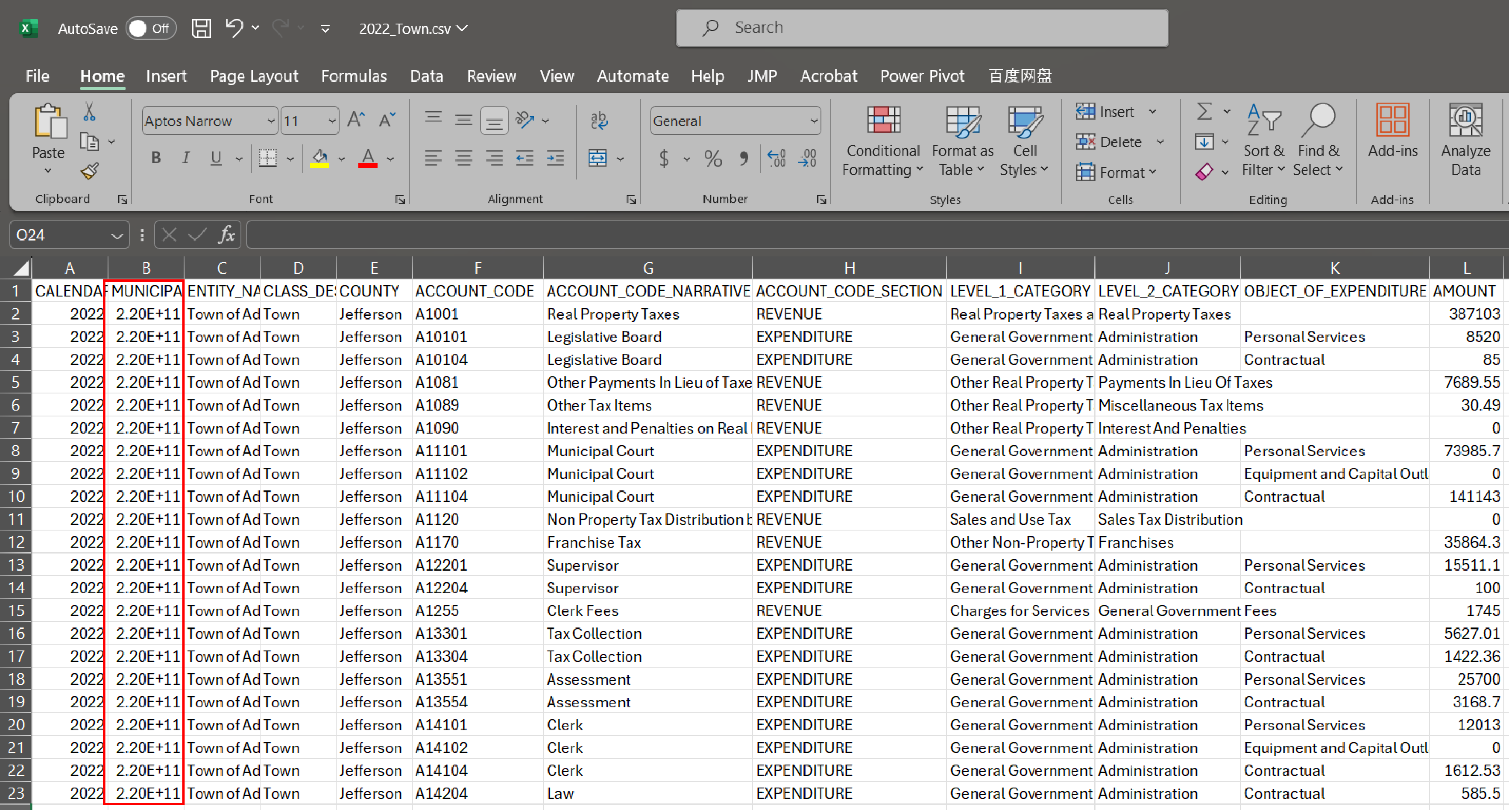
Socio-economic Data Part 2

There are a total of **40** columns of data, including **31** columns of socio-economic data and **9** columns of data on government expenditures and revenues.

	Columns:	Description:
2	ENTITY_NAME	city, county, or other governmental unit.
3	SUB_GOV_TYPE	The subtype of government the entity falls under, such as municipal, county, state, etc.
4	COUNTY	The county in which the entity is located.
5	Type of Stress	Types of financial pressures on Governments
6	GEOID	A geographic identifier
7	TOT_POP	Total population of the entity
8	GINI	Gini coefficient
9	PCT_HHwKID	Percentage of households with kids
10	AVG_HHSIZE	Average household size
11	PCT_HSGRAD	Percentage of residents with a high school diploma
12	PCT_COLLGRAD	Percentage of residents with a college degree
13	PCT_VETERAN	Percentage of residents who are veterans
14	PCT_DISABILITY	Percentage of residents with disabilities
15	PCT_FOREIGN	Percentage of residents who are foreign-born
16	PCT_UNEMPLOYED	Unemployment rate
17	PCT_SERVICEJOB	Percentage of jobs in the service sector
18	PCT_MANAGEJOB	Percentage of jobs in management
19	MED_HHINCOME	Median household income
20	PCT_PUBASSIST	Percentage of residents receiving public assistance
21	PCT_NOHEALTHINS	Percentage of residents without health insurance
22	PCT_BELOWPOV	Percentage of residents below the poverty line
23	PCT_VACANTHOUSI	Percentage of vacant housing units
24	PCT_RENTER	Percentage of residents who rent their home
25	PCT_NOVEHICLE	Percentage of households without a vehicle
26	MED_HVALUE	Median home value
27	RATIO_SEX	Ratio of males to females
28	PCT_UNDER18	Percentage of the population under 18 years of age
29	PCT_OVER65	Percentage of the population over 65 years of age
30	PCT_BLACK	Percentage of the population that is Black or African American
31	PCT_ASIAN	Percentage of the population that is Asian
32	PCT_LATINO	Percentage of the population that is Latino or Hispanic
33	TRANSFER_REV_RAT	The ratio of transfer revenue (could refer to funds received from higher levels of government or other e
34	TAX_REV_RATIO	The ratio of tax revenue to some other financial metric.
35	AID_REV_RATIO	The ratio of aid revenue, likely funds received from external sources like federal or state aid.
36	SALES_REV_RATIO	The ratio of sales revenue, possibly in the context of sales tax.
37	PERSONNEL_EXP_R	The ratio of personnel expenses, likely the cost of staff and employees.
38	PERSONNEL_REV_R	The ratio of revenue to personnel costs.
39	EXP_RIGIDITY	A measure of how inflexible the spending is, indicating mandatory versus discretionary spending.
40	CASH_EXP_RATIO	The ratio of cash expenditures.
41	OP_DEFICIT	Operational deficit, which indicates spending exceeding revenue

