***Communities and Crime Data Set***

***Introduction***

The goal of this project was to predict crime rate in different communities using different regression techniques. And introduce different feature reduction techniques to get most important features which effect more towards crime. Also we have done some exploratory data analysis to see the trend in each communities.

The dataset used for this experiment is real and authentic. The dataset is acquired from UCI machine learning repository website. It is prepared using real data from socio-economic data from 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. This dataset contains a total number of 147 attributes and 2216 instances.

***Algorithm and Implementation***

After deleting our dataset has 125 features and 1 predicted variable. To begin, we chose 13 features of interest and plotted each variable vs the output variable Y (ViolentCrimesPerPop) with scatter plot to get the relation.

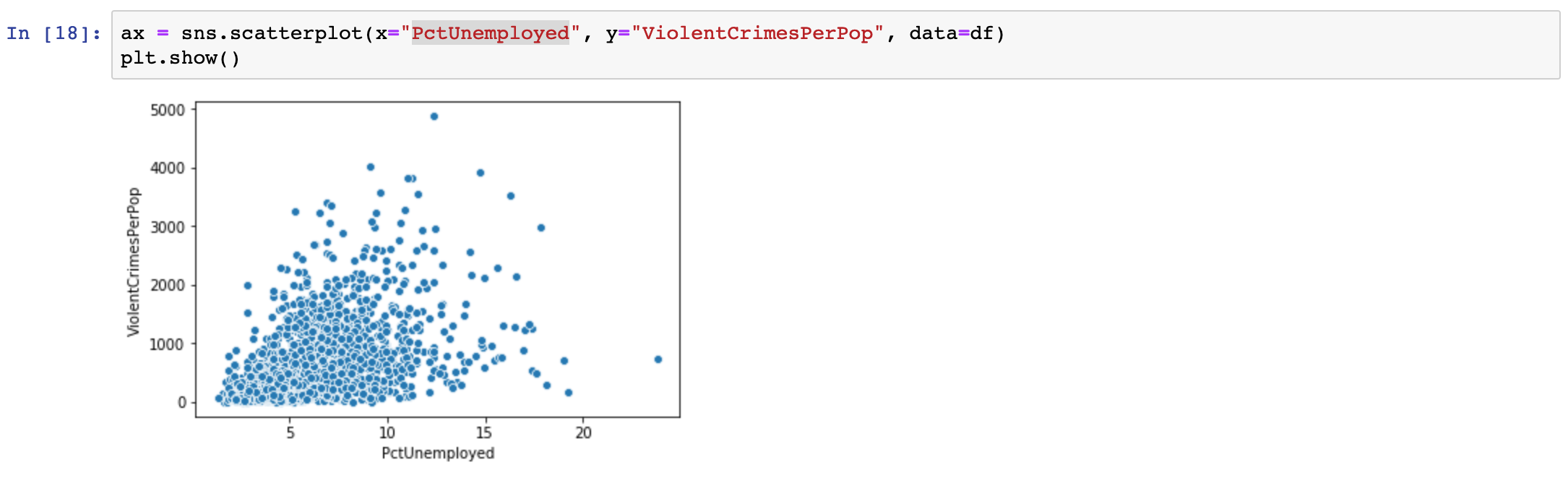


Figure [1]

* In the figure we plot the percentage of unemployment people and total number of violent crimes per 100K population.
* Here we can see the it is directly proportional to each other.

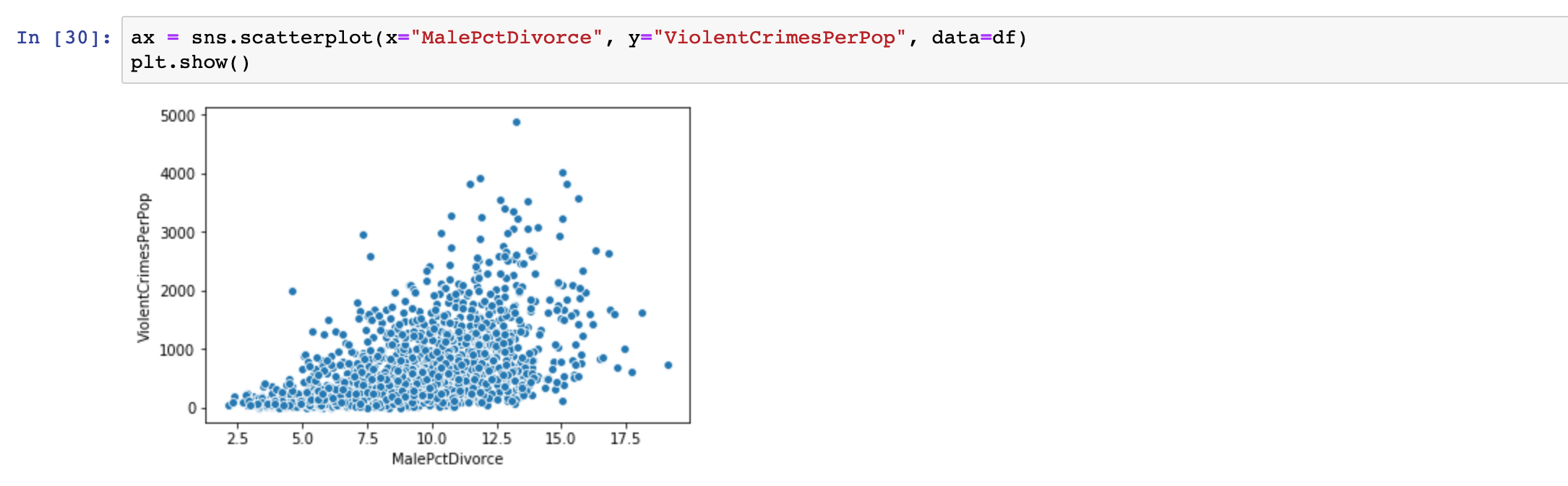


Figure [2]

* In the figure we plot the percentage of divorce male people and total number of violent crimes per 100K population.
* Here we can see the it is directly proportional to each other.

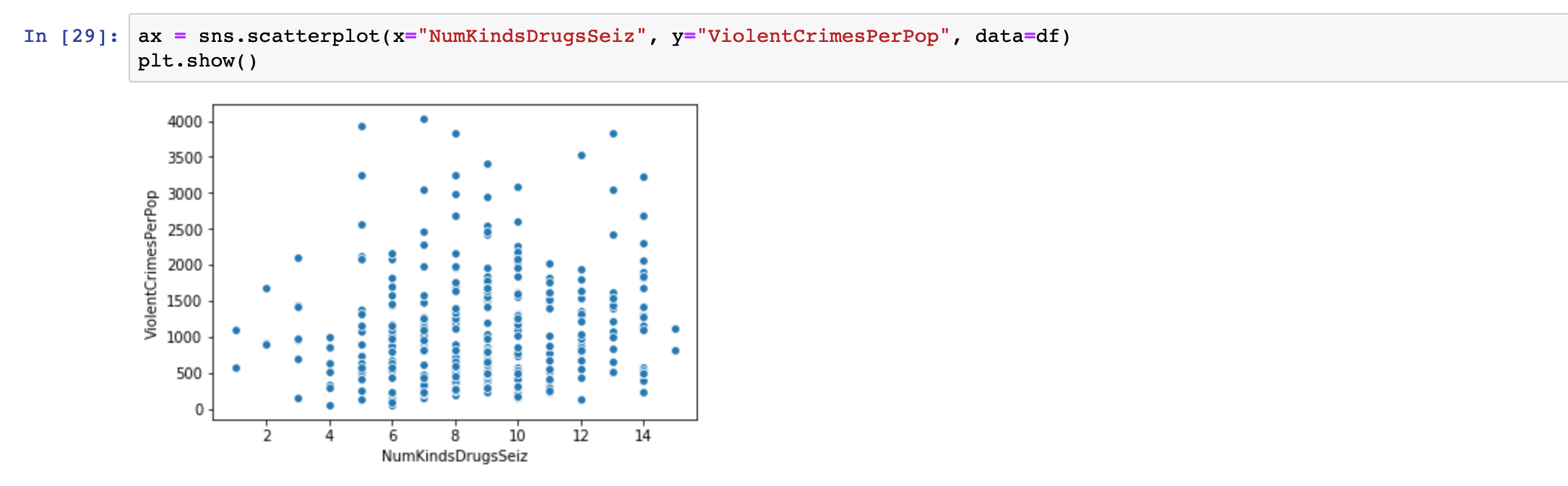


Figure [3]

* In the figure we plot the number of different drugs seized vs total number of violent crimes per 100K population.
* Here we can see the more is number of drugs siezed relatively crime rate is high there.

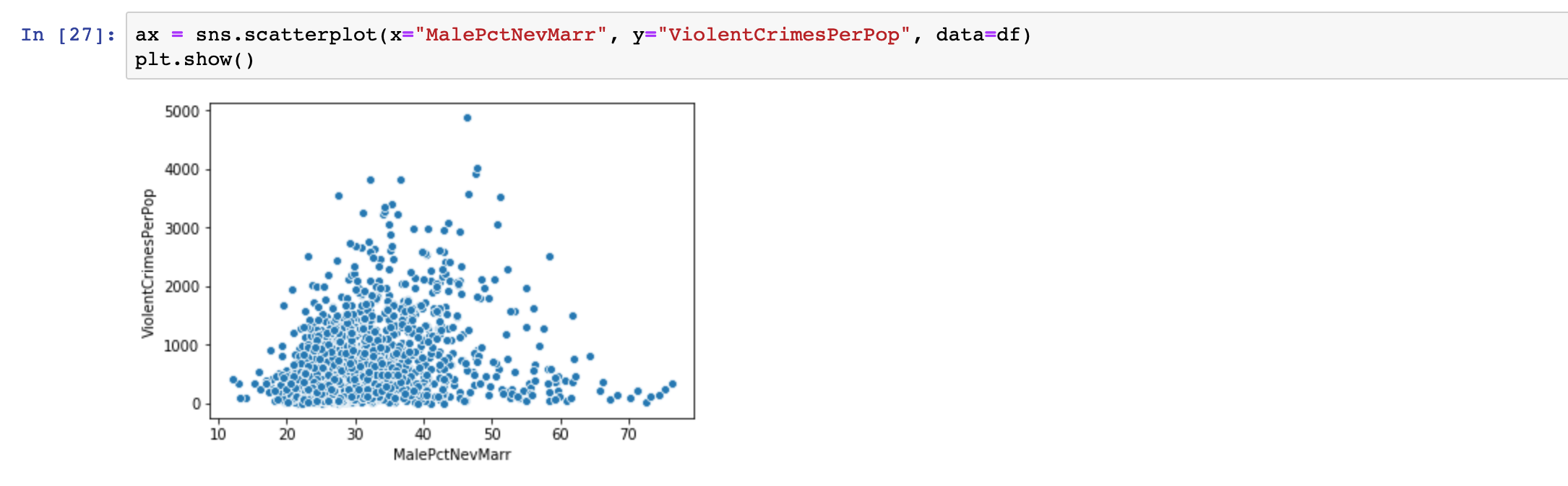


Figure [4]

* In the figure we plot the percentage never married male vs total number of violent crimes per 100K population.
* Here it is directly proportional to each other.

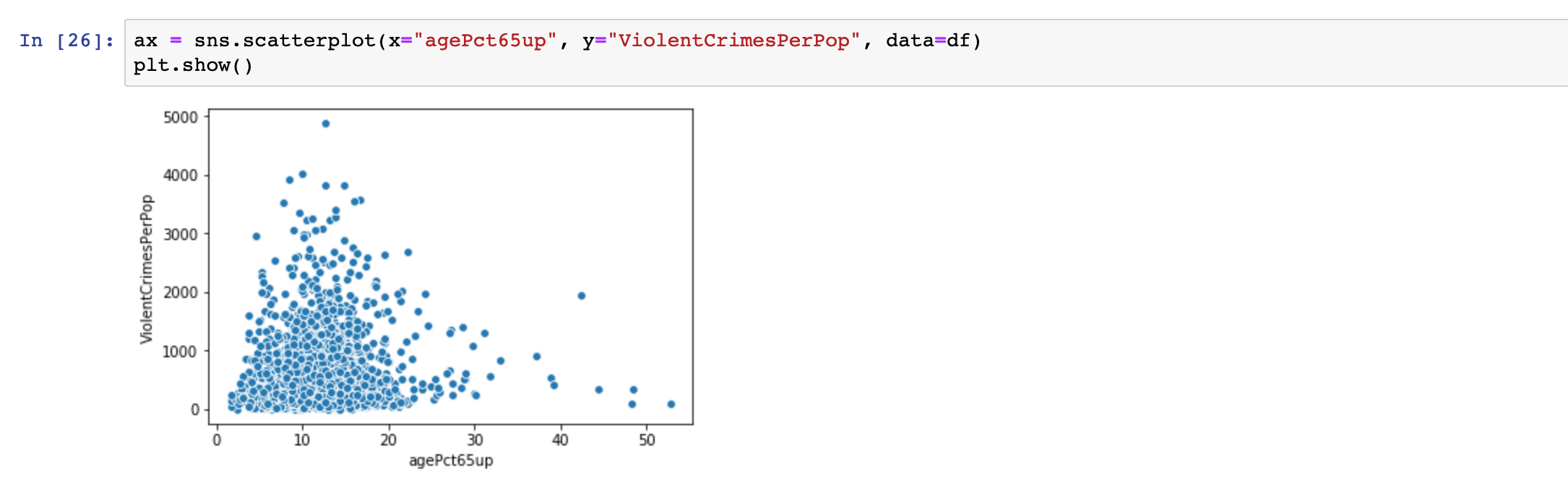


Figure [5]

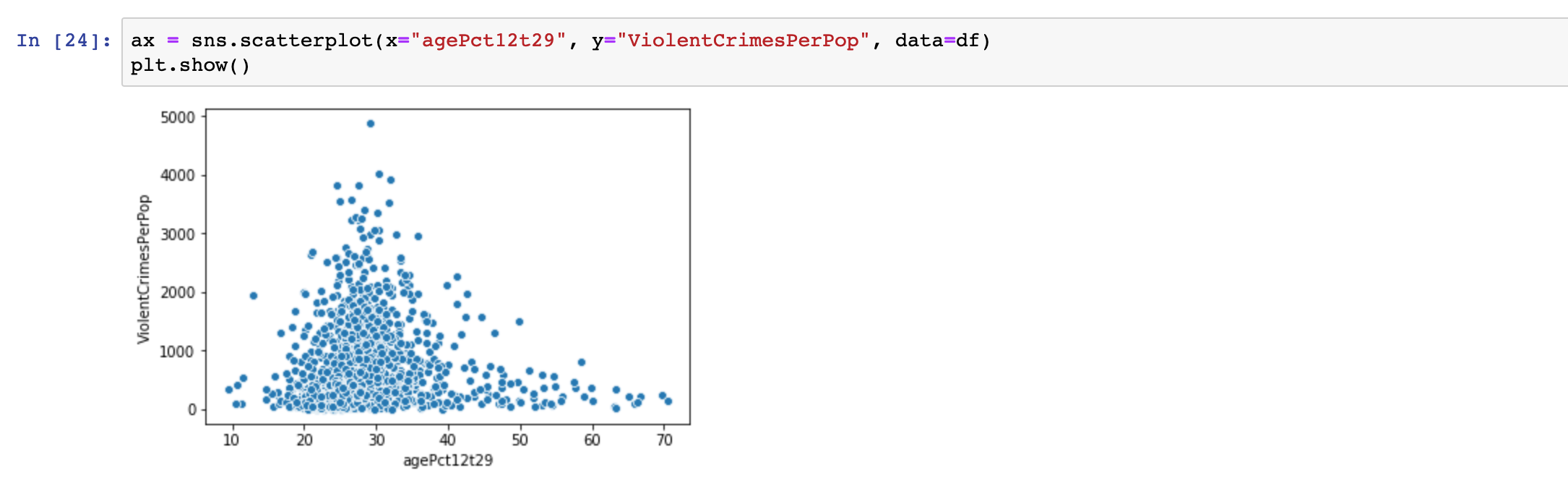
* In the figure we plot the percentage of 65 above age people vs total number of violent crimes per 100K population.
* Here we can see there is no much relation however 10 to 20 percentage this category communities have more crime rate.   
   

