Mushroom Classification

-dead or alive-

Team Cerberos

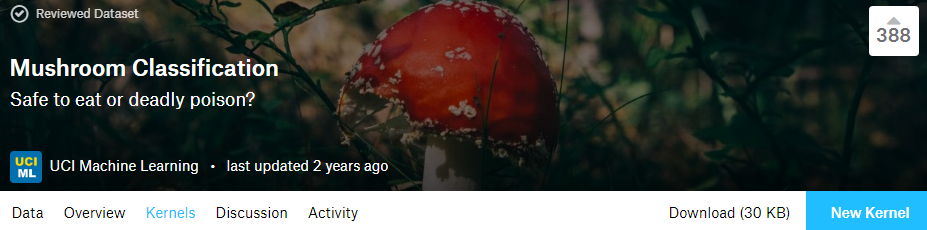
2018-06-26

1. Introduction

본격적인 프로젝트를 진행하기에 앞서 Kaggle의 Dataset을 사용해 자료의 Basic Description을 분석한 뒤 Github에 Commit하는 과정을 통해 Data의 Format 및 Github에 익숙해지는 시간을 가져보았다.

1. Data Used

Kaggle의 Dataset 중 ‘Mushroom Classification’을 사용했다. 해당 Dataset은 UCI에서 Machine Learning 교육을 위해 제공한 자료이며, 가상의 버섯 8000여개를 23개의 feature로 labeling한 text data이다.



▲Mushroom Classification Dataset

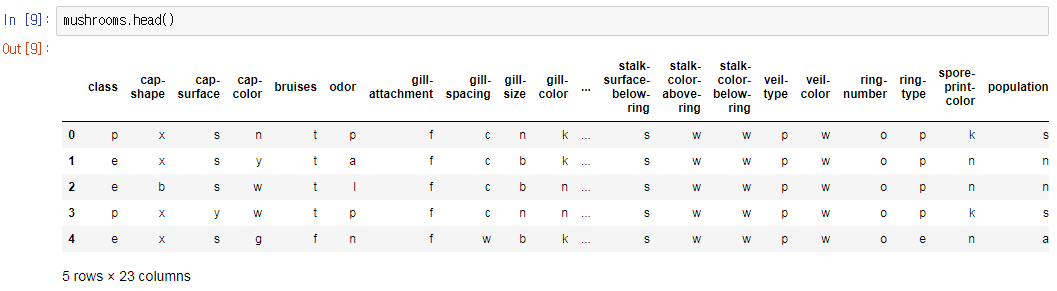
1. Methods

|  |  |
| --- | --- |
| 언어 | Python 3 |
| 환경 | Jupyter Notebook |
| 라이브러리 | Numpy, Pandas, Matplotlib |

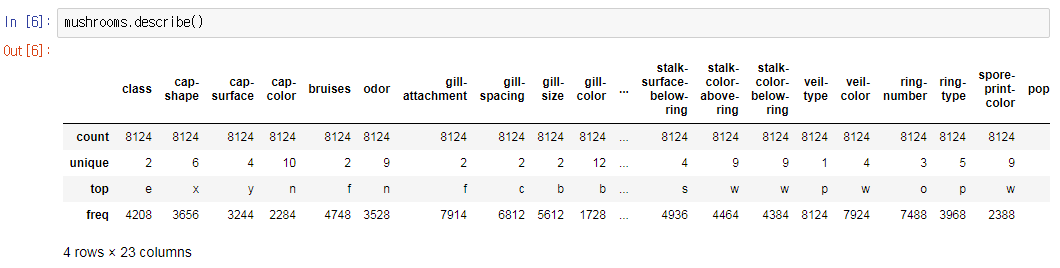
1. Results
2. **Dataset Format Analysis**
3. shape : 총 8124개의 버섯이 23개의 feature들로 labeling되어 있다.



1. head() : 각 데이터가 23개의 feature들로 구분되어 있고 각 feature는 단일값을 가진다.

