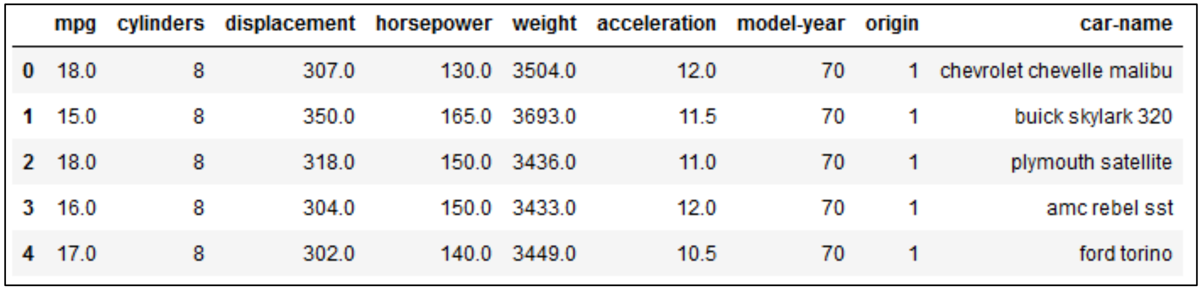
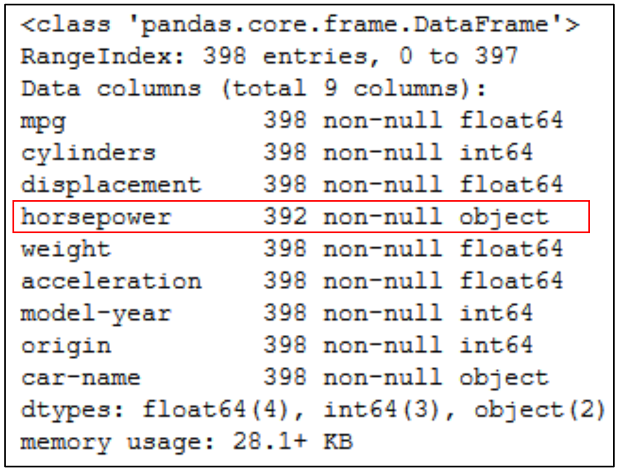
Data Understanding

- Source : http://archive.ics.uci.edu/ml/datasets/Auto+MPG
- Instances : 398
- Attributes : 8
- Goal : Regression to Predict MPG (Mile Per Gallon) or fuel consumption

