Predicting NFL Arrests

Juan Salazar

Samuel Sachnoff

Michael Silin

PREPARING THE DATA

BEFORE

	season	week_num	day_of_week	gametime_local	home_team	away_team	home_score	away_score	OT_flag	arrests	division_game
0	2011	1	Sunday	1:15:00 PM	Arizona	Carolina	28	21	1	5	n
1	2011	4	Sunday	1:05:00 PM	Arizona	New York Giants	27	31	1	6	n
2	2011	7	Sunday	1:05:00 PM	Arizona	Pittsburgh	20	32	1	9	n
3	2011	9	Sunday	2:15:00 PM	Arizona	St. Louis	19	13	0	6	y
4	2011	13	Sunday	2:15:00 PM	Arizona	Dallas	19	13	0	3	n
5	2011	14	Sunday	2:05:00 PM	Arizona	San Francisco	21	19	1	4	y
6	2011	15	Sunday	2:15:00 PM	Arizona	Cleveland	20	17	0	1	n
7	2011	17	Sunday	2:15:00 PM	Arizona	Seattle	23	20	0	4	y
8	2012	1	Sunday	1:25:00 PM	Arizona	Seattle	20	16	1	0	y
9	2012	3	Sunday	1:05:00 PM	Arizona	Philadelphia	27	6	1	12	n
10	2012	4	Sunday	1:05:00 PM	Arizona	Miami	24	21	0	4	n

AFTER

	season	week_num	day_of_week	gametime_local	home_team	away_team	home_score	away_score	OT_flag	arrests	division_game
0	2011	1	0	0	0	4	28	21	1	5	0
1	2011	4	0	1	0	20	27	31	1	6	0
2	2011	7	0	1	0	24	20	32	1	9	0
3	2011	9	0	2	0	28	19	13	1	6	1
4	2011	13	0	2	0	8	19	13	1	3	0
5	2011	14	0	3	0	26	21	19	1	4	1
6	2011	15	0	2	0	7	20	17	1	1	0
7	2011	17	0	2	0	27	23	20	1	4	1
8	2012	1	0	4	0	27	20	16	1	0	1
9	2012	3	0	1	0	23	27	6	1	12	0
10	2012	4	0	1	0	16	24	21	1	4	0
11	2012	6	0	1	0	3	19	16	1	1	0
12	2012	8	1	5	0	26	3	24	1	3	1

- Most of the games were played on Sunday
- We used Custom functions and DataFrame.transform() to convert the data to vectors
- Randomly shuffle the indices by using np.random.shuffle()
- Split the Shuffled Data Frame into 80% Training and 20% Testing sets

- Normalized training and testing data and multiply negative correlations by -1 whether in cross validation or in actual model predictions

