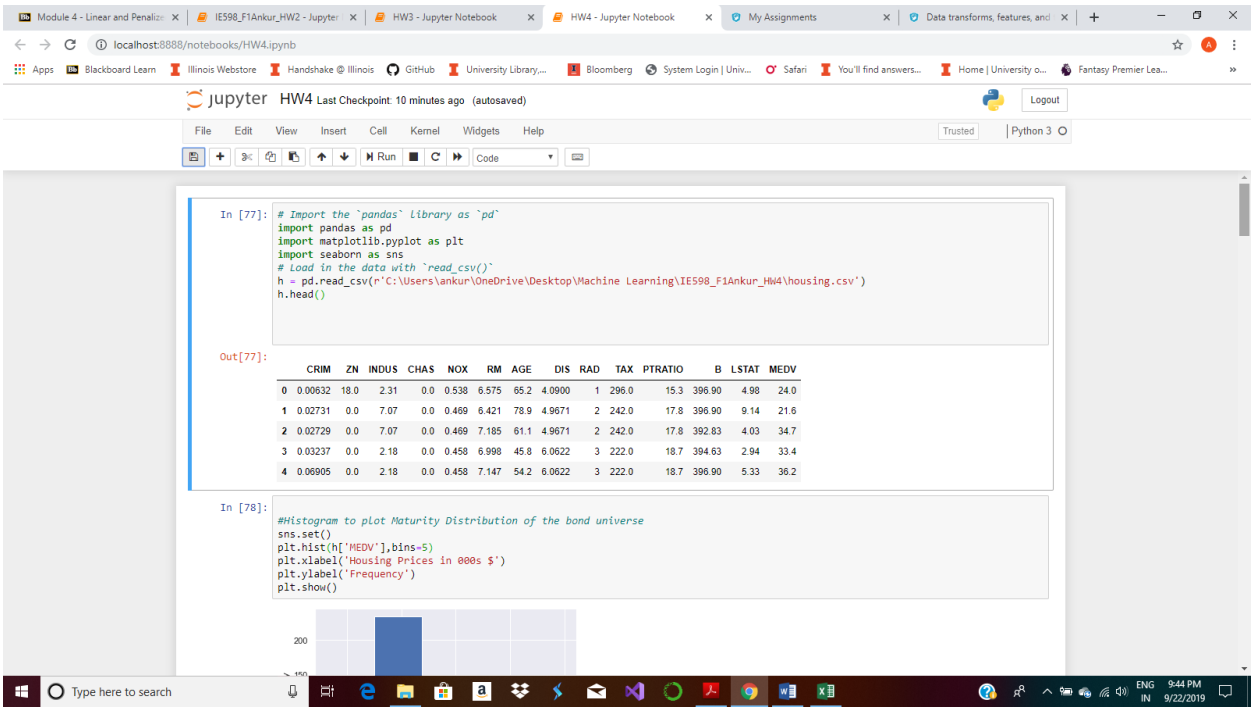
Ankur Mukherjee (ankurm3)

IE598 MLF F18

Module 4 Homework (Regression)

