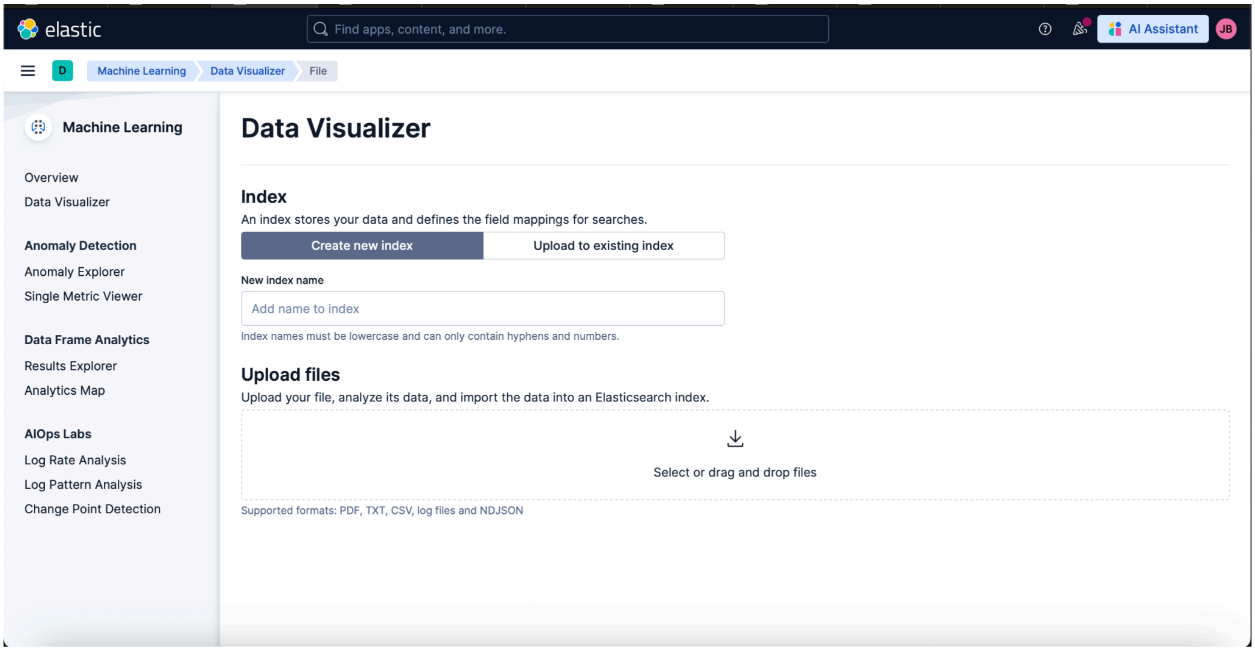
Creating a Machine Learning Model to Find out What Drives Compensation

Step 1: First, go to this link and download the salaries dataset
https://www.kaggle.com/datasets/kaggle/sf-salaries

Step 2: Then go to machine learning / data visualizer to upload the dataset that was downloaded. I recommend typing this in the search bar to pull up the path immediately without having to search and scroll around for it.



Data Visualizer

Index

An index stores your data and defines the field mappings for searches.