CALENDAF	MUNICIPA	ENTITY_NA	CLASS_DE:	COUNTY	ACCOUNT_CODE_NARRATIVE	ACCOUNT_CODE_SECTION	LEVEL_1_CATEGORY	LEVEL_2_CATEGORY	OBJECT_OF_EXPENDITURE	AMOUNT
2022	1.02E+10	City of Alba	City	Albany	Real Property Taxes	REVENUE	Real Property Taxes a	Real Property Taxes		59597300
2022	1.02E+10	City of Alba	City	Albany	Legislative Board	EXPENDITURE	General Government	Administration	Personal Services	478117
2022	1.02E+10	City of Alba	City	Albany	Legislative Board	EXPENDITURE	General Government	Administration	Contractual	29221
2022	1.02E+10	City of Alba	City	Albany	Legislative Board	EXPENDITURE	Employee Benefits	Unclassified Employe	Employee Benefits	120098
2022	1.02E+10	City of Alba	City	Albany	Other Payments In Lieu of Taxe	REVENUE	Other Real Property T	Payments In Lieu Of Taxes		19279890
2022	1.02E+10	City of Alba	City	Albany	Interest and Penalties on Real I	REVENUE	Other Real Property T	Interest And Penalties		366304
2022	1.02E+10	City of Alba	City	Albany	Non Property Tax Distribution b	REVENUE	Sales and Use Tax	Sales Tax Distribution		44862476
2022	1.02E+10	City of Alba	City	Albany	Utilities Gross Receipts Tax	REVENUE	Sales and Use Tax	Utilities Gross Receipts Tax		1589587
2022	1.02E+10	City of Alba	City	Albany	Privilege Tax on Coin Operated	REVENUE	Sales and Use Tax	Miscellaneous Use Taxes		210
2022	1.02E+10	City of Alba	City	Albany	OTB Surtax	REVENUE	Other Non-Property T	Miscellaneous Non-Property Taxes		141073
2022	1.02E+10	City of Alba	City	Albany	Franchise Tax	REVENUE	Other Non-Property T	Franchises		1165521
2022	1.02E+10	City of Alba	City	Albany	Mayor	EXPENDITURE	General Government	Administration	Personal Services	534413
2022	1.02E+10	City of Alba	City	Albany	Mayor	EXPENDITURE	General Government	Administration	Contractual	69824
2022	1.02E+10	City of Alba	City	Albany	Mayor	EXPENDITURE	Employee Benefits	Unclassified Employe	Employee Benefits	107550
2022	1.02E+10	City of Alba	City	Albany	Treasurer Fees	REVENUE	Charges for Services	General Government Fees		0
2022	1.02E+10	City of Alba	City	Albany	Clerk Fees	REVENUE	Charges for Services	General Government Fees		14164
2022	1.02E+10	City of Alba	City	Albany	Other General Departmental In	REVENUE	Charges for Services	General Government Fees		583258
2022	1.02E+10	City of Alba	City	Albany	Auditor	EXPENDITURE	General Government	Administration	Personal Services	359544

Socio-economic Data Part 2-City



	A	B	C	D	E	F	G	H	I	J	K	L
1	CALENDAI	MUNICIPA	ENTITY_NA	CLASS_DE	COUNTY	ACCOUNT_CODE	ACCOUNT_CODE_NARRATIVE	ACCOUNT_CODE_SECTION	LEVEL_1_CATEGORY	LEVEL_2_CATEGORY	OBJECT_OF_EXPENDITURE	AMOUNT
2	2022	2.20E+11	Town of Ad	Town	Jefferson	A1001	Real Property Taxes	REVENUE	Real Property Taxes a	Real Property Taxes		387103
3	2022	2.20E+11	Town of Ad	Town	Jefferson	A10101	Legislative Board	EXPENDITURE	General Government	Administration	Personal Services	8520
4	2022	2.20E+11	Town of Ad	Town	Jefferson	A10104	Legislative Board	EXPENDITURE	General Government	Administration	Contractual	85
5	2022	2.20E+11	Town of Ad	Town	Jefferson	A1081	Other Payments In Lieu of Taxe	REVENUE	Other Real Property T	Payments In Lieu Of Taxes		7689.55
6	2022	2.20E+11	Town of Ad	Town	Jefferson	A1089	Other Tax Items	REVENUE	Other Real Property T	Miscellaneous Tax Items		30.49
7	2022	2.20E+11	Town of Ad	Town	Jefferson	A1090	Interest and Penalties on Real	REVENUE	Other Real Property T	Interest And Penalties		0
8	2022	2.20E+11	Town of Ad	Town	Jefferson	A11101	Municipal Court	EXPENDITURE	General Government	Administration	Personal Services	73985.7
9	2022	2.20E+11	Town of Ad	Town	Jefferson	A11102	Municipal Court	EXPENDITURE	General Government	Administration	Equipment and Capital Outl	0
10	2022	2.20E+11	Town of Ad	Town	Jefferson	A11104	Municipal Court	EXPENDITURE	General Government	Administration	Contractual	141143
11	2022	2.20E+11	Town of Ad	Town	Jefferson	A1120	Non Property Tax Distribution b	REVENUE	Sales and Use Tax	Sales Tax Distribution		0
12	2022	2.20E+11	Town of Ad	Town	Jefferson	A1170	Franchise Tax	REVENUE	Other Non-Property T	Franchises		35864.3
13	2022	2.20E+11	Town of Ad	Town	Jefferson	A12201	Supervisor	EXPENDITURE	General Government	Administration	Personal Services	15511.1
14	2022	2.20E+11	Town of Ad	Town	Jefferson	A12204	Supervisor	EXPENDITURE	General Government	Administration	Contractual	100
15	2022	2.20E+11	Town of Ad	Town	Jefferson	A1255	Clerk Fees	REVENUE	Charges for Services	General Government Fees		1745
16	2022	2.20E+11	Town of Ad	Town	Jefferson	A13301	Tax Collection	EXPENDITURE	General Government	Administration	Personal Services	5627.01
17	2022	2.20E+11	Town of Ad	Town	Jefferson	A13304	Tax Collection	EXPENDITURE	General Government	Administration	Contractual	1422.36
18	2022	2.20E+11	Town of Ad	Town	Jefferson	A13551	Assessment	EXPENDITURE	General Government	Administration	Personal Services	25700
19	2022	2.20E+11	Town of Ad	Town	Jefferson	A13554	Assessment	EXPENDITURE	General Government	Administration	Contractual	3168.7
20	2022	2.20E+11	Town of Ad	Town	Jefferson	A14101	Clerk	EXPENDITURE	General Government	Administration	Personal Services	12013
21	2022	2.20E+11	Town of Ad	Town	Jefferson	A14102	Clerk	EXPENDITURE	General Government	Administration	Equipment and Capital Outl	0
22	2022	2.20E+11	Town of Ad	Town	Jefferson	A14104	Clerk	EXPENDITURE	General Government	Administration	Contractual	1612.53
23	2022	2.20E+11	Town of Ad	Town	Jefferson	A14204	Law	EXPENDITURE	General Government	Administration	Contractual	585.5

Socio-economic Data Part 2-Town

Table A1
Data Description.