- We used a 10-fold cross validation to create our model

- we then trained our model and predicted the testing data

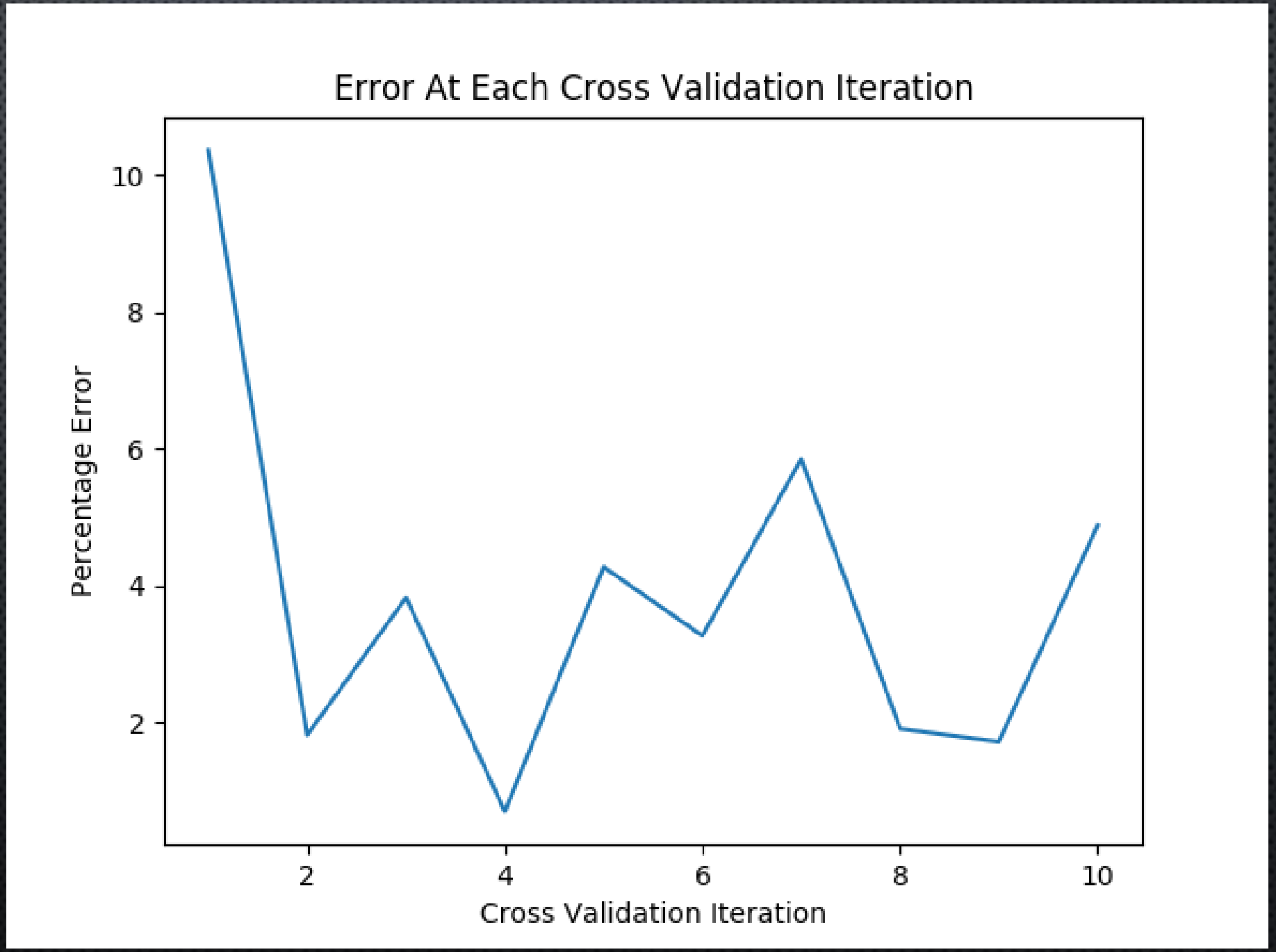
- Used a polynomial degree of 5

- we then ran our program 10 times to determine how Accurate our model was

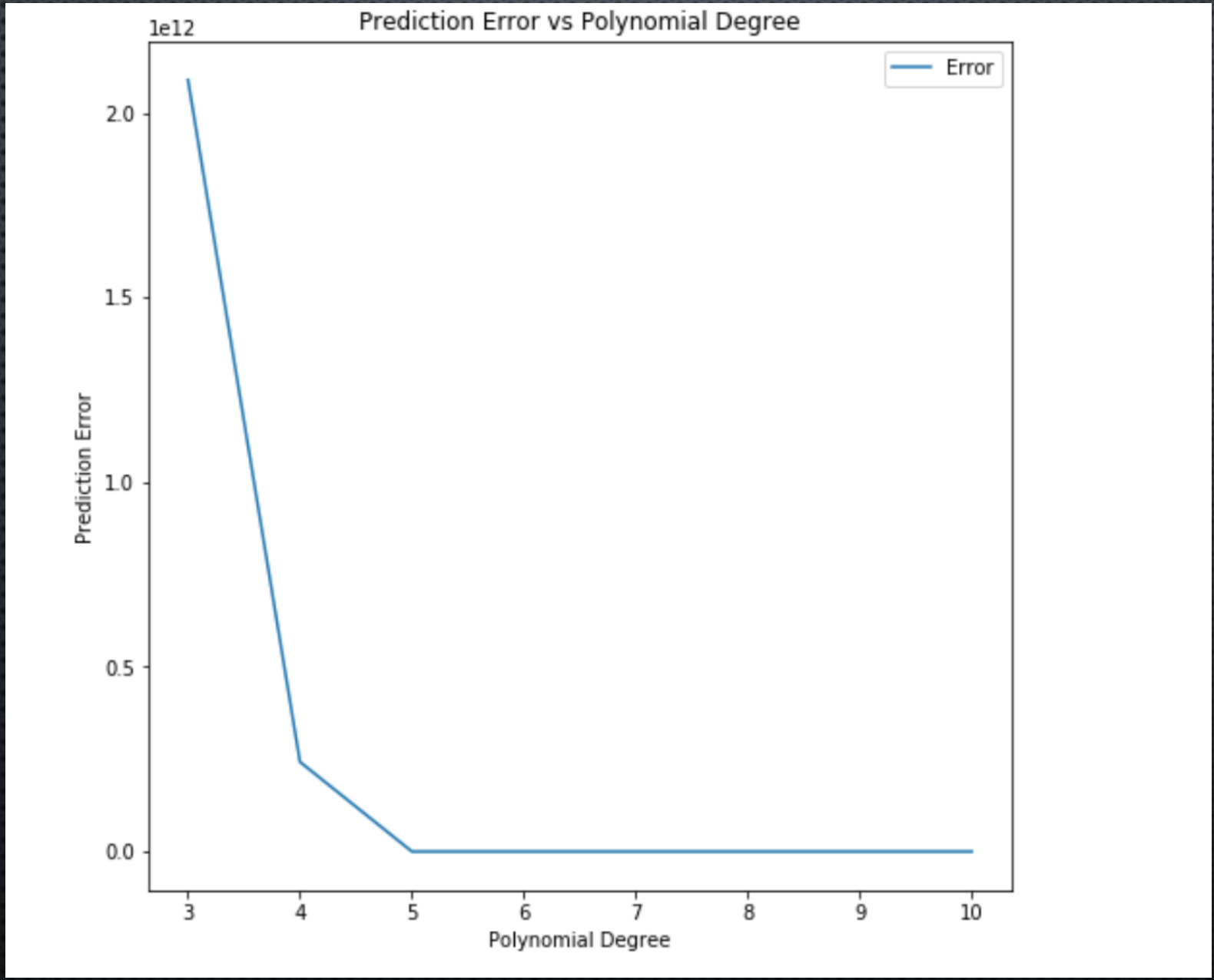
- avg error: +/- 3.54 arrests

```
def crossValidate(df1):
    errors= []
    #get the different x columns to take correlations of
    x_columns= np.array([])
    x_columns= np.append(df1.columns[0],x_columns)
    x_columns= np.append(df1.columns[2:],x_columns)
    y_data= df1['arrests']
    x_data= df1[x_columns]
    kf = KFold(n_splits=10,shuffle= True)
    sampleIndices= [5,12]
    model= make_pipeline(PolynomialFeatures(5),linear_model.LinearRegression())
    for train_index, test_index in kf.split(x_data):
        train_x= x_data.iloc[train_index]
        test_x= x_data.iloc[test_index]
        train_y= y_data.iloc[train_index]
        test_y= y_data.iloc[test_index]
        #normalize the data
        for col in x_columns:
            if train_x[col].corr(train_y)<0:
                train_x[col]=train_x.loc[:,col]*-1
                test_x[col]= test_x.loc[:,col]*-1
        train_x= (train_x-train_x.mean())/train_x.std()
        test_x= (test_x-test_x.mean())/test_x.std()
        model.fit(train_x,train_y)
        #get the error
        y_fit= model.predict(test_x)
        error= (test_y-y_fit).mean()
        errors.append(error)
    errors= np.array(errors)
    print(errors.mean())

    #Plot graph of errors
    erroSeries = pd.Series(errors, index=np.arange(1,len(errors)+1))
    for i in range(len(erroSeries)):
        if(erroSeries.iloc[i] < 0):
            erroSeries.iloc[i] *= -1
    erroSeries.plot()

    import matplotlib.pyplot as plt
    plt.ylabel("Percentage Error")
    plt.xlabel("Cross Validation Iteration")
    plt.title("Error At Each Cross Validation Iteration")

    return model
```