Figure [6]

* In the figure we plot the percentage of 12 to 29 age people vs total number of violent crimes per 100K population.
* Here we can see there is no much relation however 20 to 40 percentage this category communities have more crime rate.

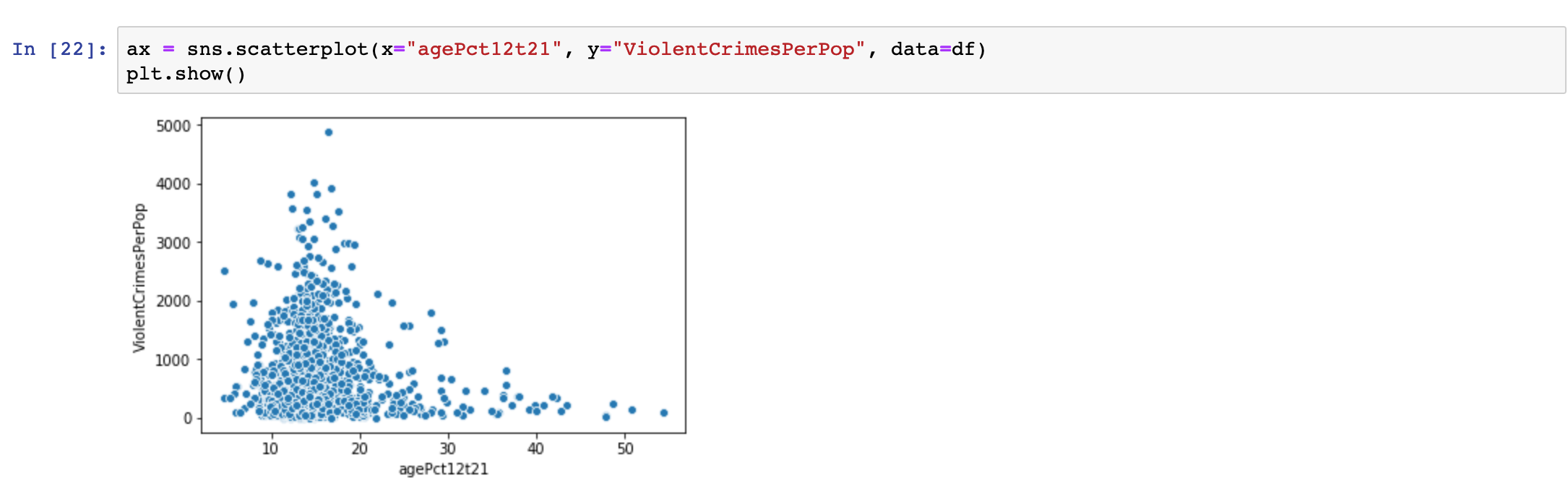


Figure [7]

* In the figure we plot the percentage of 12 to 21 age people vs total number of violent crimes per 100K population.
* Here we can see there is no much relation however 10 to 20 percentage this category communities have more crime rate.

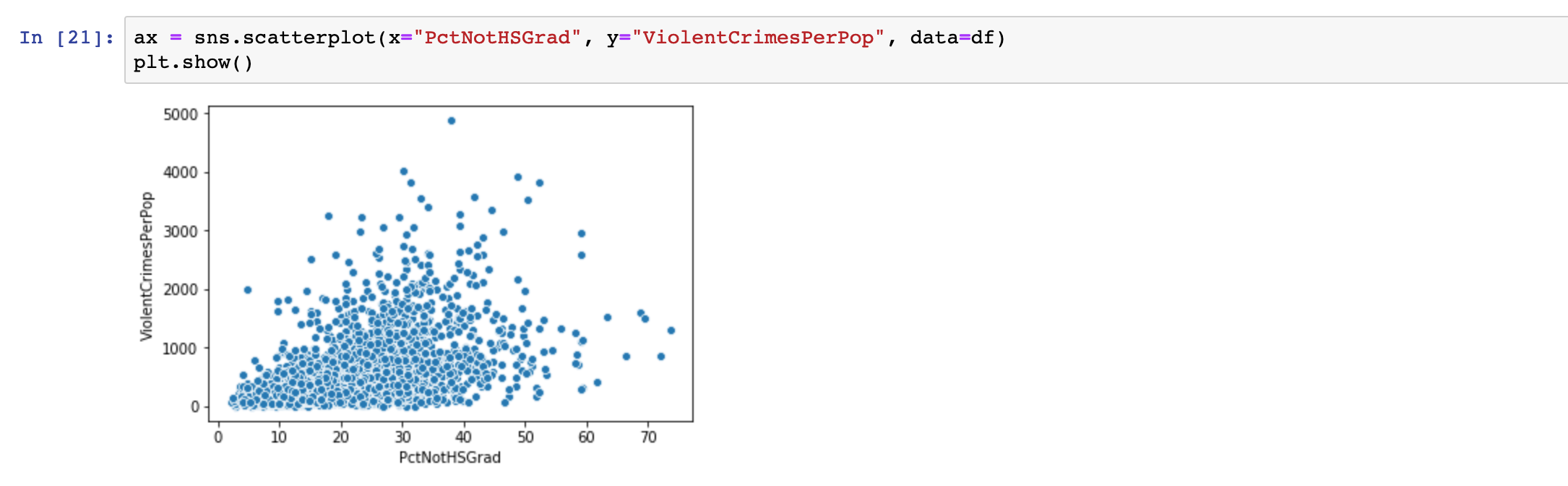


Figure [8]

* In the figure we plot the percentage after 25 years old people who are never high school graduate vs total number of violent crimes per 100K population.
* Here it is directly proportional to each other.

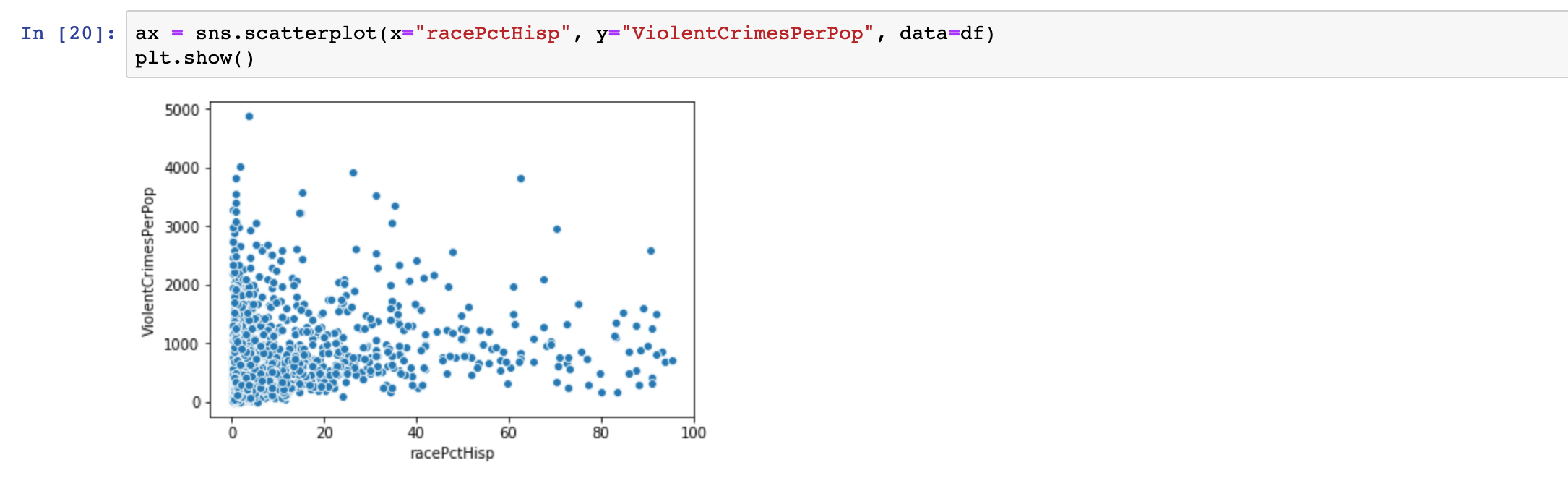


Figure [9]

* In the figure we plot the percentage of population that is of hispanic heritage vs total number of violent crimes per 100K population.
* Here it is randomly distributed so no relation here.

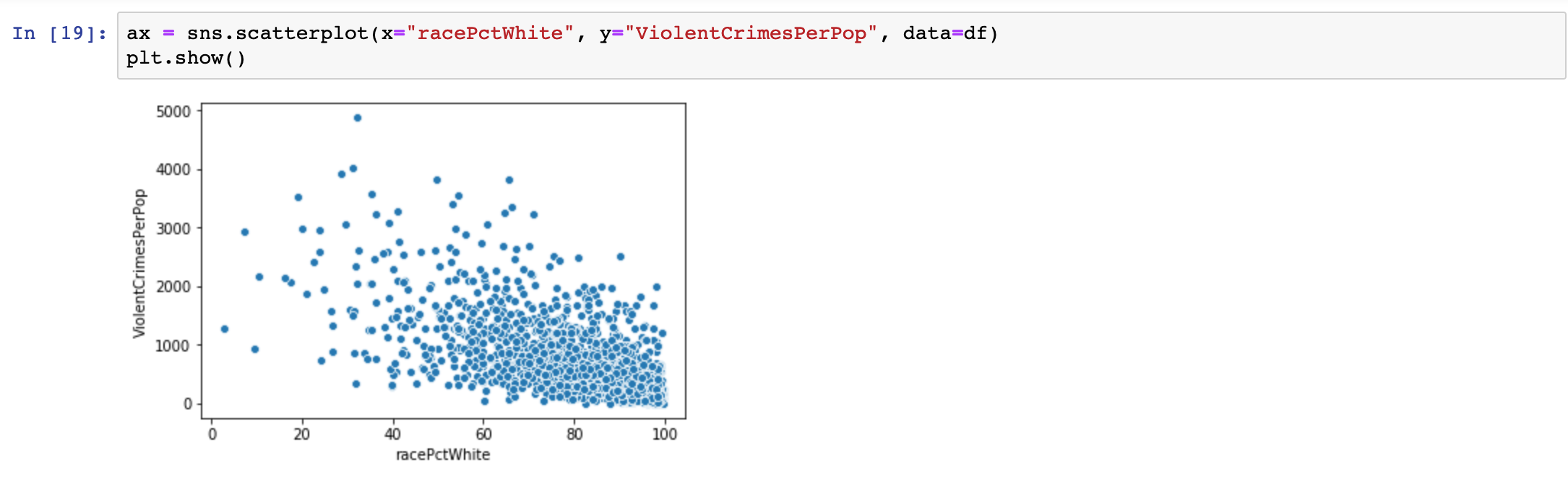


Figure 10

* In the figure we plot the percentage of caucasian people vs total number of violent crimes per 100K population.
* Here the relation is clearly inversely proportional and totally related..

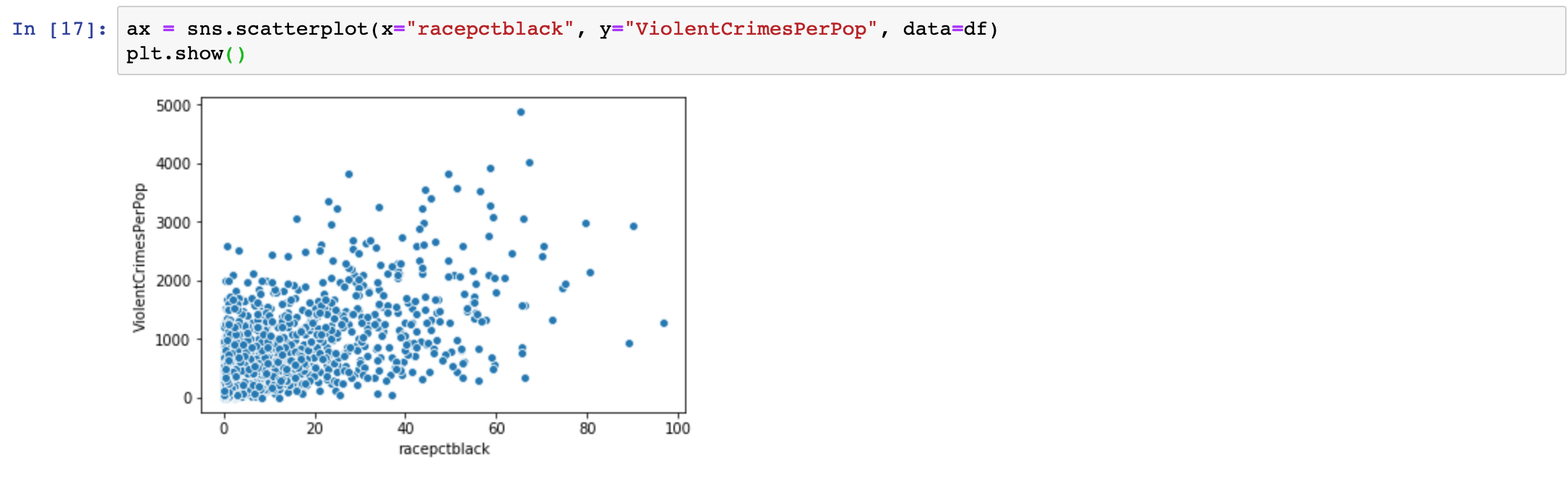


Figure [11]

* In the figure we plot the percentage of african american people vs total number of violent crimes per 100K population.
* Here the relation is clearly directly proportional and totally related..

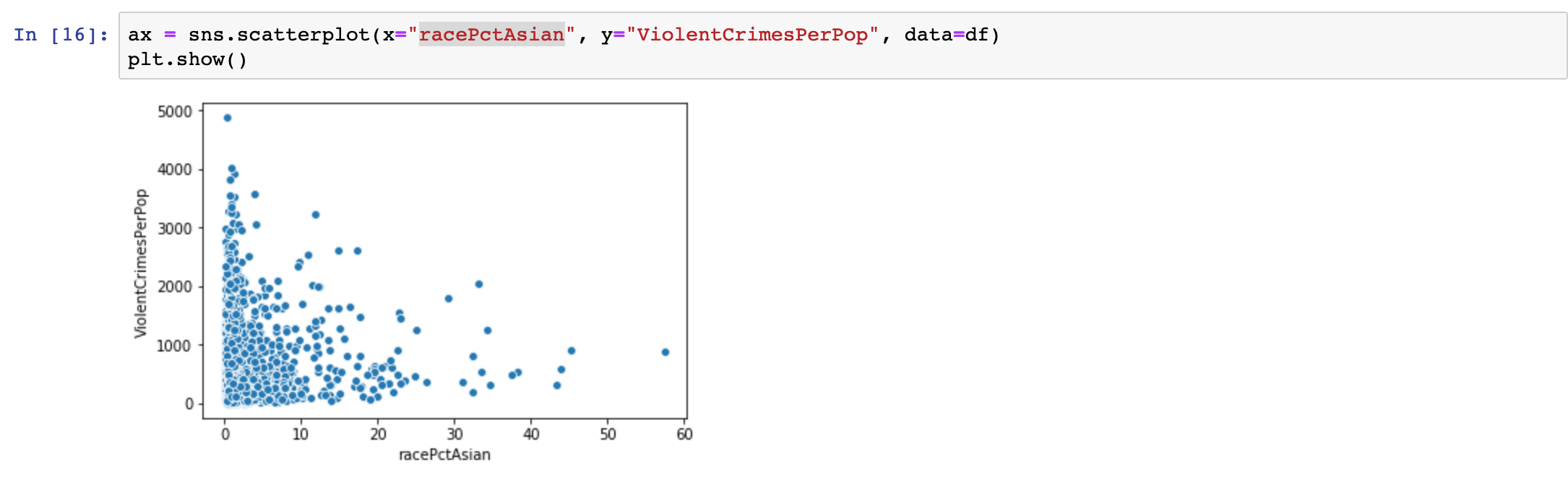


Figure [12]

* In the figure we plot the percentage of asian american people vs total number of violent crimes per 100K population.
* Here there is no relation and randomly distributed.