Financial indices	Var Name
Taxation autonomy degree: ETR / (ETR + ETS + EET). Where ETR = tributary revenues (assessments); ETS = current revenues from contributions and transfers (assessments); EET = extra-tributary revenues (assessments).	FIN Autonom 1
Degree of dependence on contributions and current transfers: ETS / (ETR + ETS + EET). Where ETR = tributary revenues (assessments); ETS = current revenues from contributions and transfers (assessments); EET = extra-tributary revenues (assessments).	FIN Autonom 2
~~Internal borrowing degree: (EET + RC + AP) / ET. Where EET = extra-tributary revenues (assessments); RC = revenues from credits (assessments); AP = sale of assets (assessments); ET = revenues (assessments).~~	~~FIN Autonom 3~~
~~Financial autonomy degree: (ETR + EET) / ET. Where ETR = tributary revenues (assessments); EET = extra-tributary revenues (assessments); ET=revenues (assessments).~~	FIN Autonom 4
Expenditure rigidity: (SP + RP) / (ETR + ETS + EET). Where SP = current expenditures on personnel (commitments); RP = loans payment (commitments); ETR = tributary revenues (assessments); ETS = current revenues from contributions and transfers (assessments); EET = extra-tributary revenues (assessments).	FIN EXP Rigit 1
Personnel expenditures / current expenditures: SP / SC. Where SP = current expenditures on personnel (commitments); SC = current expenditures (commitments).	FIN EXP Personal 1
~~External borrowing dependency. EP / ET. Where~~ EP = revenues from loans (assessments)ET = revenues (assessments).	FIN Autonom 5
Personnel expenditures / current revenues: SP / (ETR + ETS + EET). Where SP = current expenditures on personnel (commitments); ETR = tributary revenues (assessments); ETS = current revenues from contributions and transfers (assessments); EET = extra-tributary revenues (assessments).	FIN EXP Personal 2
Loan repayment expenditures / current revenues: RP / (ETCP + ETC + EET). Where RP = loans payment (commitments); ETCP = current revenues from taxes, contributions and equalizations (assessments); ETC = current transfers (assessments); EET = extra-tributary revenues (assessments)	FIN EXP Debt 1
Index of beginning debt residuals: RPI / ST. Where RPI = initial residual liabilities; ST = expenditures (commitments).	FIN Consistency 1
~~Financial indices~~	~~Var Name~~
Index of final debt residuals: Where RPF / STRPF = final residual liabilitiesST = expenditures (commitments).	FIN Consistency 2
Coverage of current and loan repayment expenditures: (ETR + ETS + EET) / (SC + RMP + RPO + RDP). Where ETR = tributary revenues (assessments); ETS = current revenues from contributions and transfers (assessments); EET = extra-tributary revenues (assessments); SC = current expenditures (commitments); RMP = repayment of capital share of mortgages and loans (commitments); RPO = repayment of bonds (commitments); RDP = repayment of capital share of long term debts (commitments).	FIN Autonom 6
Capital expenses financed by loans: (EMP + EBOC) / SCC. Where EMP = revenues from loans and mortgages (commitments); EBOC = revenues from bond issues (commitments)SCC = capital expenditures (commitments)	FIN EXP Debt 2
Collecting capacity: ETcc / ET. Where ETcc = revenues (collections in accrual accounting); ET = revenues (assessments).	FIN Autonom 7
Expenditure capacity: STcc / ST. Where STcc = expenditures (payments in accrual accounting); ST = expenditures (commitments).	FIN Autonom 8
Residual liabilities accumulation: RPC / RPI. Where RPC = residual liabilities from accrual accounting; RPI = initial residual liabilities.	FIN EXP Debt 3
Residual liabilities disposal: RPP / RPI. Where RPP = paid residual liabilities; RPI = initial residual liabilities.	FIN EXP Debt 4
Administration result / current revenues: RA / (ETR + ETS + EET). Where RA = administration result; ETR = tributary revenues (assessments); ETS = current revenues from contributions and transfers (assessments); EET = extra-tributary revenues (assessments).	FIN Deficit 1
Extra-budgetary debts / current revenues: DFB / (ETR + ETS + EET). Where DFB = extra-budgetary debts; ETR = tributary revenues (assessments); ETS = current revenues from contributions and transfers (assessments); EET = extra-tributary revenues (assessments).	FIN Deficit 2
Financing debts variation: Where DFF / DFI; DFF = final financing debts; DFI = initial financing debts.	FIN Deficit 3
Current transfers / current expenditures: TR / SC. Where TR = current transfers (commitments); SC = current expenditures (commitments).	FIN Tranfers 1
Capital transfers / capital expenses: TRCC / SCC. TRCC = capital transfers (commitments); SCC = capital expenditures (commitments).	FIN Tranfers 2
Financial indices	**Var Name**
~~Financial flows / personnel current expenditures: (ET + ST) / SP. Where ET = revenues (assessments); ST = expenditures (commitments); SP = current expenditures on personnel (commitments).~~	~~FIN Tranfers 3~~
Sale of assets / current expenditures (percentage values): AP / SC * 100. Where AP = sale of assets (assessments); SC = current expenditures (commitments).	FIN Sells 1
Disposal / accumulation of debt residuals: RPP / RPC. Where RPP = paid residual liabilities; RPC = residual liabilities from accrual accounting.	FIN Tranfers 4
External expenditures / financial resources: (SC + SCC - SCagc - SCCagc) / ET. Where SC = current expenditures (commitments); SCC = capital expenditures (commitments); SCagc = current expenditures for administration, management and control (commitments); SCCagc = capital expenditures for administration, management and control (commitments); ET = revenues (assessments).	FIN Tranfers 5

(*continued on next page*)

Reference:

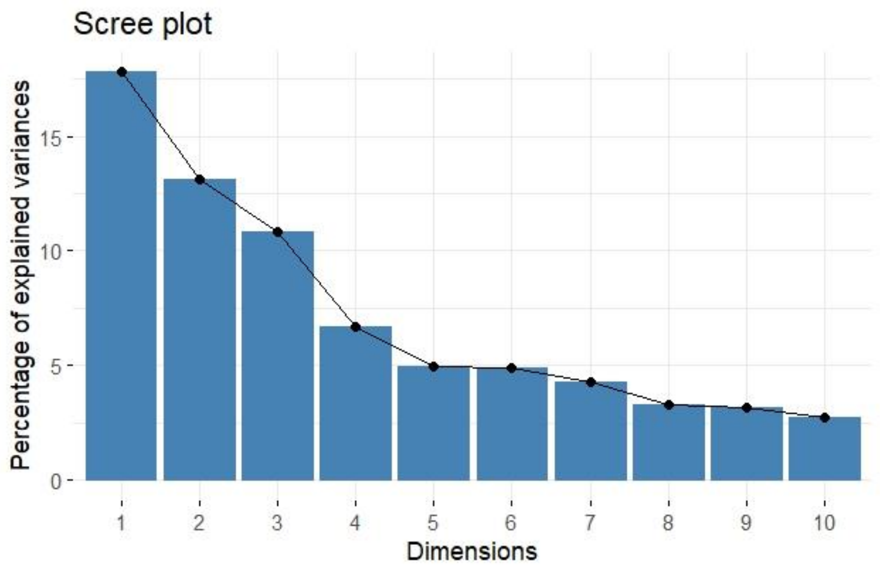
Antulov-Fantulin, N., Lagravinese, R., & Resce, G. (2021).
Predicting bankruptcy of local government:
A machine learning approach.
Journal of Economic Behavior & Organization, 183, 681-699.
https://doi.org/10.1016/j.jebo.2021.01.014

Andini, M., Ciani, E., de Blasio, G., D'Ignazio, A., & Salvestrini, V. (2018). Targeting with machine learning:
An application to a tax rebate program in Italy.
Journal of Economic Behavior & Organization, 156, 86-102.
https://doi.org/10.1016/j.jebo.2018.09.010