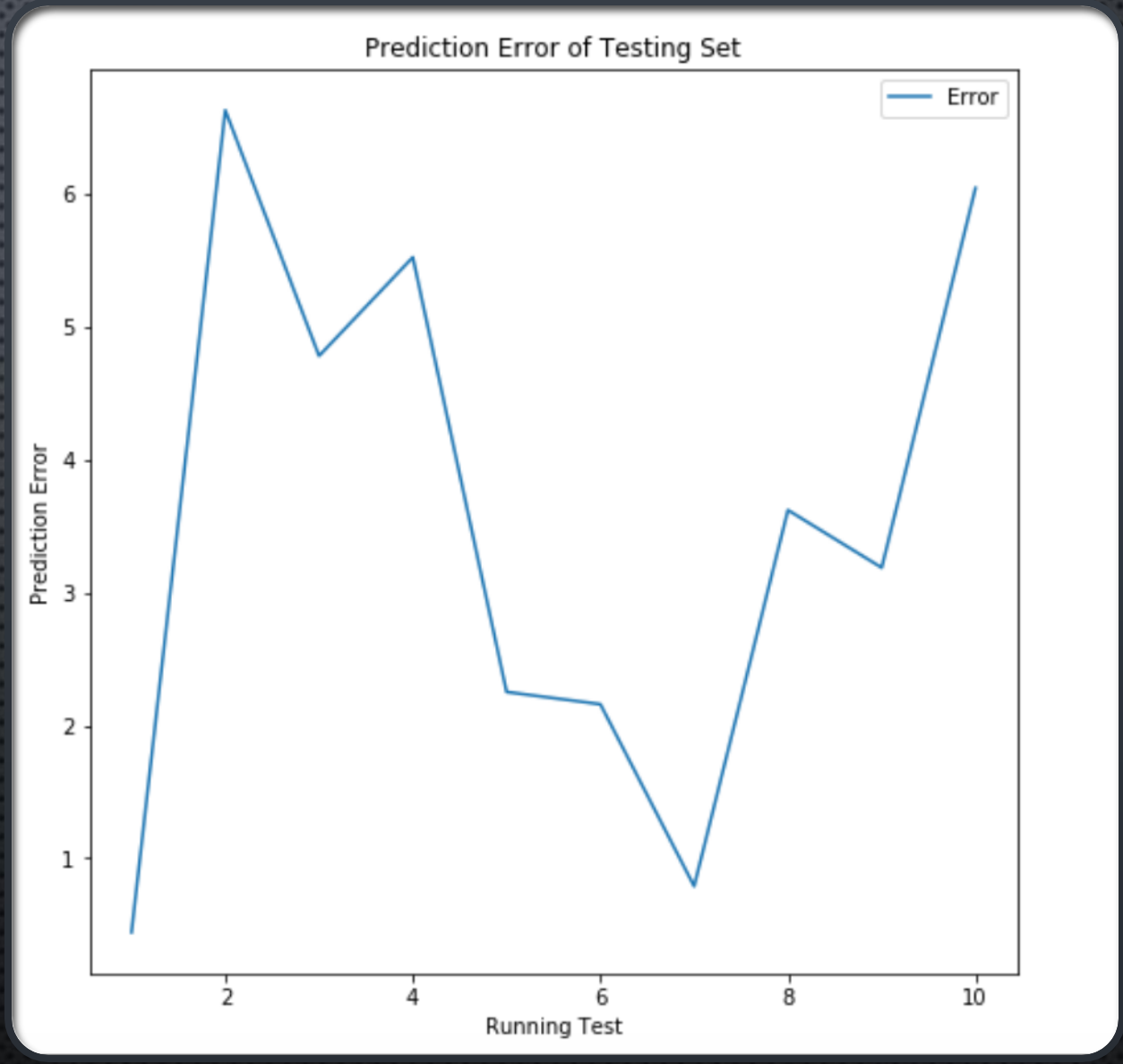
CROSS VALIDATION ERROR