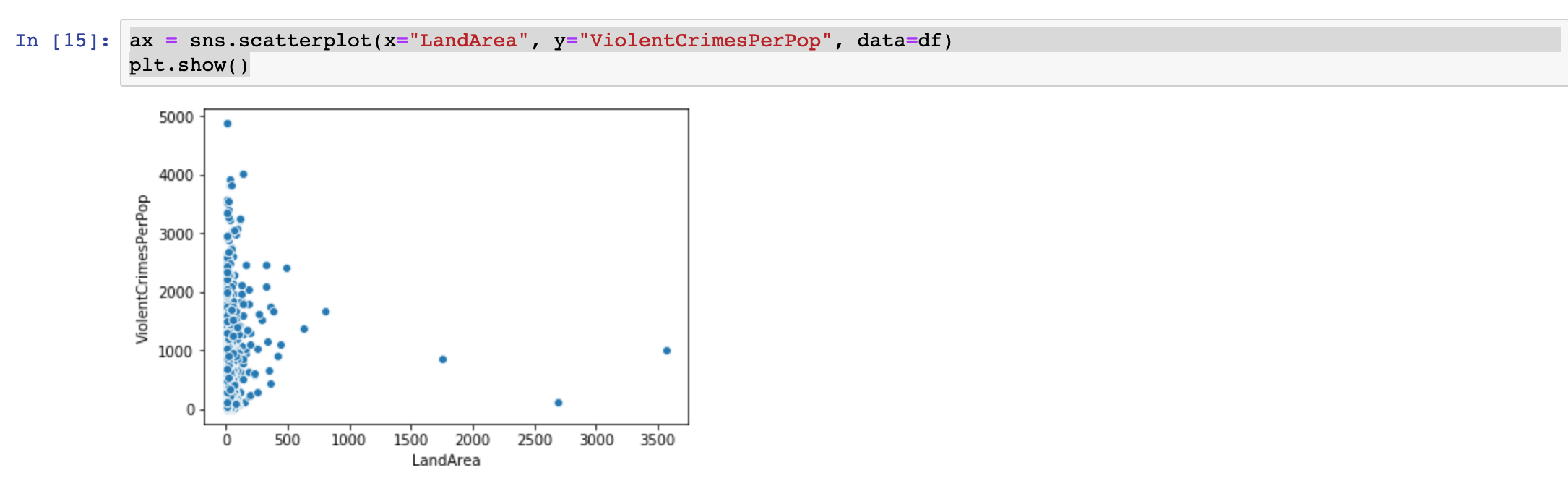


Figure [13]

* In the figure we plot the percentage of african american people vs total number of violent crimes per 100K population.
* Here there is no relation at all.

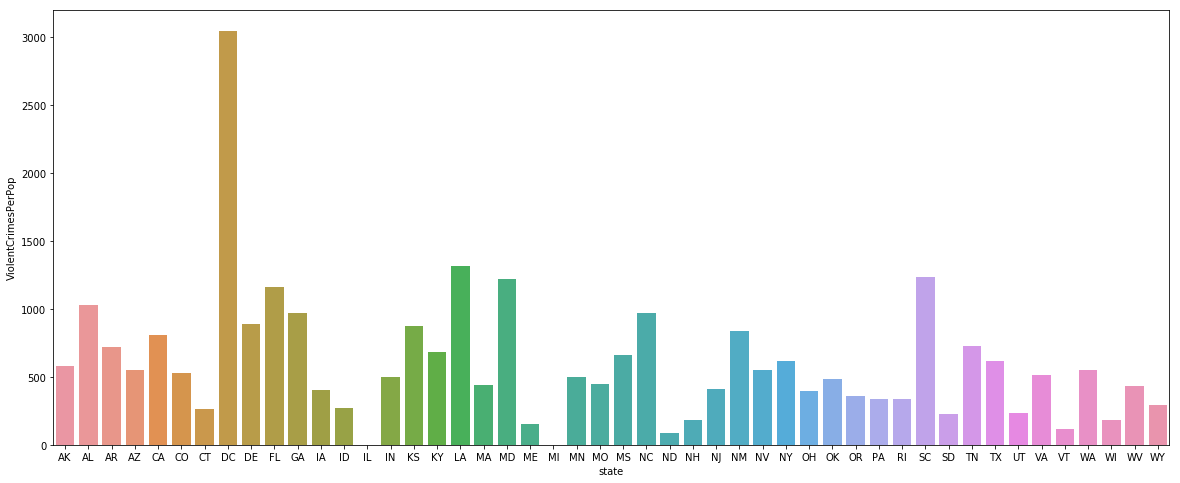


Figure [14]

* In the figure we plot bar graph for area vs total number of violent crimes per 100K population.
* Here DC area communities have the highest crime rate.

**Handling missing values** **& categorical variable** :

Given figure is the missing value percentage in our total dataset.

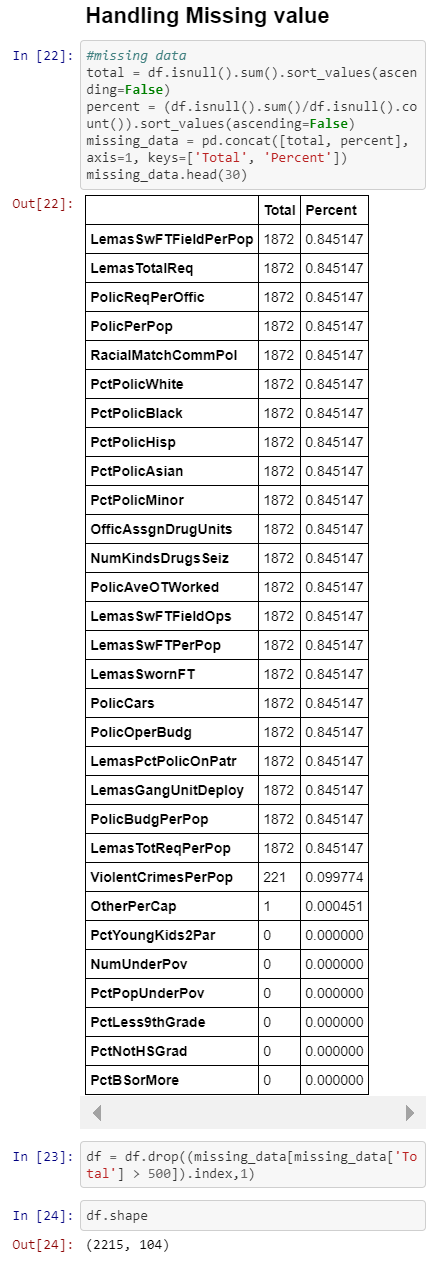


Figure [15]

Above, we find the missing values with total percentage missing . The columns with more than 500 missing values were deleted.

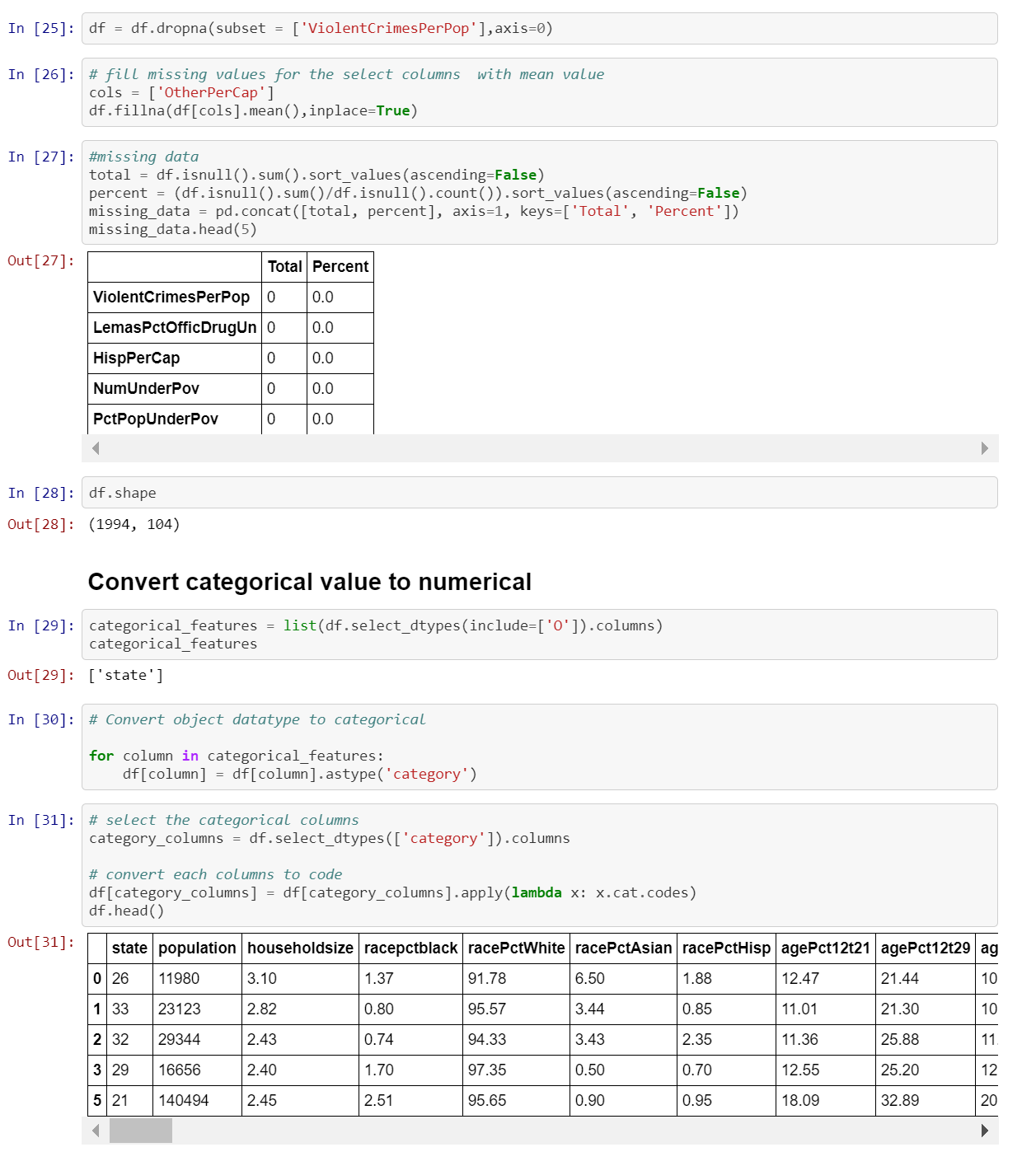


Figure [16]

Next, for the feature, ‘OtherPerCap’, the mean value replaced the only null value.

Categorical values are converted to numerical (States -> assigned numbers)

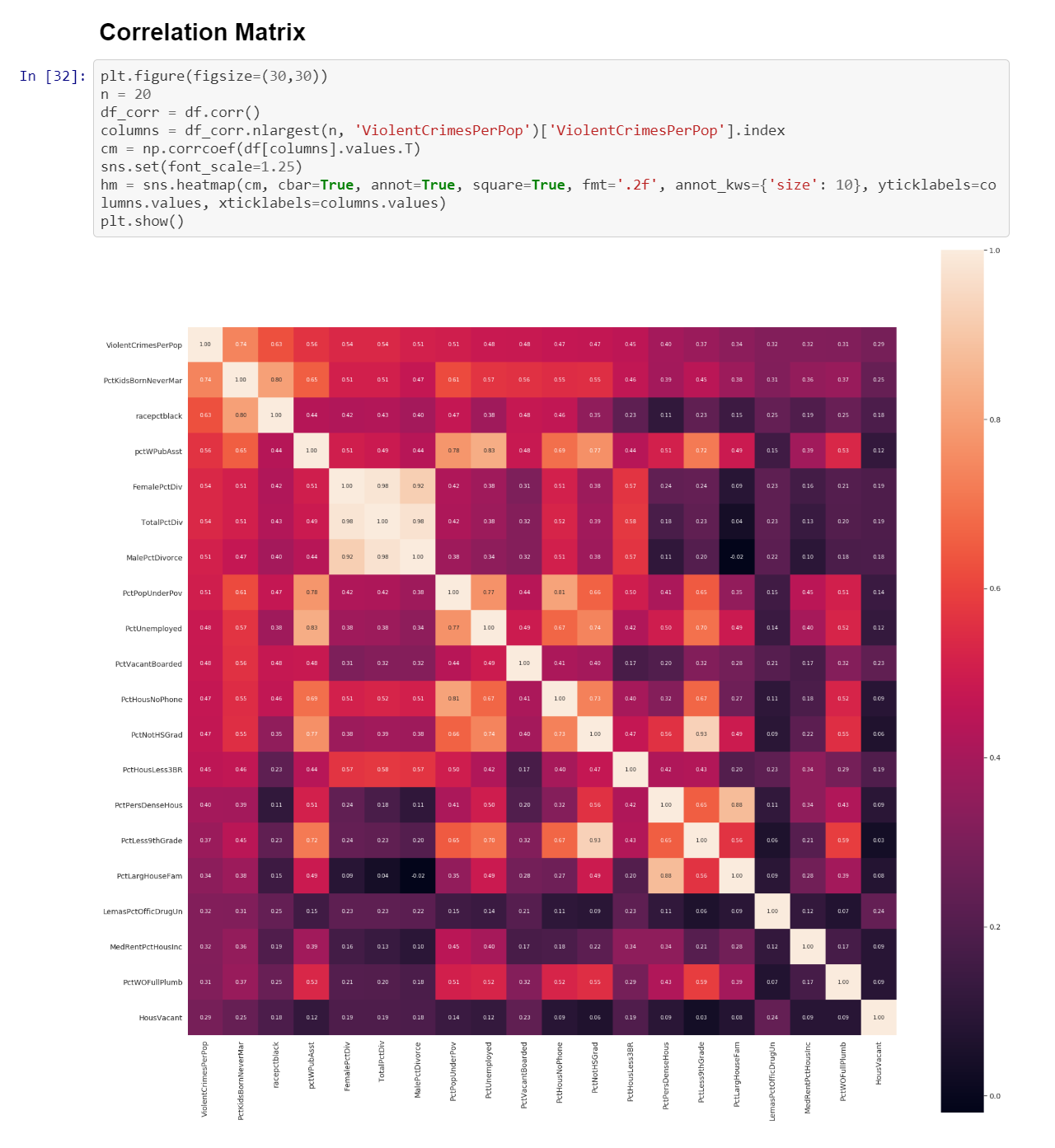
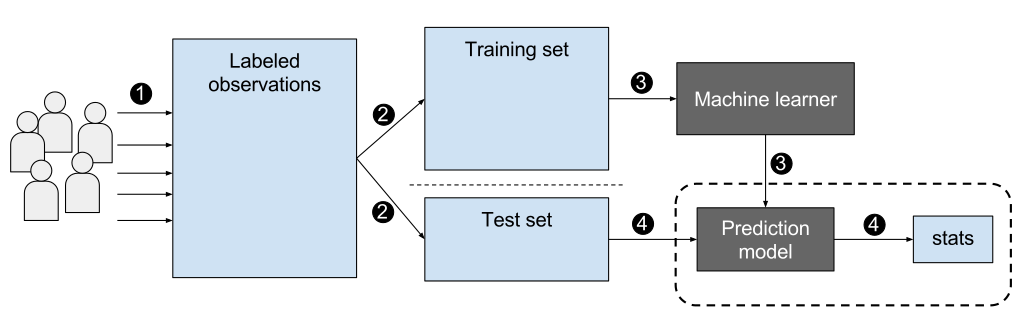


Figure [17]

Correlation matrix created using the pandas library. The correlation matrix describes the dependency of variable with respect to each other.