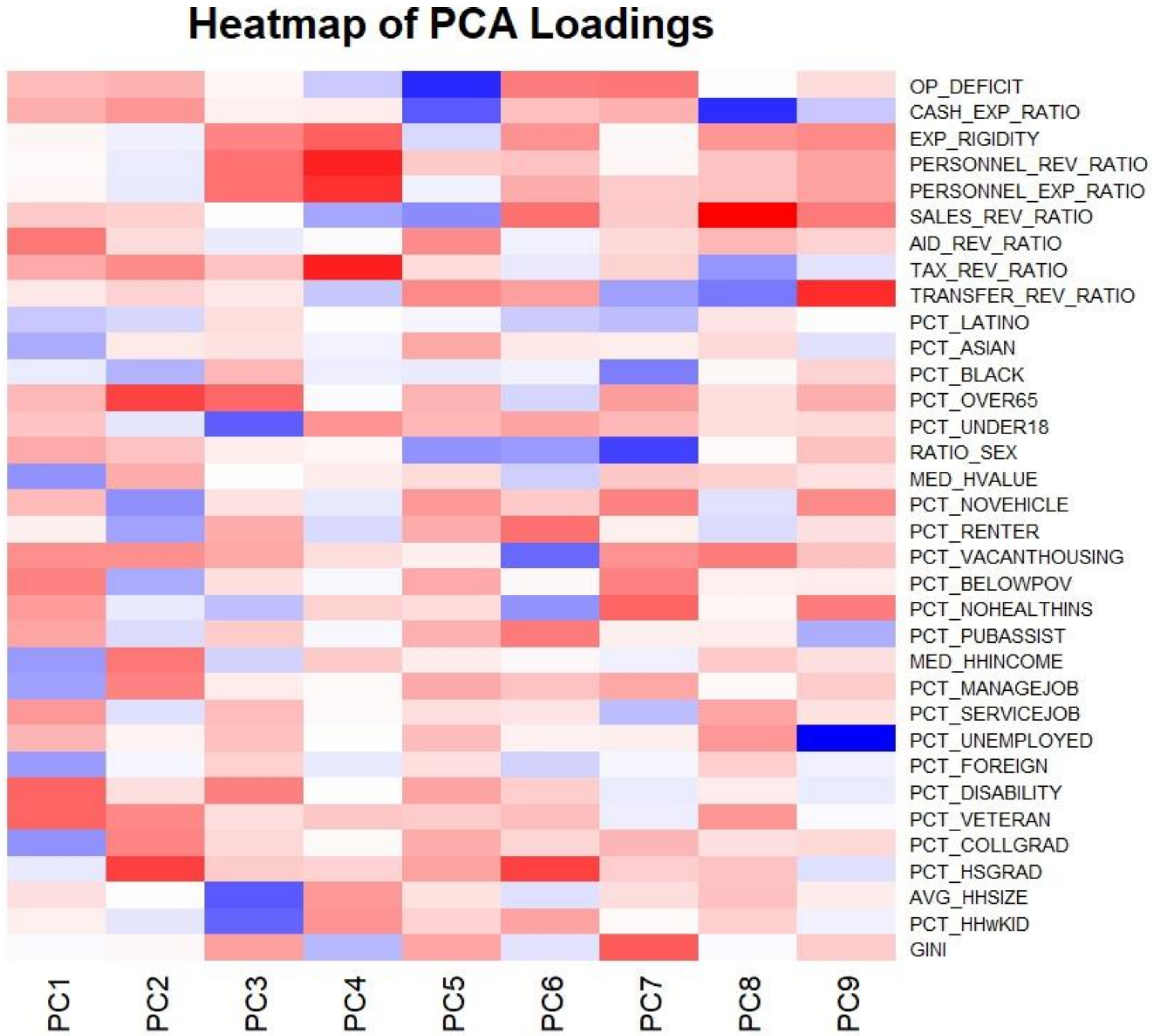
Eigenvalue

```
 [1] 2.4605303 2.1098002 1.9175070 1.5056505 1.2966833 1.2922117 1.2017656 1.0592413 1.0330939 0.9652547 0.9319844
[12] 0.9027830 0.8840435 0.8538179 0.8312294 0.8010810 0.7378520 0.7221161 0.6692395 0.6583851 0.6175670 0.6074880
[23] 0.5911344 0.5627623 0.5484907 0.5289529 0.5070013 0.4797227 0.4532001 0.3722642 0.3410412 0.3289154 0.2918900
[34] 0.1489258
> # So we chooose 9 components
```

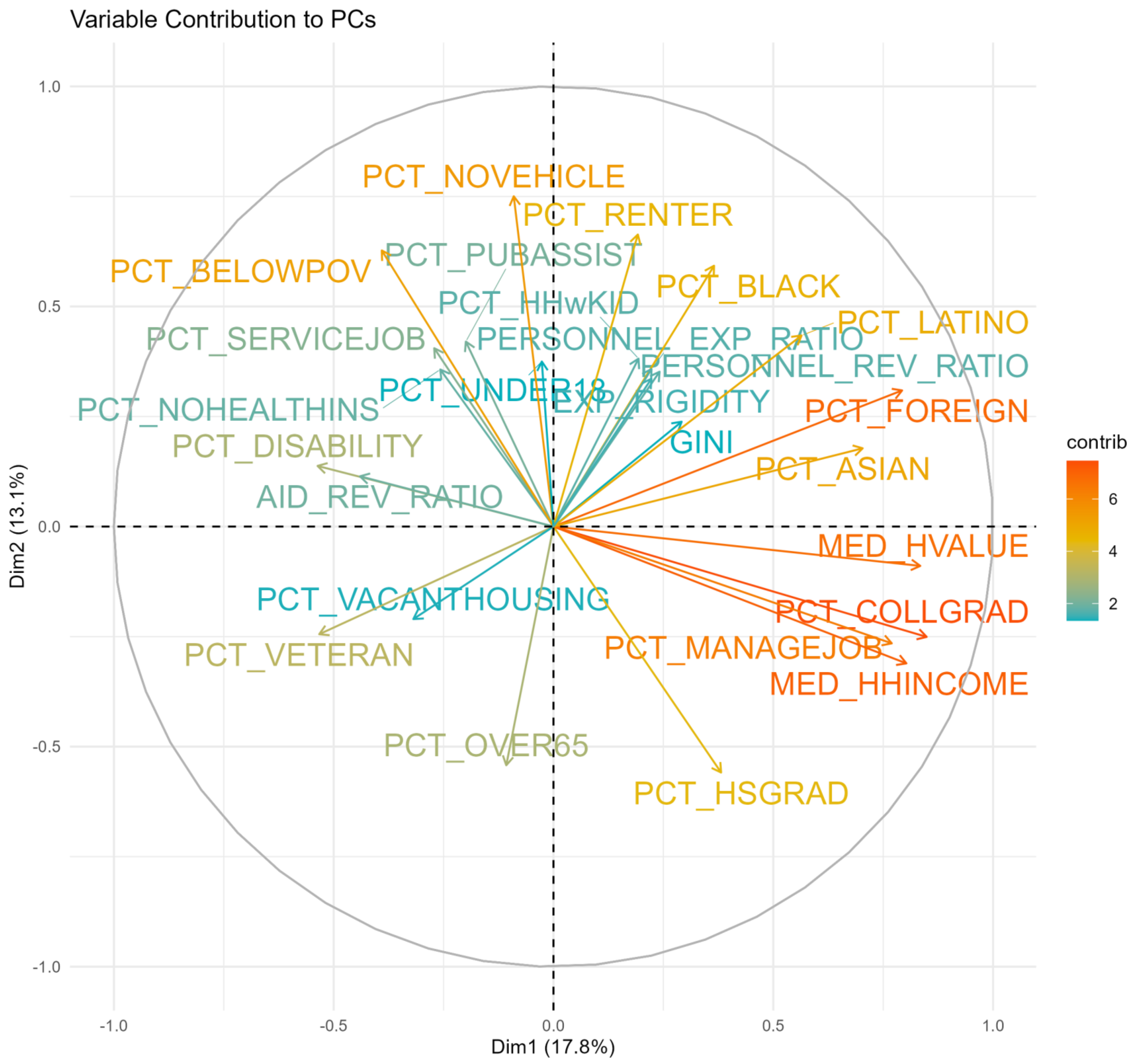
```
> cum_var_explained
 [1] 0.1780650 0.3089843 0.4171265 0.4838024 0.5332550 0.5823671 0.6248448 0.6578445 0.6892352 0.7166386
[11] 0.7421856 0.7661566 0.7891429 0.8105842 0.8309061 0.8497805 0.8657930 0.8811298 0.8943028 0.9070520
[21] 0.9182693 0.9291234 0.9394011 0.9487158 0.9575641 0.9657933 0.9733536 0.9801222 0.9861631 0.9902390
[31] 0.9936599 0.9968418 0.9993477 1.0000000
```



- **Eigenvalue > 1**
- **The first nine Cumulative Explained Variances add up to 68.9%**
- **Therefore, I selected 9 PCA components**



Which variables play an important role in the different PCA components



This graph shows the relationship between the variables and the first two principal components derived from the PCA, revealing how demographic and socioeconomic factors (such as percent race and median income) affect the variability of the data. The direction and length of the arrows indicate the correlation and effect of each variable on the components: for example, variables pointing in similar directions are positively correlated. While the figure does not show data points, often this visualization identifies clusters and outliers, providing a snapshot of the data structure and underlying patterns in the dataset. We can also discover the contribution of different variables from the shades of color.

Why use the Random Forest model?

Random Forest is a popular machine learning algorithm that's often used for classification and regression tasks for several reasons:

1. Feature Importance: Random Forest provides insights into feature importance, indicating which variables have the most influence on the prediction. This is particularly useful when working with PCA components, as it helps to understand which components are most relevant for the outcome.
2. Minimal Preprocessing: Unlike some algorithms that require data normalization or scaling, Random Forest can perform well without such preprocessing steps. This is advantageous when working with PCA scores, which are already scaled.
3. Robustness: Random Forest is less prone to overfitting than decision trees because it builds multiple trees and merges their results. This robustness is beneficial when dealing with complex datasets.
4. Flexibility: It can handle both numerical and categorical data, so you don't have to transform categorical variables into dummy variables.