Part 1: Exploratory Data Analysis

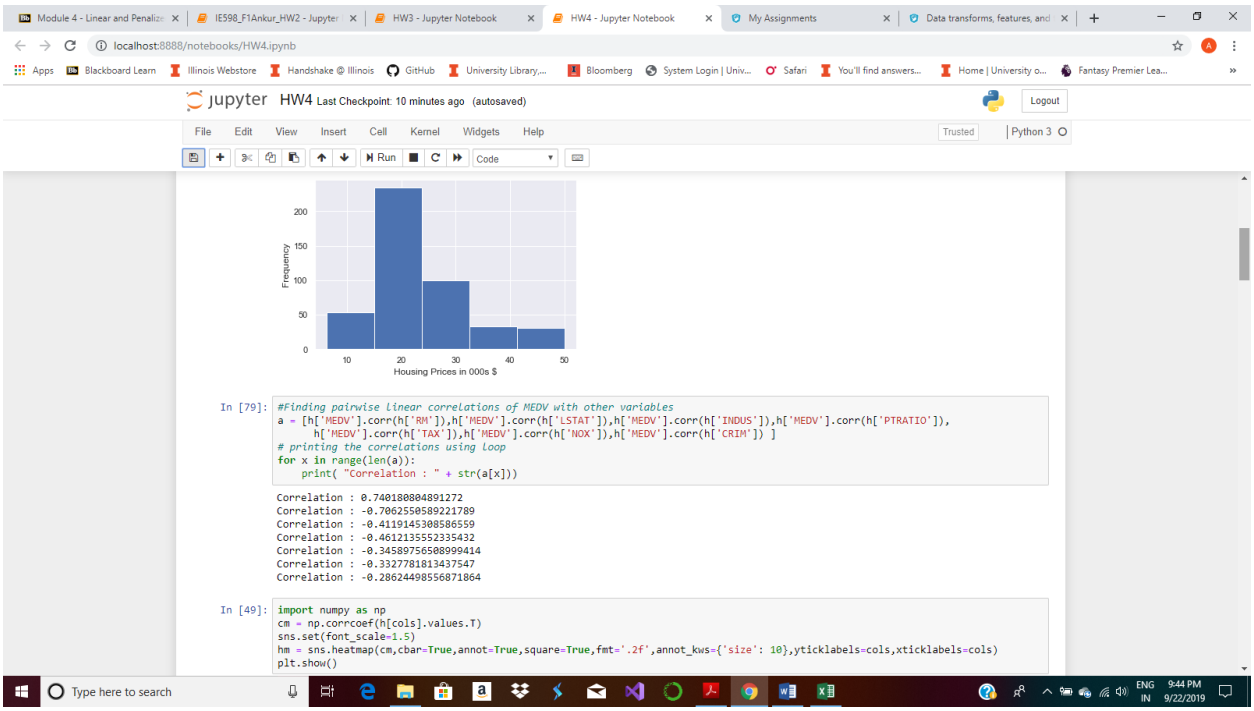


In [77]:
```
# Import the `pandas` library as `pd`
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# Load in the data with `read_csv()`
h = pd.read_csv(r'C:\Users\ankur\OneDrive\Desktop\Machine Learning\IE598_F1Ankur_HW4\housing.csv')
h.head()
```

Out[77]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.90	5.33	36.2

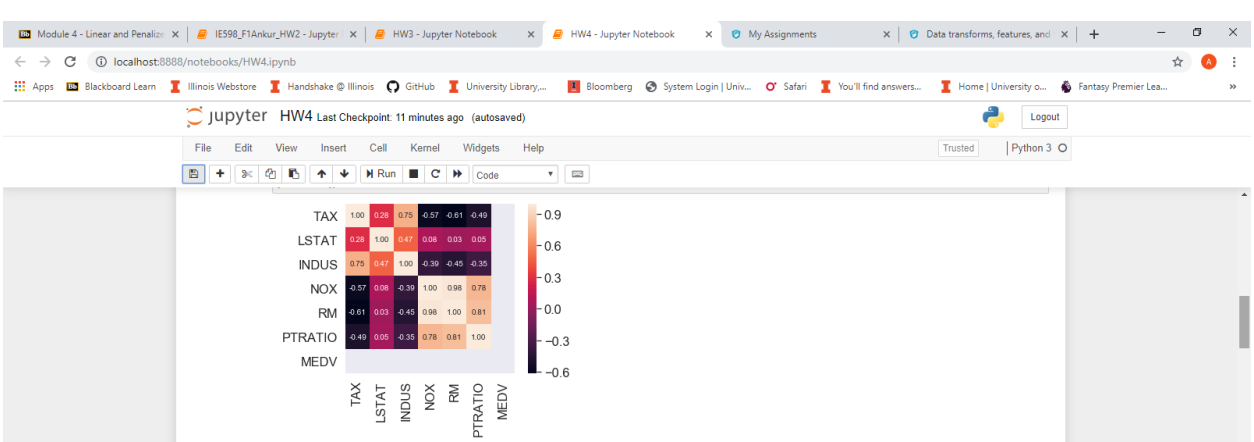
In [78]:
```
#Histogram to plot Maturity Distribution of the bond universe
sns.set()
plt.hist(h['MEDV'],bins=5)
plt.xlabel('Housing Prices in 000s $')
plt.ylabel('Frequency')
plt.show()
```



