Predictive Modeling for Employee Terminations

Tom Rice

Goal

This analysis will attempt to predict when an employee will terminate, which will help the HR department with employee turnover by developing strategies to fill upcoming vacant roles. This analysis is divided into three parts: descriptive analysis, diagnosis analysis, and predictive analysis. This analysis will use binary classification, as there can only be two outcomes determining whether an employee is active or terminated. The concepts covered in this project include data munging, splitting the dataset into test, validation and training sets, descriptive analytics, diagnostic analytics, and predictive analytics, supervised learning, feature selection, logistic regression, and supervised neural networks.

Data Munging

The data must be cleaned and prepared for the machine learning analysis before it can be performed. This researcher took the following steps to prepare the data for analysis.

1. For privacy, the name and employee columns were replaced with Row ID columns.
2. The effective date row was removed as it was not needed for analysis.
3. Location and Department IDs are numerical values that needed to be combined so that the data in all locations and departments was uniform.
4. The location description column was munged, so all the location descriptions were uniform and correct.
5. Normalized the Pay group column so all pay groups are numerical values. Daisy Elementary had an experimental all-year-round program that changed their pay group designations to include alpha numeric characters.
6. The building column had missing values. The building column is tied to the location column, so the missing values were filled in based on the populated values of the location.
7. The building column was divided into eight categories. Program schools are also considered high schools, so their values change to 10. The following tables illustrate the numerical codes given to building descriptions.

Building_Code	Description
10	High Schools
20	Middle Schools
30	Elementary Schools
50	Other Schools
60	Transporation
70	Multi-School
80	Other Locations
90	To be Determined

8. The date fields were corrupted, so this researcher had to update the fields to show the correct date. The following image shows the corrupted fields and data missing from the BUILDING column.

M	N	O	P	Q
POSITION	BUILDING	HIRE_DT	BIRTHDAT	HCS_P
3225		1/0/1900	1/0/1900	S
3249		1/0/1900	1/0/1900	P
1790		1/0/1900	1/0/1900	S
3110		1/0/1900	1/0/1900	S
3490		1/0/1900	1/0/1900	P
3491		1/0/1900	1/0/1900	S
1131		1/0/1900	1/0/1900	P
1023		1/0/1900	1/0/1900	P
1023		1/0/1900	1/0/1900	P
1023		1/0/1900	1/0/1900	P
2738		1/0/1900	1/0/1900	P
1023		1/0/1900	1/0/1900	P
1023		1/0/1900	1/0/1900	P
3069		1/0/1900	1/0/1900	P
1023		1/0/1900	1/0/1900	P
4502	40	1/0/1900	1/0/1900	S
3224	40	1/0/1900	1/0/1900	S
2109		1/0/1900	1/0/1900	S
3664	40	1/0/1900	1/0/1900	S
6053	40	1/0/1900	1/0/1900	S

9. The Termination Date was missing from the original dataset and was added for analysis.
10. Hire Date and Termination date were subtracted based on year to determine Years of Service
11. Age was determined by subtracting their Termination date or today's day if still active by their birthday.
12. Adjust NULL Date rows to show blank, but in the calculations, use the current date to get years of service
13. Create an age column by subtracting today's date from the employees' age.
14. Categorized the Positions into ten broad categories. The following tables illustrate the numerical codes given to employee positions.

POS_CAT	Description
1	Teachers
2	Principals
3	Assistant Principals
4	Custodians
5	Food Service
6	District Admins
7	Non-Teaching Prof
8	Support Staff
9	Bus Drivers
10	Board Members

15. Termination reasons were given numerical categories. The following tables illustrate the numerical codes given to employee termination reasons.

Code	Termination Description
1	Active
2	Death
3	Retire
4	Unknown
5	Leaving Profession
6	Another Position in Education
7	Family Reasons
8	Temporary Agreements
9	Declined Position
10	Health Issues
11	Dismissal
12	Moving
13	Cert Issues

16. Adjusted employee pay grades are to be numerical, and normalized pay grades are to match the current HCS salary scale.

17. The employee's steps were normalized to reflect the proper years of service. The district uses different numbers for ROTC instructors (Step 97) and new teachers (Step 99).
18. This researcher adjusted the binary employee status field to represent 0 for terminated and 1 for active employees.
19. Removed 5 Duplicate Rows

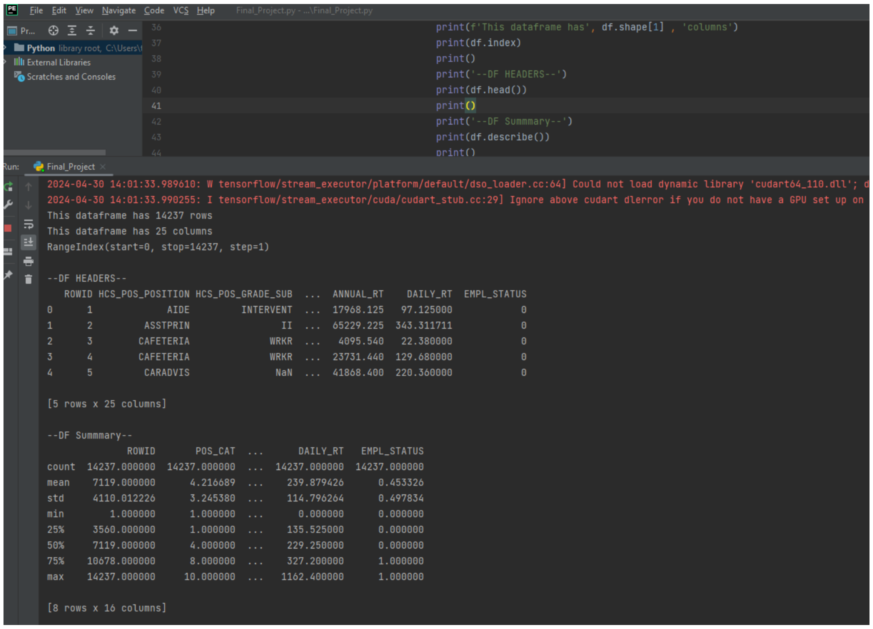
Importing Python Packages

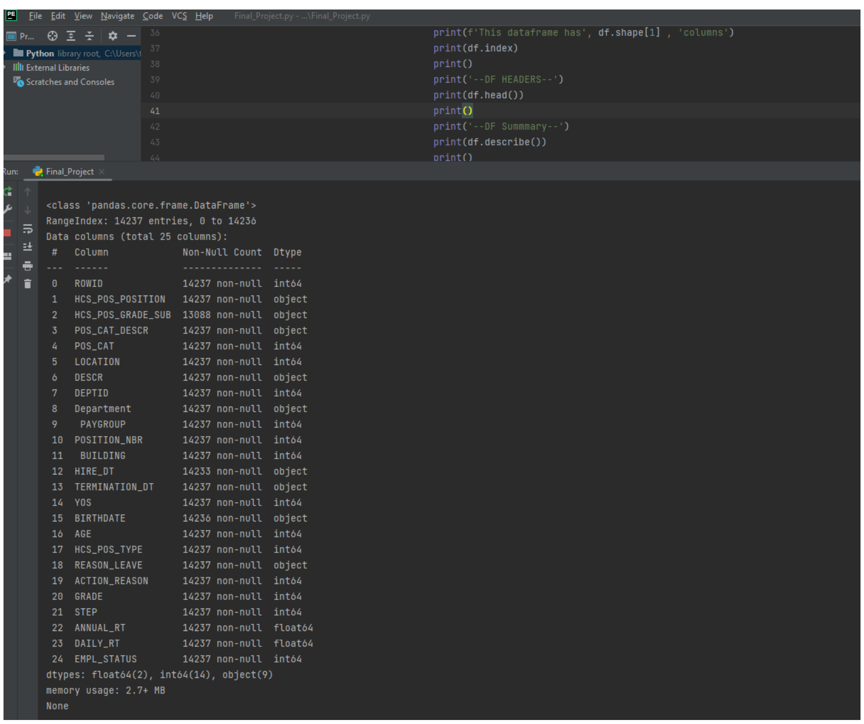
After the data was cleaned and prepared for analysis, the environment for this analysis was prepared in Python by importing the required packages. The following image shows the packages needed for this analysis.

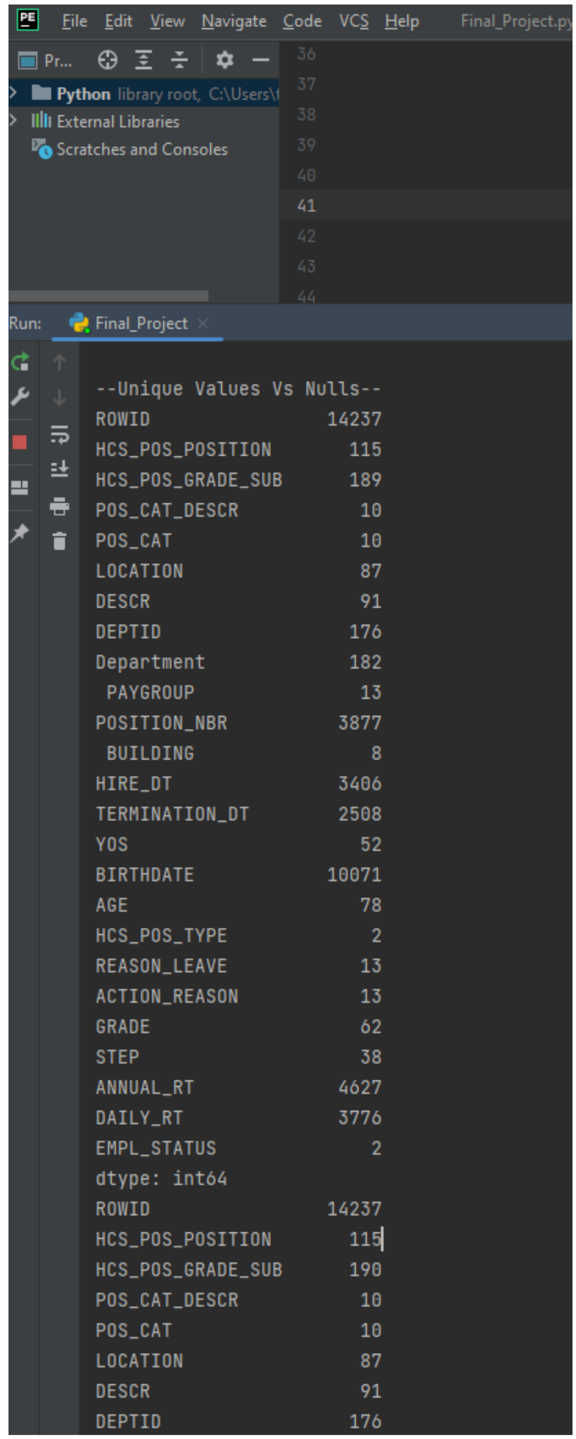
```
##Employee Termination Analysis
## By: Tom Rice
## IST 679 - Big Data Analytics

##Libraries to start with
################################
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import scipy
from scipy import stats
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.preprocessing import StandardScaler, MinMaxScaler, OrdinalEncoder, OneHotEncoder
from sklearn.compose import make_column_transformer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.impute import SimpleImputer
from sklearn.pipeline import make_pipeline
from pycaret.classification import *
##############################
```

Descriptive Analysis

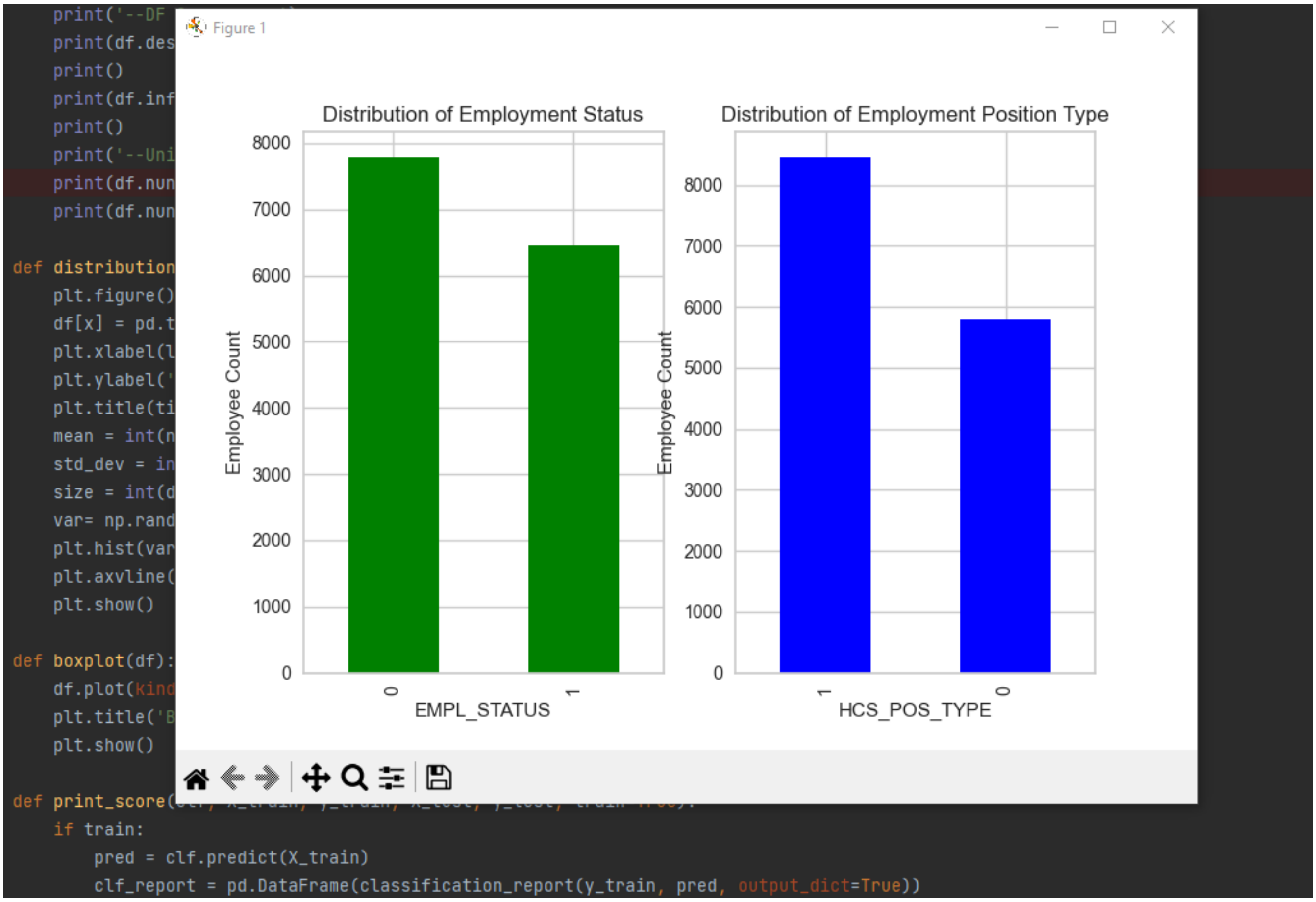
With the environment ready, the data was imported into a Python data frame using Pandas. The data was munged by applying numerical designations to all numerical fields in the data frame. This researcher created a function to run a series of descriptive functions to help understand how this researcher shaped the data frame and provide summary statistics for each field. The following images show how the descriptive functions run successfully, summarizing the data frame.