Random Forest allows for an evaluation of the importance of each principal component (PC1~PC9) in classifying the type of financial stress, providing a way to understand which components have the strongest relationships with the outcome variable.

- Accuracy: The model achieved perfect accuracy (1.0 or 100%), which means every prediction made by the model matched the actual class labels exactly.

- Kappa Statistic: Also showing a value of 1, which indicates perfect agreement between the predictions and the actual classifications, far beyond what would be expected by chance. This is particularly impressive and suggests that the model is exceptionally well-fitted to the data.

- Prevalence: Shows how common each category is in your dataset. The prevalence of No Designation is very high at approximately 0.9897 (98.97%), which suggests that data is highly imbalanced towards this category.

- Detection Rate: Matches the prevalence, indicating that the model captures all actual instances of each category.

```
> print(rf_confusion_matrix)
Confusion Matrix and Statistics

                            Reference
Prediction                    Moderate Fiscal Stress No Designation Significant Fiscal Stress Susceptible Fiscal Stress
  Moderate Fiscal Stress                           1              0                         0                         0
  No Designation                                   0            862                         0                         0
  Significant Fiscal Stress                        0              0                         1                         0
  Susceptible Fiscal Stress                        0              0                         0                         7

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9958, 1)
    No Information Rate : 0.9897
    P-Value [Acc > NIR] : 0.0001178

                  Kappa : 1
```

```
Confusion Matrix and Statistics

                             Reference
Prediction                    Moderate Fiscal Stress No Designation Significant Fiscal Stress Susceptible Fiscal Stress
  Moderate Fiscal Stress                           1              0                         0                         0
  No Designation                                   0            862                         0                         0
  Significant Fiscal Stress                        0              0                         1                         0
  Susceptible Fiscal Stress                        0              0                         0                         7

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9958, 1)
    No Information Rate : 0.9897
    P-Value [Acc > NIR] : 0.0001178

                  Kappa : 1
```

Addressing Data Imbalance

- Perfect Performance Metrics: While the displayed metrics suggest that the model performs exceptionally well, such perfect statistics often warrant a deeper investigation, particularly in real-world scenarios where perfect classification is rare.
- Data Imbalance: The very high prevalence of the No Designation category and perfect classification metrics might indicate that the model's performance could be skewed by an imbalanced dataset. In such cases, the model might be overfitting to the majority class and might not perform as well on a more balanced or unseen dataset.
- Generalization Ability: Given the skewed data distribution, there's a risk that the model might not generalize well to new, unseen data, especially those from minority classes. Testing the model on a separate, ideally more balanced test set would be crucial to truly gauge its predictive power.
- Model Validation: Despite the high performance on the training set, it's important to validate the model on an independent set or use techniques like cross-validation to assess model stability and robustness. This helps in understanding whether the model's predictions are reliable and applicable in practical scenarios.

Solutions:

- **Oversampling:** I first calculated the maximum class size from my dataset using the **'max()'** function on the **'table()'** of class counts. and then created a function, **'oversample_class()'**, which oversamples the data for each class to match this maximum class size. This was done by replicating the rows of each class proportionally until the target size was reached. The oversampled data for each class was then combined into a single balanced dataset.
- **Custom Proportions:** In a subsequent section, I defined a total number of samples and specific proportions for each class. Using these proportions, I recalculated the target sizes for each class and applied the oversampling process again to achieve these specified distributions. The data was then shuffled to ensure randomness using **'sample()'** function with a set seed for reproducibility.
- **Random Forest Model:** After balancing the dataset, I applied a Random Forest classifier from the "randomForest" package. This model was trained on the balanced data, which included predictors for each type of fiscal stress. The model used 500 trees (ntree = 500).

```r
# Calculate the size of the largest class to know how many samples each class should have
class_counts <- table(pca_scores_cleaned$Type.of.Stress)
max_class_size <- max(class_counts)

# Function to oversample a specific class
oversample_class <- function(data, class_name, target_size) {
  class_data <- data[data$Type.of.Stress == class_name, ]
  replicate_times <- ceiling(target_size / nrow(class_data))
  oversampled_data <- class_data[rep(1:nrow(class_data), times = replicate_times), ]
  return(oversampled_data[1:target_size, ])
}

# Applying the oversampling function to each class
oversampled_data_list <- lapply(names(class_counts), function(class_name) {
  oversample_class(pca_scores_cleaned, class_name, max_class_size)
})
```

```r
# Combine all oversampled data into one dataframe
balanced_data <- do.call(rbind, oversampled_data_list)

# Check the new class distribution
table(balanced_data$Type.of.Stress)
rf_model <- randomForest(Type.of.Stress ~ ., data = balanced_data, ntree = 500)
print(rf_model)
rf_predicted_labels <- predict(rf_model, balanced_data)
rf_confusion_matrix <- confusionMatrix(rf_predicted_labels, balanced_data$Type.of.Stress)

# Print the confusion matrix
print(rf_confusion_matrix)

# Print overall accuracy from the confusion matrix
accuracy <- rf_confusion_matrix$overall['Accuracy']
print(paste("Accuracy of the Random Forest model:", accuracy))
```

```
Confusion Matrix and Statistics

                           Reference
Prediction                   Moderate Fiscal Stress No Designation Significant Fiscal Stress Susceptible Fiscal Stress
  Moderate Fiscal Stress                        600              0                         0                         0
  No Designation                                  0            500                         0                         0
  Significant Fiscal Stress                       0              0                       500                         0
  Susceptible Fiscal Stress                       0              0                         0                       400

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9982, 1)
    No Information Rate : 0.3
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
```

```
# Set the desired proportion for each class
desired_proportions <- c("Moderate Fiscal Stress" = 0.3,
                         "No Designation" = 0.25,
                         "Significant Fiscal Stress" = 0.25,
                         "Susceptible Fiscal Stress" = 0.2)
```

Reset the proportion for each class0.3: 0.35: 0.25: 0.2

```
Confusion Matrix and Statistics

                           Reference
Prediction                  Moderate Fiscal Stress No Designation Significant Fiscal Stress Susceptible Fiscal Stress
  Moderate Fiscal Stress                       862              0                         0                         0
  No Designation                                 0            862                         0                         0
  Significant Fiscal Stress                      0              0                       862                         0
  Susceptible Fiscal Stress                      0              0                         0                       862

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9989, 1)
    No Information Rate : 0.25
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
```