**Flow diagram:**

****

**Modeling :**



Figure [18]

Above, We took 103 features after cleaning the dataset and transformation.

we import the train test split library, initialize normalization function, and drop ‘state’ from data set because it is categorical for normalization. Normalization function is then ran on the numerical features. Data set is then split into two groups, the training set and testing set. The model is then ran using the OLS regression function using the statsmodel library.

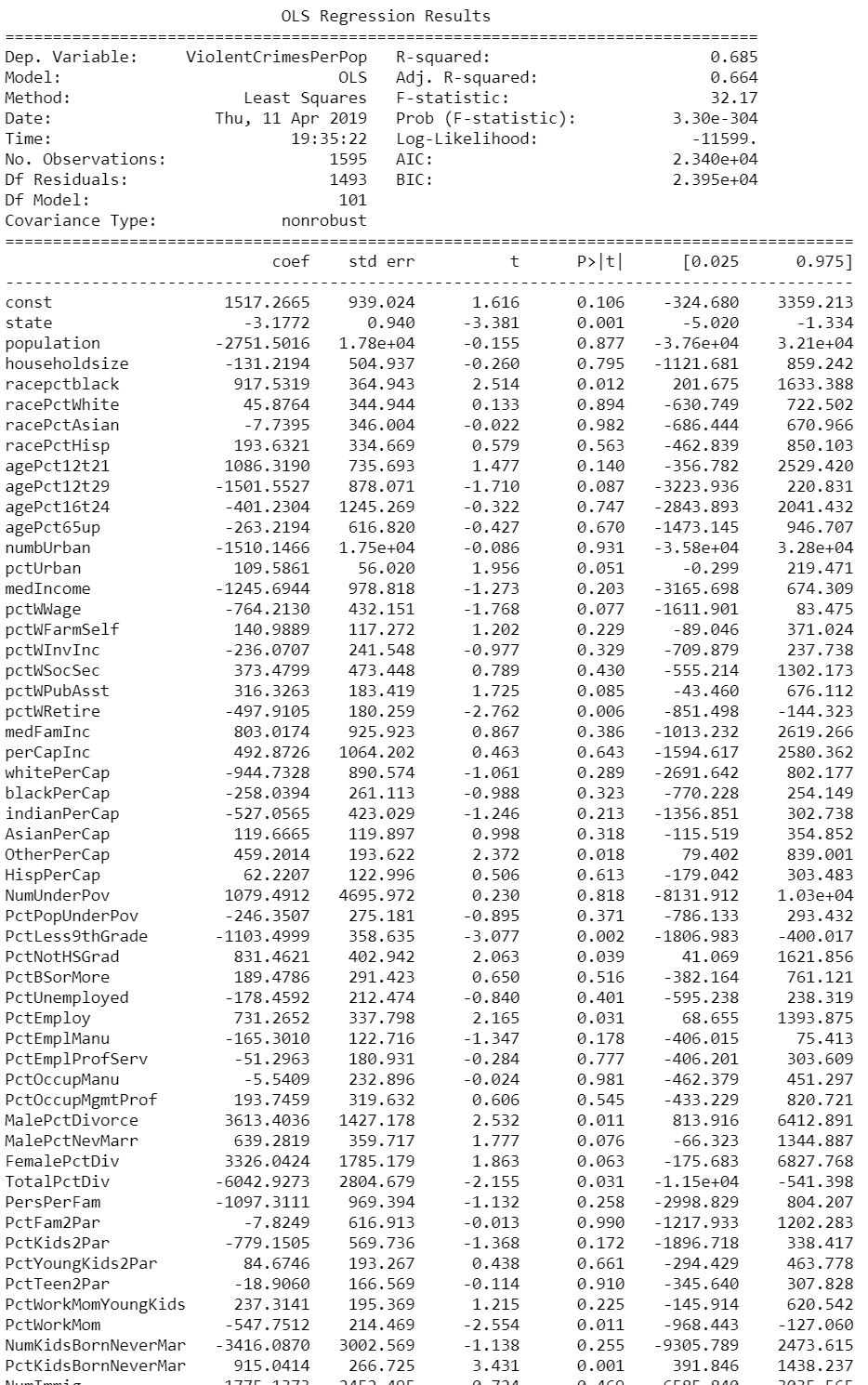


Figure [18]

These are the results of the model being ran. And p-value and z-value along with weight values.

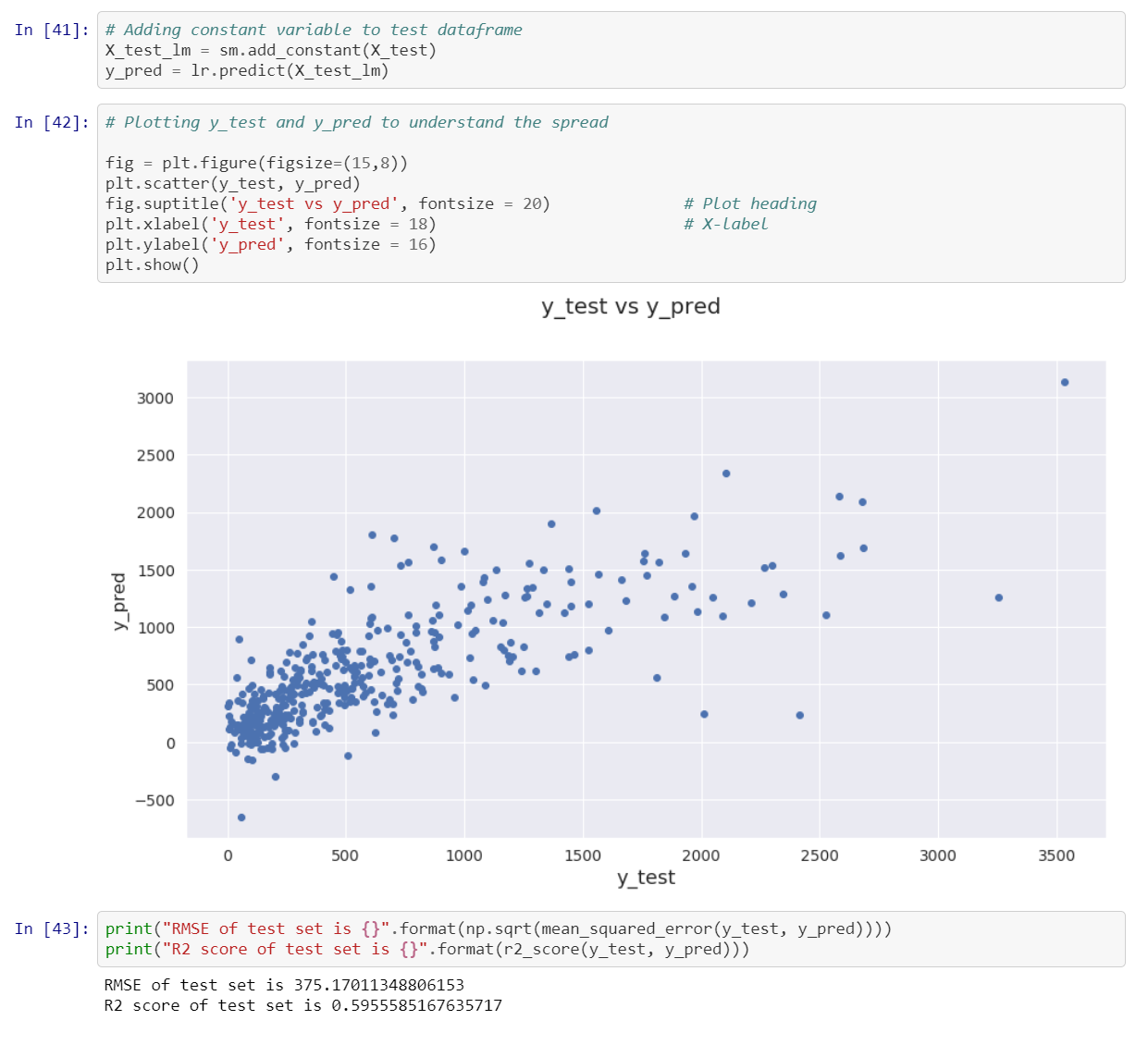


Figure [19]

First, we predict the test set using our model using statsmodels library.

The above plot is our predicted value vs. true value.

The RMSE results shows the error between true value and predicted value.

The R^2 results shows that are its greater than 0.5 so it is pretty good model for starter.

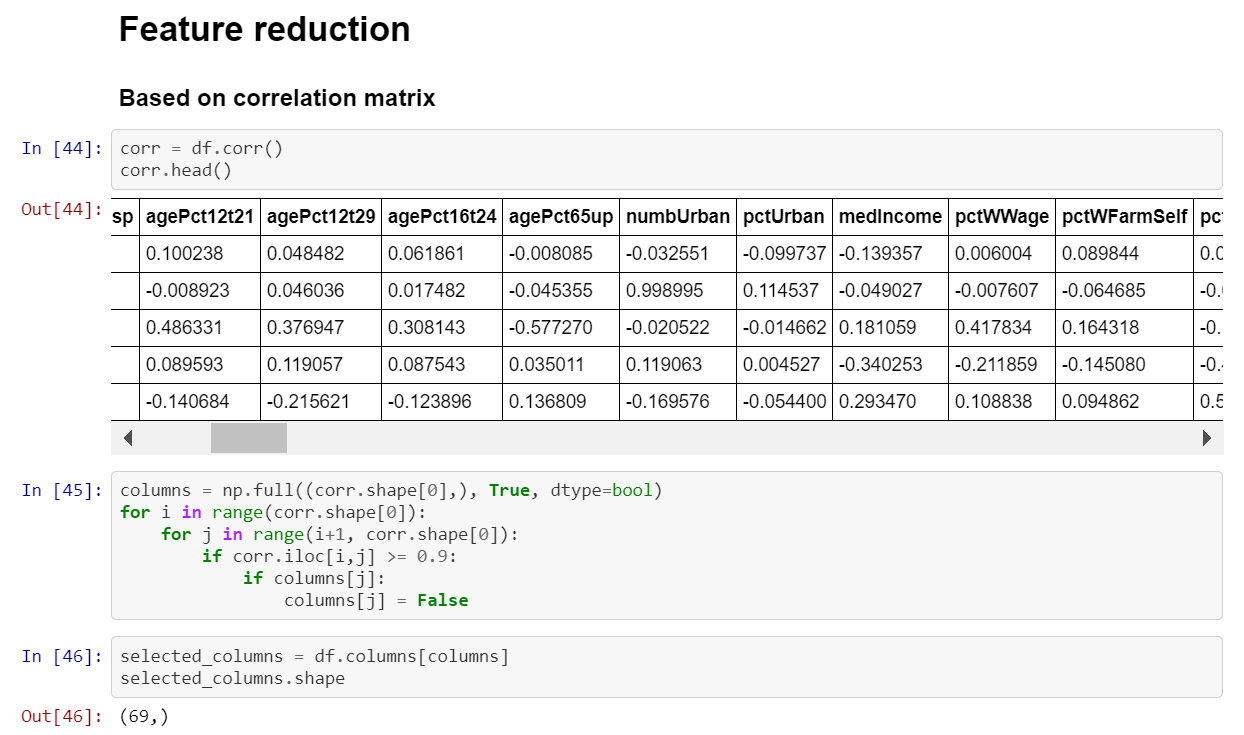


Figure [20]

In line 44 above, the good features are being selected based off the correlation matrix. The top features are kept while the bad features are deleted. (Only features with a correlation higher than .9 are saved)

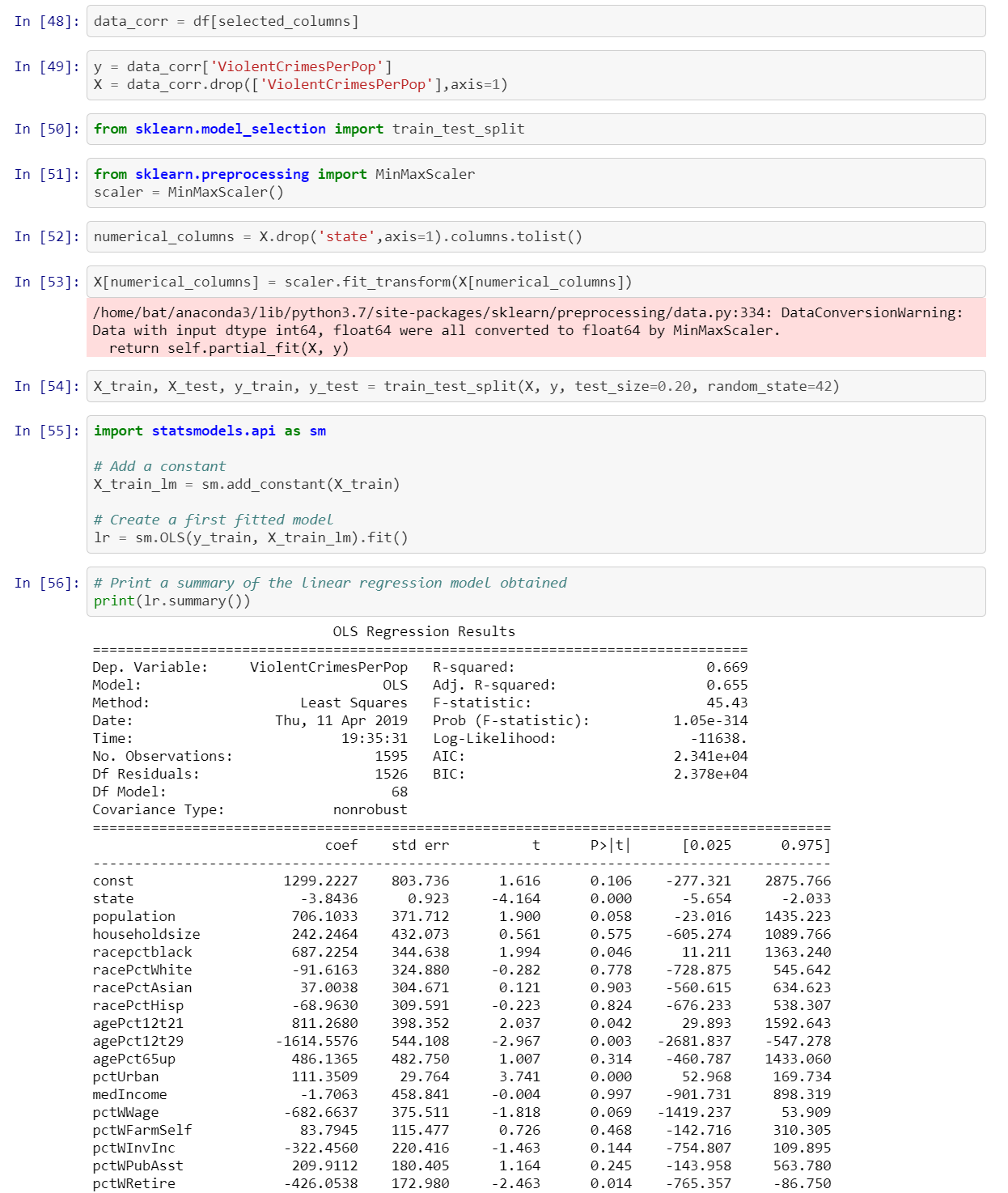


Figure [21]

Above, we import the train test split library, initialize normalization function, Normalization function is then ran on the numerical features. Data set is then split into two groups, the training set and testing set. The model is then ran using the OLS regression function using the statsmodel library.