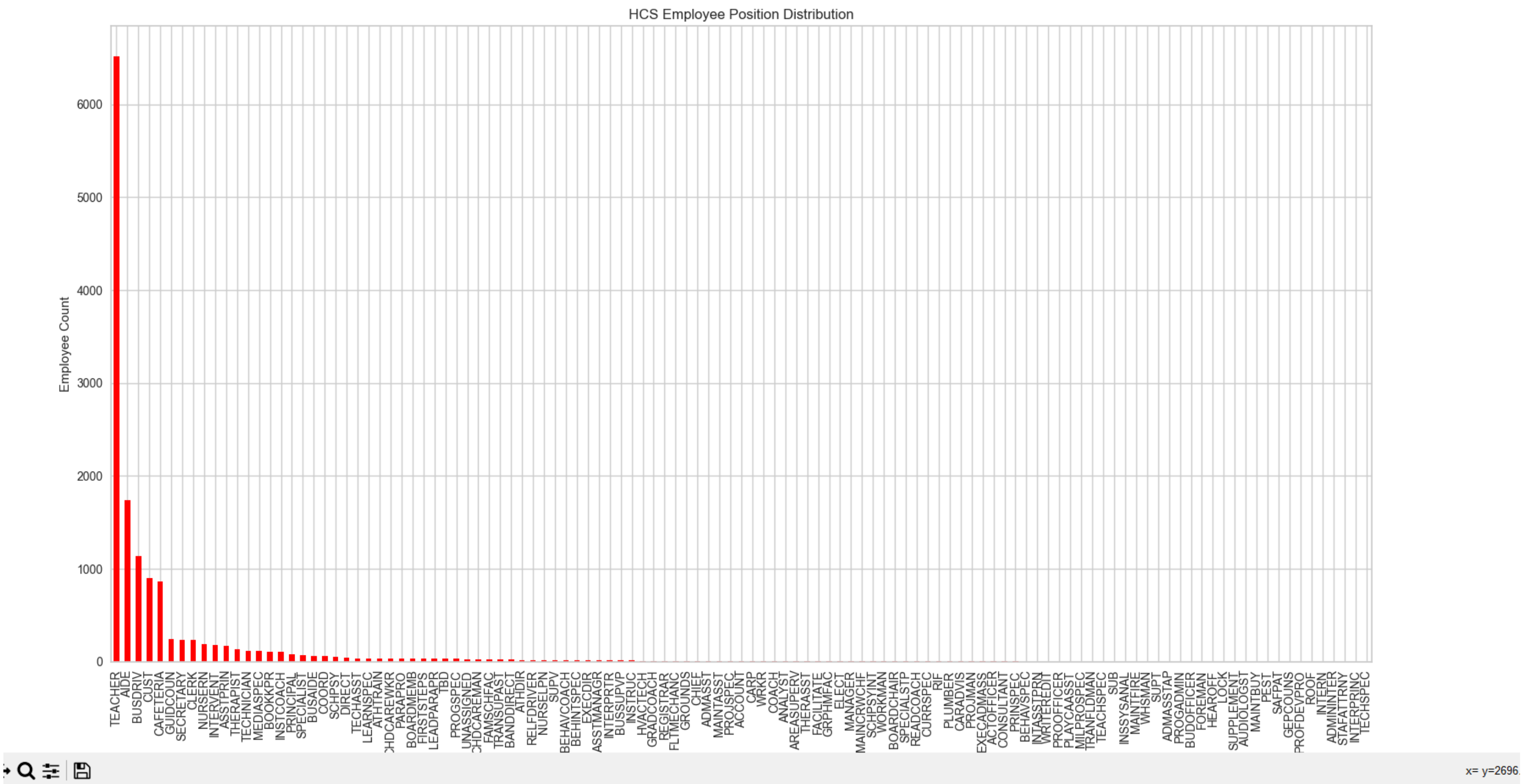


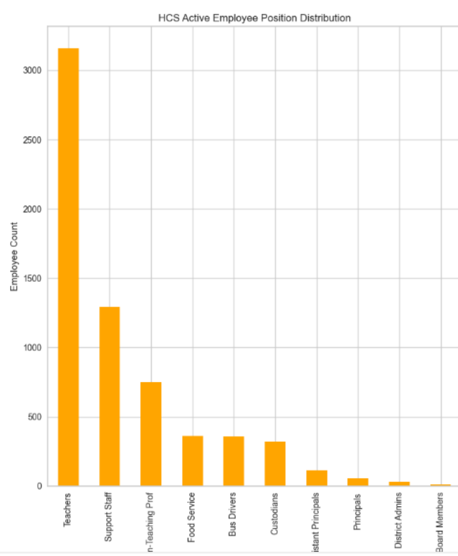


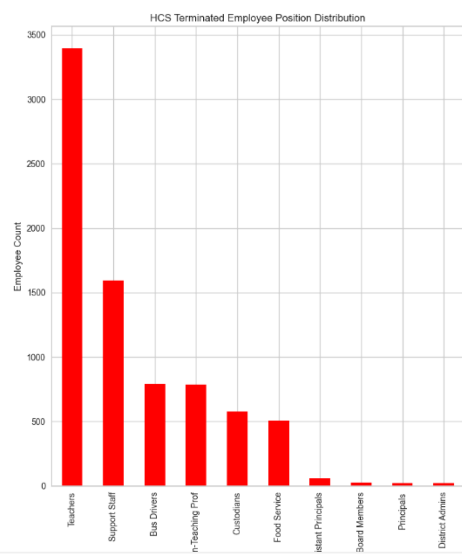
After the program lists the data frame's summary statistics, the program did not reveal any noticeable outliers. The analysis continued by looking at a distribution of the number of terminated employees compared to those still active and the number of professional and support staff. While there is a slight imbalance in the number of terminated employees compared to active employees, this researcher chose not to remove any records. If this researcher determines there needs to be a balance between the number of active (1) and terminated (0) employees, this researcher would remove the older termination records, keeping the more recent records for analysis. The following image also shows an imbalance between professional (1) and support (0) employees.



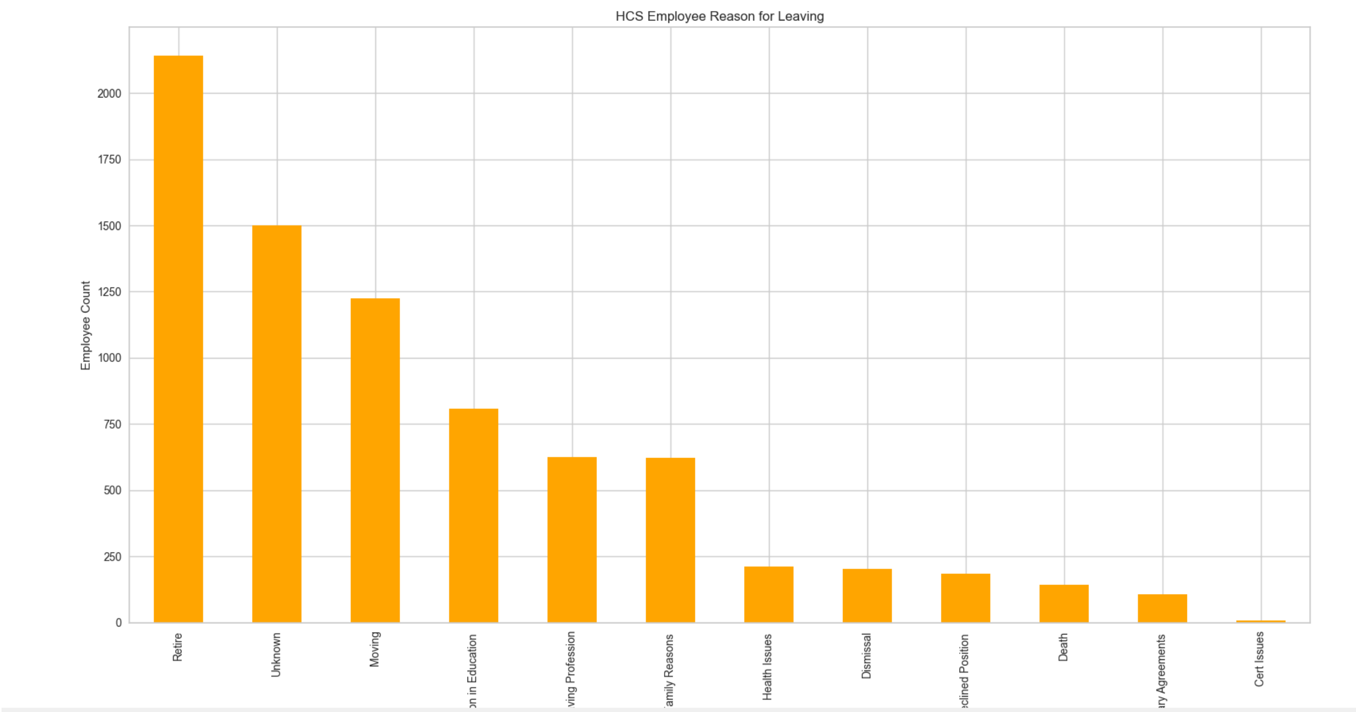
The following images show the distribution of employees based on their position type and position category. The chart shows a large imbalance in the number of teacher records compared to other employee types. The top four employee types after Teachers are aides, bus drivers, custodians, and cafeteria staff. This researcher added the option to focus specific position types if needed.