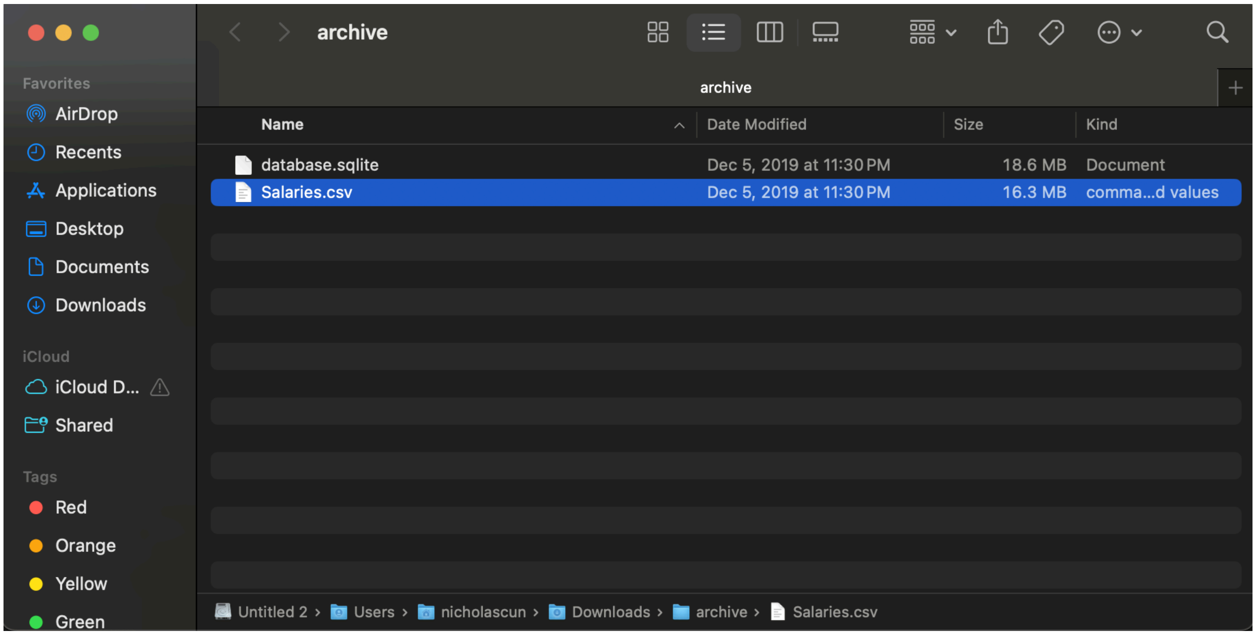
Index names must be lowercase and can only contain hyphens and numbers.

Upload files

Upload your file, analyze its data, and import the data into an Elasticsearch index.

Supported formats: PDF, TXT, CSV, log files and NDJSON

Make sure to choose the .csv file



Name	Date Modified	Size	Kind
database.sqlite	Dec 5, 2019 at 11:30 PM	18.6 MB	Document
Salaries.csv	Dec 5, 2019 at 11:30 PM	16.3 MB	comma...d values

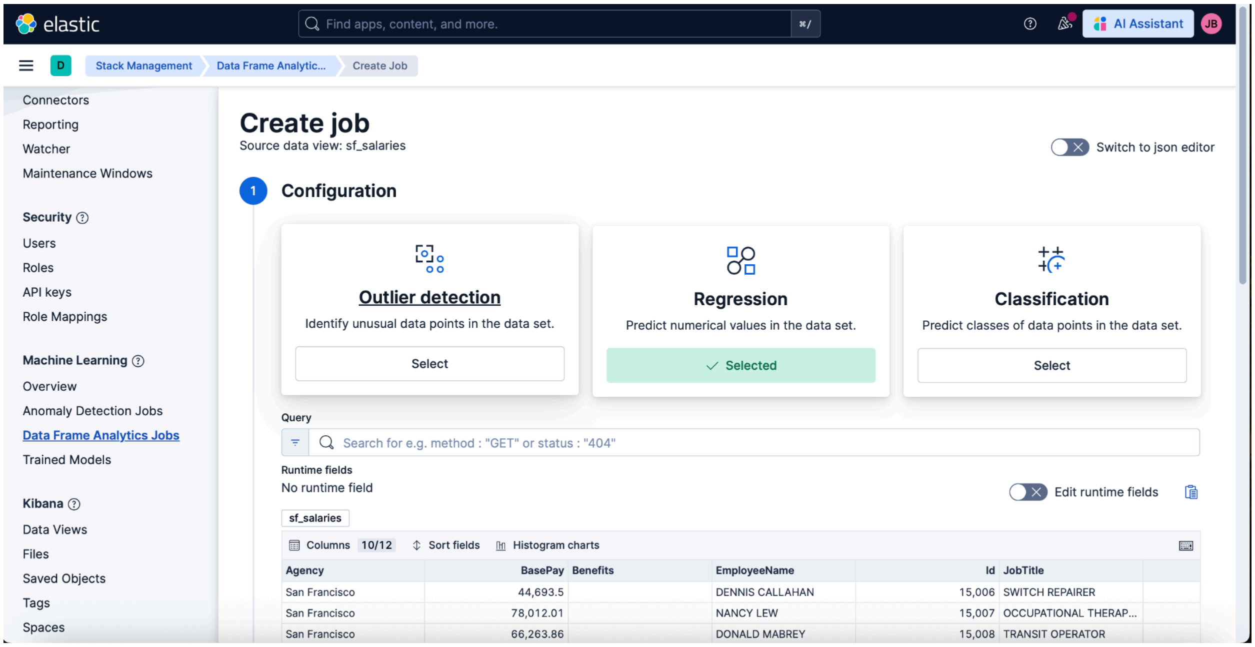
Untitled 2 › Users › nicholascun › Downloads › archive › Salaries.csv