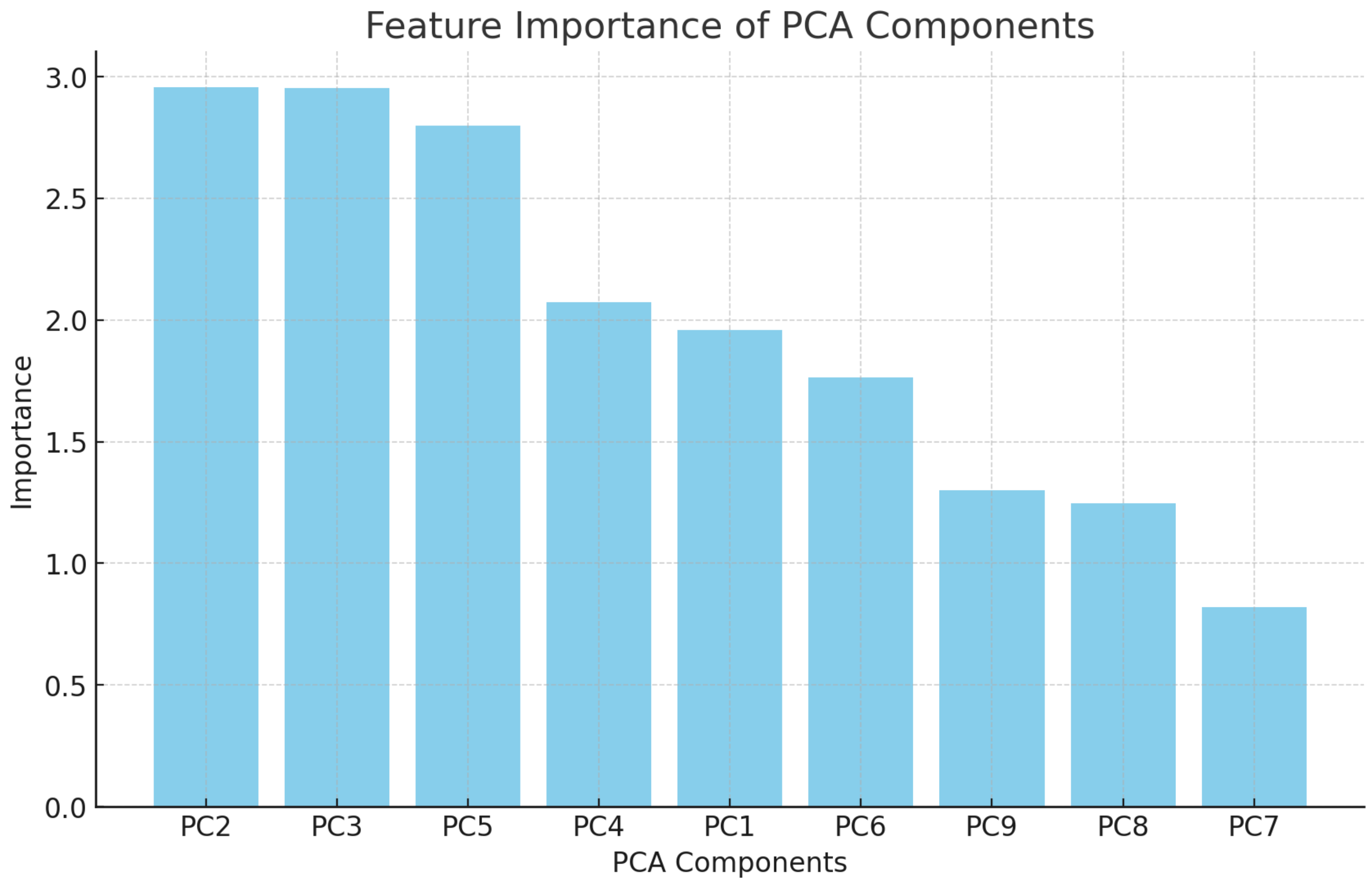
Reset the proportion for each class 1: 1: 1: 1

```
Confusion Matrix and Statistics

                             Reference
Prediction                     Moderate Fiscal Stress No Designation Significant Fiscal Stress Susceptible Fiscal Stress
  Moderate Fiscal Stress                          250              0                         0                         0
  No Designation                                    0            250                         0                         0
  Significant Fiscal Stress                         0              0                       250                         0
  Susceptible Fiscal Stress                         0              0                         0                       250

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9963, 1)
    No Information Rate : 0.25
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
```

```
# Set the desired proportion for each class
desired_proportions <- c("Moderate Fiscal Stress" = 0.25,
                         "No Designation" = 0.25,
                         "Significant Fiscal Stress" = 0.25,
                         "Susceptible Fiscal Stress" = 0.25)
```

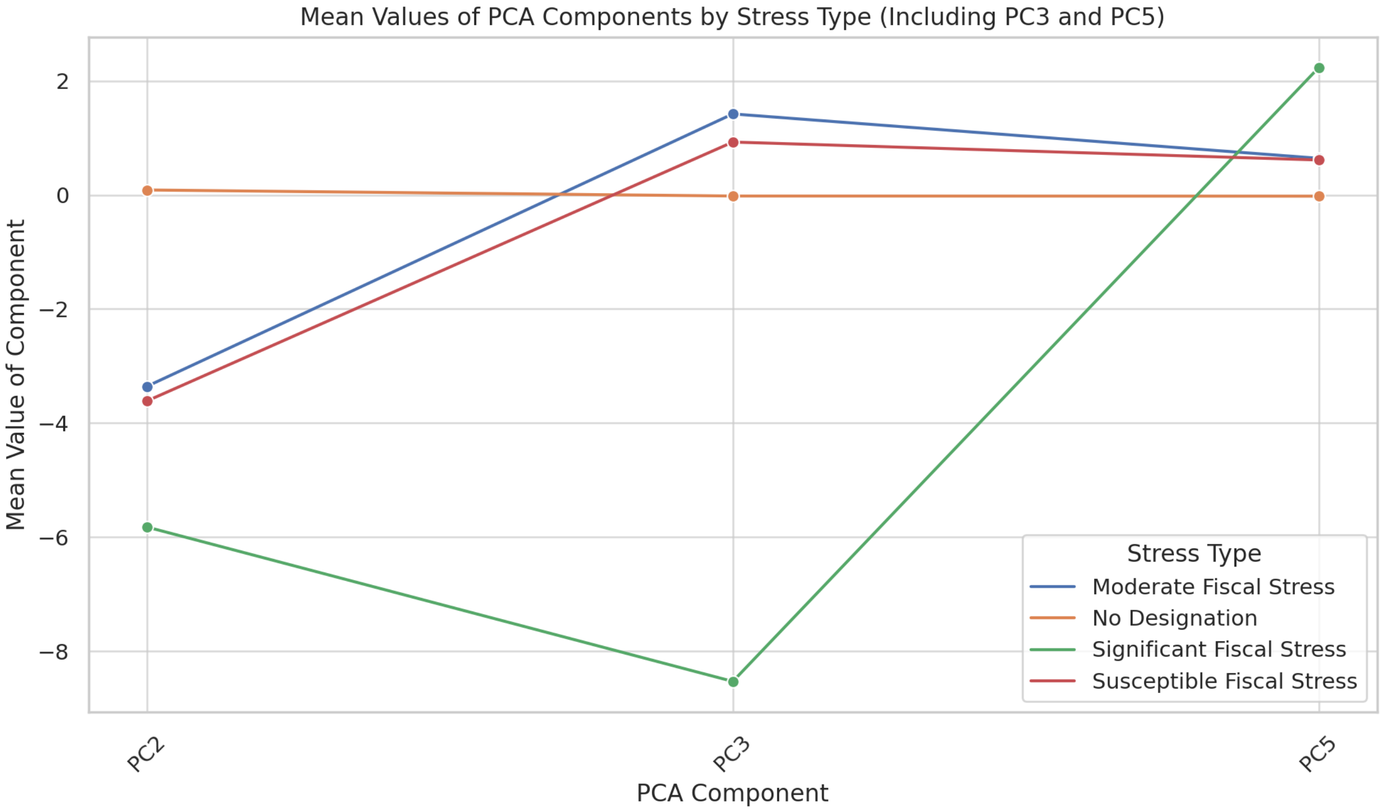
Reset the proportion for each class 0.25: 0.25: 0.25: 0.25

```
> print(importance_sorted)
      PC2       PC3       PC5       PC4       PC1       PC6       PC9       PC8       PC7 
2.9560060 2.9542147 2.7997600 2.0721122 1.9585937 1.7650616 1.3012136 1.2452337 0.8201995 
```



- For PC2 and PC3, where higher values are associated with "No Designation,“ this is considered a **positive** relationship with the "No Designation" outcome. In other words, as PC2 and PC3 increase, the likelihood of the "No Designation" category being the outcome also increases.

- For PC5, where higher values are associated with "Significant Financial Stress," this indicates a **positive** relationship with the "Significant Financial Stress" outcome. That means as the PC5 increases, so does the probability of the outcome being "Significant Financial Stress".