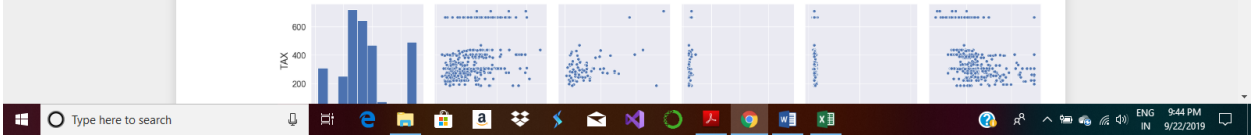
In [79]:
```
#Finding pairwise linear correlations of MEDV with other variables
a = [h['MEDV'].corr(h['RM']),h['MEDV'].corr(h['LSTAT']),h['MEDV'].corr(h['INDUS']),h['MEDV'].corr(h['PTRATIO']),
     h['MEDV'].corr(h['TAX']),h['MEDV'].corr(h['NOX']),h['MEDV'].corr(h['CRIM']) ]
# printing the correlations using loop
for x in range(len(a)):
    print( "Correlation : " + str(a[x]))
```

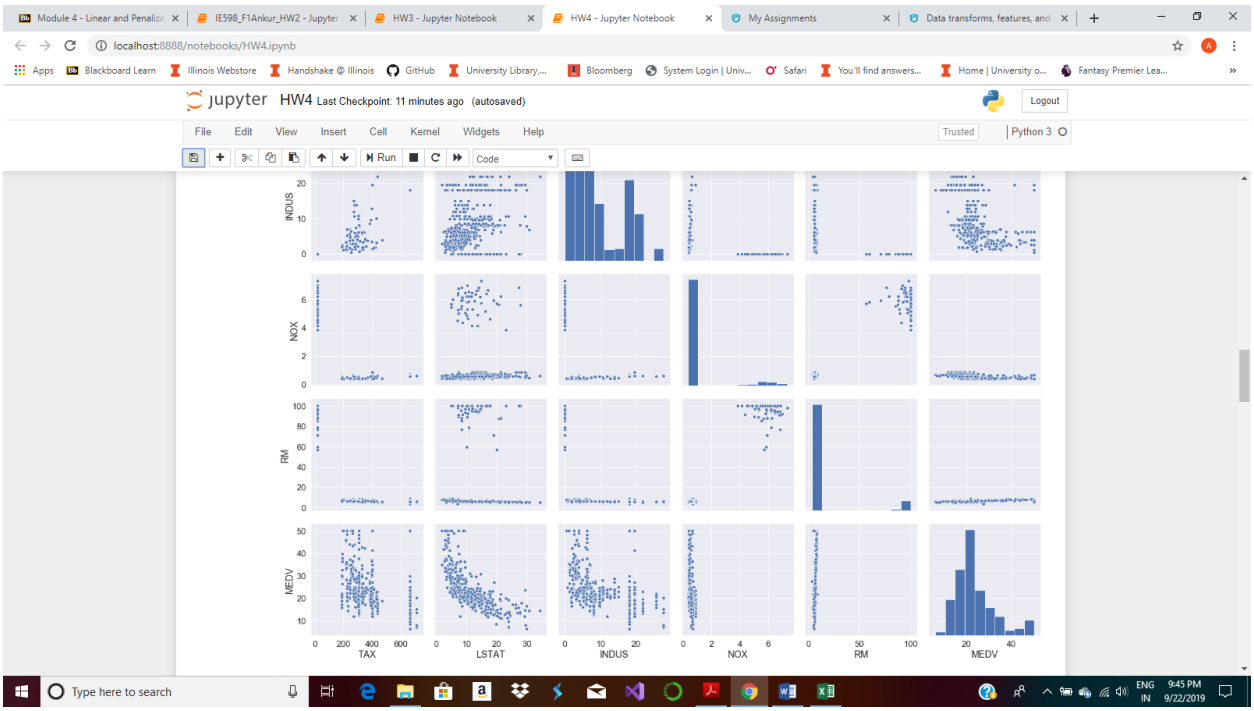
```
Correlation : 0.740180804891272
Correlation : -0.7062550589221789
Correlation : -0.4119145308586559
Correlation : -0.4612135552335432
Correlation : -0.34589756508999414
Correlation : -0.3327781813437547
Correlation : -0.28624498556871864
```

