# Data Mining Group Report

第七組 陳子新 洪小媛 ## Dataset Source:<https://www.kaggle.com/netzuel/pokmon-go-dataset-15-generations>

## Data Preprocessing

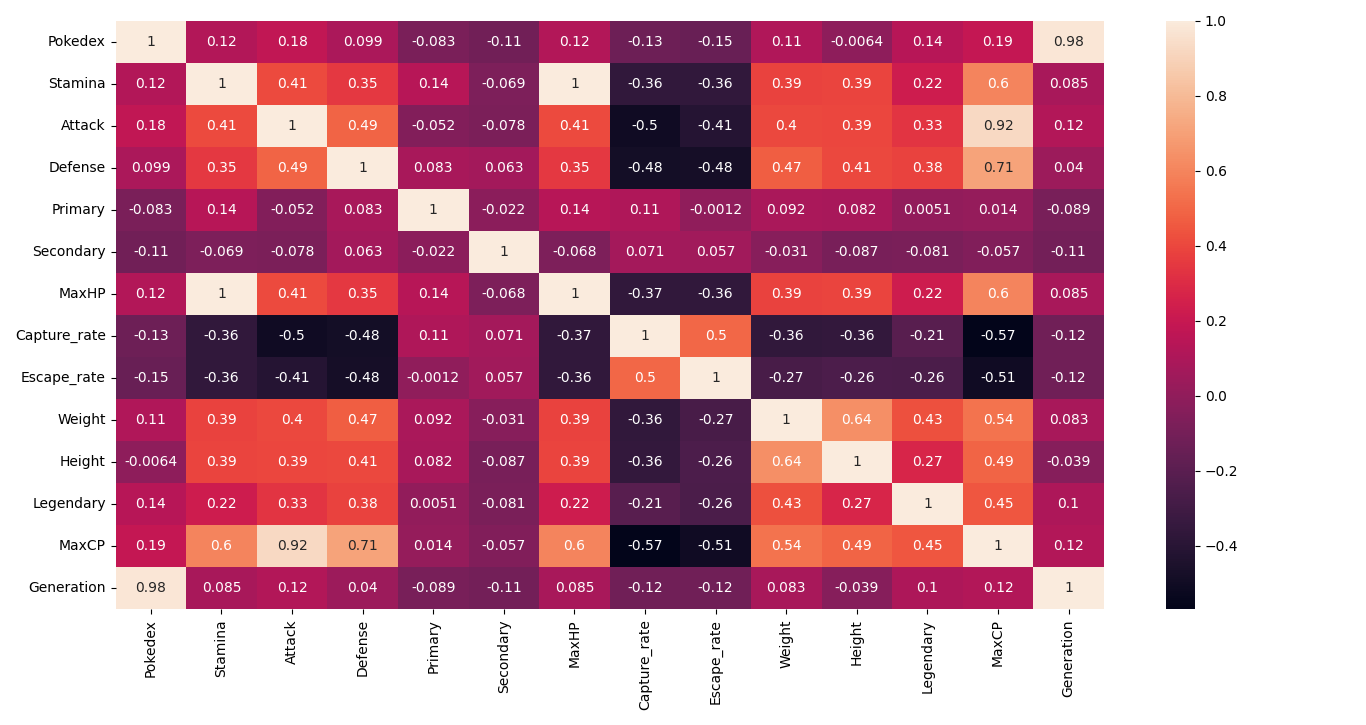
決定取用的feature前，先身對features進行correlation的分析，由於Primary、Secondary和Lengendary的類型皆為string，因此使用以下function轉換為數字的表示法:

# Convert string attributes to categorical attributes  
df['Primary'] = df['Primary'].astype('category').cat.codes  
df['Secondary'] = df['Secondary'].astype('category').cat.codes  
df['Legendary'] = df['Legendary'].astype('category').cat.codes

接著使用seanborn的heatmap畫出correlation matrix:

corr\_matrix = df.corr()  
sn.heatmap(corr\_matrix, annot=True)  
plt.show()

### Correlation Matrix



### Feature Selection

選擇和Capture\_ratecorrelation絕對值大於等於0.4的features Attack|Defense|Escape\_rate|MaxCP —:|:—:|:—:|:— -0.501|-0.484|0.499|-0.569

X = df['Attack','Defense','Escape\_rate','MaxCP']

但是後來發現這樣跑出來的accuracy不是很好，於是我們嘗試把correlation的threshold降到0.2，基本上除了一些nominal和ordinal(correlation不高)的feature外，其他9個feature都放進去了

X = df.drop(['Name','Pokedex','Primary','Secondary','Legendary','Capture\_rate'], axis='columns')

最後跑出來的accuracy如下，選擇9個feature的效果比較好，推測是這樣才能提供足夠多的資訊，因此我們選擇9個feature

# Accuracy with only 4 features  
0.8769230769230769  
# Accuracy with 9 features  
0.9076923076923077

## Target

我們希望預測的資料為Capture\_rate，由於它本身是連續的資料，因此需要做discretization，經過實驗後，我們發現使用equal width的分數比equal frequency高出很多，推測應該是因為Capture\_rate不是常態分佈，而equal width比較能呈現原本的資料分佈，因此我們選擇使用**equal width**的方式去做discretization，

以下為分別使用equal width和equal frequency跑出的accuracy:

# Accuracy when using equal width  
0.9076923076923077  
# Accuracy when using equal frequency  
0.6

### 分析

* use equal width instead of equal frequency on Capture\_rate,因為capture\_rate的分佈不平均，若用equal frequency會失去原本的分佈
* feature用全部的效果比較好，應該是因為這樣能提供的資訊比較多

## 結果比較

* 子新
  + Decision Tree
  + 0.9076923076923077
* 小媛
  + SVM