Pre-processing Data



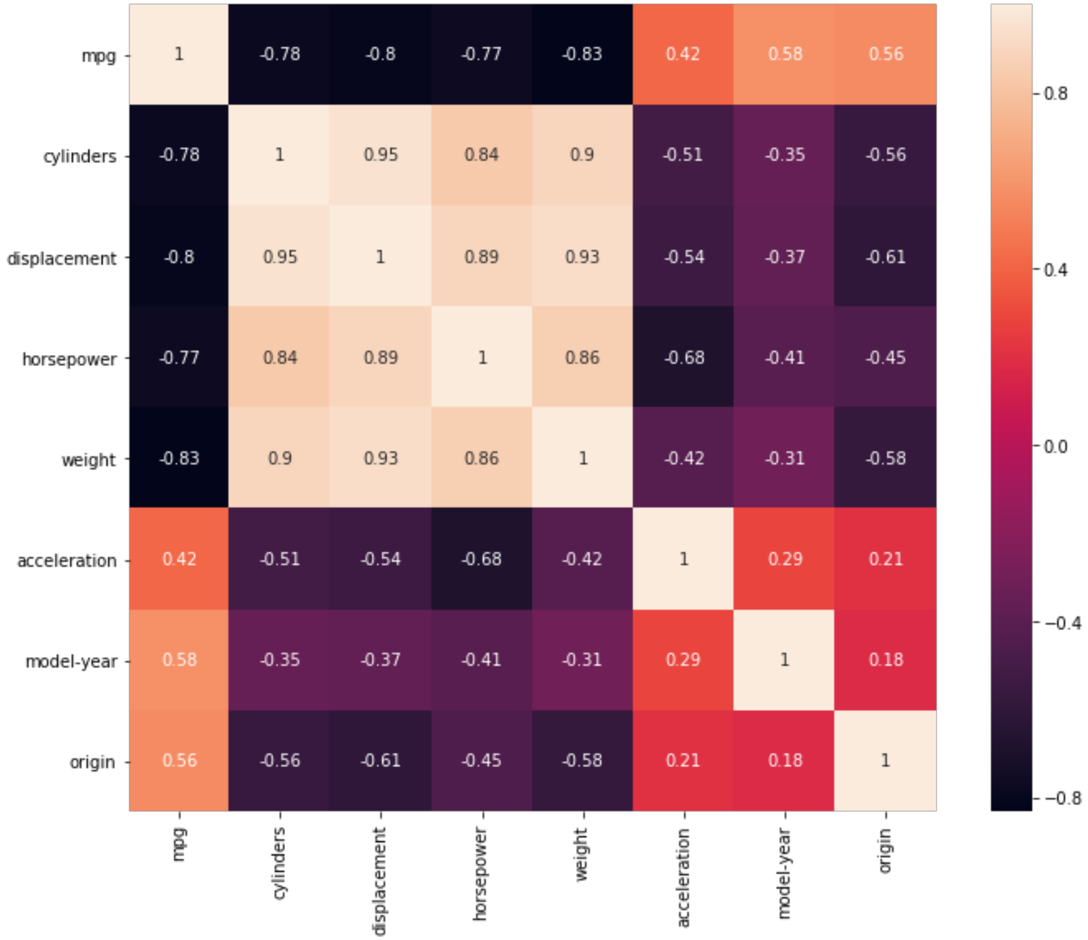
	mpg	cylinders	displacement	horsepower	weight	acceleration	model-year	origin	car-name
0	18.0	8	307.0	130.0	3504.0	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165.0	3693.0	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150.0	3436.0	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150.0	3433.0	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140.0	3449.0	10.5	70	1	ford torino