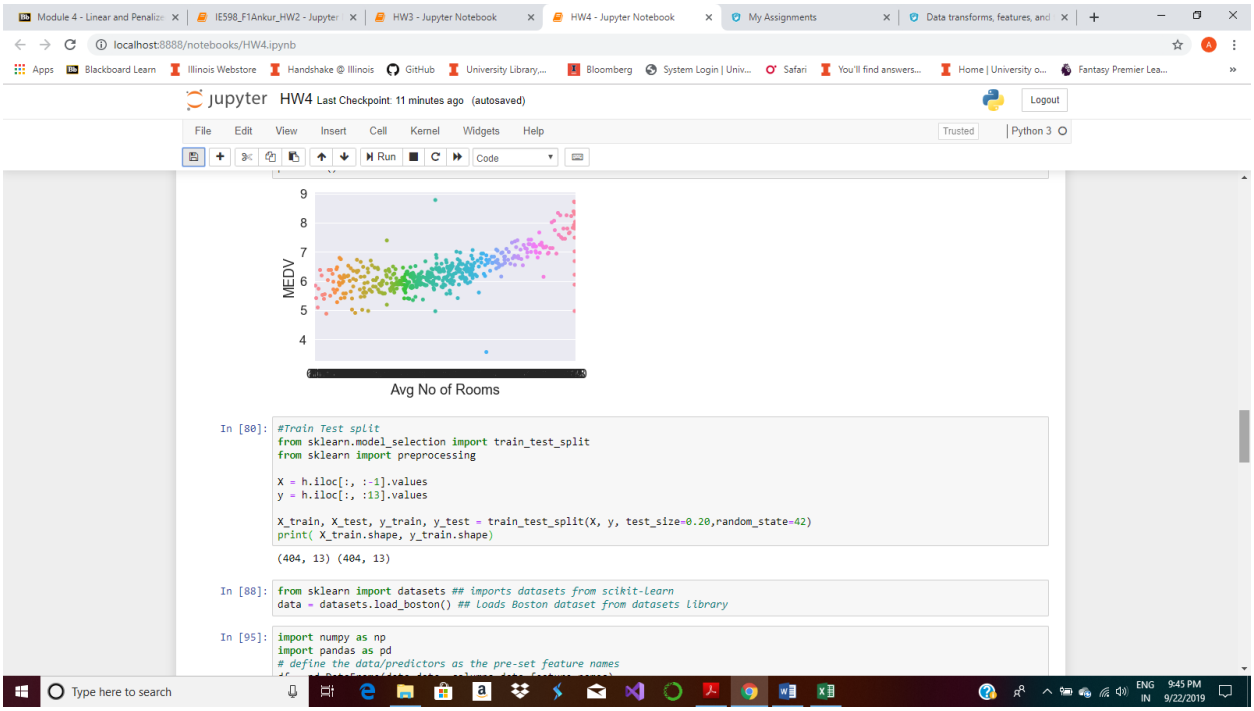
In [49]:
```
import numpy as np
cm = np.corrcoef(h[cols].values.T)
sns.set(font_scale=1.5)
hm = sns.heatmap(cm,cbar=True,annot=True,square=True,fmt='.2f',annot_kws={'size': 10},yticklabels=cols,xticklabels=cols)
plt.show()
```