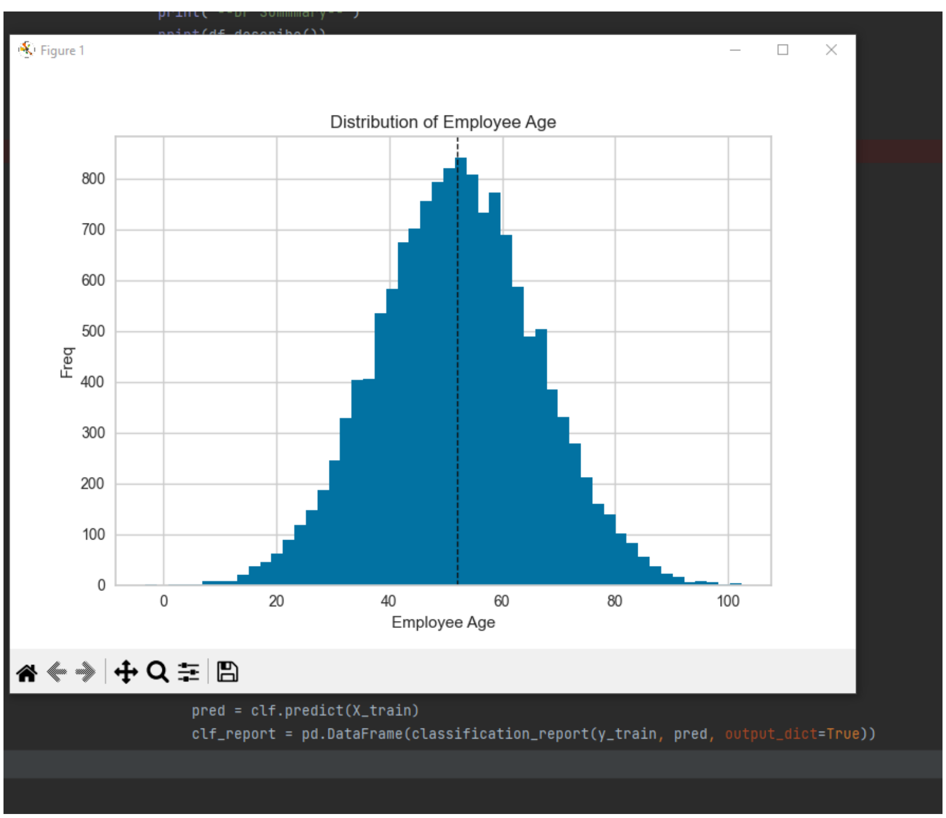




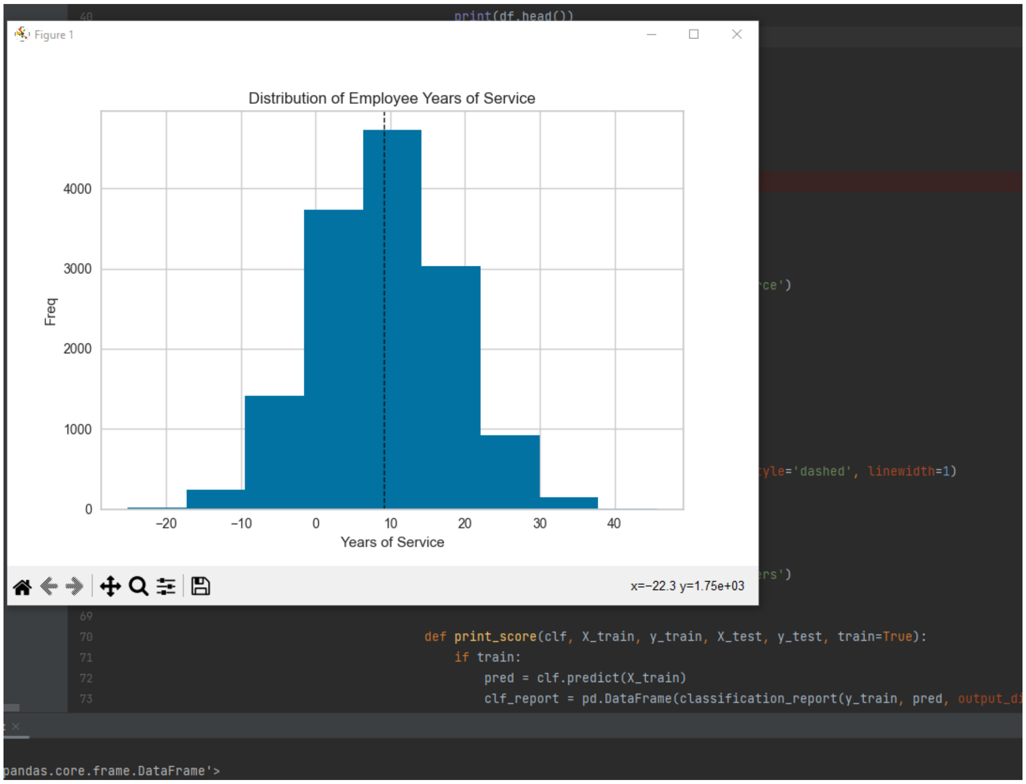
The following images show the distribution of employees based on their position type and position category. The chart shows a significant imbalance in teacher records compared to other employee types. The top four employee types after Teachers are aides, bus drivers, custodians, and cafeteria staff. This researcher added the option to focus on specific position types if needed.

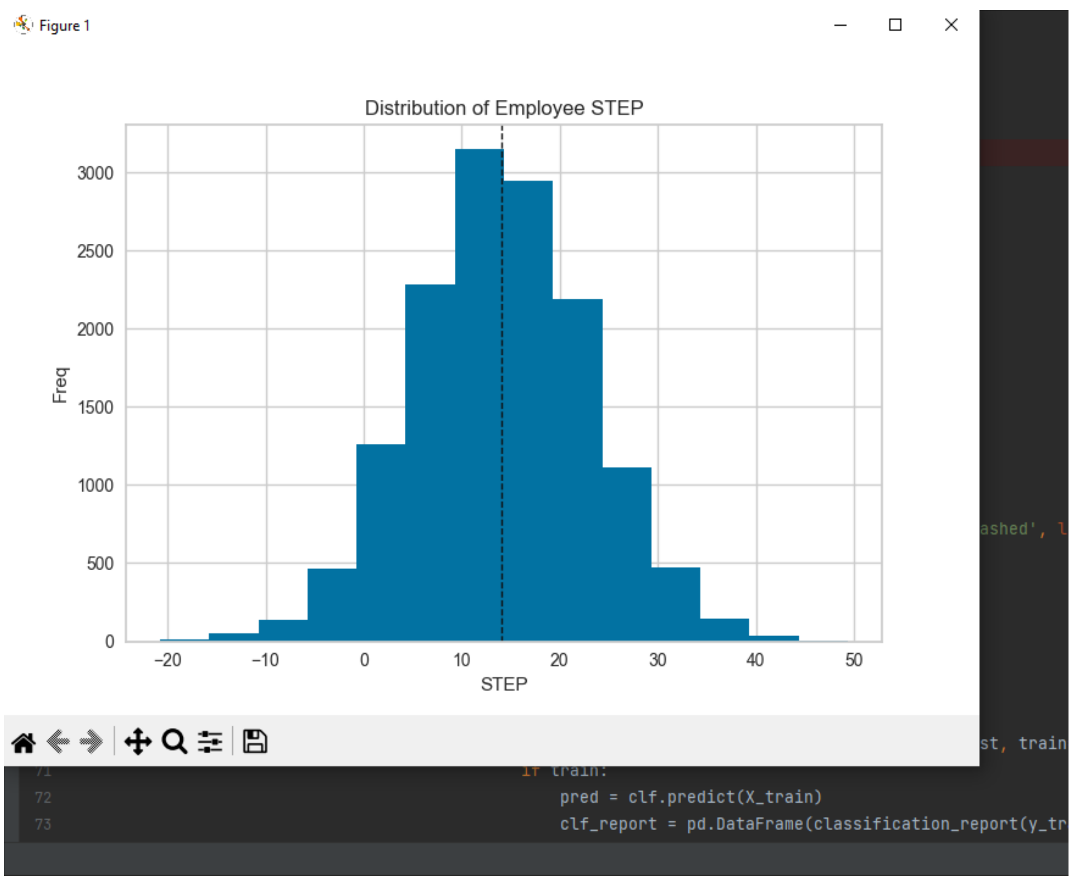


The following image shows employee distribution by age. The distribution shows that the average age for employees is 51, indicating an older workforce.

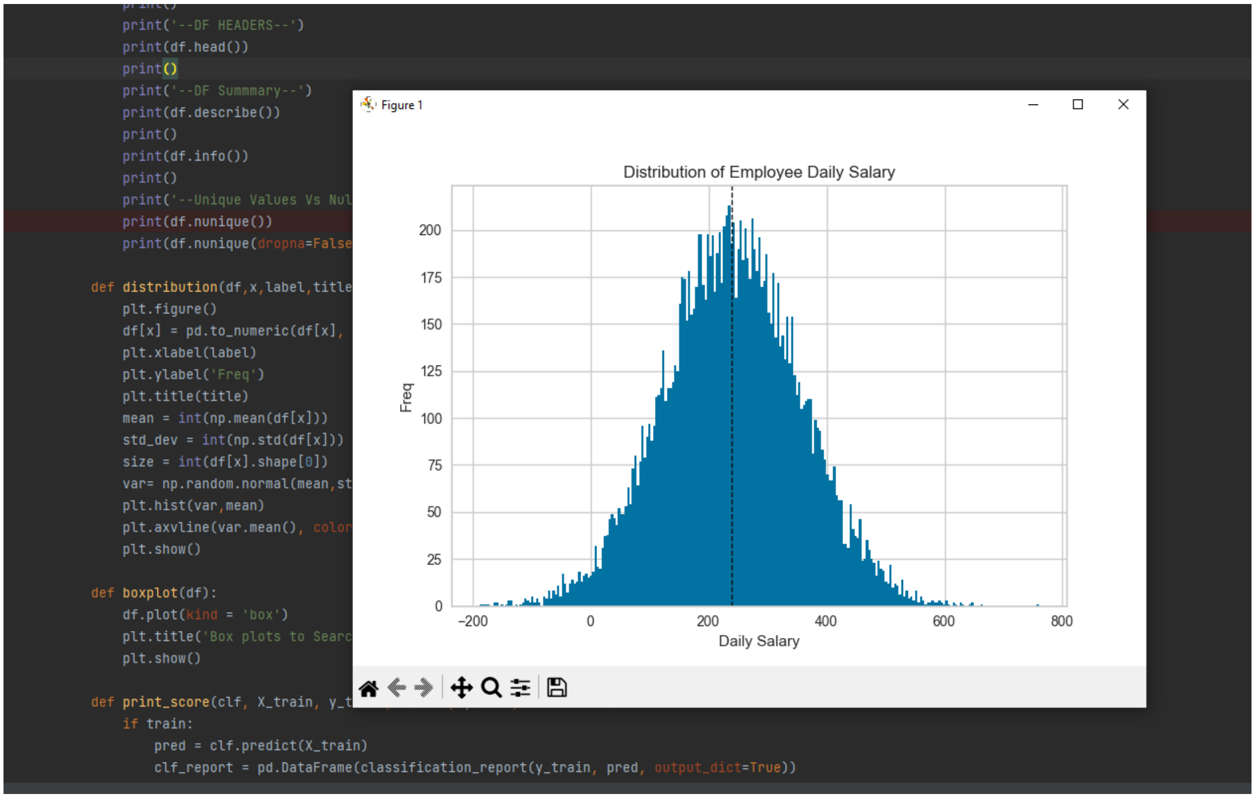


The following images show employee distribution by years of service and steps. Steps are similar designations to years of service, so these charts are expected to show identical results. If these charts were not comparable, an error would be flagged somewhere in the data.