Step 3: Then go to Dev Tools and run these two functions in order to include job title in the ML model. Having the original JobTitle in it's text form will produce an error in the ML model, so we need to make a keyword version of it to use instead.

```
PUT sf_salaries_fixed
{
 "mappings": {
   "properties": {
     "JobTitle": {
       "type": "text",
       "fields": {
         "keyword": {
           "type": "keyword"
         }
       }
     }
   }
 }
}
```

```
POST _reindex
{
 "source": {
   "index": "sf_salaries"
 },
 "dest": {
   "index": "sf_salaries_fixed"
 }
}
```

Step 4: Create a data frame analytics regression model. In the search bar go to Machine Learning / Data Frame Analytics Jobs. Then choose regression.



Create job

Source data view: sf_salaries

Switch to json editor

1 Configuration

Outlier detection	Regression	Classification
Identify unusual data points in the data set.	Predict numerical values in the data set.	Predict classes of data points in the data set.
Select	✓ Selected	Select

Query

Search for e.g. method : "GET" or status : "404"

Runtime fields

No runtime field

Edit runtime fields

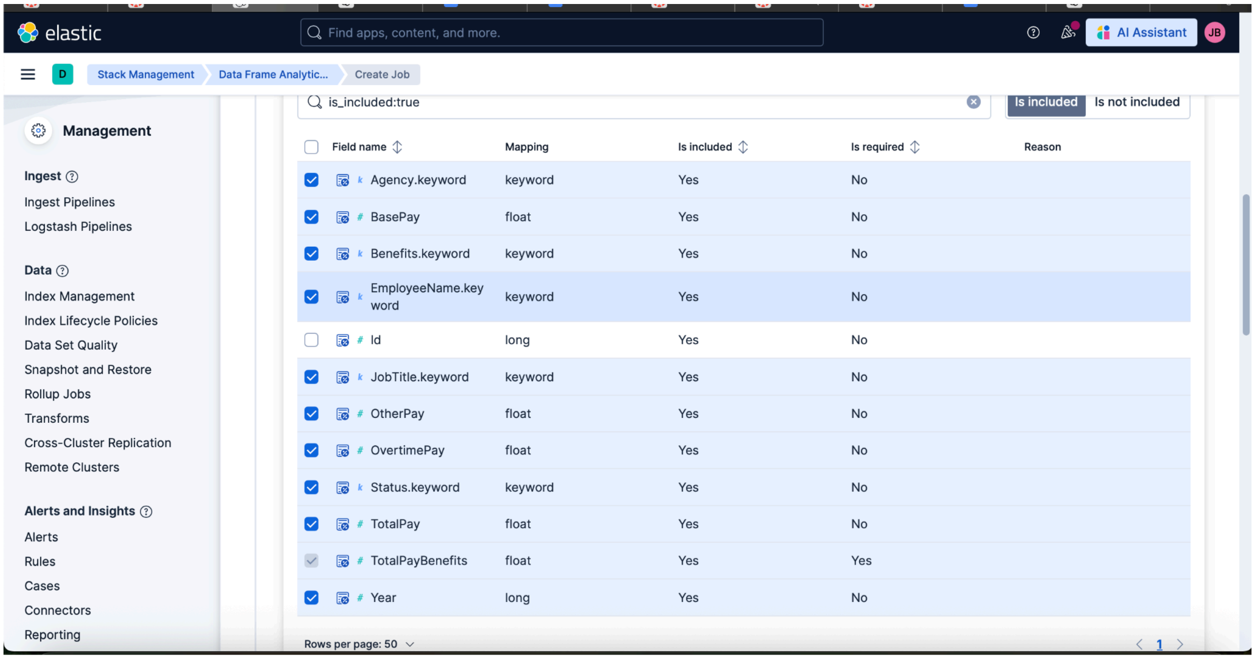
sf_salaries

Columns 10/12 Sort fields Histogram charts

Agency	BasePay	Benefits	EmployeeName	Id	JobTitle
San Francisco	44,693.5		DENNIS CALLAHAN	15,006	SWITCH REPAIRER
San Francisco	78,012.01		NANCY LEW	15,007	OCCUPATIONAL THERAP...
San Francisco	66,263.86		DONALD MABREY	15,008	TRANSIT OPERATOR

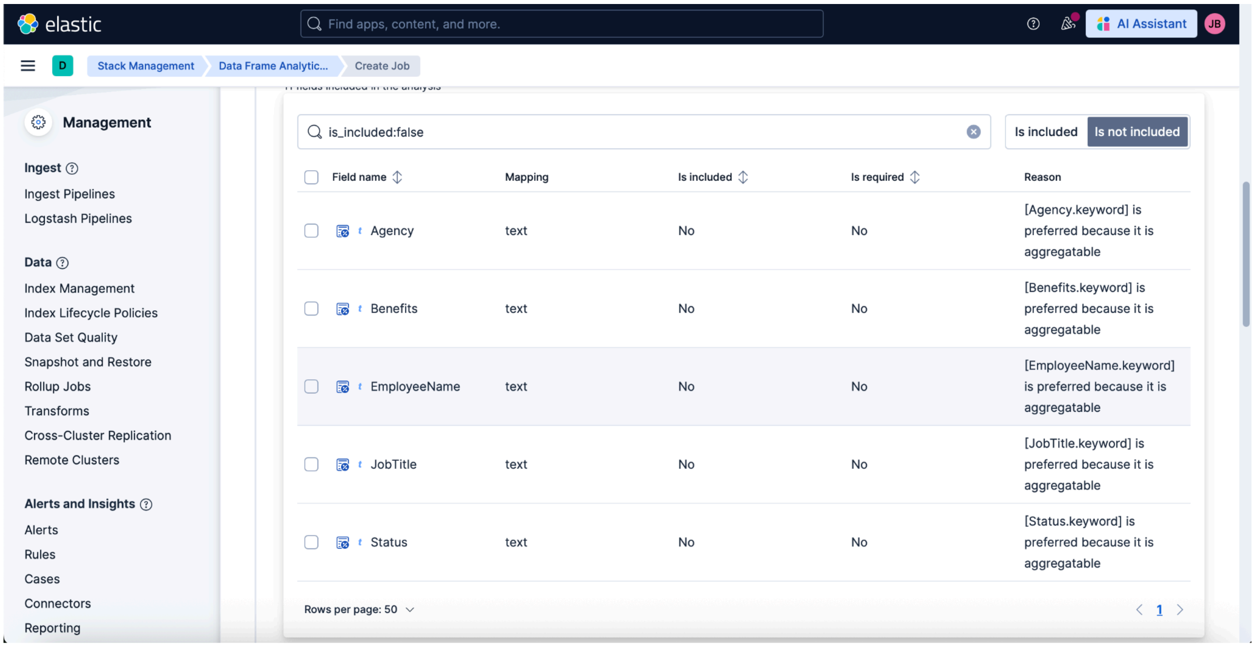
Step 5: Select TotalPayBenefits as the Dependant variable.