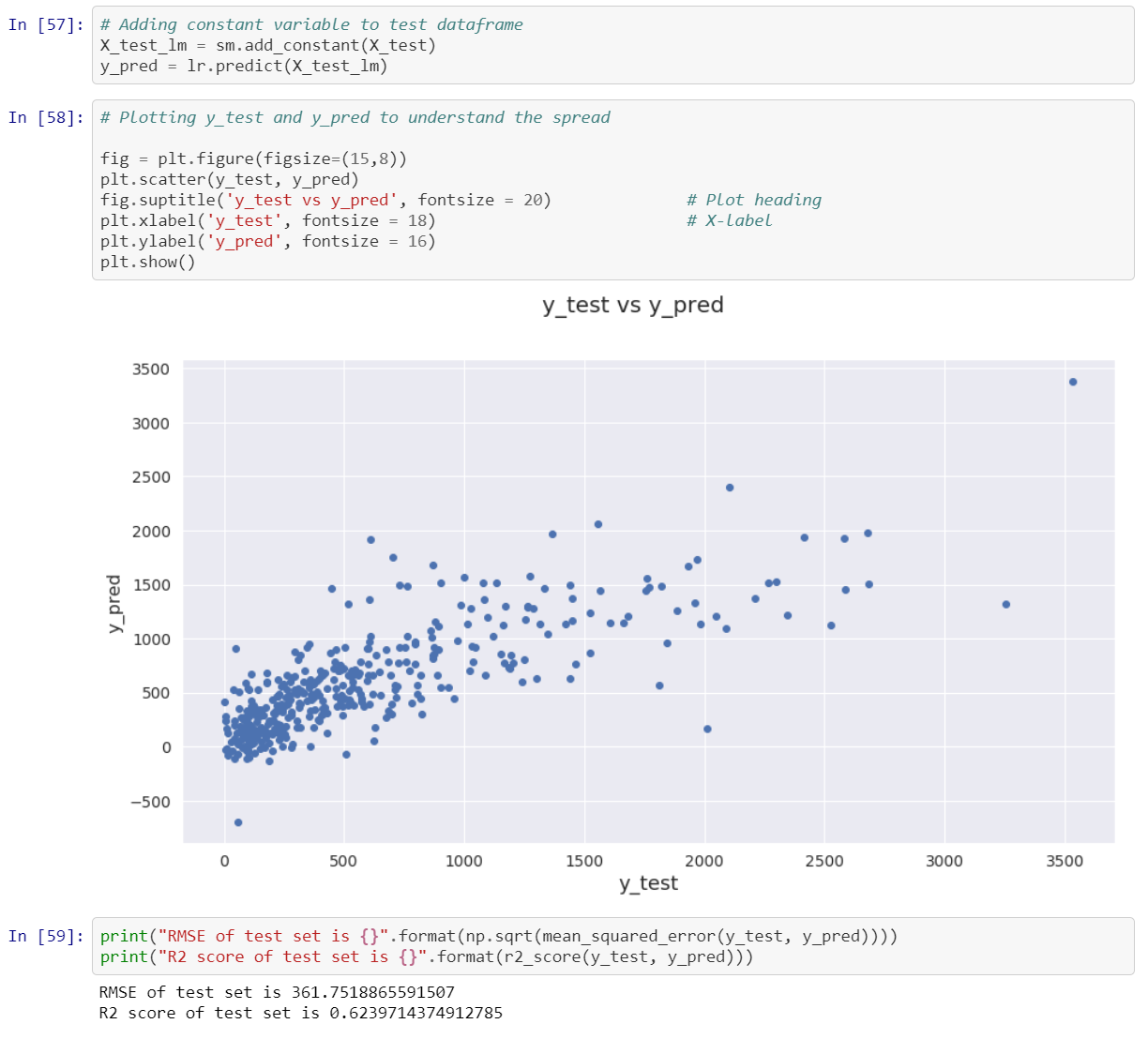


Figure [22]

Again like before, we predict the test set using our model using statsmodels library.

The above plot is our predicted value vs. true value.

The RMSE results shows the error between true value and predicted value.

The R^2 results shows that the score increase since last time by .03 so the model keeps getting better than last one.

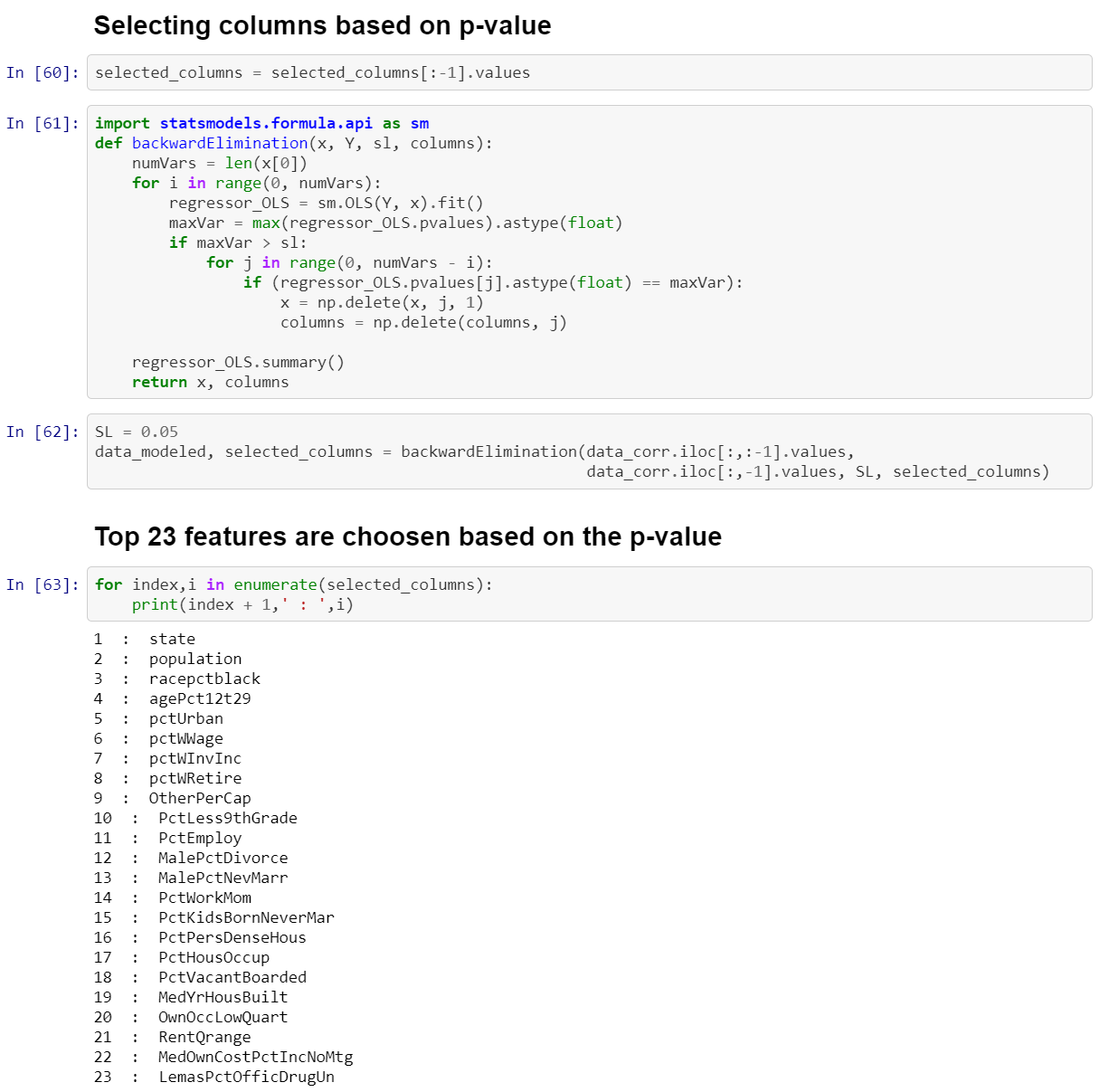


Figure [23]

Based on the P values, features were reduced based on p value.

Now above, we are parsing through the data and eliminating features based off p values. This only leaves our top 23 features.

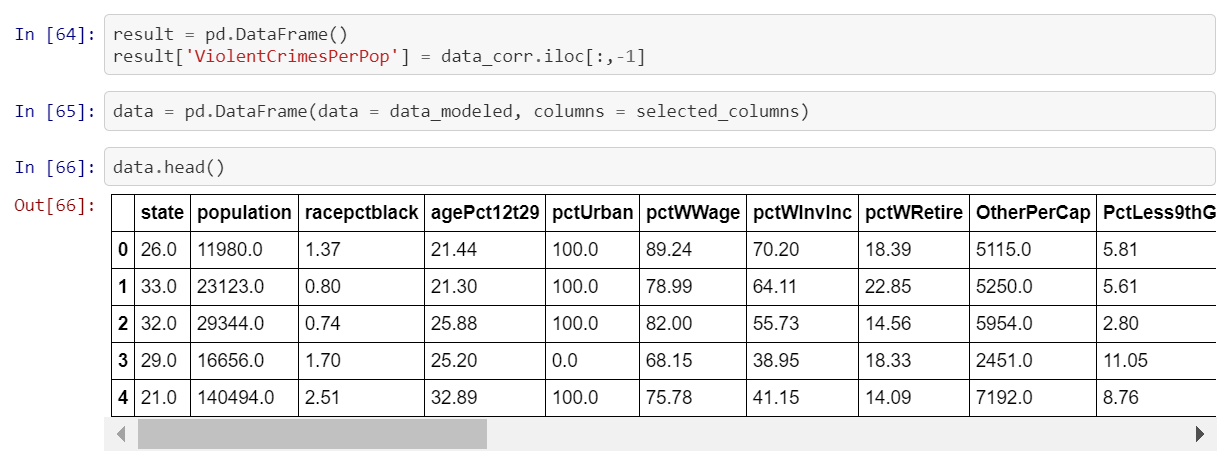


Figure [24]

Here in line 64 we are extracting the predicted variable and in line 65 we are extracting all the rows based on the selected features.

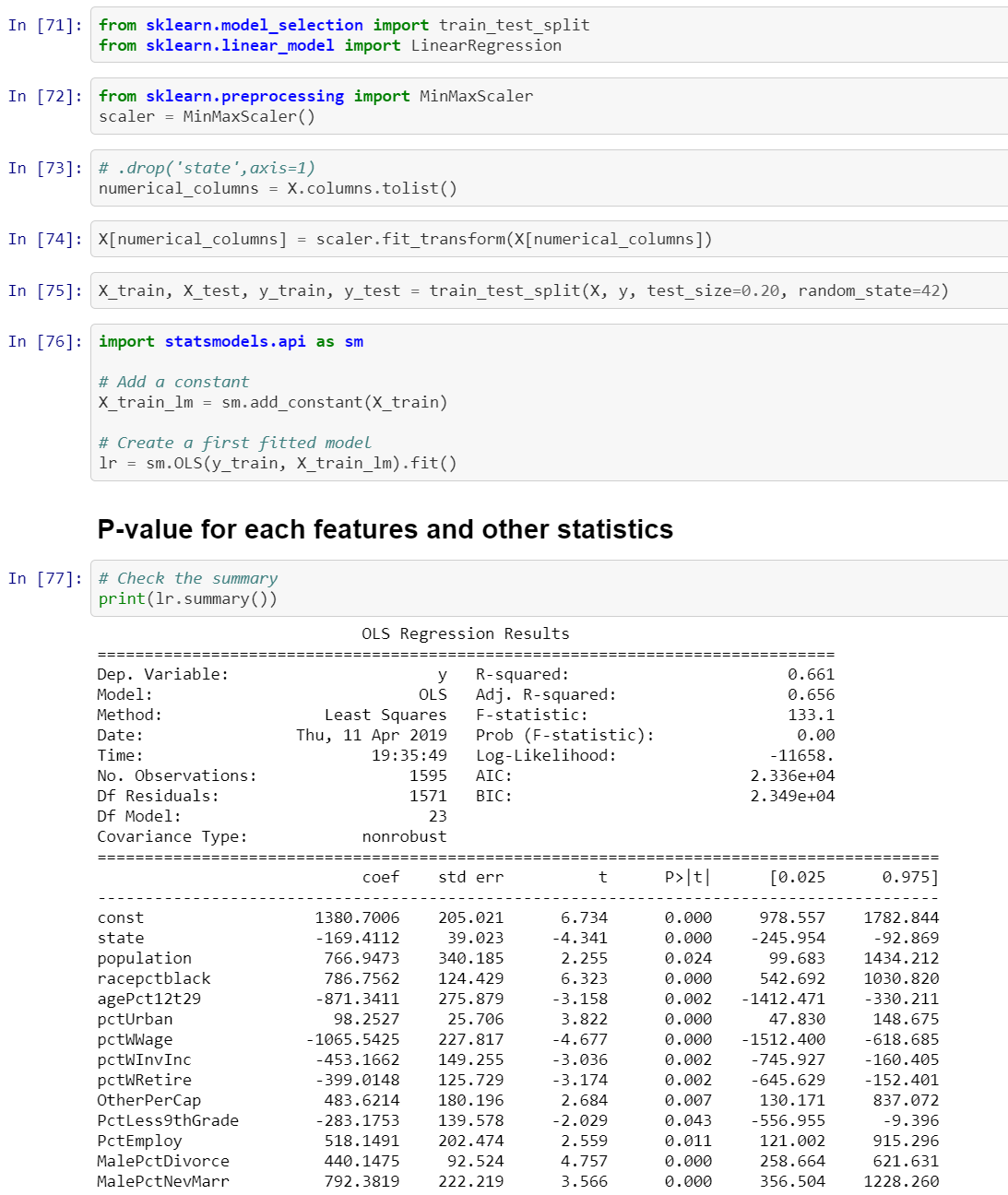


Figure [25]

Above, we import the train test split library, initialize normalization function, Normalization function is then ran on the features. Data set is then split into two groups, the training set and testing set. The model is then ran using the OLS regression function using the statsmodel library.