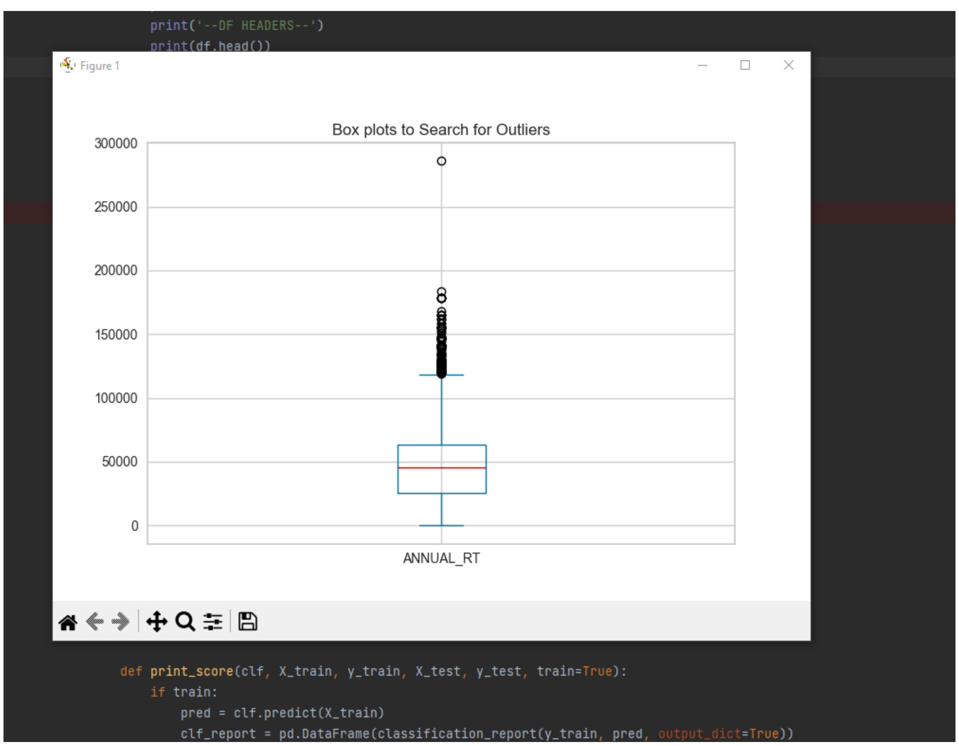




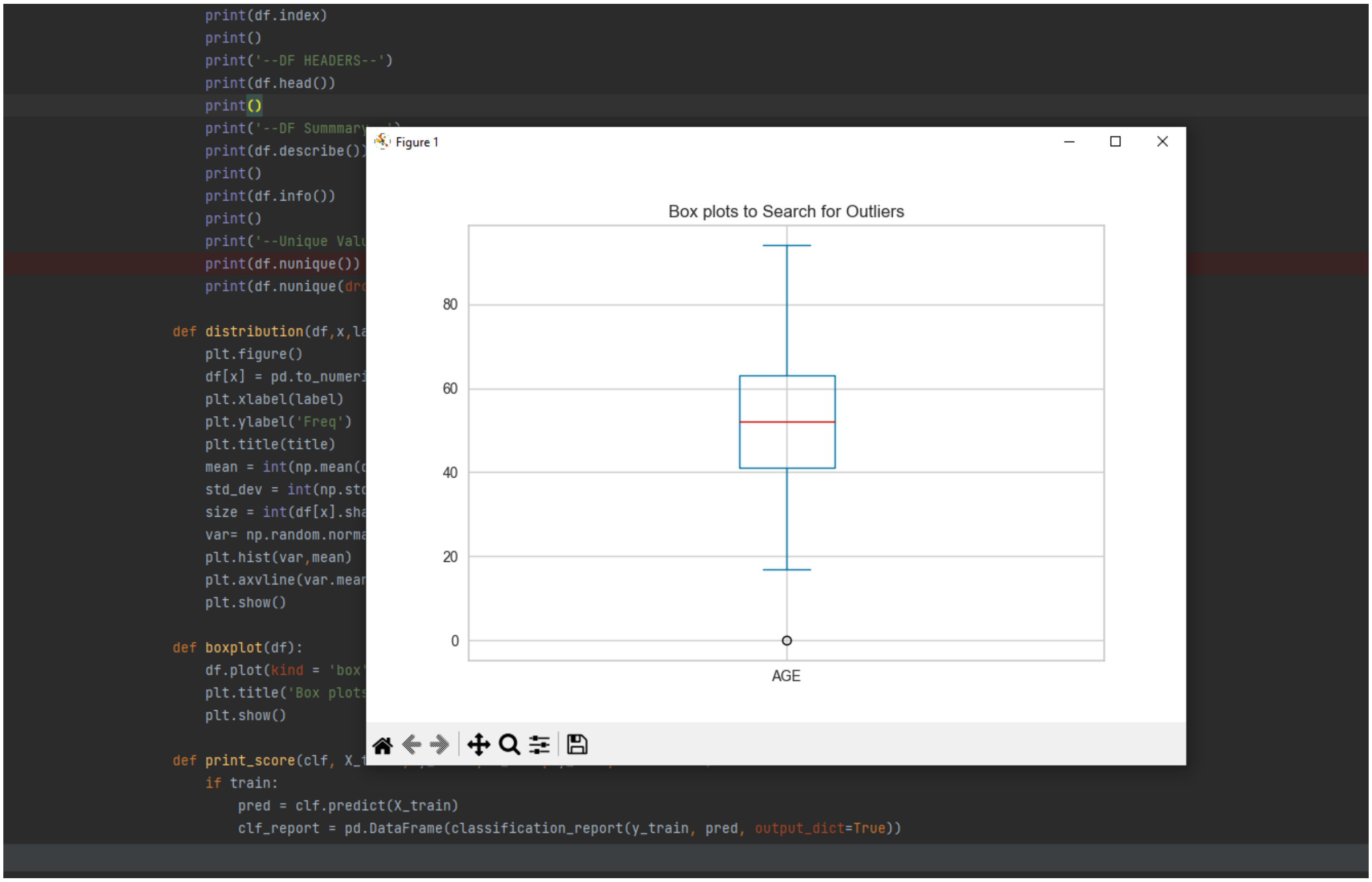
The following image shows the distribution of the employees' daily salaries. Some outliers can be seen at the extremities of the distribution. While these can be removed, this researcher has included them in this analysis.



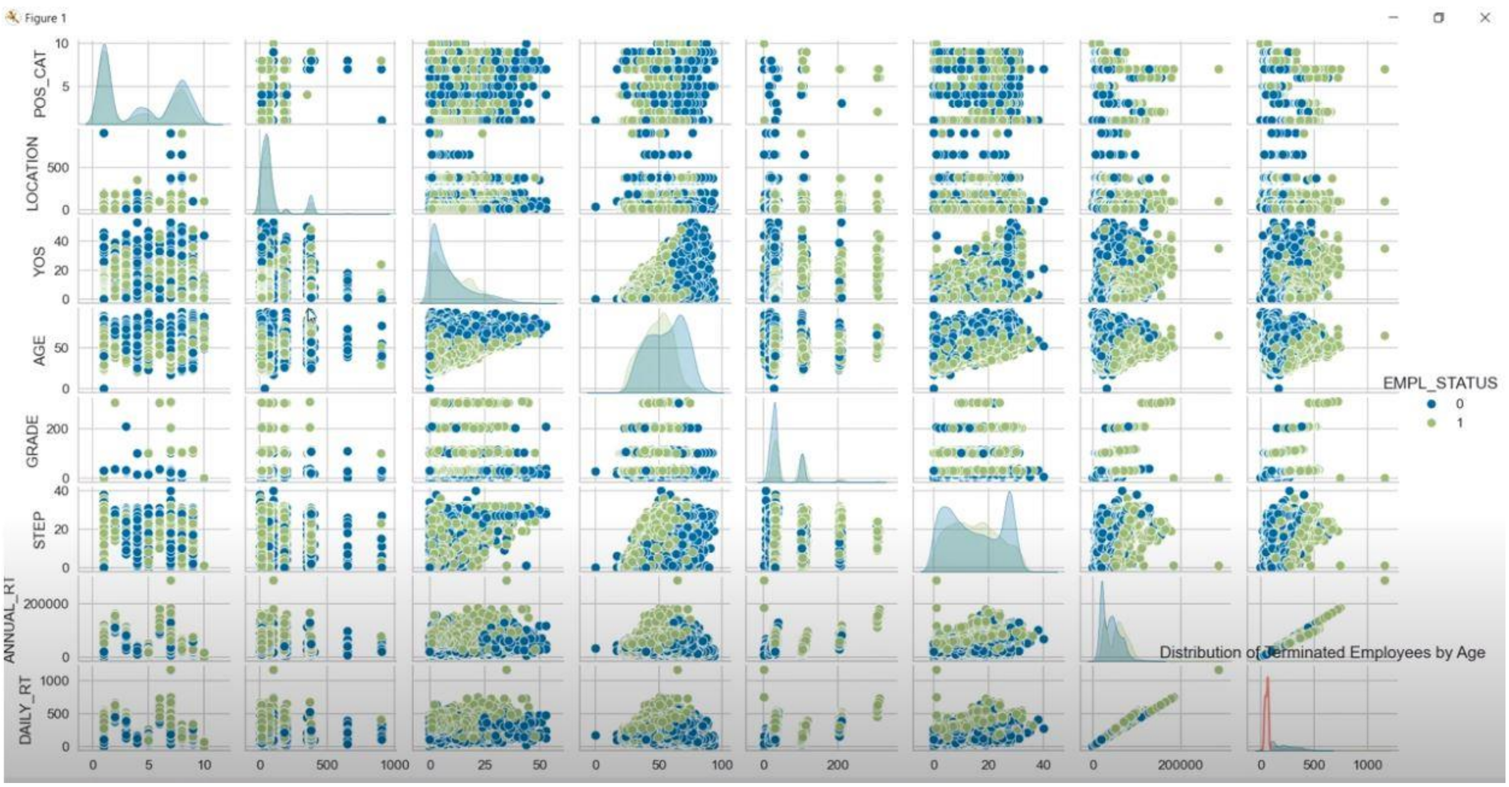
The following image shows outlier salaries for employees, indicating a severe imbalance in the district's pay practices. The outlier salaries are for the principal, district administrators, and superintendent.



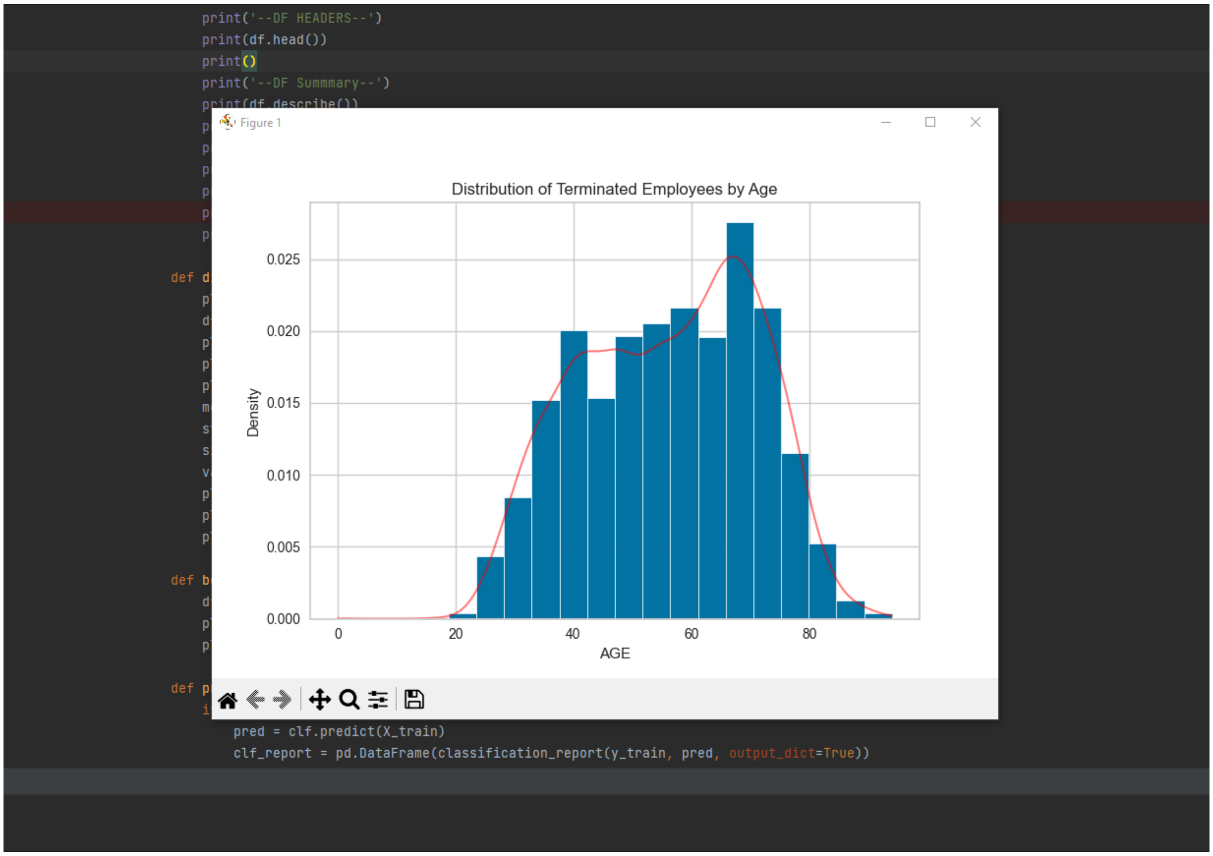
When looking for outliers for employee age, the program found an outlier for an employee who was 0 years old. Upon further research, this researcher discovered that the employee needed to provide Horry County with information, as the district terminated the employee before beginning work. This record will be an outlier but included in this analysis as issues with filling out new hire paperwork could indicate terminations..



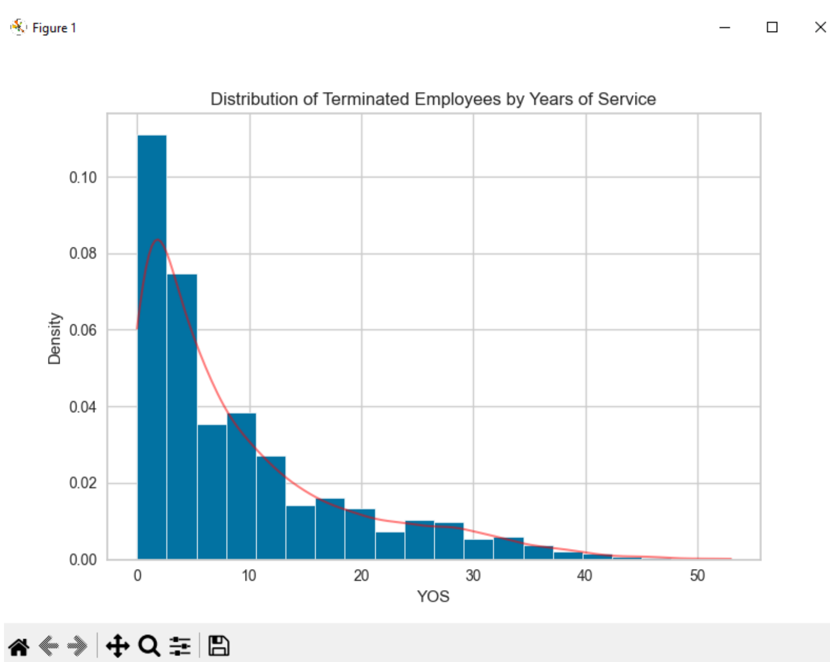
The following image displays a pair plot created by the program to help this researcher visualize any hidden trends when comparing the different variables in the dataset. This technique helps determine what type of machine learning algorithm should be used as this researcher can see logistic, linear, and other types of nonlinear relationships among the variables. This researcher can use these charts to validate any hypothesis regarding the variables' relationships and check for data inconsistencies. Pair plotting is a valuable technique for visualizing and describing the relationships of many variables at one time.