Step 6: Uncheck id, EmployeeName, and Benefits.keyword from the “is included” section. This will make it easier on the ML model and safer since it has less unnecessary variables to work with. Don’t forget to uncheck JobTitle aswell since it will give you an error if you do not. We will be using JobTitle.keyword instead.



is_included:true

☐	Field name	Mapping	Is included	Is required	Reason
☑	Agency.keyword	keyword	Yes	No	
☑	BasePay	float	Yes	No	
☑	Benefits.keyword	keyword	Yes	No	
☑	EmployeeName.keyword	keyword	Yes	No	
☐	Id	long	Yes	No	
☑	JobTitle.keyword	keyword	Yes	No	
☑	OtherPay	float	Yes	No	
☑	OvertimePay	float	Yes	No	
☑	Status.keyword	keyword	Yes	No	
☑	TotalPay	float	Yes	No	
☑	TotalPayBenefits	float	Yes	Yes	
☑	Year	long	Yes	No	



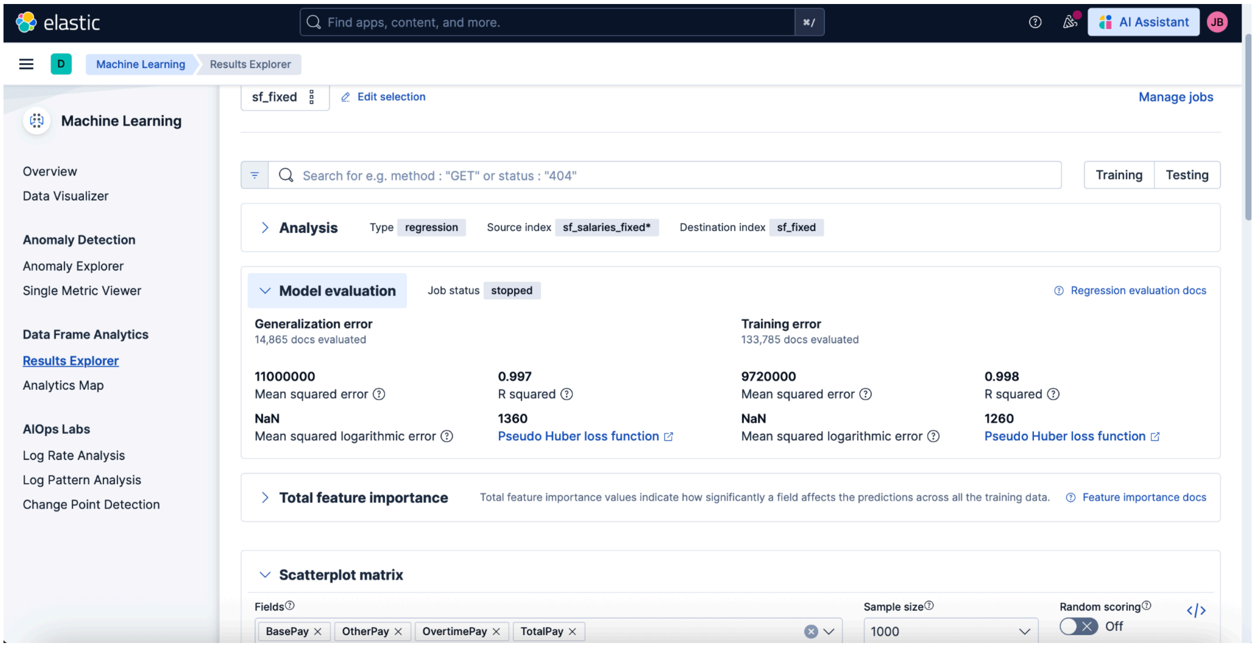
is_included:false

☐	Field name	Mapping	Is included	Is required	Reason
☐	Agency	text	No	No	[Agency.keyword] is preferred because it is aggregatable
☐	Benefits	text	No	No	[Benefits.keyword] is preferred because it is aggregatable
☐	EmployeeName	text	No	No	[EmployeeName.keyword] is preferred because it is aggregatable