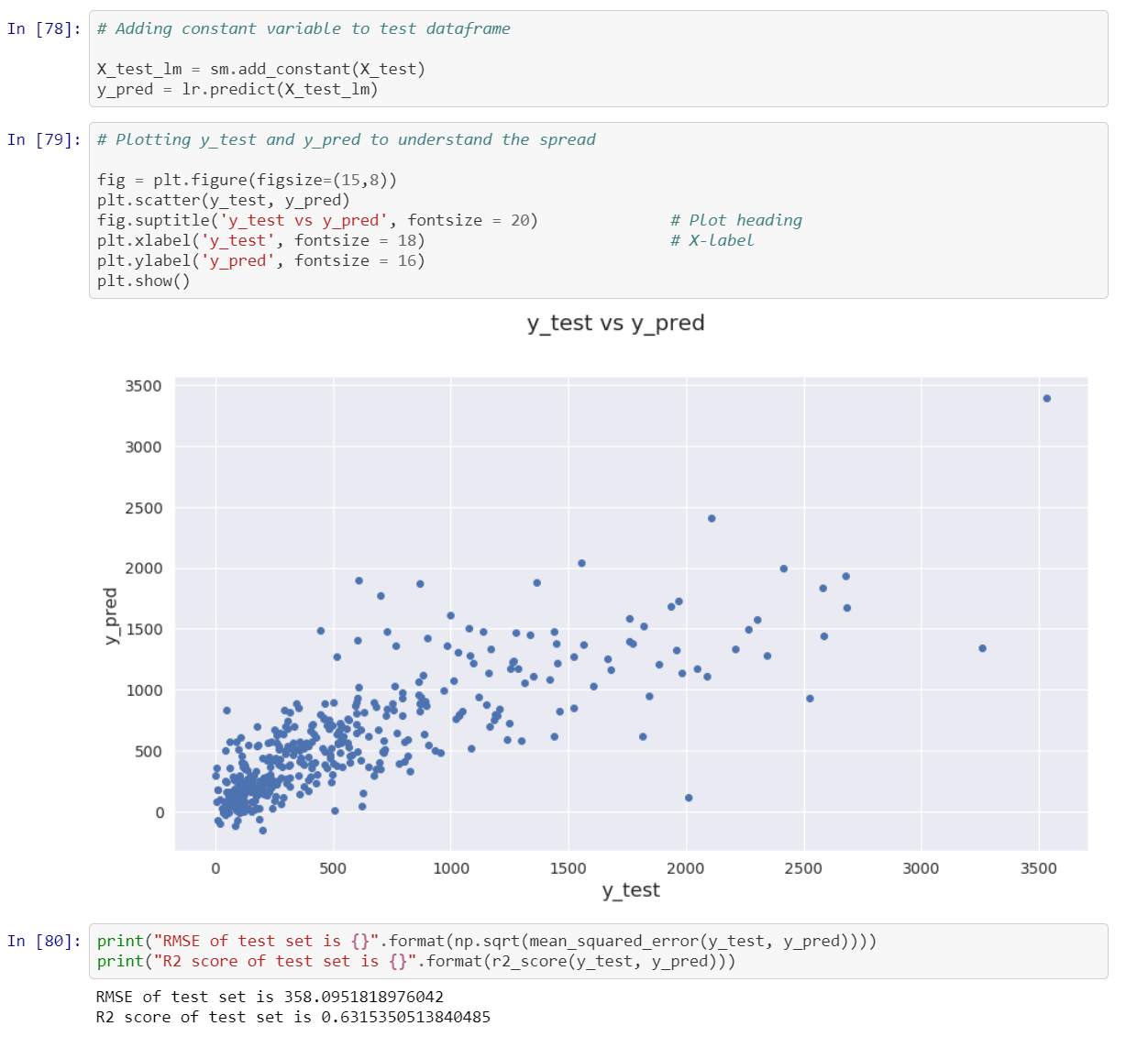


Figure [26]

Lastly, again like before, we predict the test set using our model using statsmodels library.

The above plot is our predicted value vs. true value.

The RMSE results shows the error between true value and predicted value.

The R^2 results shows that the score increase since last time by .01 so the model keeps getting better than last one.

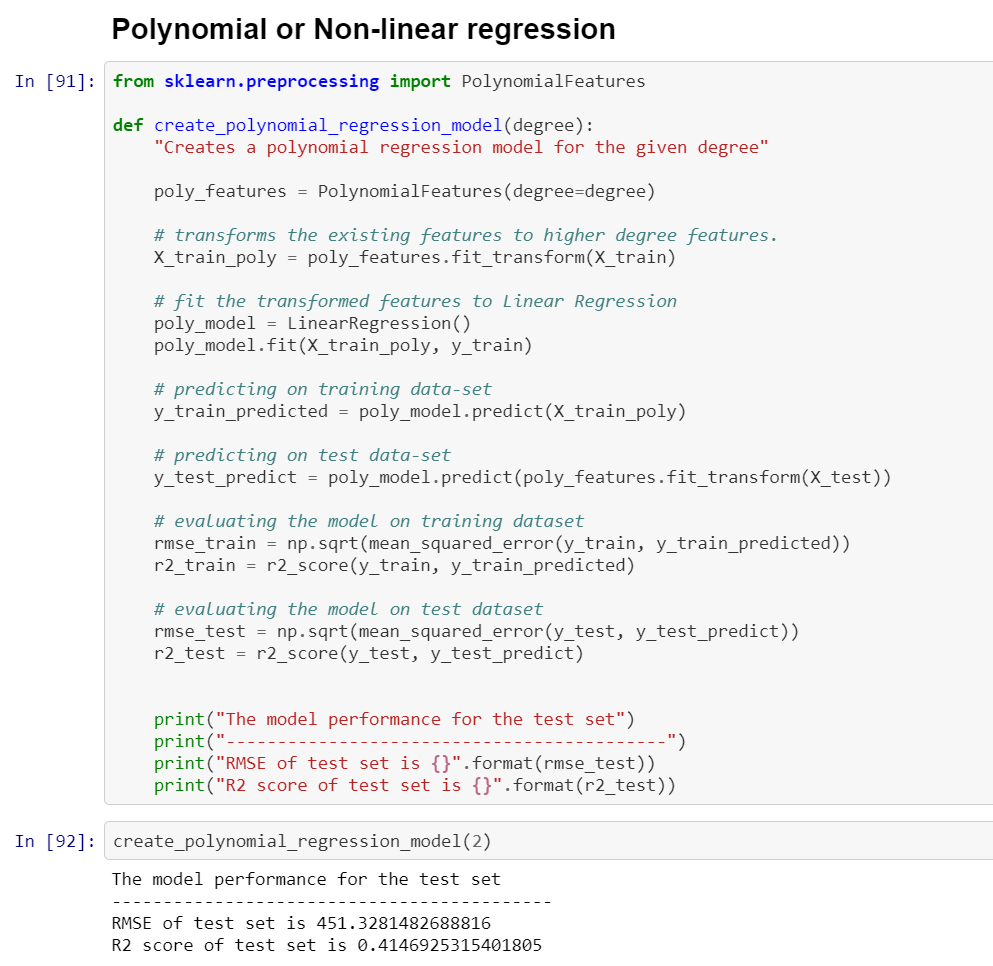
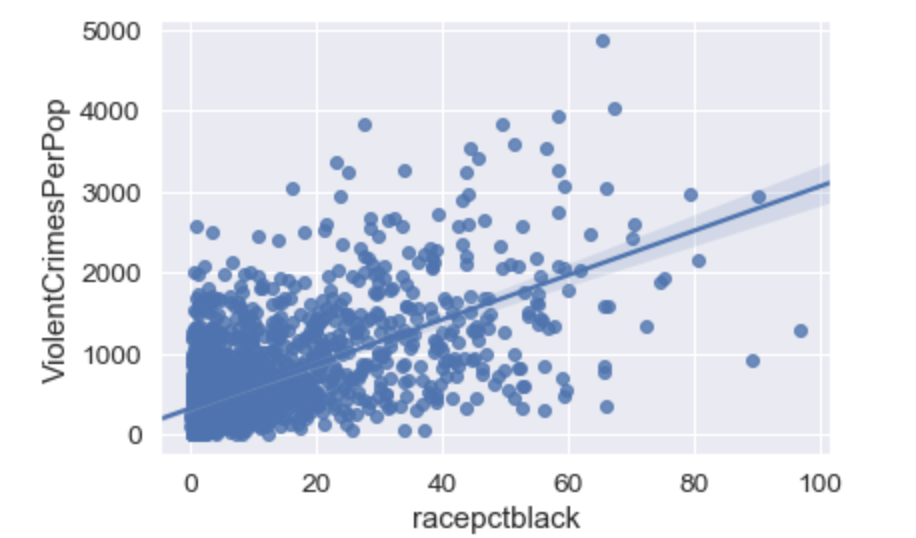


Figure [27]

We started with defining the function create\_polynomial\_regression\_model with degree 2.

Since our dataset describes only linear relationships, the results got worse when taking into account nonlinear data.

*Some Residual plots and Line of Best Fit plots*



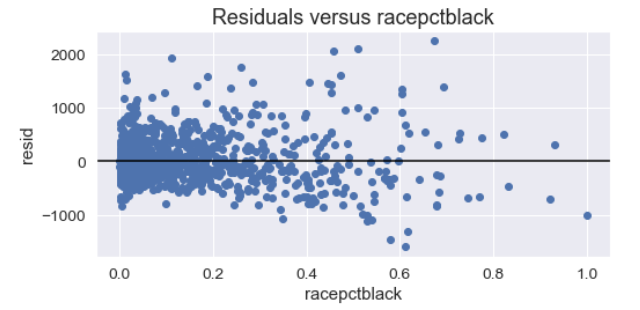
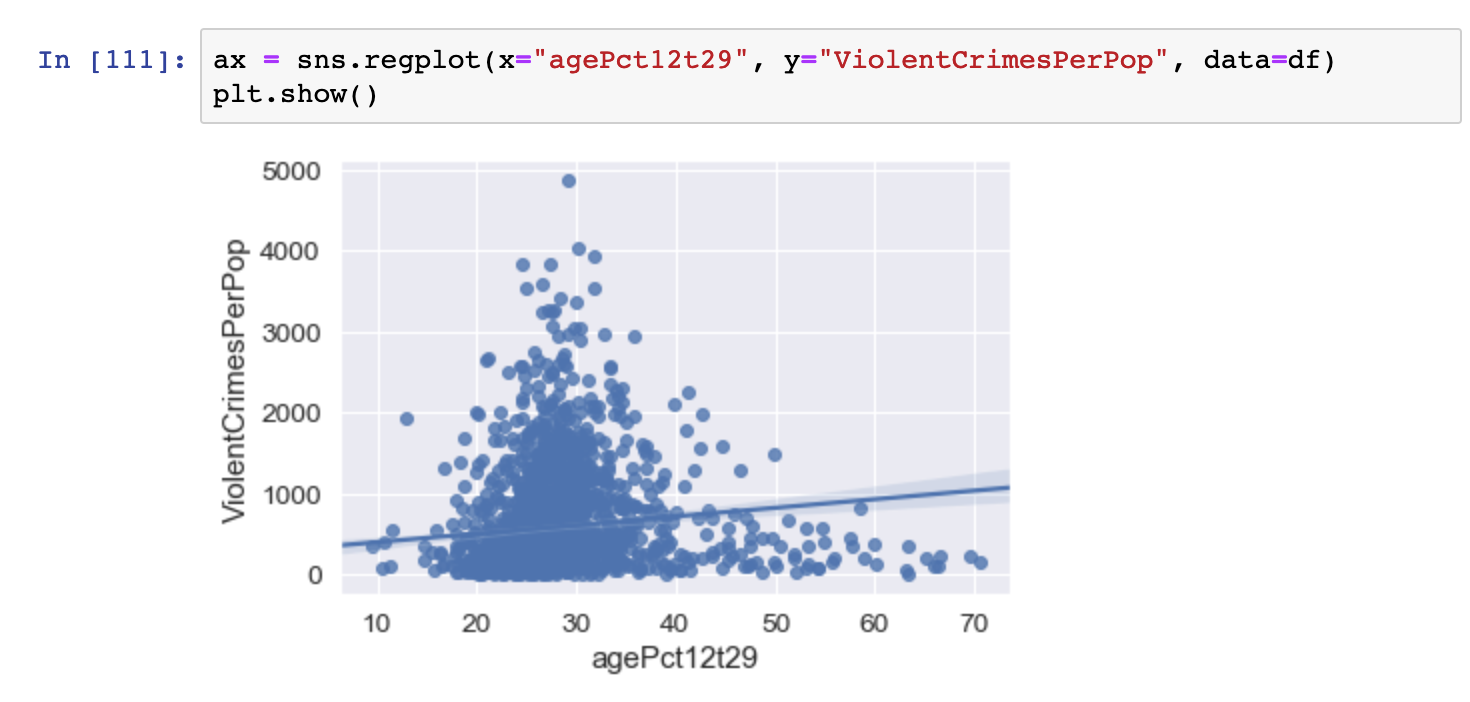


Figure [28]

The independent variable black race percentage shows a direct relationship with violent crime as the slope of the plots is positive.

As all points are normally distributed in each side of the residual plot.

The residual plot appears to have a constant variance scattered around zero.



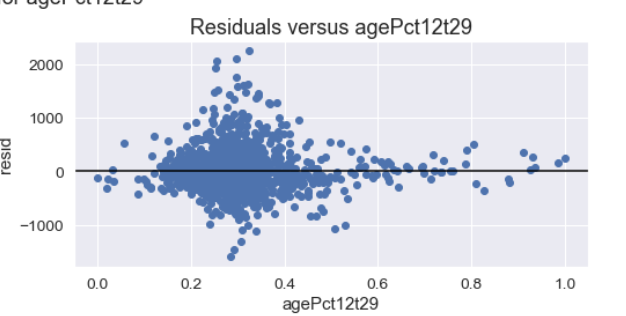
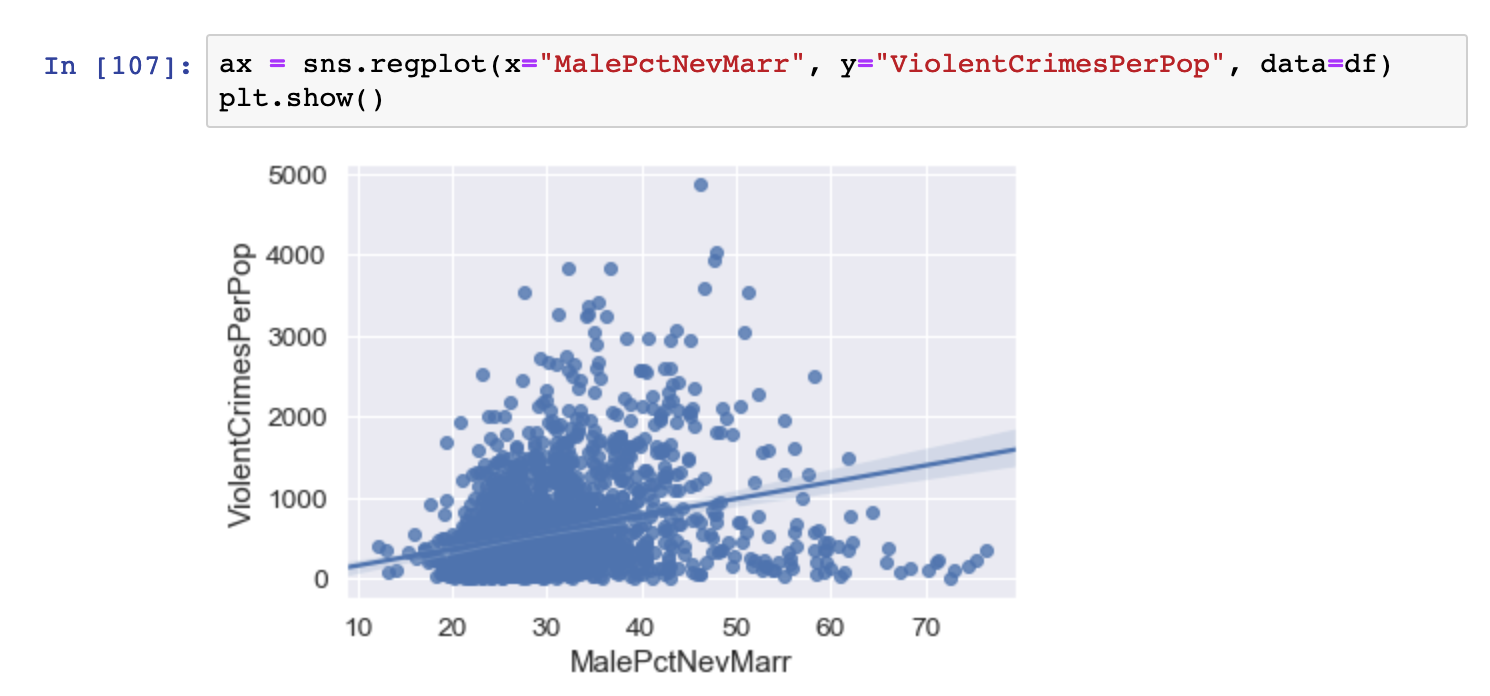


Figure [29]

The independent variable age percentage 12 to 29 shows a nonlinear relationship with violent crime as the slope of the plots is inconsistent.

As all points are not normally distributed in the residual plot.

The residual plot appears to not have a constant variance scattered. But the mean looks like it is zero.