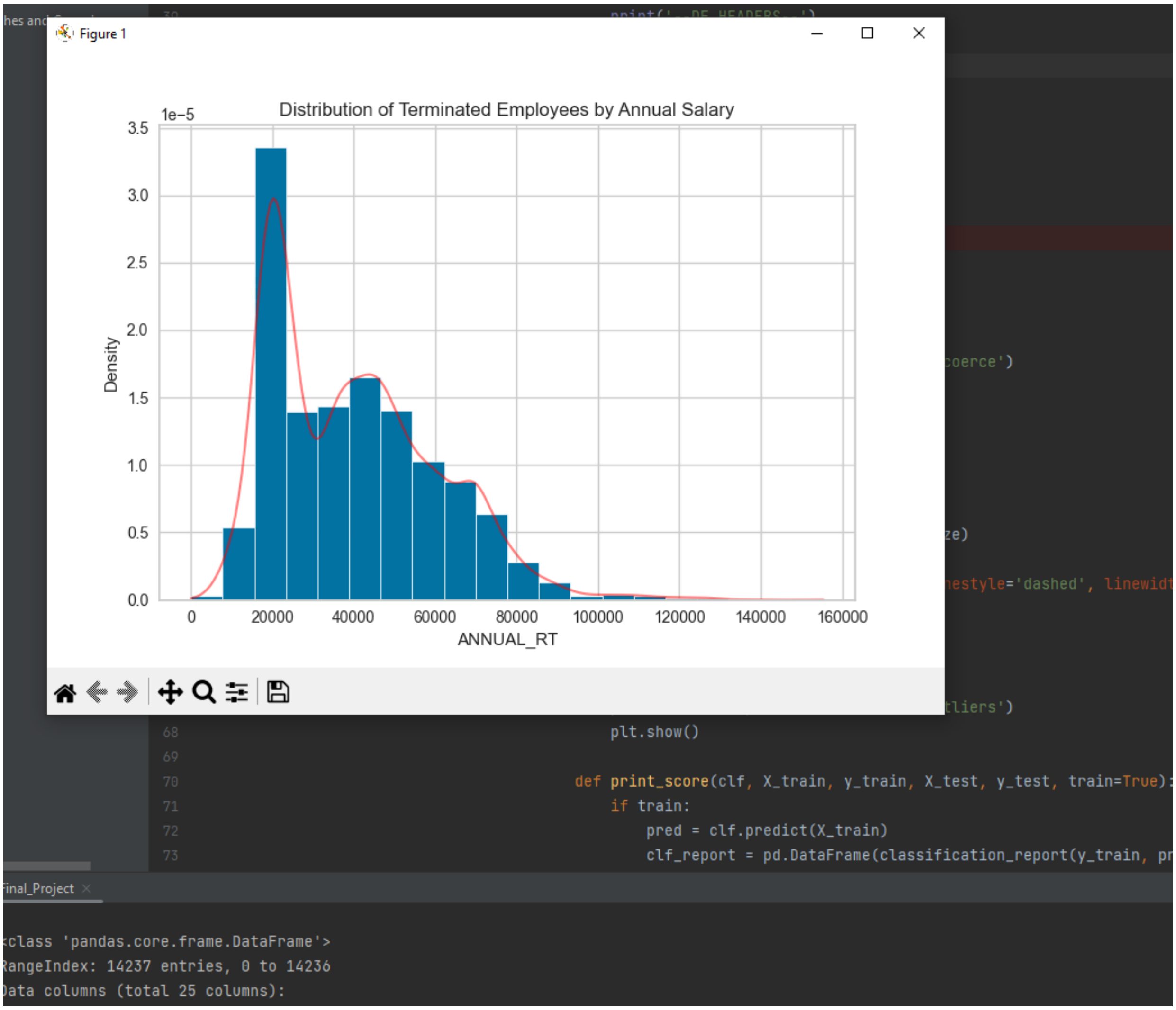
The following image shows the distribution of the employees who have been terminated by age. There is a large spike in terminations among employees between the ages of 35 and 40 and when they are over 60.



The following image shows the distribution of terminated employees' years of service. There is a significant spike in terminations with employees who have served less than five years. While retirement is the more frequent cause of termination, this is interesting as this chart indicates most employees who terminate leave within five years of working for the district. This researcher also developed the hypothesis that retirement is the dominating reason for employees' termination when they have over ten years of service.



The following image shows the distribution of employees terminated by their annual salary. There is a significant spike in terminations for employees who earn less than $30,000 annually. The chart is interesting as it can relate to the outlier chart when looking at the wide variance in employee salaries.

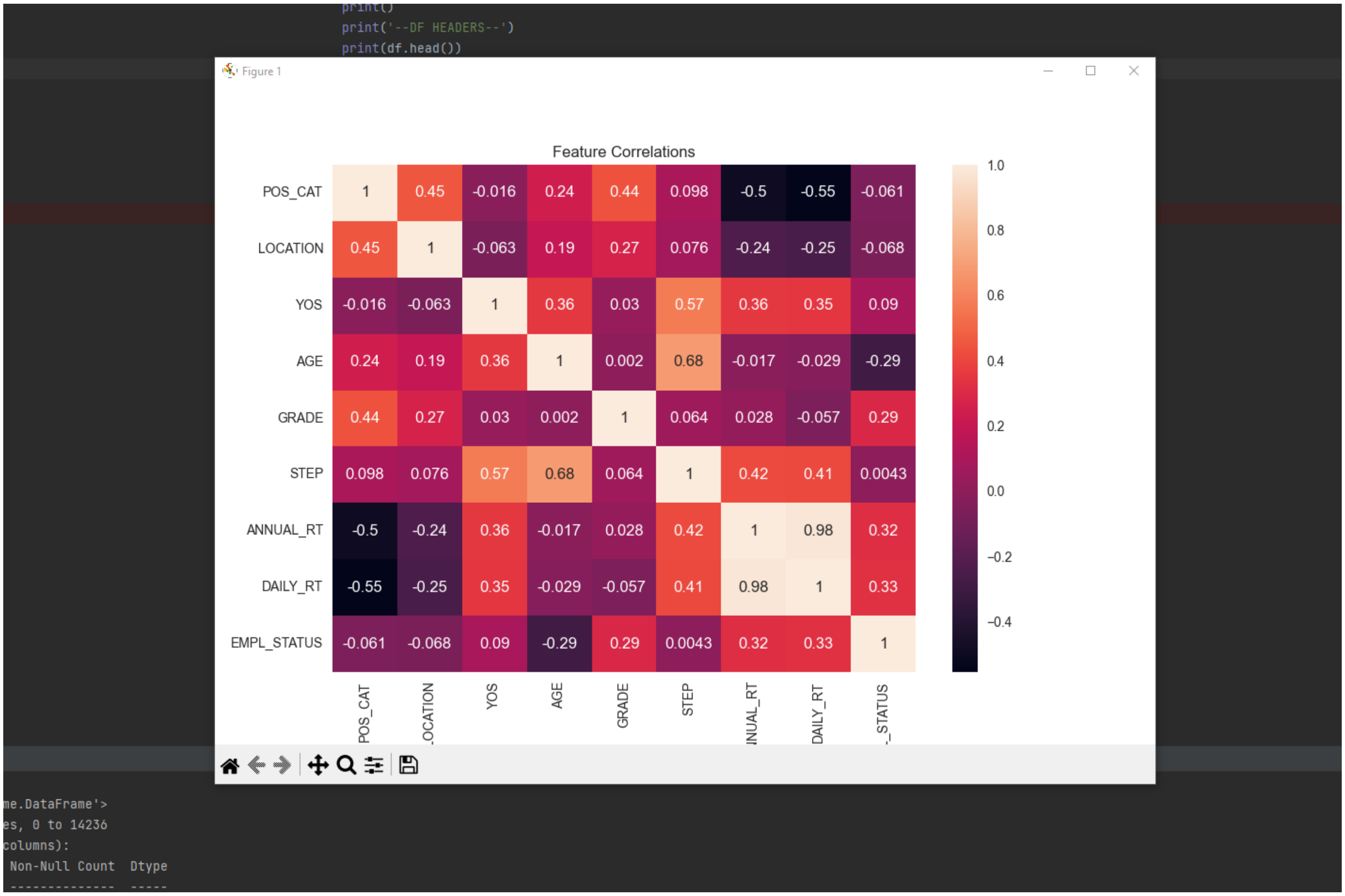


Descriptive Analysis Summary

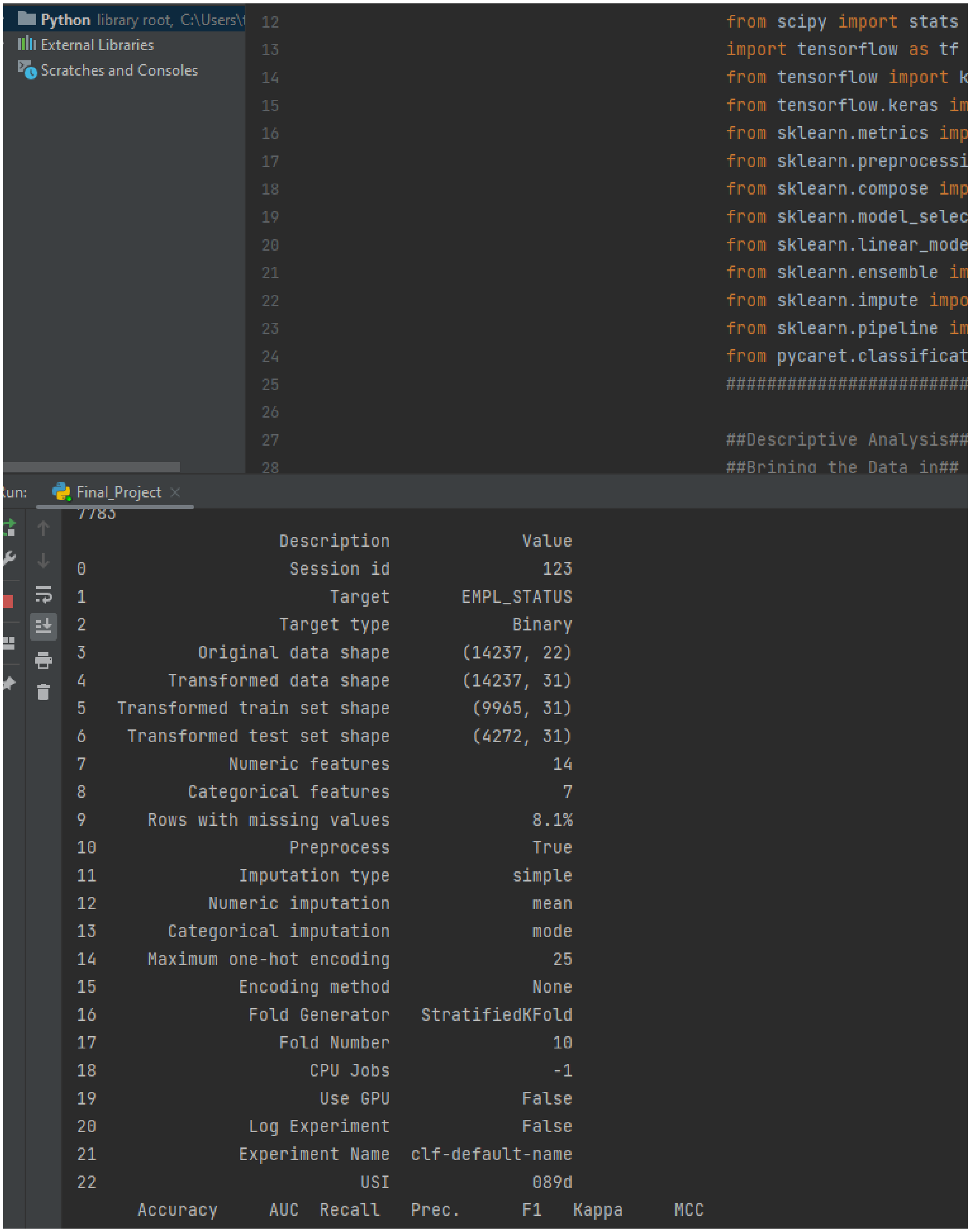
Based on the historical data provided, the descriptive analysis described what factors lead to employee terminations. The study showed that an employee's age may significantly affect terminations. As employees age, they retire, and younger employees are more likely to shift jobs and careers. There is a significant spike in employees who leave before five years of service, indicating there may be a lack of support for new employees. The analysis also described that unknown causes are the second leading cause for terminations, which could suggest that employees are disgruntled. The absence of data indicates many terminated employees do not want to fill out the resignation paperwork as they leave the district. Custodial, cafeteria, and bus drivers have higher turnover than professional employees, indicating low pay and job difficulties could be leading causes for employees leaving before retirement. There is also a significant variation in employee salaries, which may be a component that leads to employees terminating before retirement.

Diagnostic Analysis

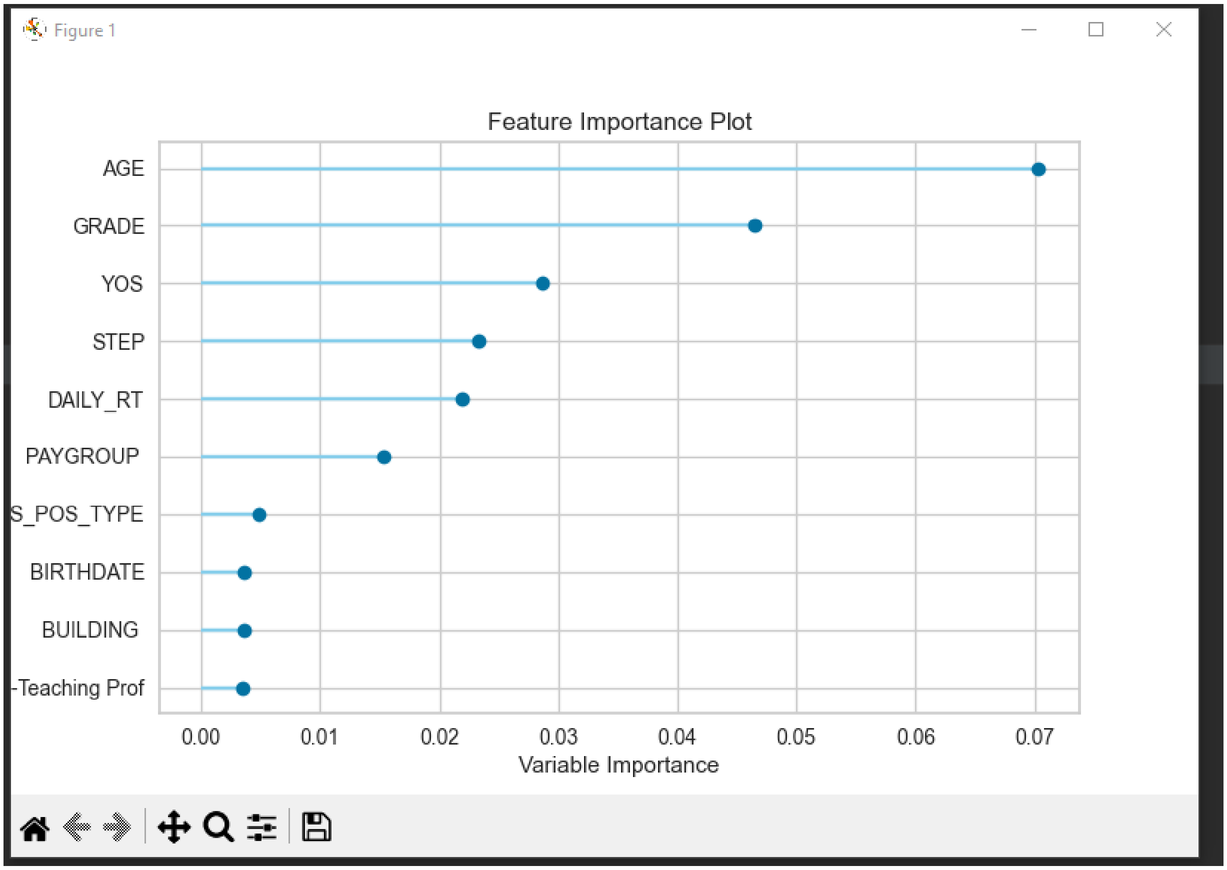
The next stage of this analysis is to build on the previous descriptive analysis and identify key variables that would affect employee terminations. The following images show feature correlations, creating the basis for the diagnostic analysis. Daily rate, annual salary, and employee pay grade all have stronger positive correlations to employment status, while age has a significant negative correlation to employee status. The position category has a strong negative correlation to daily and annual rates.