☐	JobTitle	text	No	No	[JobTitle.keyword] is preferred because it is aggregatable
☐	Status	text	No	No	[Status.keyword] is preferred because it is aggregatable

Step 7: Set the training percent to 90

Step 8: Start the job and wait for all the phases to finish up.

Step 9: Once all phases are completed and the model has stopped running, go to Data Frame analytics / Results Explorer to view the results of the ML model.



sf_fixed

Analysis Type regression Source index sf_salaries_fixed* Destination index sf_fixed

Model evaluation Job status stopped

Generalization error 14,865 docs evaluated		Training error 133,785 docs evaluated	
11000000 Mean squared error	0.997 R squared	9720000 Mean squared error	0.998 R squared
NaN Mean squared logarithmic error	1360 Pseudo Huber loss function	NaN Mean squared logarithmic error	1260 Pseudo Huber loss function

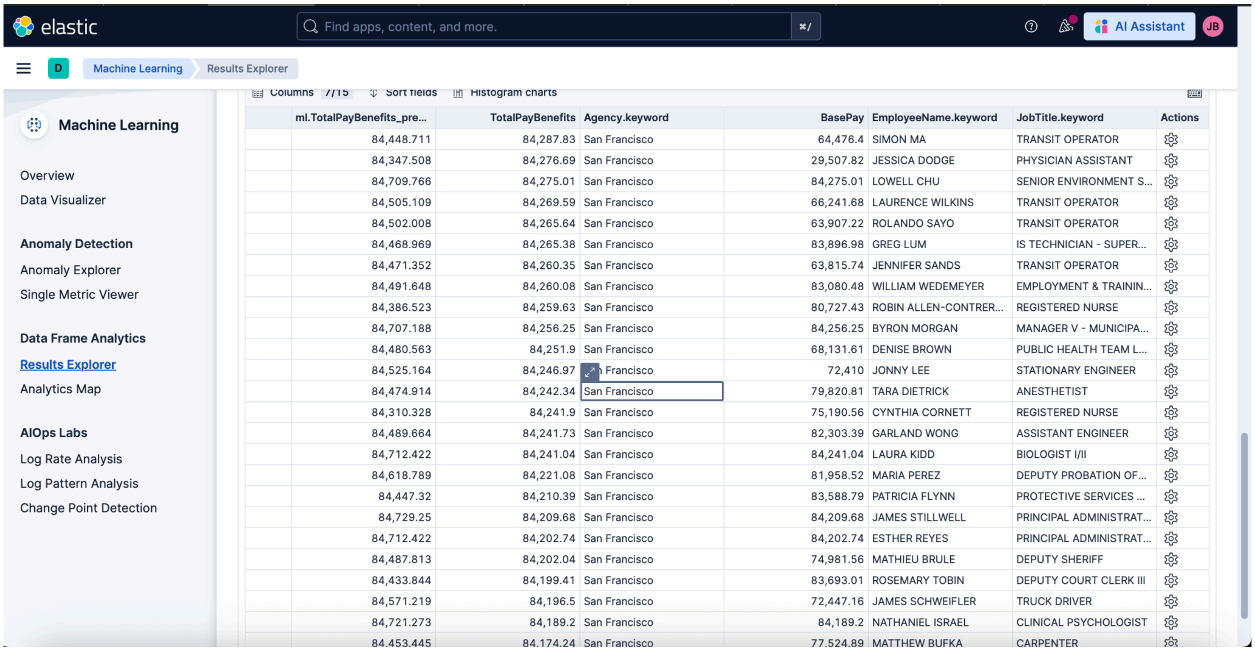
Total feature importance Total feature importance values indicate how significantly a field affects the predictions across all the training data.