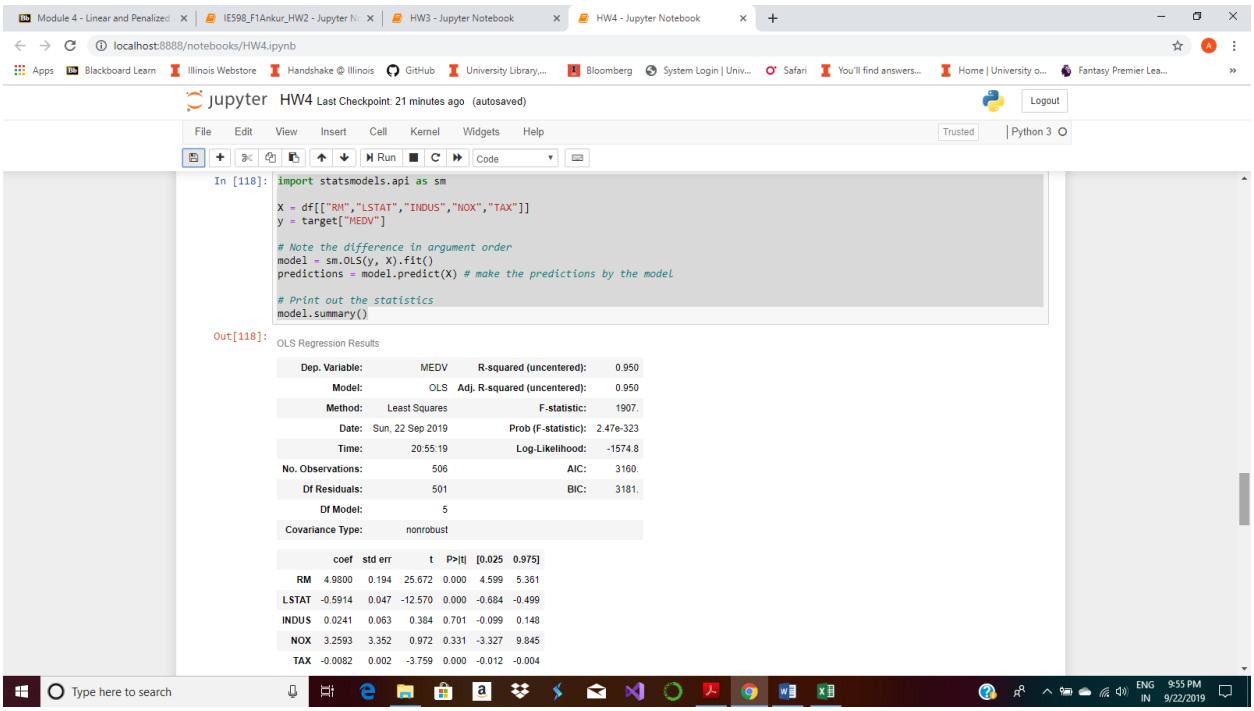
In [50]:
```
#Scatter plot for the above variables which show strong +ve and -ve correlation
import matplotlib.pyplot as plt
import seaborn as sns

cols = ['TAX','LSTAT', 'INDUS', 'NOX', 'RM','MEDV']
sns.pairplot(h[cols], size=3.5)
plt.tight_layout()
plt.show()
```