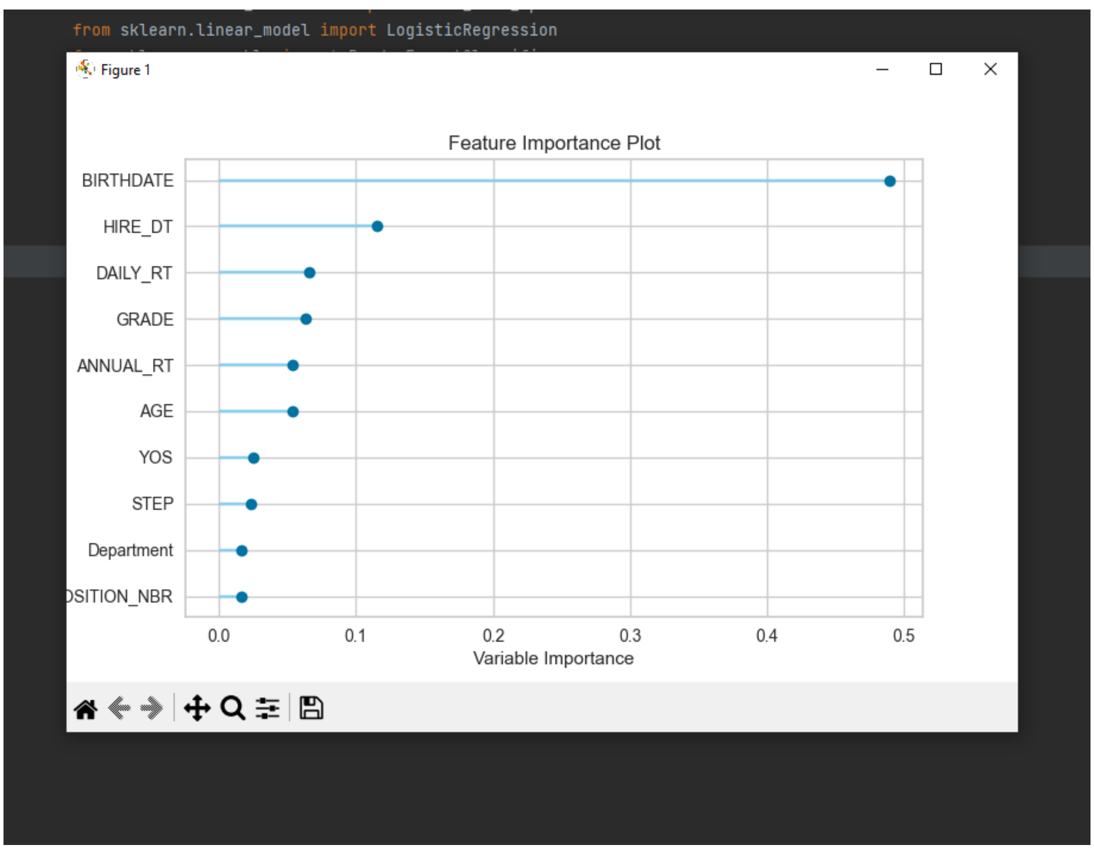
The Python Package PyCaret validated essential variables for the following diagnostic analysis. As this analysis will use logistic regression and random forest decision trees, PyCaret created two models to determine significant variables using logistic regression and random forests. The package also provided summary statistics identifying 14 numeric features, seven categorical features, and 8.1% of the rows in the dataset that were missing data. The created models indicated that age, pay grade, years of service, step, daily rate, and annual salary are significant variables, which supports the correlation analysis above.



Feature Significance for using Logistic Regression



Feature Significance for using Random Forest



Diagnostic Analysis Summary

The diagnostic analysis showed that employee age is the most significant variable in determining when an employee will terminated followed by employee grade and step, years of service and their daily rate. The random forest model was identical to the logistic regression model showing employee birthday, hire date, daily rate, and employee grade as significant variables. Both models support age, years of service, employee grade and step, and salary as important features to determine if an employee will terminate. Employee grade is related to position category as each positon is assigned a pay grade. These features will be used in the following predictive analysis.

Predictive Analysis

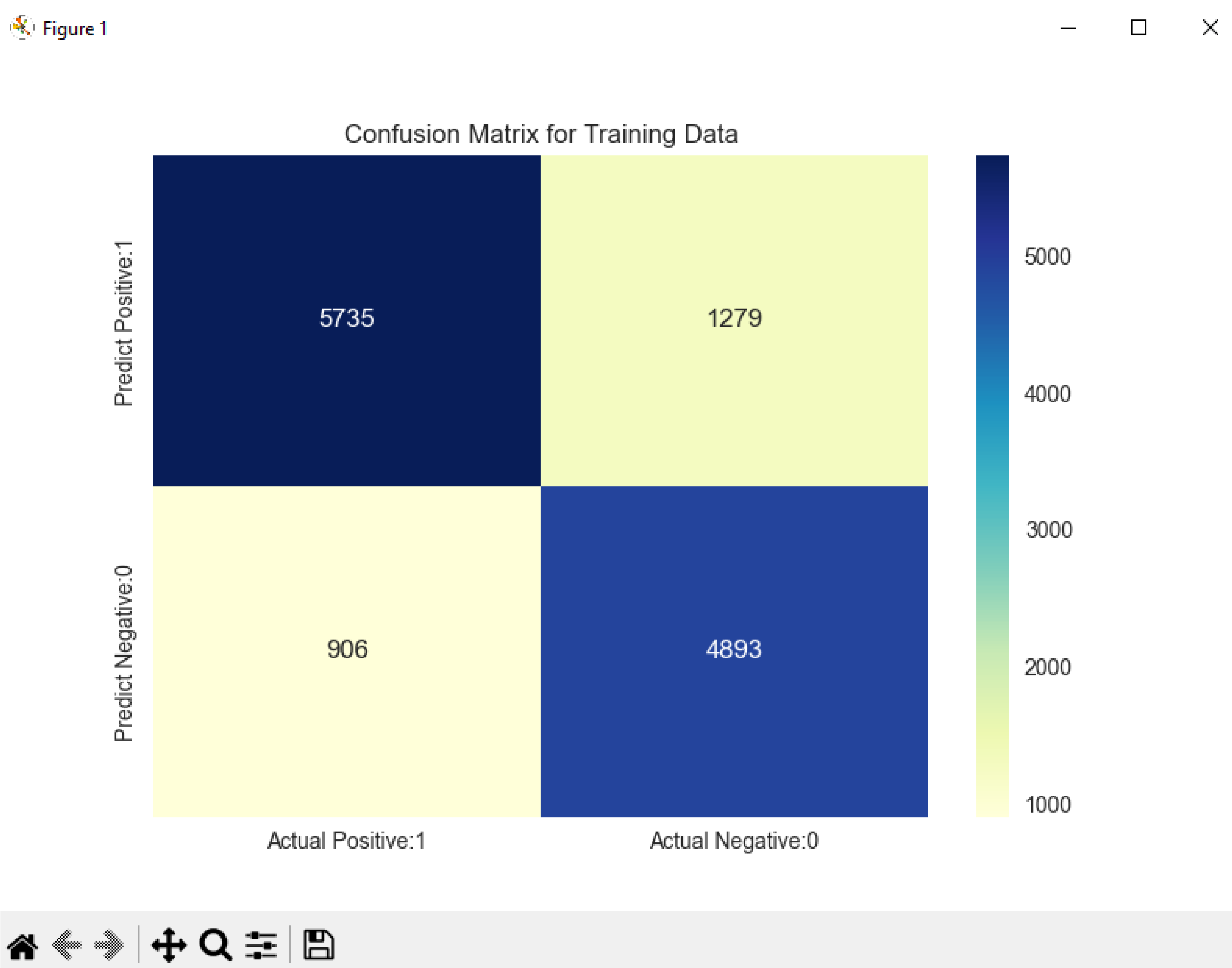
After this researcher identified significant variables, this researcher used the machine learning algorithms logistic regression, random forest model, and supervised neural network to compare employee termination predictions based on selected variables.

The data was split into training and test sets using a 90% to 10% split and was normalized using a standard scaler. The following image shows the data being split into training and test data before running the model.

```
#SETTING UP THE ANALYSIS
X = df_corr.drop(['EMPL_STATUS'],axis = 1)  #Features chosen from our Descirptive Analysis
y = df['EMPL_STATUS']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=15)

num_columns = ['POS_CAT','LOCATION','YOS','AGE','GRADE','STEP','ANNUAL_RT','DAILY_RT']

scale = make_column_transformer(
    (MinMaxScaler(), num_columns),
    (StandardScaler(), num_columns),
    remainder='passthrough'
)

X_train = scale.fit_transform(X_train)
X_test = scale.transform(X_test)
```

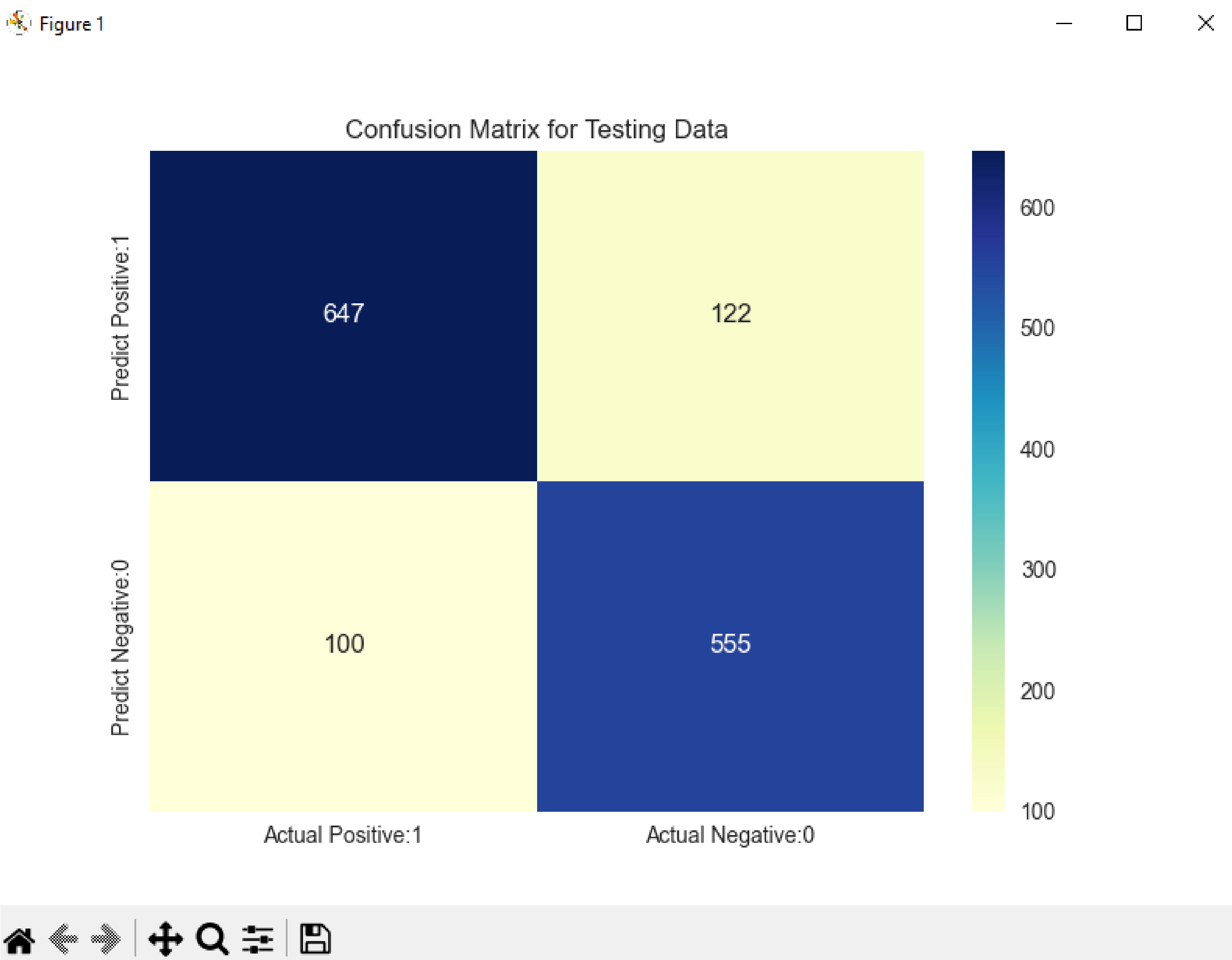
Logistic Regression Results with Training Data using a Standard Scaler

```
Results Logistic Regression With Training Data
Training Result:
================================================

_______________________________________________
CLASSIFICATION SUMMARY:
                     0            1  accuracy     macro avg  weighted avg
precision     0.863575     0.792774   0.82947      0.828174      0.831531
recall        0.817650     0.843766   0.82947      0.830708      0.829470
f1-score      0.839985     0.817476   0.82947      0.828730      0.829798
support    7014.000000  5799.000000   0.82947  12813.000000  12813.000000
Accuracy Score: 82.95%
```