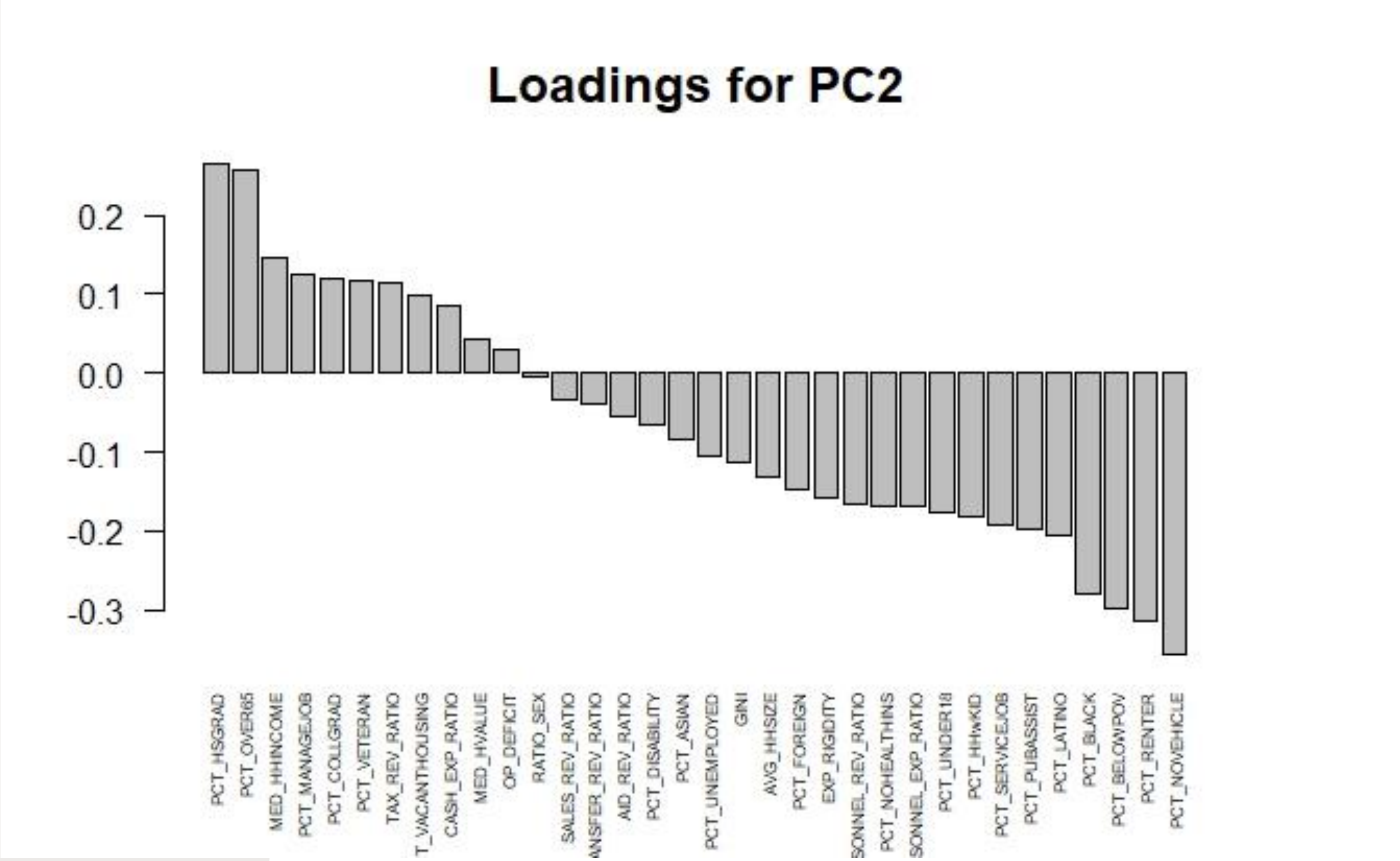


PC2, PC3, and PC5 play an important role in the model so I'm curious who are the variables inside PC2, PC3, PC5?

- Positive Factors: Higher values of **PCT_HSGRAD** (percentage of high school graduates), **PCT_OVER65** (percentage of the population over 65 years old), and **MED_HHINCOME** (median household income)

- Negative Factors: **PCT_NOVEHICLE**(Percentage of households without a vehicle)
- **PCT_RENTER**(Percentage of residents who rent their home)
- **PCT_BELOWPOV**(Percentage of residents below the poverty line)

- The more of these positive factors, the lower the risk of a fiscal crisis.

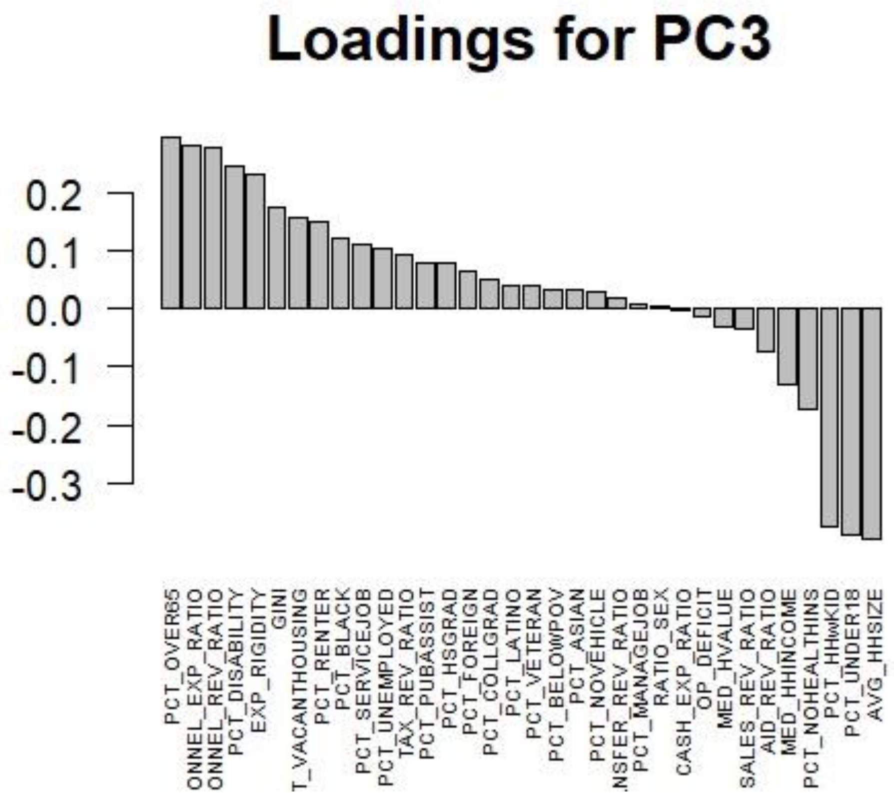



```
> print(sorted_pc2_loadings)
      PCT_HSGRAD          PCT_OVER65        MED_HHINCOME        PCT_MANAGEJOB         PCT_COLLGRAD
     0.264891569         0.257086530         0.147367343          0.125802943          0.118839031
     PCT_VETERAN       TAX_REV_RATIO  PCT_VACANTHOUSING       CASH_EXP_RATIO           MED_HVALUE
     0.115979530         0.113636801         0.099663418          0.086388495          0.042385351
      OP_DEFICIT           RATIO_SEX     SALES_REV_RATIO   TRANSFER_REV_RATIO        AID_REV_RATIO
     0.030588221        -0.005026525        -0.032783228         -0.040014700         -0.054098032
  PCT_DISABILITY           PCT_ASIAN      PCT_UNEMPLOYED                 GINI           AVG_HHSIZE
    -0.065272159        -0.084260368        -0.104270623         -0.112891247         -0.130769596
     PCT_FOREIGN        EXP_RIGIDITY PERSONNEL_REV_RATIO      PCT_NOHEALTHINS  PERSONNEL_EXP_RATIO
    -0.146597829        -0.159225036        -0.166742559         -0.169104379         -0.169660640
     PCT_UNDER18          PCT_HHwKID      PCT_SERVICEJOB        PCT_PUBASSIST           PCT_LATINO
    -0.177727769        -0.180835739        -0.192049812         -0.199175610         -0.205743371
       PCT_BLACK        PCT_BELOWPOV          PCT_RENTER        PCT_NOVEHICLE
    -0.280869791        -0.297311931        -0.314473959         -0.355665387
```