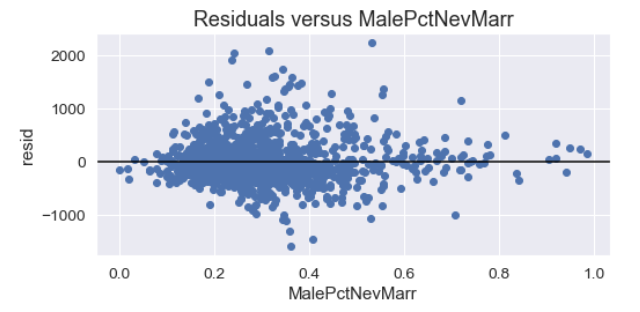
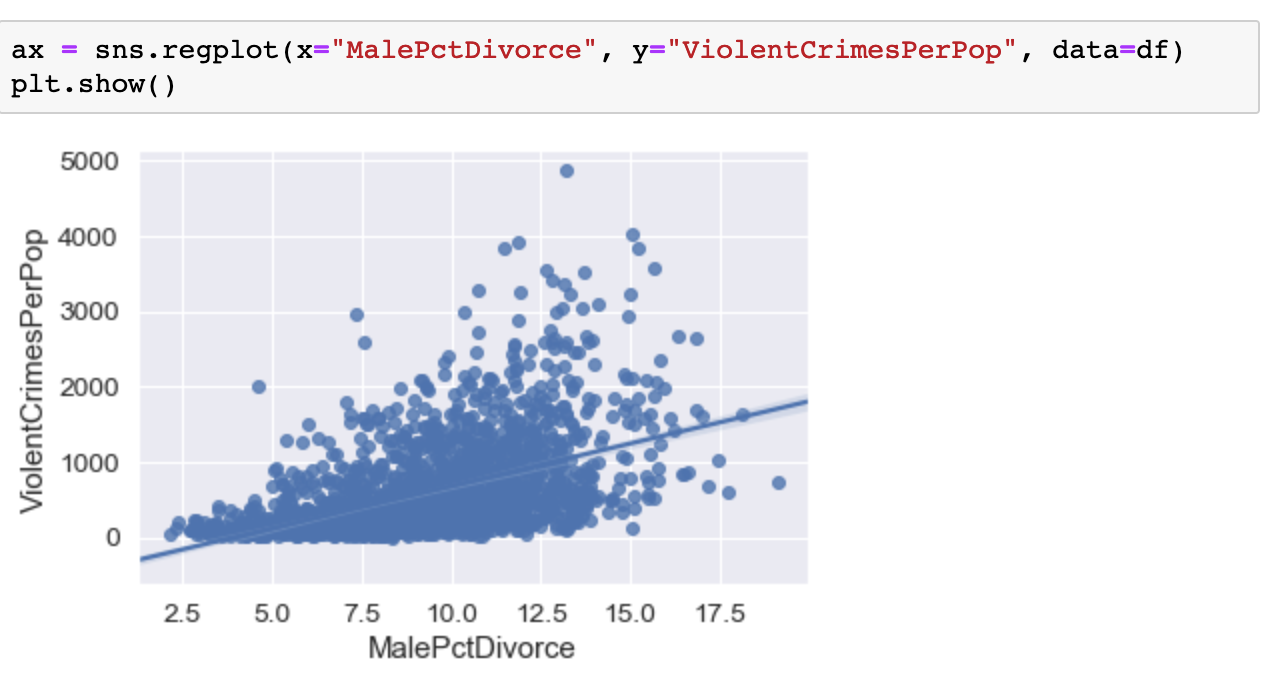


Figure [30]

The independent variable Male percent Never Married shows a direct relationship with violent crime as the slope of the plots is positive.

As all points are normally distributed in each side of the residual plot.

The residual plot appears to have a constant variance scattered around zero.



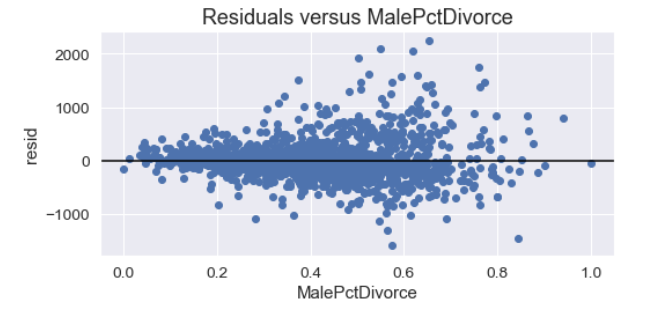


Figure [31]

The independent variable Male Divorced percentage shows a direct relationship with violent crime as the slope of the plots is positive.

As all points are normally distributed in each side of the residual plot.

The residual plot appears to have a constant variance scattered around zero.

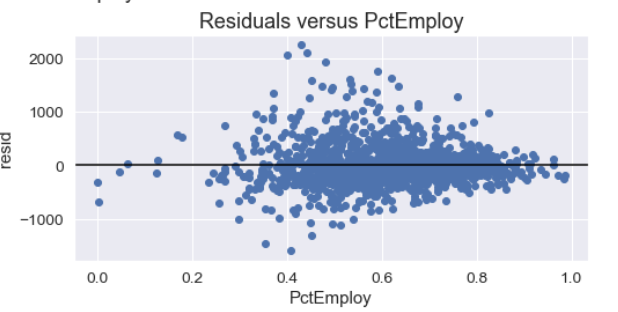
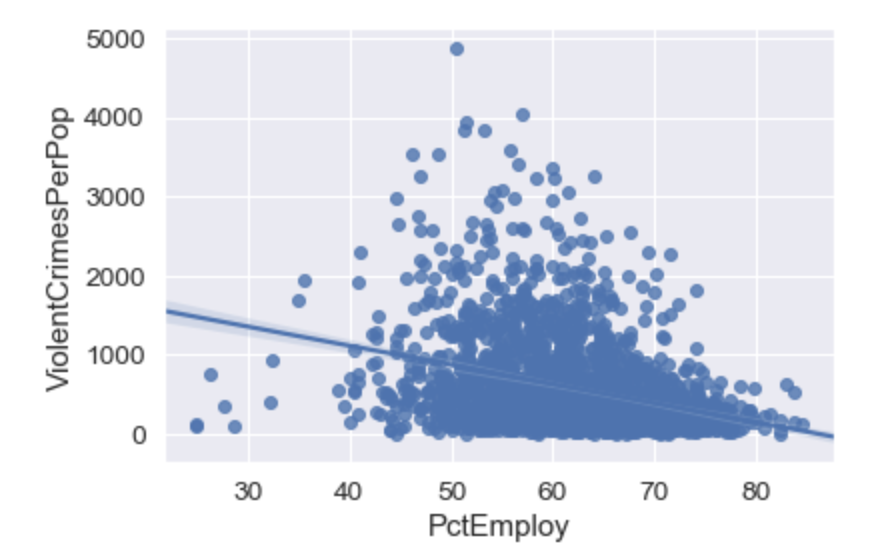


Figure [32]

The independent variable employment percentage shows an indirect relationship with violent crime as the slope of the plots is negative.

As all points are normally distributed in each side of the residual plot.

The residual plot appears to not have a constant variance. The mean is still zero.

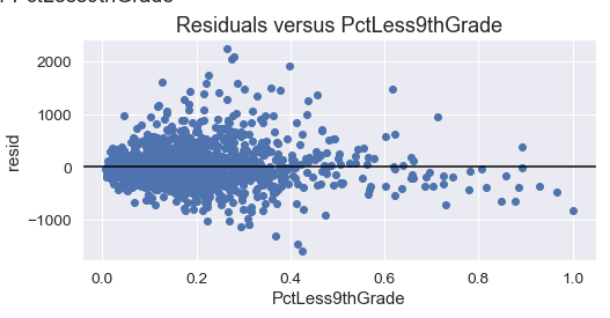
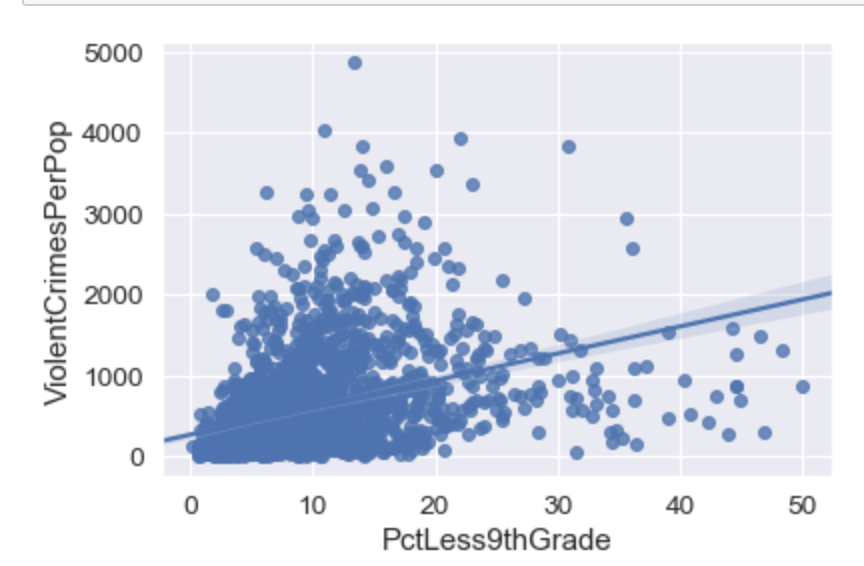


Figure [33]

The independent variable less than 9th grade percentage shows a direct relationship with violent crime as the slope of the plots is positive.

As all points are normally distributed in each side of the residual plot.

The residual plot appears to have a constant variance scattered around zero.



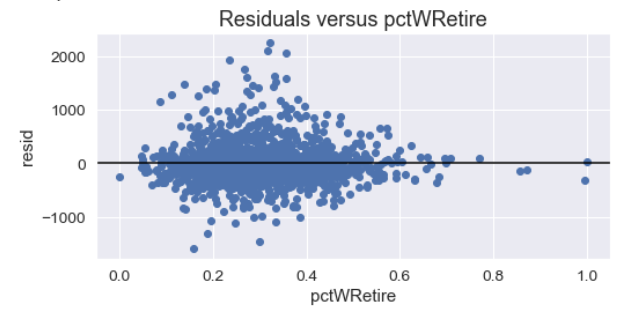
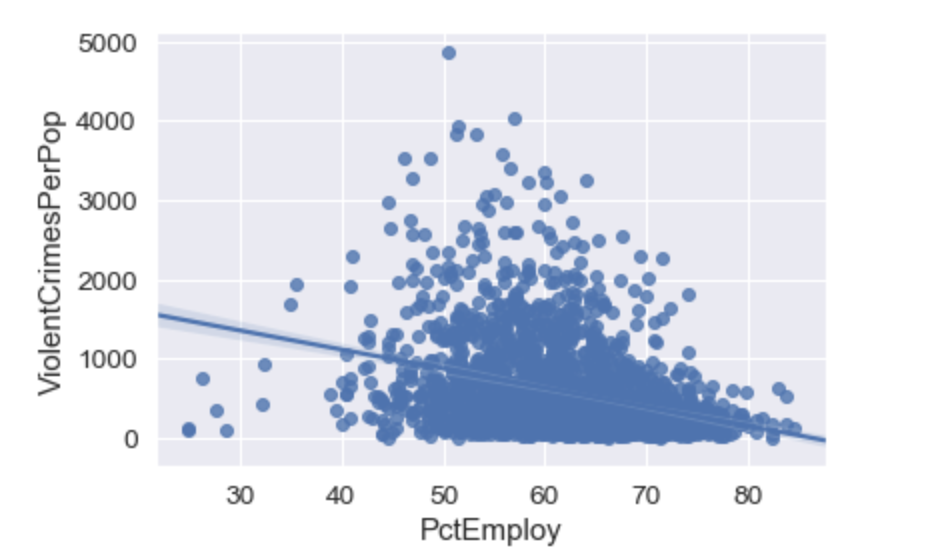


Figure [34]

The independent variable percent retired shows a nonlinear relationship with violent crime as the slope of the plots is inconsistent.

As all points are normally distributed in the residual plot.

The residual plot appears to not have a constant variance scattered. But the mean looks like it is zero.



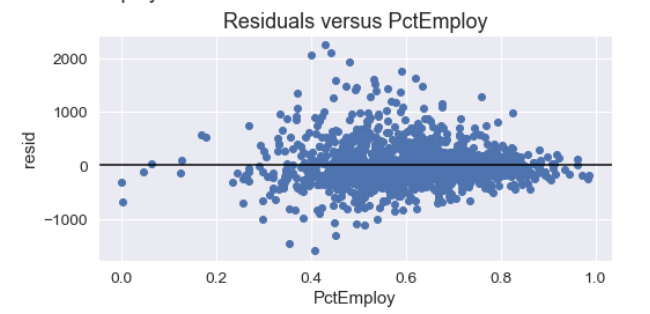


Figure [35]

The independent variable employment percentage shows an indirect relationship with violent crime as the slope of the plots is negative.

As all points are normally distributed in each side of the residual plot.

The residual plot appears to have a constant variance. The mean is zero

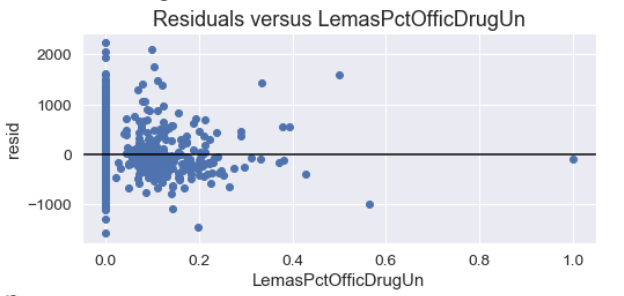


Figure [36]

The independent variable LemasPctOfficDrugshows a no clear relationship with violent crime.

As all points are not normally distributed in each side of the residual plot.

The residual plot appears to not have a constant variance. Mean is still scattered around zero.

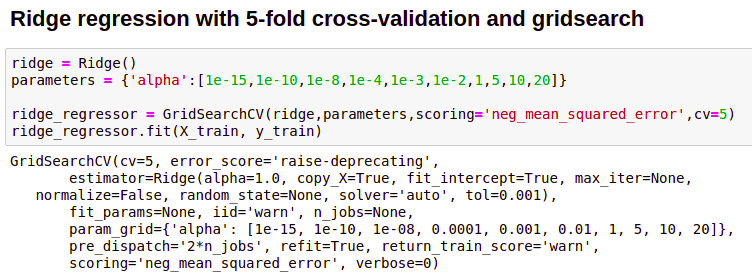


Figure [37]

**Algorithm, Implementation and Results**

We start **cross validation with 5 fold**

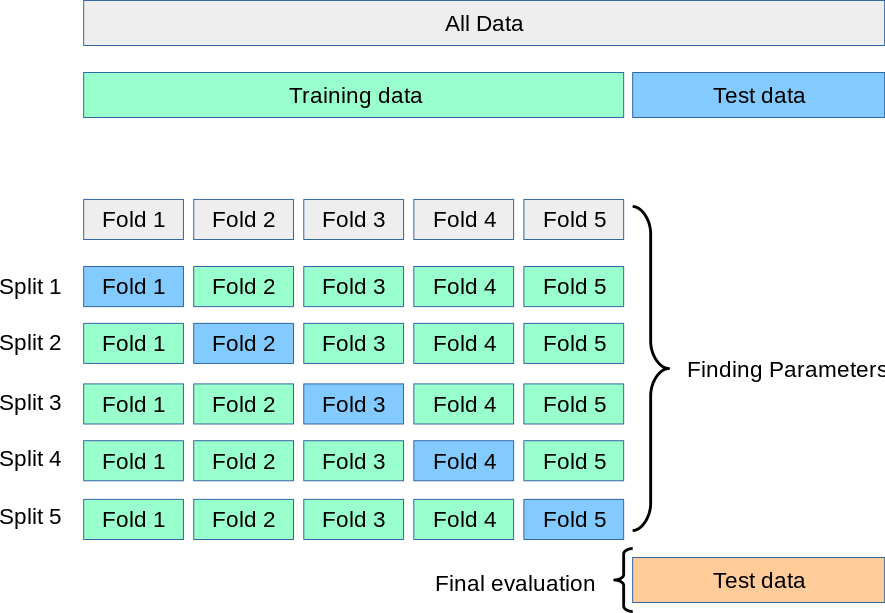
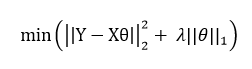


Figure [38]

Five equal parts split differently and iterated 5 times. Last time in regular linear regression, one iteration of an 80 to train 20 to split was implemented. EX: A-B-C-D-E, A-D for training and E for testing, we repeat these steps where B-E is for training and A is tested as well using 5 fold. As shown above in the diagram, essentially 5 iterations are tested times for a result of test data for different five mse respectively. Finally we take a mean of the mean squared errors.