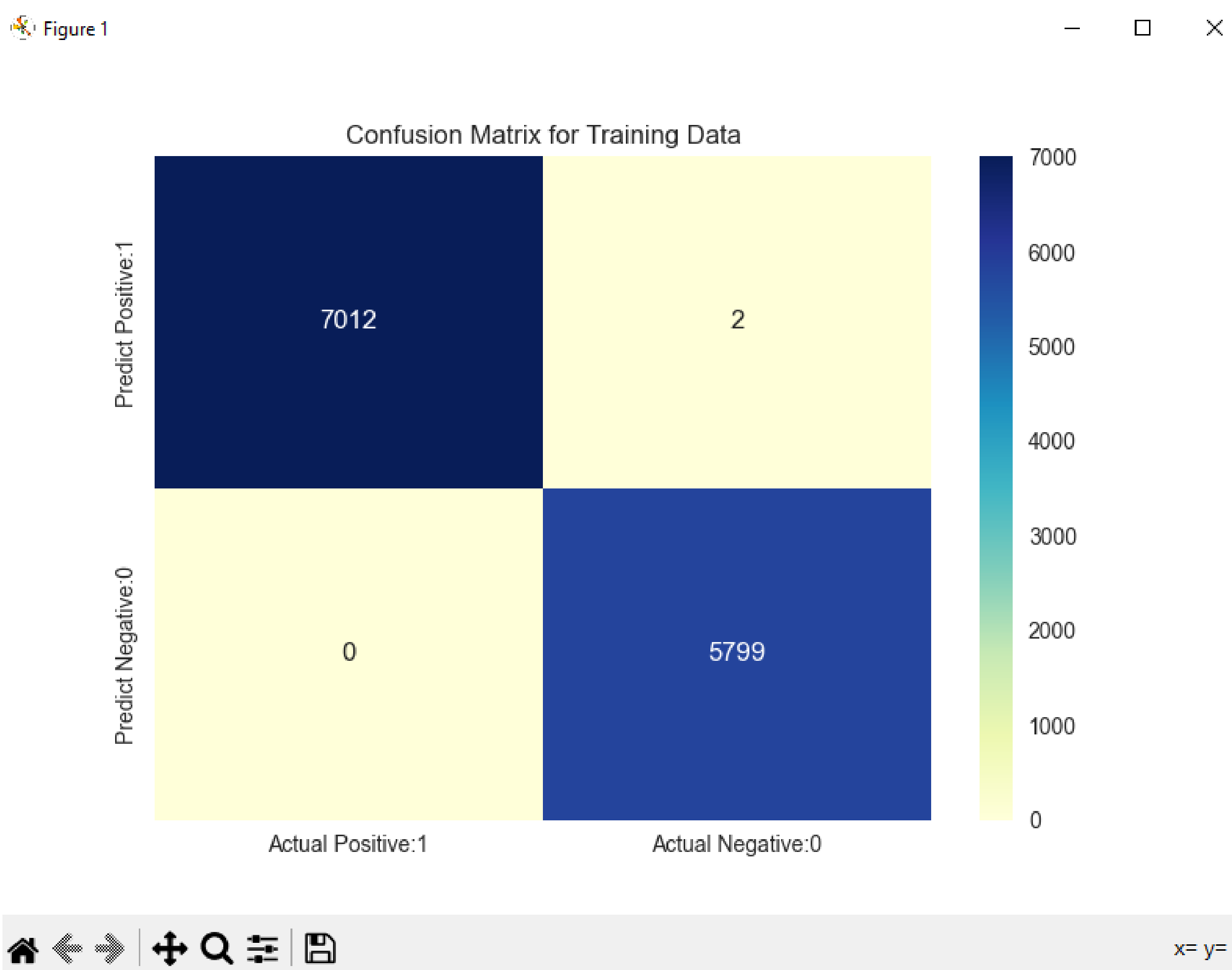
L Logistic Regression Results with Testing Data using a Standard Scaler

```
Results Logistic Regression With Testing Data
Testing Results:
================================================
CLASSIFICATION SUMMARY:
                    0           1  accuracy    macro avg  weighted avg
precision    0.866131    0.819793  0.844101     0.842962      0.844817
recall       0.841352    0.847328  0.844101     0.844340      0.844101
f1-score     0.853562    0.833333  0.844101     0.843448      0.844257
support    769.000000  655.000000  0.844101  1424.000000   1424.000000
_______________________________________________
Accuracy Score: 84.41%
```



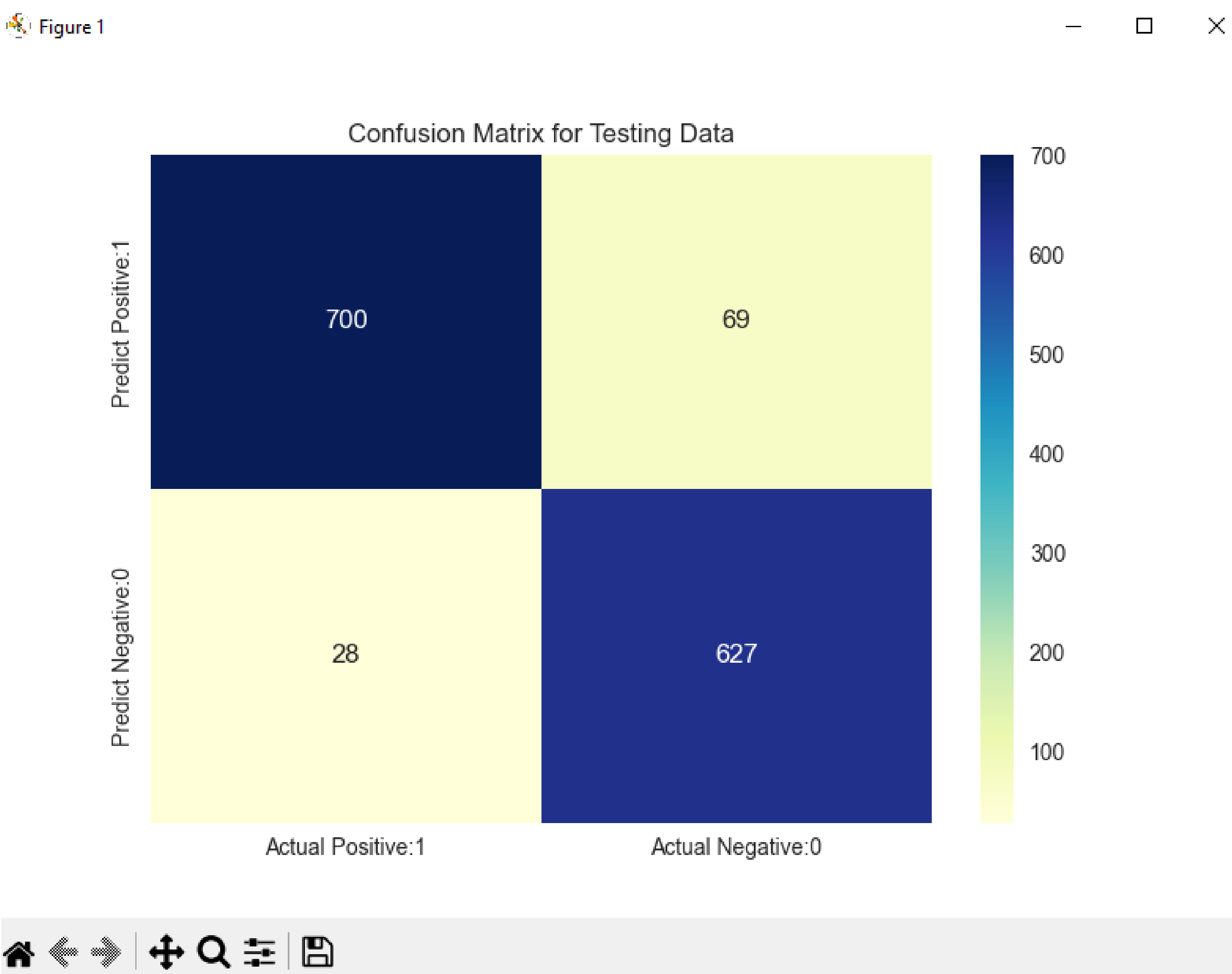
Random Forest Classifier Results with Training Data based on 1000 estimates

```
Results Random Forest Classifer with Training Data
Training Result:
================================================

-----------------------------------------------
CLASSIFICATION SUMMARY:
                      0            1  accuracy     macro avg  weighted avg
precision      1.000000     0.999655  0.999844      0.999828      0.999844
recall         0.999715     1.000000  0.999844      0.999857      0.999844
f1-score       0.999857     0.999828  0.999844      0.999842      0.999844
support     7014.000000  5799.000000  0.999844  12813.000000  12813.000000
Accuracy Score: 99.98%

-----------------------------------------------
```



Random Forest Classifier Results with Testing Data based on 1000 estimates

```
Results Random Forest Classifer with Testing Data
Testing Results:
================================================
CLASSIFICATION SUMMARY:
                    0           1  accuracy    macro avg  weighted avg
precision    0.961538    0.900862  0.931882     0.931200      0.933629
recall       0.910273    0.957252  0.931882     0.933762      0.931882
f1-score     0.935204    0.928201  0.931882     0.931703      0.931983
support    769.000000  655.000000  0.931882  1424.000000   1424.000000

_______________________________________________
Accuracy Score: 93.19%
```