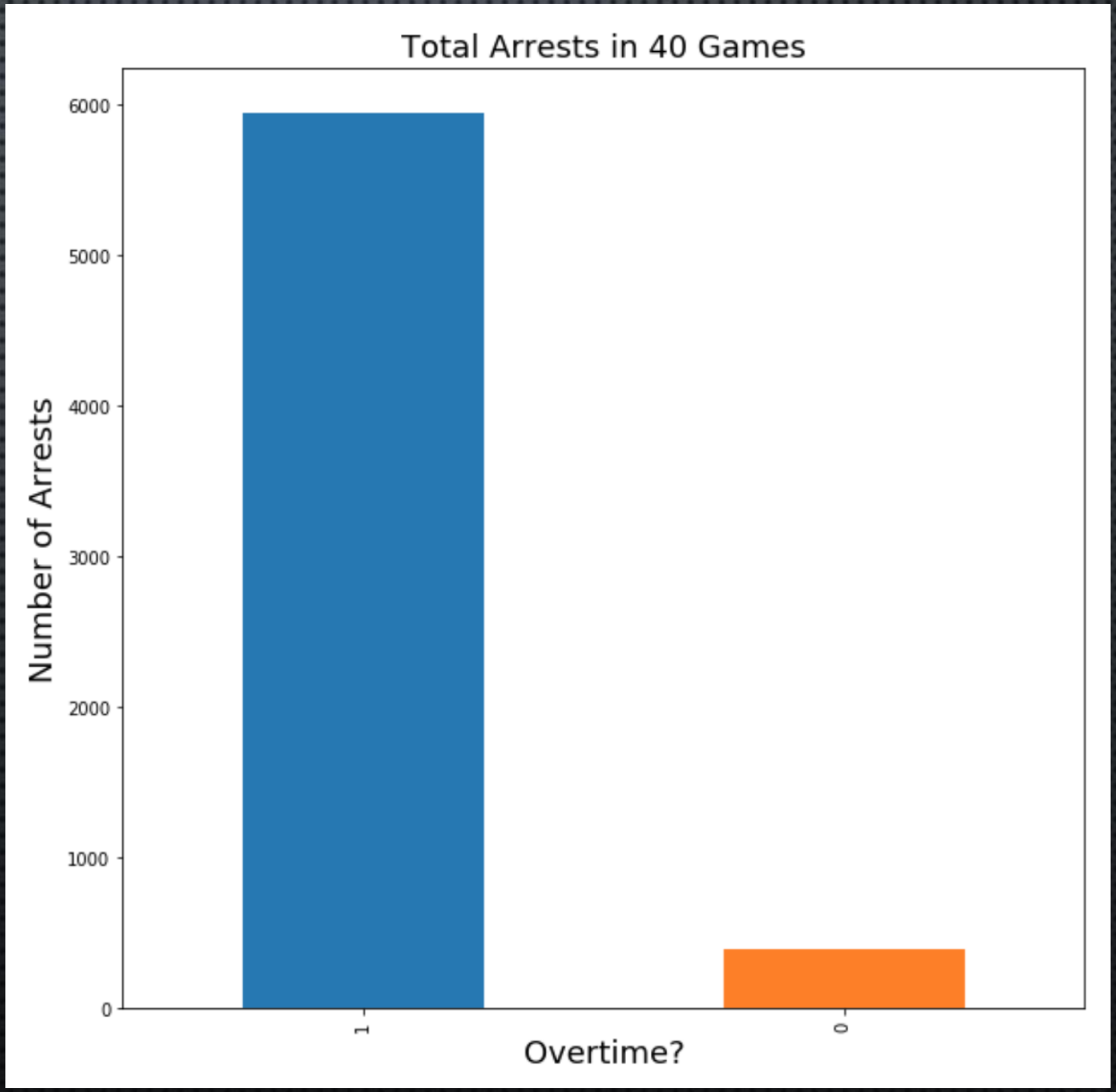
Errors of different Poly Degrees



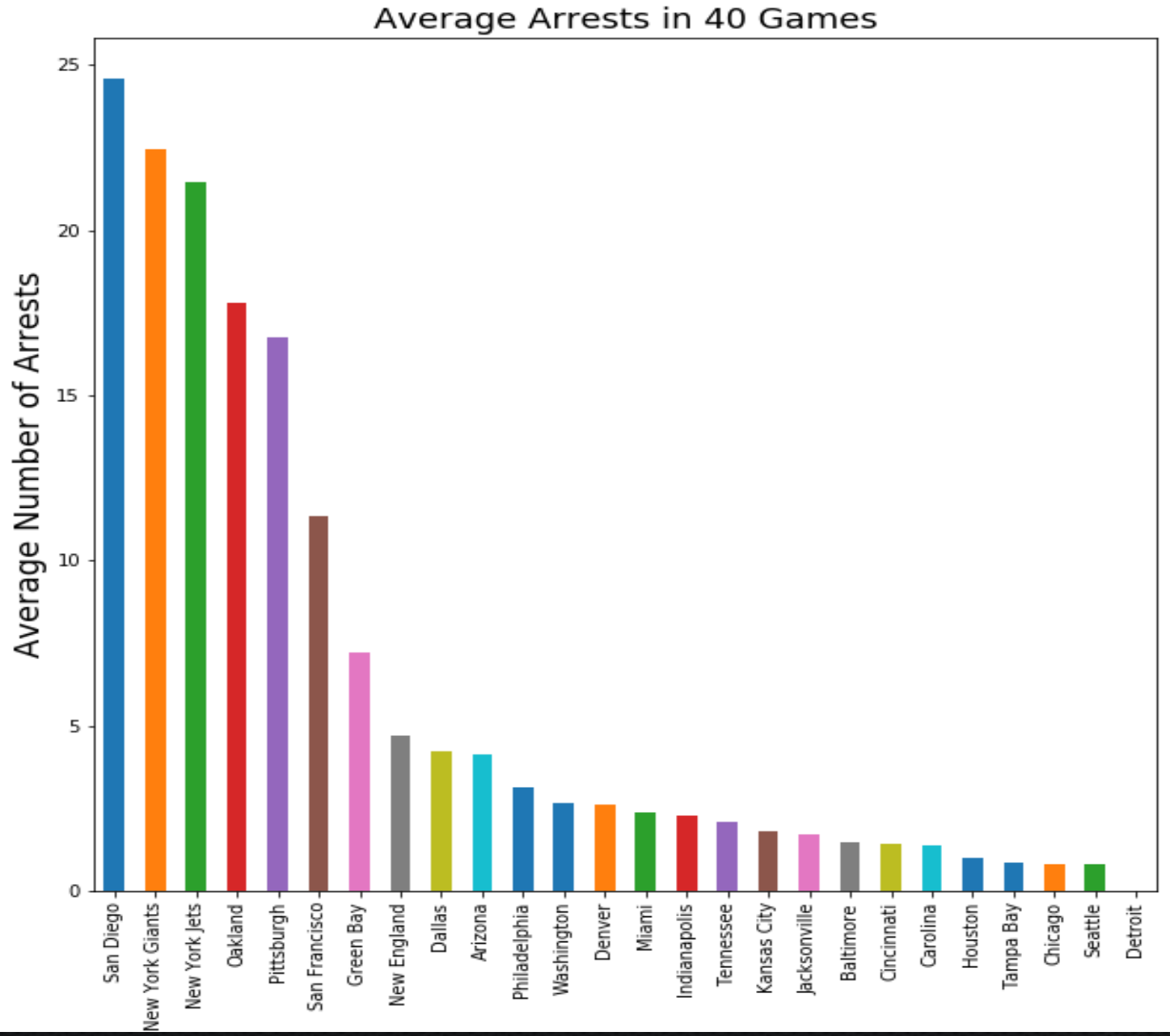
#PREDICTION ERRORS AFTER RUNNING THE PROGRAM 10 DIFFERENT TIMES