***Ridge Regression-***

Equation =

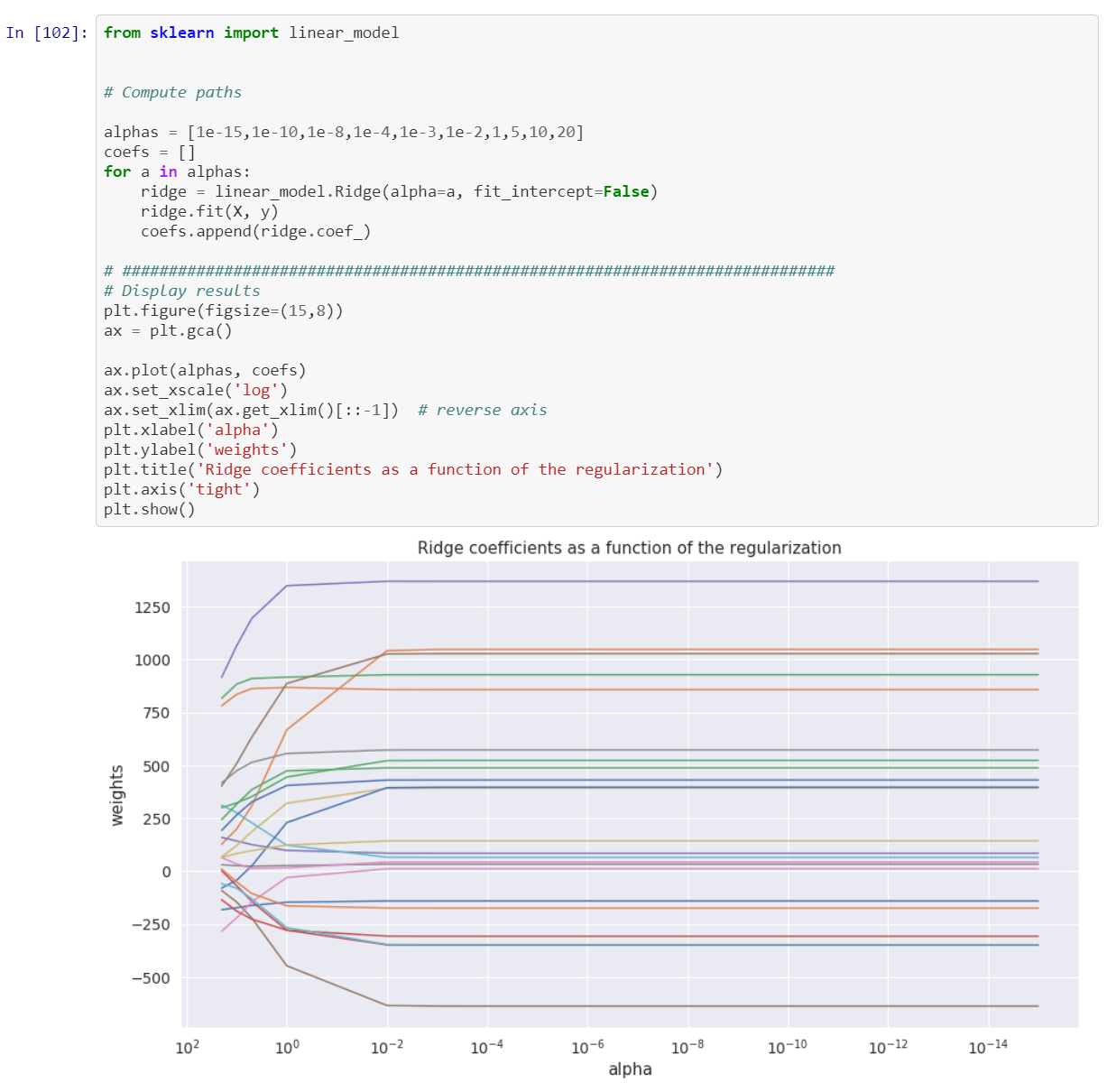


Figure [39]

Coefficients are placed into an array, and we loop through the array for alpha values. Testing on different alphas for ridge regression, we want an optimized alpha value that is least error. In our application we want the lowest mse.

We initialize the ridge regression parameters, in the the end we are appending (adding to list) the coefficients.

Alpha = lambda in equation above

The graph depicts when alpha is added to theta and with this the weight will be more, theta is larger when alpha is added to it. As alpha increases, the magnitude of the coefficients decrease.



Figure [40]

We are getting 50 alphas generated from the range 10^-4 to 10^1 equal distance apart.

We initialize the value ridge regression model. Next we begin iterating through each value that is in alpha\_space, we get the r2 scores based on the 5 fold cross validation. Standard deviation of the r2 scores are appended to the ridge\_scores\_std. Then we plot the results.

In the plot above, you can see that we plot CV Score with Std Error on the Y axis while alpha is on the X axis. The blue line represents the r2 score, the blue shade represents the standard deviation. You can tell that the optimized values of alpha are near 10^0 which yield a higher r2 score.

Max r2 score is **0.642**

Least r2 score is **0.628**

You can see that the best value of alpha was 1 which is equal to 10^0. Below that, you can see the MSE score.

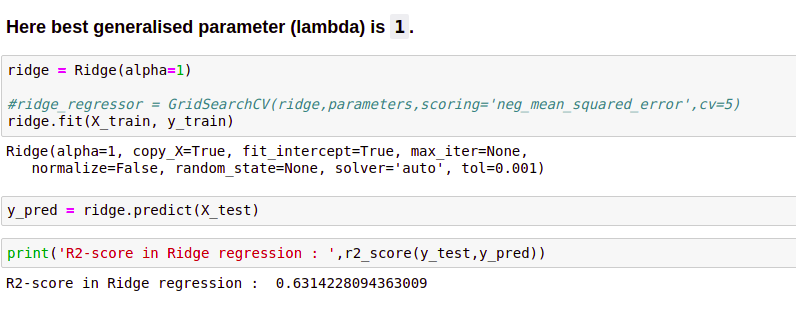


Figure [41]

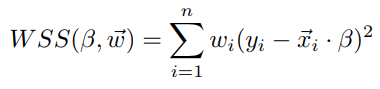
Here at line 120 is where we initialize ridge regression using the optimized alpha value. Next we train our model and generate a r2 score.

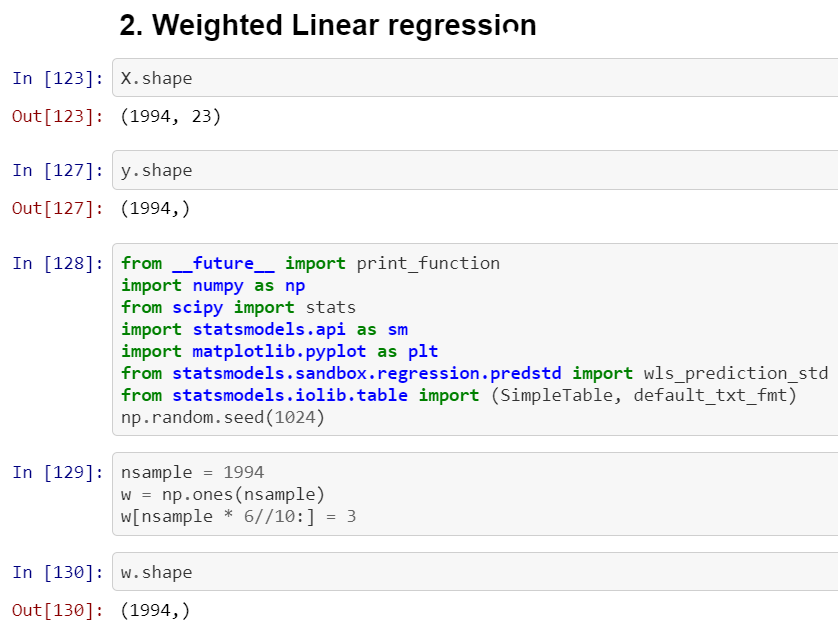
Our r2 score of .655 tells us that this model is good. In our previous model from milestone one, we got a score of .R2 score of test set is 0.6314.

**Theory :**

Weighted linear regression is when we minimize the weighted sum of squares rather than the residual sum of squares as done in regular linear regression.

The equation for this is shown below.





**Weighted Least Squares Estimation** is useful to us as it enables us to view the behavior of the random errors in models and that can then be influenced to distinguishing linear or nonlinear parameters. The nonnegative constants tells us the precision of that particular data point with respect to its placement, and this can show how it acts as a contribution to our observation.

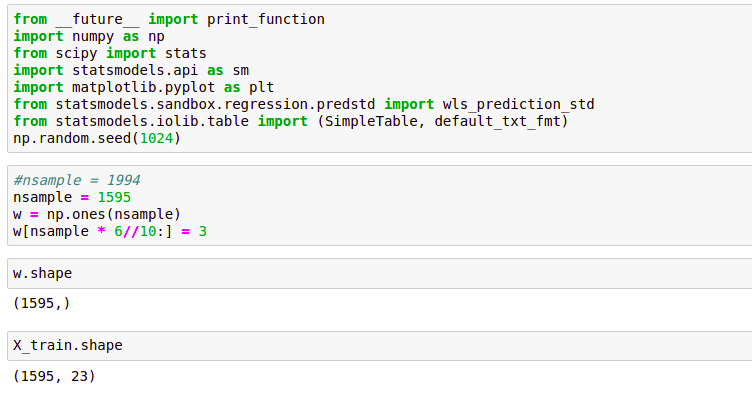


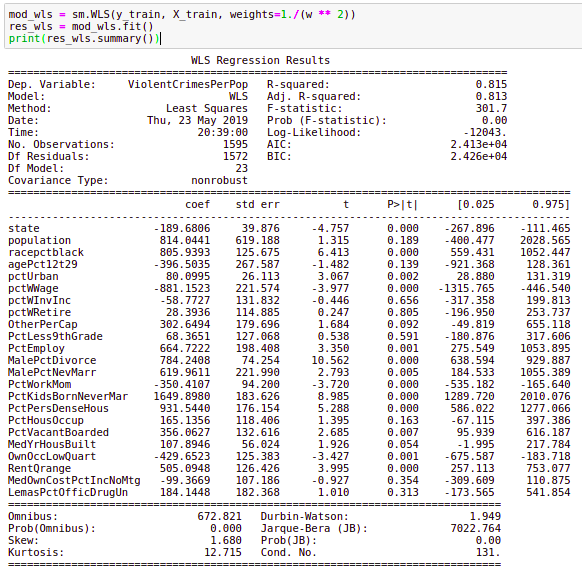
Figure [42]

Here we import the library

Statsmodels.sandbox.regression.predstd import wls\_prediction\_std for the wls\_prediction\_std function.

N sample is the number of samples of training data. W is an array n sample long with all entries equal to 1. Next we set 60% of W equal to 3.

### WLS is the true variance ratio of heteroscedasticity. W is the standard deviation of the error. The WLS variable calls for the weights being proportional to the inverse of the error variance. We initiate the wls model and fit the model prior to printing it.



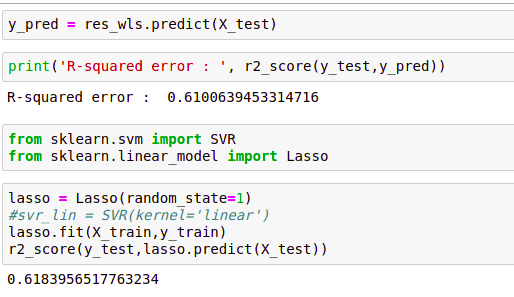


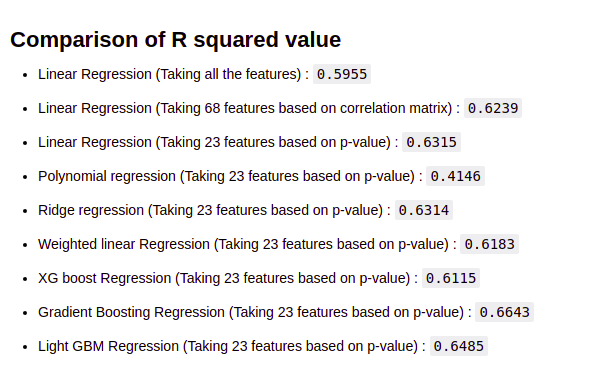
Figure [43]

Here we can tell that weighted least squares estimation is the best as the r2 value is 0.6183.

**Ensemble model Regression :**

We tried 3 ensemble model like XG boost, Gradient Boosting and Light GBM model. From which Gradient Boosting outperform all the other model with r2 score 0.6643.

**Model Comparison :**



**Top Feature Ranking :**

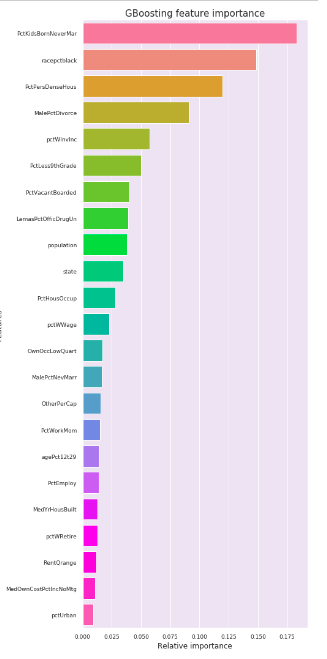
****

Figure [44]

This figure implicates the feature importance by lightgbm model with their ranking of importance. From this we can figure out what factor is more responsible towards the prediction.

**Conclusions**

The theoretical score of r2 which is perfect max score is 1, but practically by the end we reached a value of 0.6643. This is the maximum r2 score from comparing all our previous models tried in all the model. The maximum ‘perfect’ score of r2 is 1. Also we got top 23 features with ranking which has most effect on crimes.

From the figure [44] we can conclude that feature like State, percentage of black people,population, Never married male people percentage, age of people from 12 to 29,income of people, employment percentage take major role in violent crime in a community.

By utilising these factors we can minimise the crime in a community and take appropriate action to reduce crime for a community.

**Future Work :**

* We can optimise our algorithm by using cross validation and grid search method to find optimum parameter hence increase the r2 score.
* We can remove outliers to get more accurate result.
* We can transform our data to fit normal distribution.

**Works Cited**

<https://scikit-learn.org/stable/modules/cross_validation.html>

<https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/?fbclid=IwAR27WSkwIOWvX9xgfh_o3Yhz2lnHU9PH50II3fhTWMdHjzPvo9bvXsIKCrw>

<https://www.stat.cmu.edu/~cshalizi/350/lectures/18/lecture-18.pdf>

<https://www.numpy.org/>

<https://www.scipy.org/>

<https://www.statsmodels.org/stable/index.html>

<https://matplotlib.org/>

<https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime>

<https://www.kaggle.com/kkanda/analyzing-uci-crime-and-communities-dataset>

<http://www.numpy.org/>

<https://scikit-learn.org/stable/>

<https://pandas.pydata.org/>

<https://seaborn.pydata.org/>

<https://matplotlib.org/>

<https://www.statsmodels.org/stable/index.html>