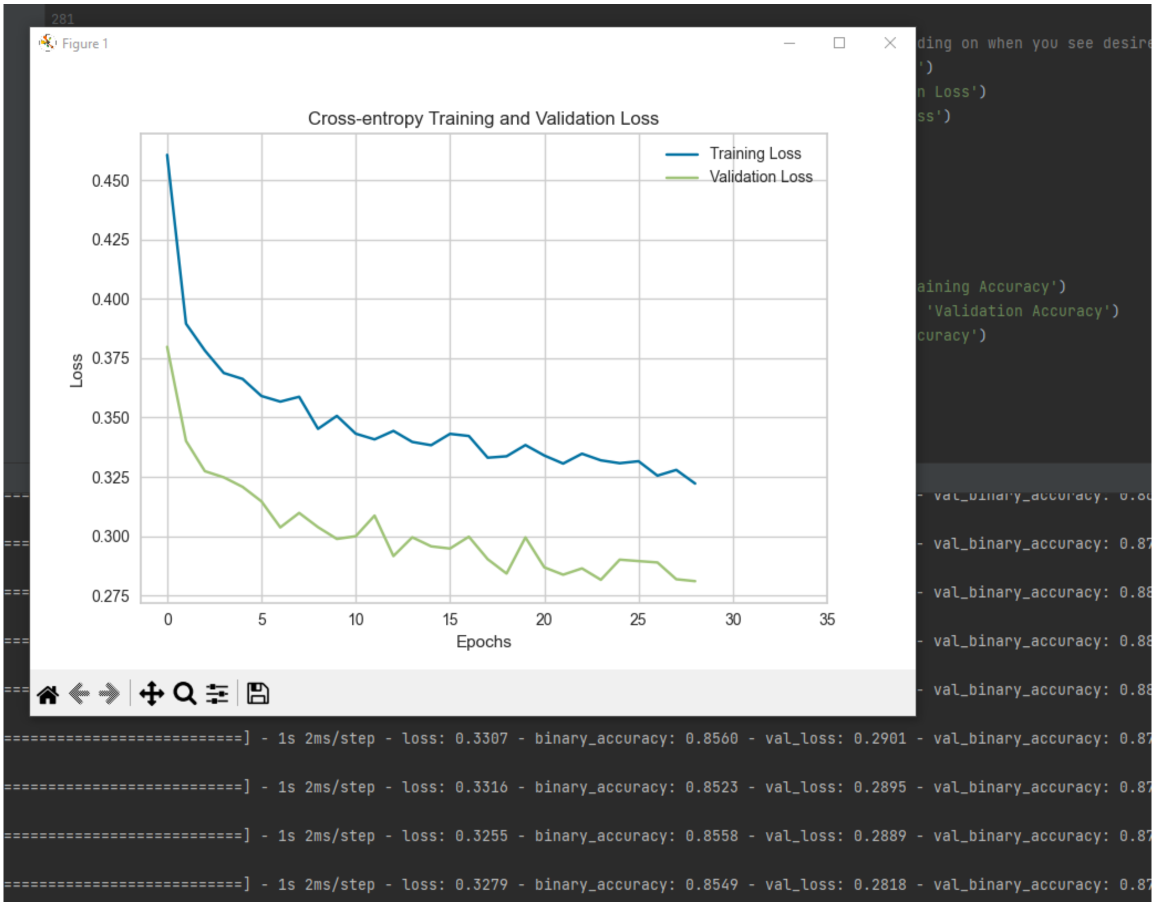


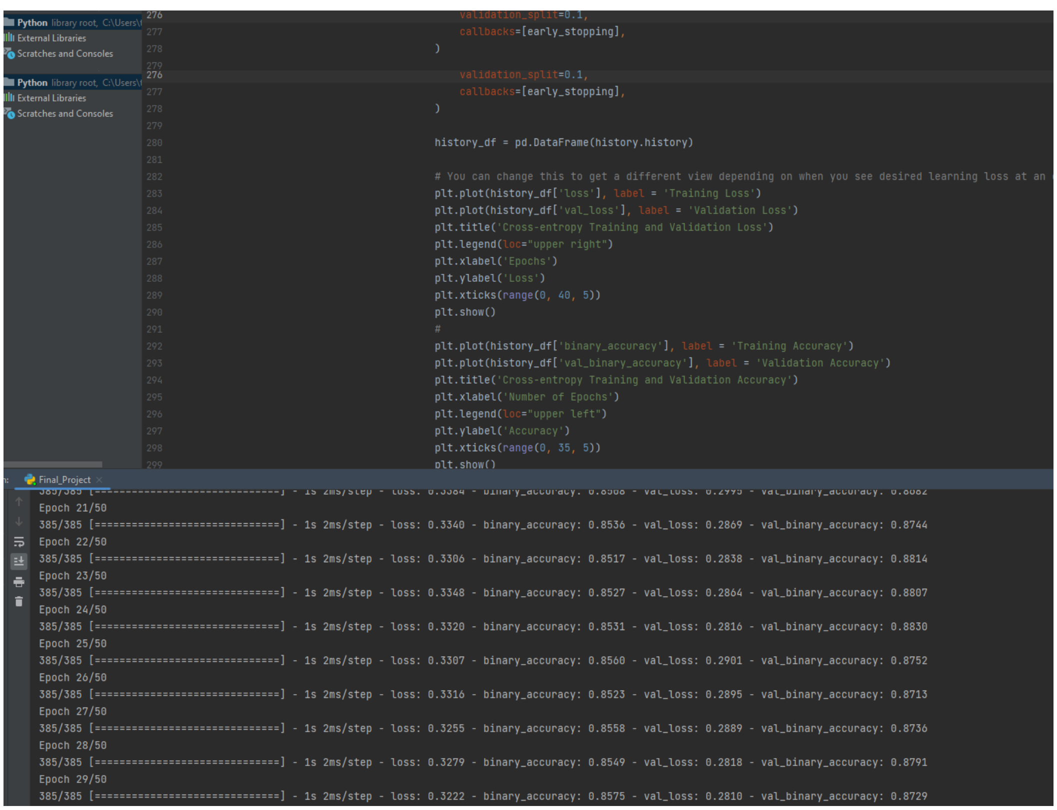
Supervised Neural Network

This researcher created a supervised Neural Network using binary cross-entropy to validate the logistic regression and random forest classifier results. This researcher split the dataset using 90% for training data and 10% for testing data. Of the 90% training data, 10% was used for validation data when training the model. The images below show the structure of the neural network, the format for the training, testing, and validation data, and the successful completion of training the network.

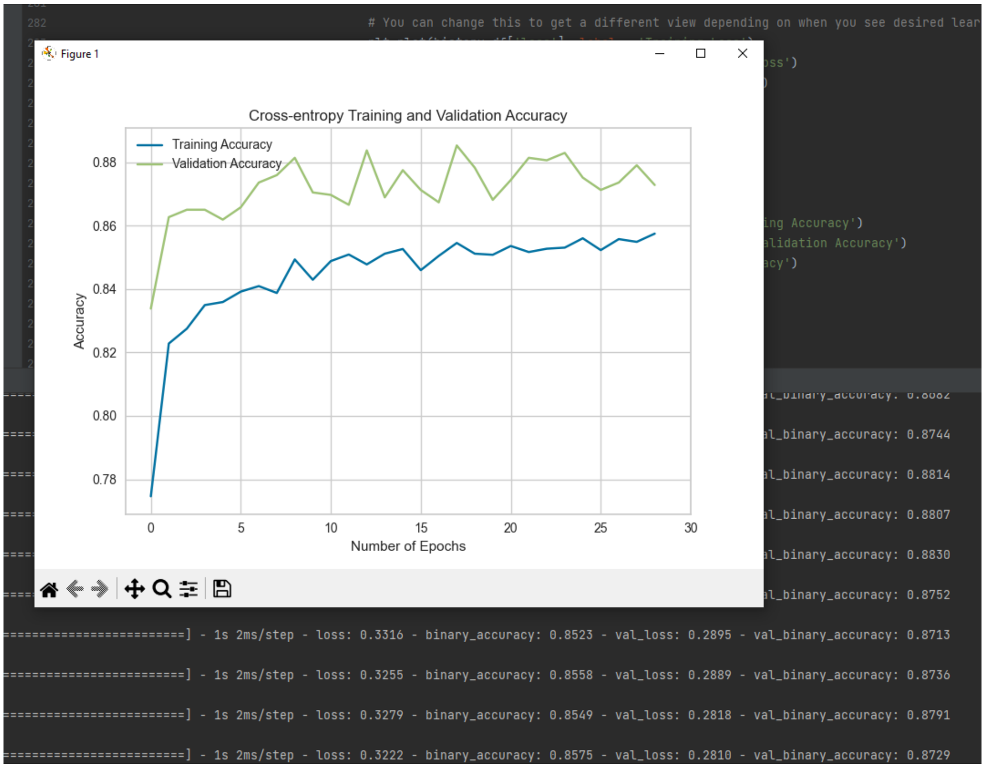
```
##From our X features in the Diagnostic Analyticis
X1_train, X1_test, y1_train, y1_test = train_test_split(X,y,stratify=y, test_size=0.1, random_state=15)
X1_train = scale.fit_transform(X1_train)
X1_test = scale.transform(X1_test)

input_shape = [X1_train.shape[1]]

model = keras.Sequential([
    layers.BatchNormalization(input_shape=input_shape),
    layers.Dense(32,activation='relu'),
    layers.Dense(32,activation='relu'),
    layers.Dense(1,activation='sigmoid')])

model.compile(optimizer='adam',loss='binary_crossentropy',metrics=['binary_accuracy'])

early_stopping = keras.callbacks.EarlyStopping(
    patience=5,
    min_delta=0.001,
    restore_best_weights=True,
)
history = model.fit(
    X1_train, y1_train,
    batch_size=30,
    epochs=50,
    validation_split=0.1,
    callbacks=[early_stopping],
)
```

When analyzing the loss function the final loss value mimics that of the logistic regression but it appears to be under fitting with its high loss value. This is an indication more data is needed to use this model.



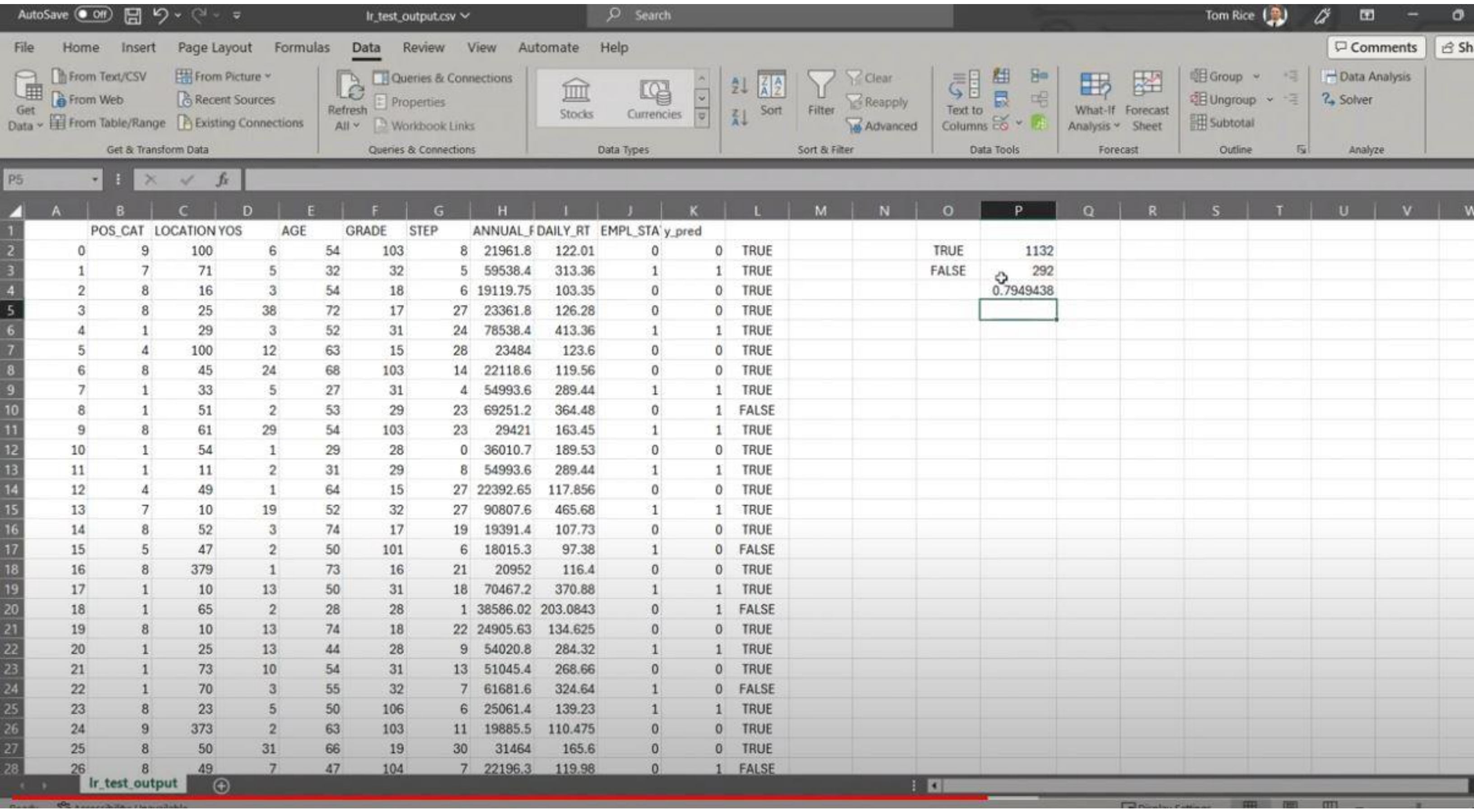


The model's accuracy mimics that of the logistic regression used earlier.



Making Predictions

After comparing all three machine learning algorithms, this researcher recommends using a logistic regression for this dataset. While the random forest classifier appears to be the most accurate, this researcher needs to scrutinize its high accuracy to verify that it accurately predicts employee terminations or if some noise affects the results. The supervised neural network would be ideal if this dataset were significant. Still, the loss function indicates it is under-fitting and needs more data or complex data to perform optimally. The following image displays the test data output from the data set combined with the logistic regression predictions. Creating an output like this allows researchers to scrutinize the model's results. Without using a standard scaler, the logistic regression model could predict employee terminations with 79.49% accuracy.



	POS_CAT	LOCATION	YOS	AGE	GRADE	STEP	ANNUAL_F	DAILY_RT	EMPL_STA	y_pred	
0	9	100	6	54	103	8	21961.8	122.01	0	0	TRUE
1	7	71	5	32	32	5	59538.4	313.36	1	1	TRUE
2	8	16	3	54	18	6	19119.75	103.35	0	0	TRUE
3	8	25	38	72	17	27	23361.8	126.28	0	0	TRUE
4	1	29	3	52	31	24	78538.4	413.36	1	1	TRUE
5	4	100	12	63	15	28	23484	123.6	0	0	TRUE
6	8	45	24	68	103	14	22118.6	119.56	0	0	TRUE
7	1	33	5	27	31	4	54993.6	289.44	1	1	TRUE
8	1	51	2	53	29	23	69251.2	364.48	0	1	FALSE
9	8	61	29	54	103	23	29421	163.45	1	1	TRUE
10	1	54	1	29	28	0	36010.7	189.53	0	0	TRUE
11	1	11	2	31	29	8	54993.6	289.44	1	1	TRUE
12	4	49	1	64	15	27	22392.65	117.856	0	0	TRUE
13	7	10	19	52	32	27	90807.6	465.68	1	1	TRUE
14	8	52	3	74	17	19	19391.4	107.73	0	0	TRUE
15	5	47	2	50	101	6	18015.3	97.38	1	0	FALSE
16	8	379	1	73	16	21	20952	116.4	0	0	TRUE
17	1	10	13	50	31	18	70467.2	370.88	1	1	TRUE
18	1	65	2	28	28	1	38586.02	203.0843	0	1	FALSE
19	8	10	13	74	18	22	24905.63	134.625	0	0	TRUE
20	1	25	13	44	28	9	54020.8	284.32	1	1	TRUE
21	1	73	10	54	31	13	51045.4	268.66	0	0	TRUE
22	1	70	3	55	32	7	61681.6	324.64	1	0	FALSE
23	8	23	5	50	106	6	25061.4	139.23	1	1	TRUE
24	9	373	2	63	103	11	19885.5	110.475	0	0	TRUE
25	8	50	31	66	19	30	31464	165.6	0	0	TRUE
26	8	49	7	47	104	7	22196.3	119.98	0	1	FALSE

TRUE	1132
FALSE	292
	0.7949438