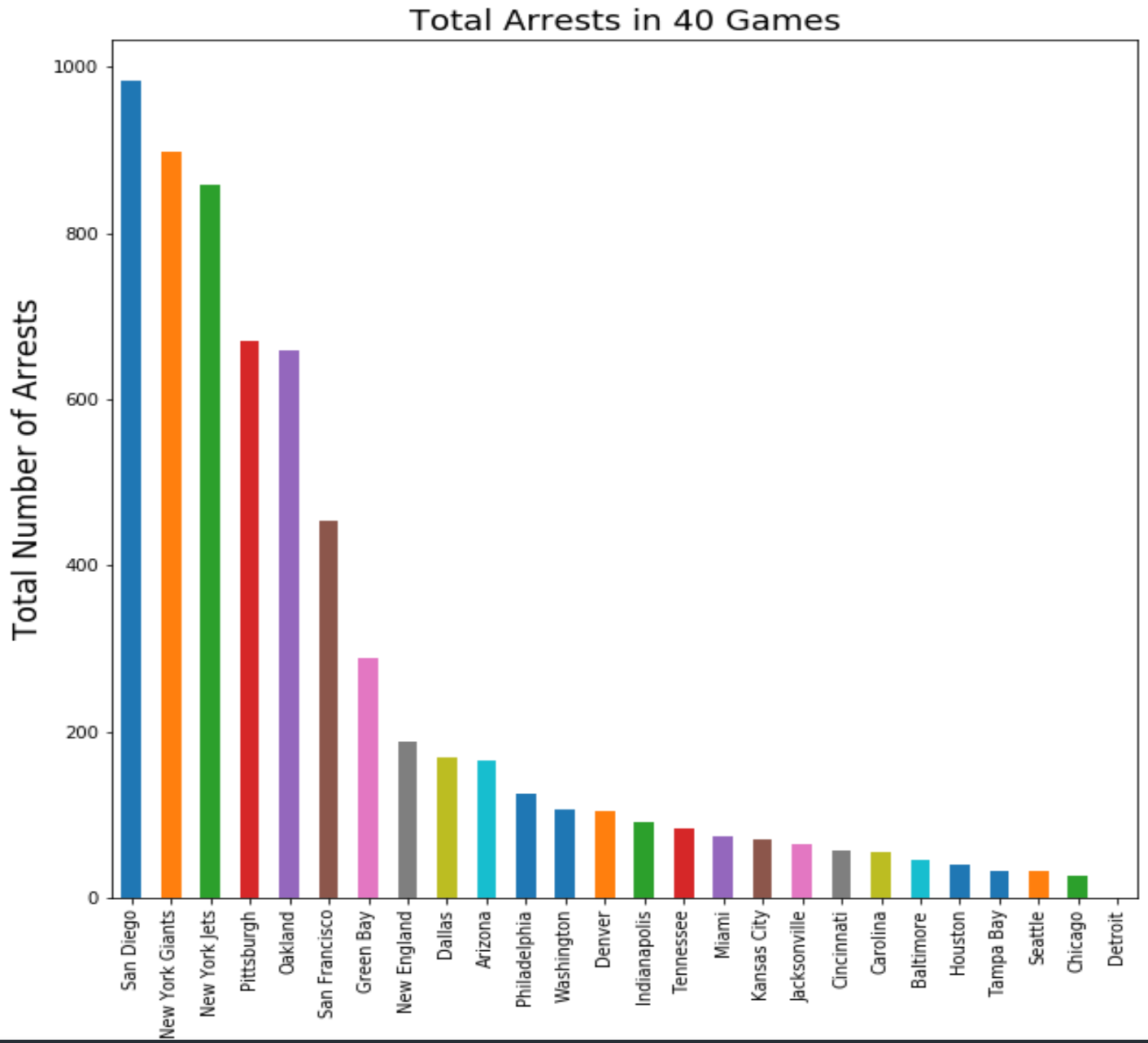


Total Arrests grouped by Overtime flag



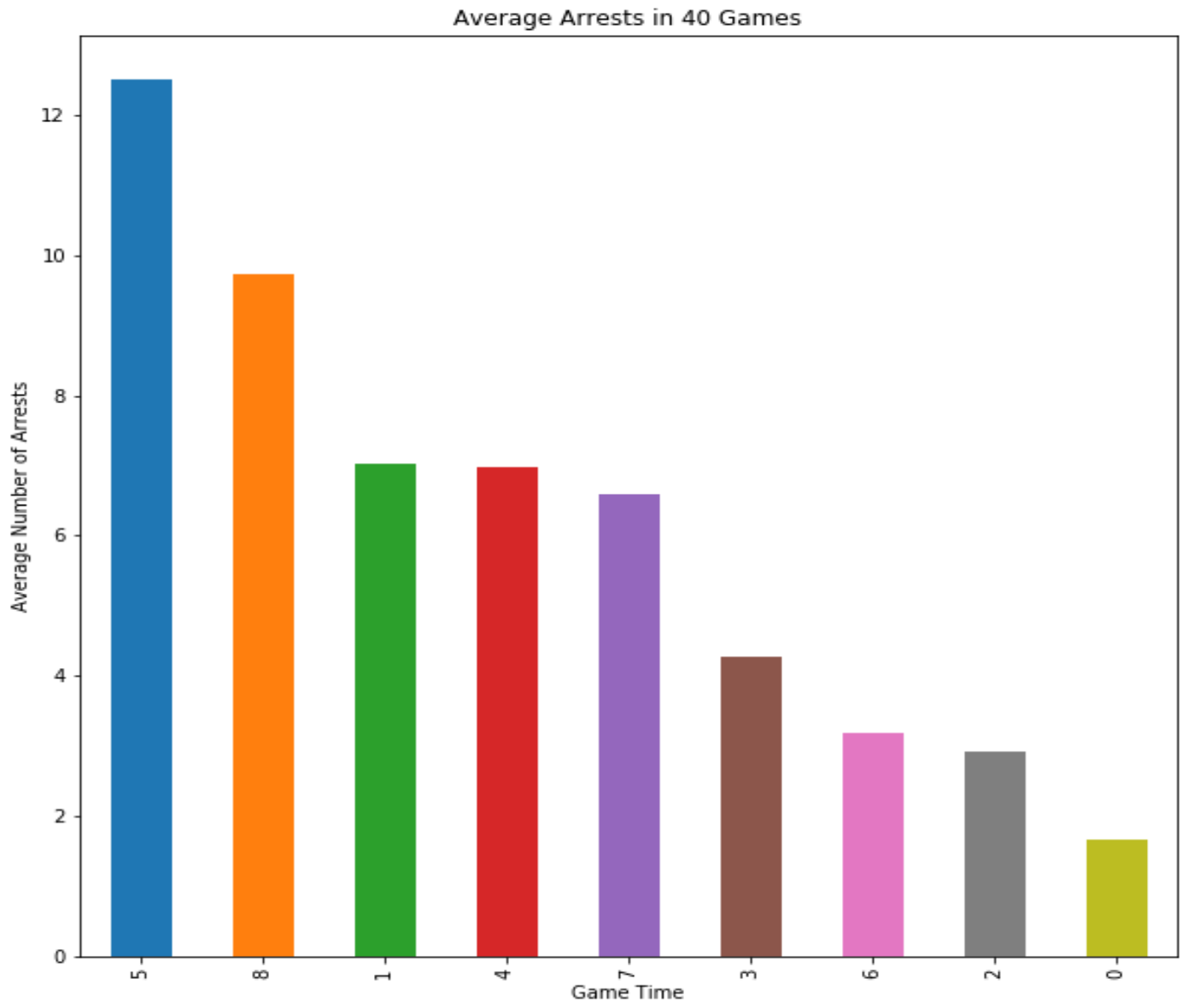
Average Arrests grouped By Team





TOTAL ARRESTS GROUPED BY TEAM

Average Arrests grouped by Game Time



OBSTACLES

- Sklearn K-fold cross validation did not automatically normalize and mult negative correlations by –1
- Was getting chained assignment warning message when doing cross validation when mult negative correlations by -1

WHAT WE LEARNED!!!

- Cross validation
- Using sklearn for linear regression
- Graphng data
- Full picture of class
- San Diego has the highest number of arrests(Almost 1000 in 40 games)