Scatterplot matrix

Fields: BasePay, OtherPay, OvertimePay, TotalPay

Sample size: 1000

Random scoring: Off



ml.TotalPayBenefits_pre...	TotalPayBenefits	Agency.keyword	BasePay	EmployeeName.keyword	JobTitle.keyword	Actions
84,448.711	84,287.83	San Francisco	64,476.4	SIMON MA	TRANSIT OPERATOR	⚙
84,347.508	84,276.69	San Francisco	29,507.82	JESSICA DODGE	PHYSICIAN ASSISTANT	⚙
84,709.766	84,275.01	San Francisco	84,275.01	LOWELL CHU	SENIOR ENVIRONMENT S...	⚙
84,505.109	84,269.59	San Francisco	66,241.68	LAURENCE WILKINS	TRANSIT OPERATOR	⚙
84,502.008	84,265.64	San Francisco	63,907.22	ROLANDO SAYO	TRANSIT OPERATOR	⚙
84,468.969	84,265.38	San Francisco	83,896.98	GREG LUM	IS TECHNICIAN - SUPER...	⚙
84,471.352	84,260.35	San Francisco	63,815.74	JENNIFER SANDS	TRANSIT OPERATOR	⚙
84,491.648	84,260.08	San Francisco	83,080.48	WILLIAM WEDEMEYER	EMPLOYMENT & TRAININ...	⚙
84,386.523	84,259.63	San Francisco	80,727.43	ROBIN ALLEN-CONTRER...	REGISTERED NURSE	⚙
84,707.188	84,256.25	San Francisco	84,256.25	BYRON MORGAN	MANAGER V - MUNICIPA...	⚙
84,480.563	84,251.9	San Francisco	68,131.61	DENISE BROWN	PUBLIC HEALTH TEAM L...	⚙
84,525.164	84,246.97	n Francisco	72,410	JONNY LEE	STATIONARY ENGINEER	⚙
84,474.914	84,242.34	San Francisco	79,820.81	TARA DIETRICK	ANESTHETIST	⚙
84,310.328	84,241.9	San Francisco	75,190.56	CYNTHIA CORNETT	REGISTERED NURSE	⚙
84,489.664	84,241.73	San Francisco	82,303.39	GARLAND WONG	ASSISTANT ENGINEER	⚙
84,712.422	84,241.04	San Francisco	84,241.04	LAURA KIDD	BIOLOGIST I/II	⚙
84,618.789	84,221.08	San Francisco	81,958.52	MARIA PEREZ	DEPUTY PROBATION OF...	⚙
84,447.32	84,210.39	San Francisco	83,588.79	PATRICIA FLYNN	PROTECTIVE SERVICES ...	⚙
84,729.25	84,209.68	San Francisco	84,209.68	JAMES STILLWELL	PRINCIPAL ADMINISTRAT...	⚙
84,712.422	84,202.74	San Francisco	84,202.74	ESTHER REYES	PRINCIPAL ADMINISTRAT...	⚙
84,487.813	84,202.04	San Francisco	74,981.56	MATHIEU BRULE	DEPUTY SHERIFF	⚙
84,433.844	84,199.41	San Francisco	83,693.01	ROSEMARY TOBIN	DEPUTY COURT CLERK III	⚙
84,571.219	84,196.5	San Francisco	72,447.16	JAMES SCHWEIFLER	TRUCK DRIVER	⚙
84,721.273	84,189.2	San Francisco	84,189.2	NATHANIEL ISRAEL	CLINICAL PSYCHOLOGIST	⚙
84,453.445	84,174.24	San Francisco	77,524.89	MATTHEW BUFKA	CARPENTER	⚙

********************This concludes the SF Salaries Tutorial*********************