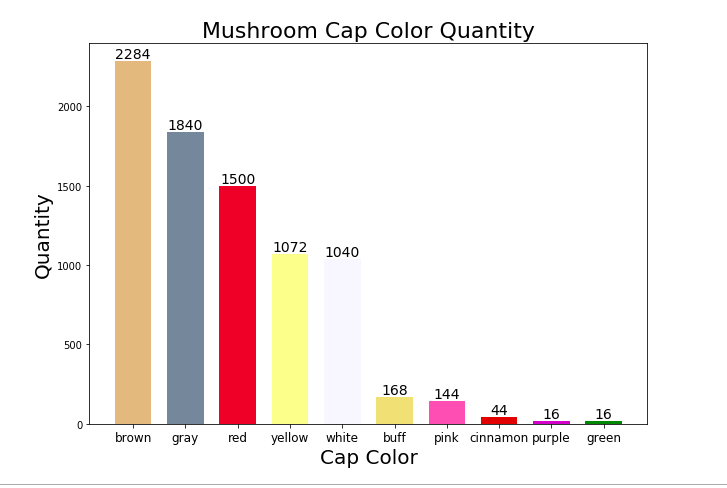


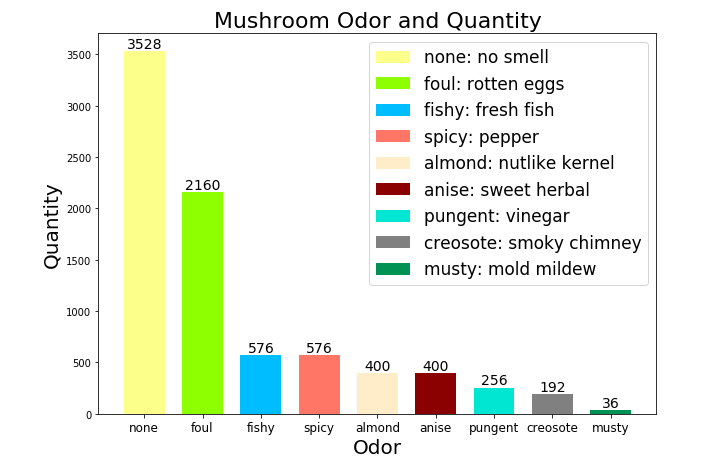
1. describe() : describe 명령어를 통해 각 feature들이 몇 종류의 값으로 구분되는지, 가장 많이 나타난 값은 무엇인지, 그리고 가장 많이 나타난 값의 횟수가 몇인지 확인할 수 있다.



1. **Histogram**

대표적으로 cap-color와 odor를 Histogram으로 시각화했다.

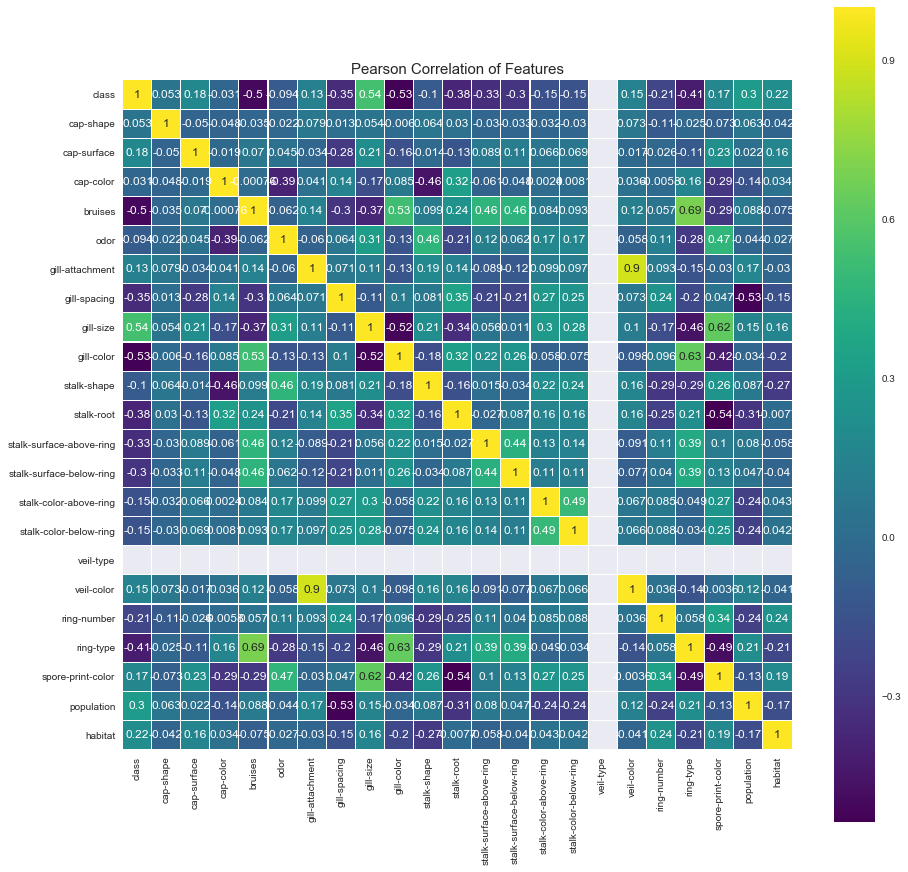




1. **Correlation**

23개의 Feature간의 상관관계를 Correlation Matrix를 이용해 표현했다. 파란색 계열은 관계의 정도가 0에 가깝기 때문에 서로 거의 영향을 미치지 않고, 노란색 및 보라색 계열은 각각 양과 음의 상관관계를 가진다는 의미이기 때문에 서로 많은 영향을 미친다는 것을 시각적으로 한 눈에 볼 수 있다.

양의 상관관계는 한 feature가 높아지면 해당 feature도 높아지는 관계이며, 음의 상관관계는 한 feature가 높아지면 해당 feature는 반대로 낮아지는 관계임을 의미한다.



1. Conclusions

오늘 한 일 : 버섯 데이터를 Jupyter Notebook에 올리고 Pandas와 Matplotlib를 이용해 데이터의 모양과 특성을 여러 가지 방법으로 분석했음

내일 할 일 : 버섯 데이터를 학습시켜 임의의 특성을 가진 버섯이 식용 버섯인지 독버섯인지 가려내는 프로그램을 작성