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 398 entries, 0 to 397
Data columns (total 9 columns):
mpg             398 non-null float64
cylinders       398 non-null int64
displacement    398 non-null float64
horsepower      392 non-null object
weight          398 non-null float64
acceleration    398 non-null float64
model-year      398 non-null int64
origin          398 non-null int64
car-name        398 non-null object
dtypes: float64(4), int64(3), object(2)
memory usage: 28.1+ KB
```

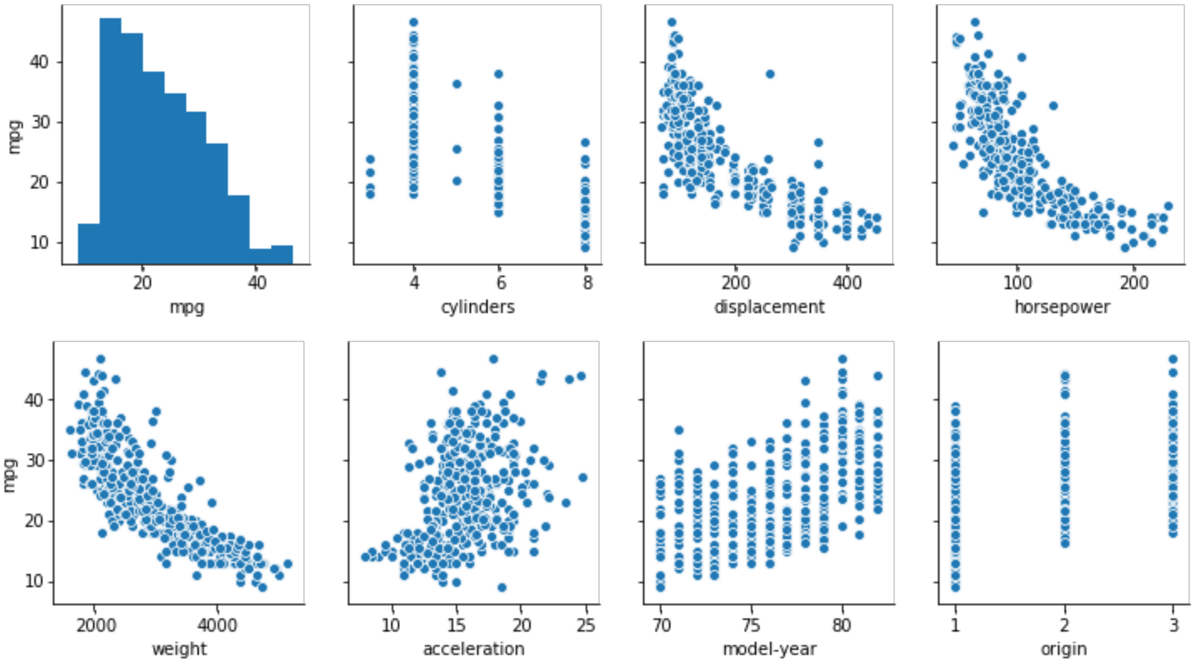
- There are 6 missing value for horsepower column, in this case i’m using mean to replace the missing value

Data Exploration (1)



- I’m pretty straight forward for this step, i jumped in to plot heatmap based on attributes correlation.
- From this heatmap, we can conclude that cylinders, displacement, horsepower, and weight have strong negative correlation to MPG

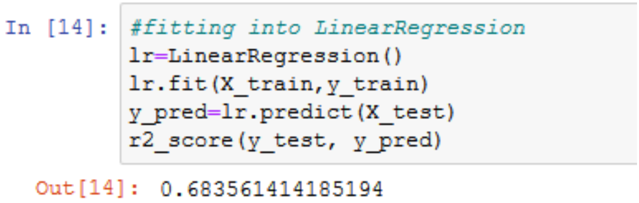
Data Exploration (2)



- Then i pairplot the data to visualize the correlation between each attributes.
- I highlighted all attributes correlation to mpg
- From this plot, it is strengthened my conclusion about cylinders, displacement, horsepower, and weight correlation to mpg
- Based on this plot, i will make machine learning model with only 4 attributes above

Modelling

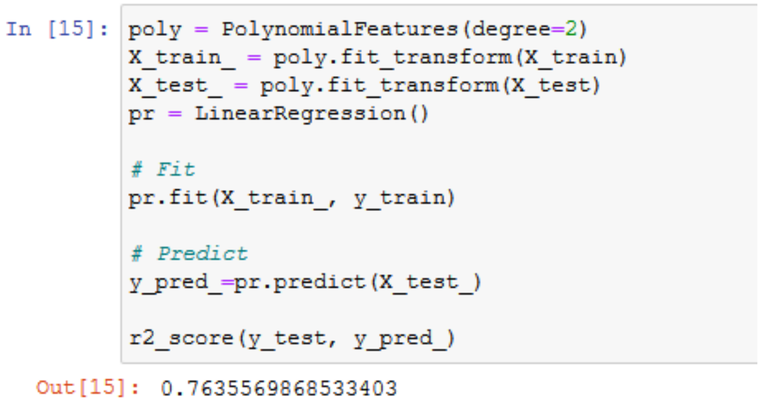
- I splitted the data to 80% of train data and 20% of test data.
- I choose 2 simple model of regression :
  - Linear Regression
  - Polynomial Regression
- Those 2 models from scikit-learn library
- Evaluation for this 2 models using R-squared method, and here the result :



```
In [14]: #fitting into LinearRegression
         lr=LinearRegression()
         lr.fit(X_train,y_train)
         y_pred=lr.predict(X_test)
         r2_score(y_test, y_pred)

Out[14]: 0.683561414185194
```

Linear Regression



```
In [15]: poly = PolynomialFeatures(degree=2)
         X_train_ = poly.fit_transform(X_train)
         X_test_ = poly.fit_transform(X_test)
         pr = LinearRegression()

         # Fit
         pr.fit(X_train_, y_train)

         # Predict
         y_pred_=pr.predict(X_test_)

         r2_score(y_test, y_pred_)

Out[15]: 0.7635569868533403
```

Polynomial Regression

Conclusion

- MPG values have strong correlation to the cylinders, displacement, horsepower, and weight values.
- Polynomial Regression is the best model so far with R-squared score : 0.76