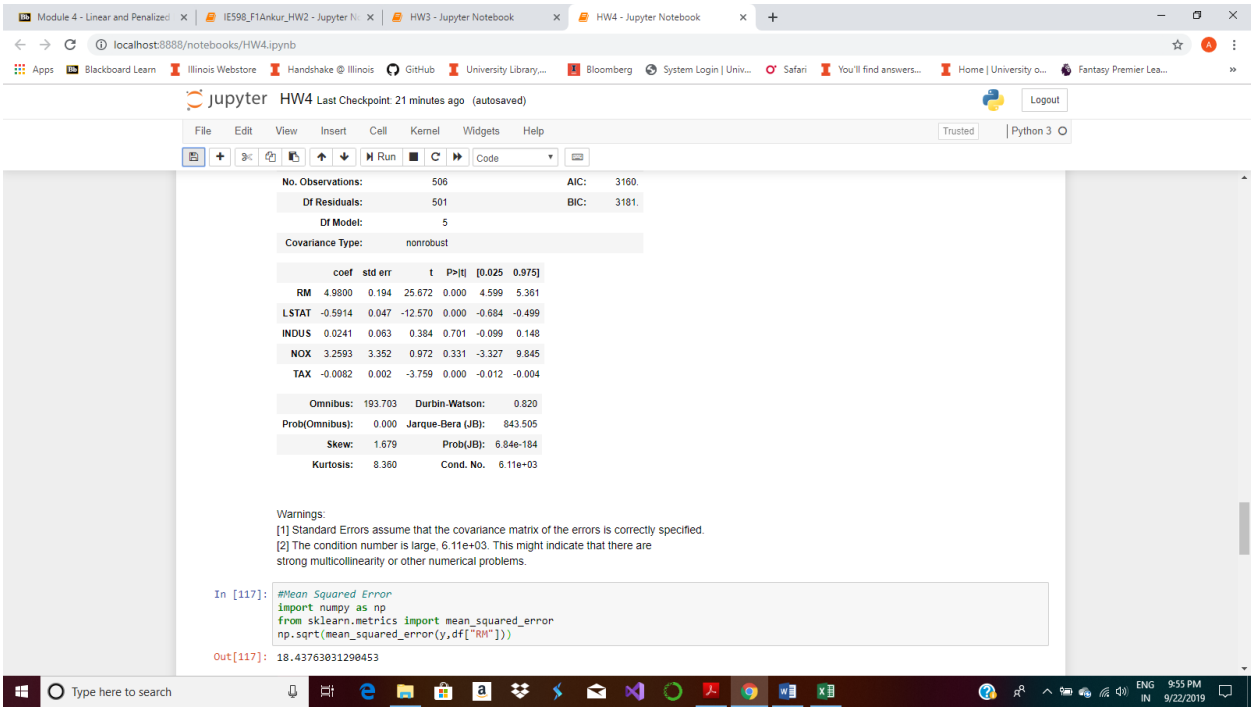




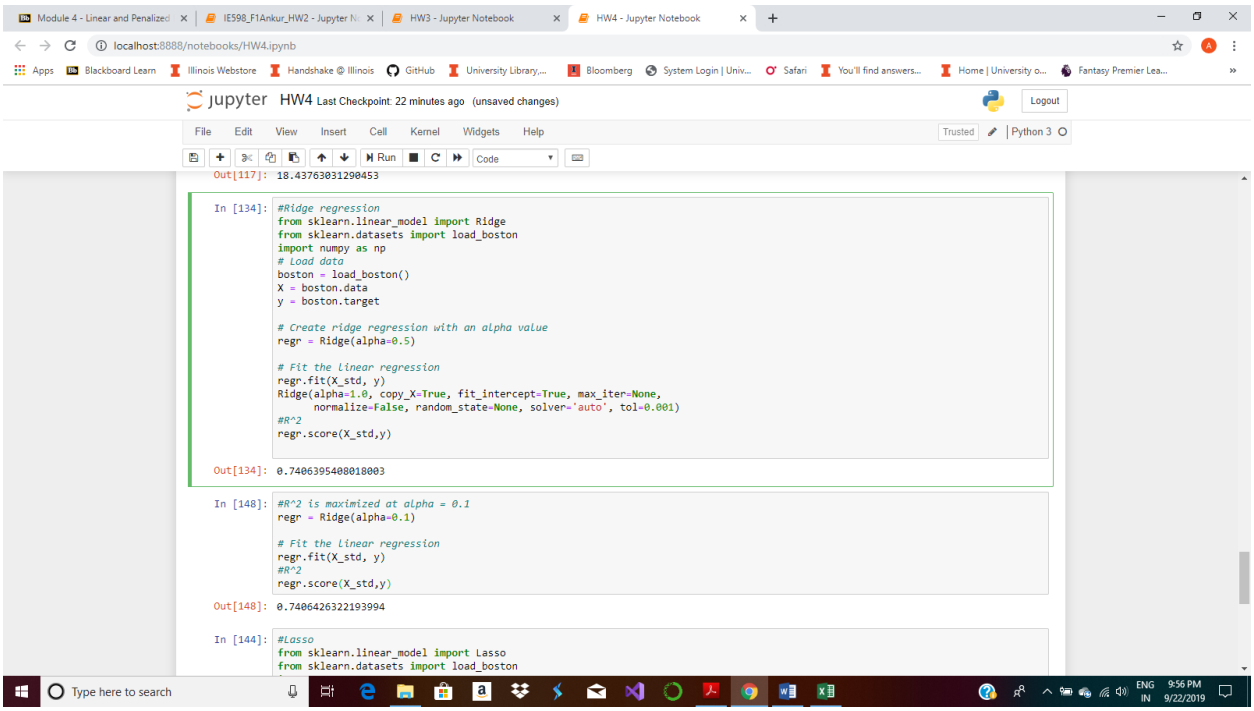


Part 2: Linear regression





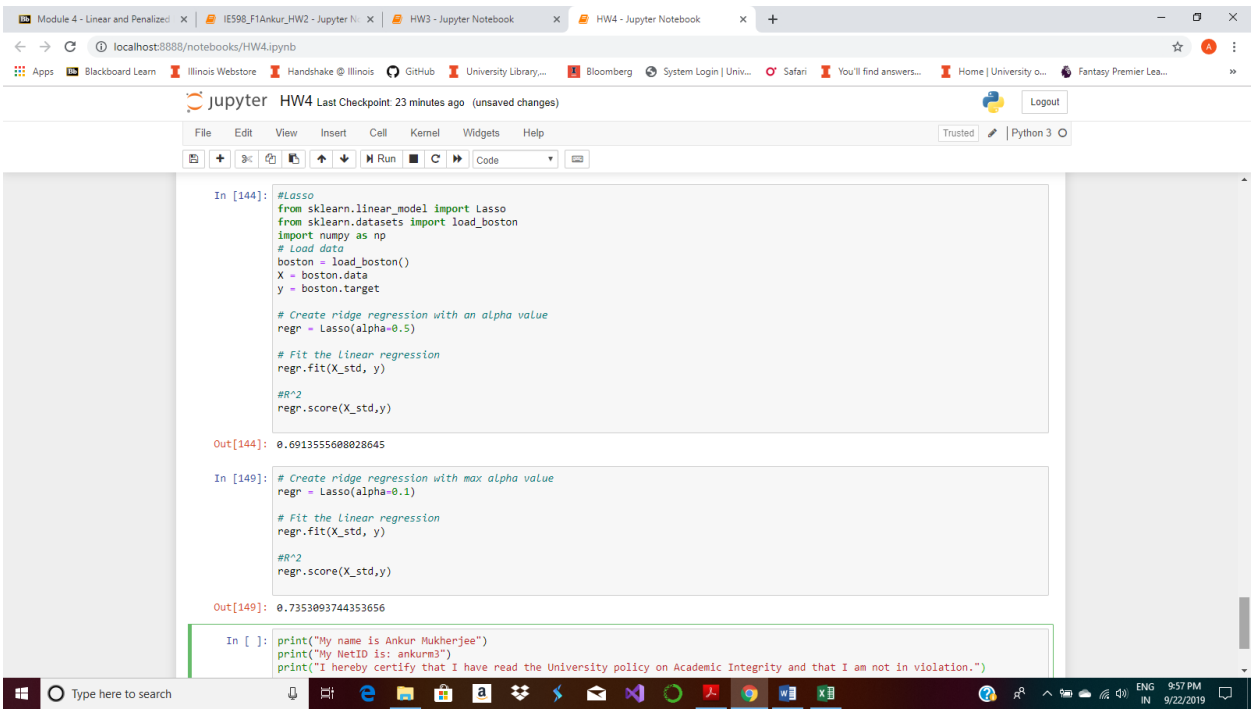
Part 3.1: Ridge regression



```
Out[117]: 18.43763031290453

In [134]: #Ridge regression
from sklearn.linear_model import Ridge
from sklearn.datasets import load_boston
import numpy as np
# Load data
boston = load_boston()
X = boston.data
y = boston.target

# Create ridge regression with an alpha value
regr = Ridge(alpha=0.5)

# Fit the linear regression
regr.fit(X_std, y)
Ridge(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=None,
      normalize=False, random_state=None, solver='auto', tol=0.001)
#R^2
regr.score(X_std,y)

Out[134]: 0.7406395408018003

In [148]: #R^2 is maximized at alpha = 0.1
regr = Ridge(alpha=0.1)

# Fit the linear regression
regr.fit(X_std, y)
#R^2
regr.score(X_std,y)

Out[148]: 0.7406426322193994

In [144]: #Lasso
from sklearn.linear_model import Lasso
from sklearn.datasets import load_boston
```

Coefficient of Determination(R^2) seems to max out as alpha decreases to 0.1

Part 3.2: LASSO regression



```
In [144]: #Lasso
from sklearn.linear_model import Lasso
from sklearn.datasets import load_boston
import numpy as np
# Load data
boston = load_boston()