- Positive Factors: **PCT_OVER65** (percentage of the population over 65 years old) **PERSONNEL_EXP_RATIO** (personnel expenditure ratio) and **PERSONNEL_REV_RATIO (**personnel revenue ratio)

- Negative Factors: **AVG_HHSIZE**(Average household size) **PCT_UNDER18**(Percentage of the population under 18 years of age)
- **PCT_HHwKID**(Percentage of households with kids)

The more of these positive factors, the lower the risk of a fiscal crisis.




```
> print(sorted_pc3_loadings)
         PCT_OVER65 PERSONNEL_EXP_RATIO PERSONNEL_REV_RATIO      PCT_DISABILITY        EXP_RIGIDITY                GINI
        0.294185757         0.280535377         0.274710654         0.243922656         0.230059741         0.172984458
   PCT_VACANTHOUSING          PCT_RENTER           PCT_BLACK      PCT_SERVICEJOB      PCT_UNEMPLOYED       TAX_REV_RATIO
        0.154942859         0.148538570         0.120656432         0.110461370         0.101583983         0.093278156
       PCT_PUBASSIST          PCT_HSGRAD         PCT_FOREIGN        PCT_COLLGRAD          PCT_LATINO         PCT_VETERAN
        0.078321679         0.076935573         0.062528913         0.051511194         0.040345807         0.038874396
        PCT_BELOWPOV           PCT_ASIAN       PCT_NOVEHICLE  TRANSFER_REV_RATIO       PCT_MANAGEJOB           RATIO_SEX
        0.033751771         0.031620040         0.029921030         0.017105939         0.009202450         0.004532039
      CASH_EXP_RATIO          OP_DEFICIT          MED_HVALUE     SALES_REV_RATIO       AID_REV_RATIO        MED_HHINCOME
       -0.001355422        -0.014665960        -0.032497837        -0.035165318        -0.074918240        -0.129693332
     PCT_NOHEALTHINS          PCT_HHwKID         PCT_UNDER18          AVG_HHSIZE
       -0.171458799        -0.373978722        -0.386437848        -0.395430925
```

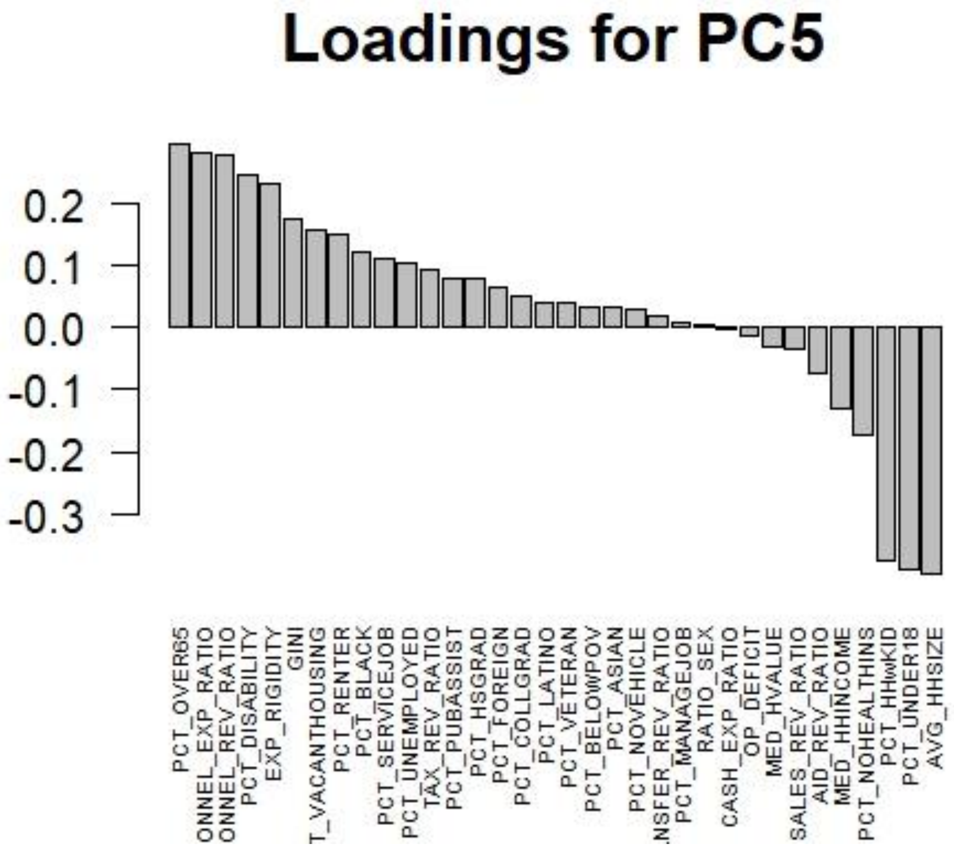
- Positive Factors: **TRANSFER_REV_RATIO** (transfer revenue ratio) **AID_REV_RATIO** (aid revenue ratio)

- Negative Factors:

CASH_EXP_RATIO (Cash Expenditure Ratio) **OP_DEFICIT** (Operating Deficit)

SALES_REV_RATIO(The ratio of sales revenue, possibly in the context of sales tax)

Attention! The more of these **Negative** factors, the lower the risk of a fiscal crisis.




```
> print(sorted_pc5_loadings)
TRANSFER_REV_RATIO      AID_REV_RATIO        PCT_NOVEHICLE          PCT_HSGRAD     PCT_DISABILITY               GINI
       0.141845648        0.136885767          0.110637102         0.089703784        0.085145672        0.081012790
         PCT_ASIAN       PCT_MANAGEJOB        PCT_BELOWPOV          PCT_RENTER       PCT_COLLGRAD      PCT_PUBASSIST
       0.076496137        0.071572951          0.070686000         0.068677074        0.065990855        0.058049492
        PCT_OVER65         PCT_UNDER18      PCT_UNEMPLOYED  PERSONNEL_REV_RATIO        PCT_VETERAN         PCT_HHwKID
       0.050567743        0.039863827          0.032345280         0.001729561       -0.003359259       -0.020331272
     TAX_REV_RATIO     PCT_NOHEALTHINS          MED_HVALUE      PCT_SERVICEJOB        PCT_FOREIGN         AVG_HHSIZE
      -0.033650207       -0.035528028         -0.036391644        -0.039883656       -0.045686435       -0.050643196
      MED_HHINCOME    PCT_VACANTHOUSING          PCT_LATINO  PERSONNEL_EXP_RATIO          PCT_BLACK       EXP_RIGIDITY
      -0.070906461       -0.077901278         -0.129625176        -0.139314451       -0.149246434       -0.187642216
         RATIO_SEX     SALES_REV_RATIO      CASH_EXP_RATIO          OP_DEFICIT
      -0.337525830       -0.356474815         -0.461513277        -0.569441968
```

- **Summarize** the factors for PC2, PC3, and PC5. We can learn that the following factors have an impact on government fiscal crisis:
- **PCT_HSGRAD** (percentage of high school graduates), **PCT_OVER65** (percentage of the population over 65 years old), and **MED_HHINCOME** (median household income) **PCT_NOVEHICLE**(Percentage of households without a vehicle)
- **PCT_RENTER**(Percentage of residents who rent their home)
- **PCT_BELOWPOV**(Percentage of residents below the poverty line)
- **PERSONNEL_EXP_RATIO** (personnel expenditure ratio) and **PERSONNEL_REV_RATIO (**personnel revenue ratio) **AVG_HHSIZE**(Average household size) **PCT_UNDER18**(Percentage of the population under 18 years of age)
- **PCT_HHwKID**(Percentage of households with kids)
- **TRANSFER_REV_RATIO** (transfer revenue ratio) **AID_REV_RATIO** (aid revenue ratio)
- **CASH_EXP_RATIO** (Cash Expenditure Ratio) **OP_DEFICIT** (Operating Deficit)
- **SALES_REV_RATIO**(The ratio of sales revenue, possibly in the context of sales tax)

Focusing on these data will help in anticipating the government's fiscal crisis.

Thank you!