X = boston.data
y = boston.target

# Create ridge regression with an alpha value
regr = Lasso(alpha=0.5)

# Fit the linear regression
regr.fit(X_std, y)

#R^2
regr.score(X_std,y)

Out[144]: 0.6913555608028645

In [149]: # Create ridge regression with max alpha value
regr = Lasso(alpha=0.1)

# Fit the linear regression
regr.fit(X_std, y)

#R^2
regr.score(X_std,y)

Out[149]: 0.7353093744353656

In [ ]: print("My name is Ankur Mukherjee")
print("My NetID is: ankurm3")
print("I hereby certify that I have read the University policy on Academic Integrity and that I am not in violation.")
```

Coefficient of Determination(R^2) seems to max out as alpha decreases to 0.1

Part 4: Conclusions

1) The linear regression model has a adj R^2 of 0.95. This means that 95% of the variation in MEDV is explained by the combination of RM, INDUS,NOX,LSTAT and TAX variables

2) Some of the coefficients have low t-stats but The f-statistics is high which means together they can explain the variation but not individually. This indicates ***multicollinearity***

3) MEDV is not normally distributed with high kurtosis(fat tails), right positive skewed with a few high outliers – few houses costing more.

Part 5: Appendix

https://github.com/ankurmukherjeeuiuc/IE598_F1Ankur_HW4