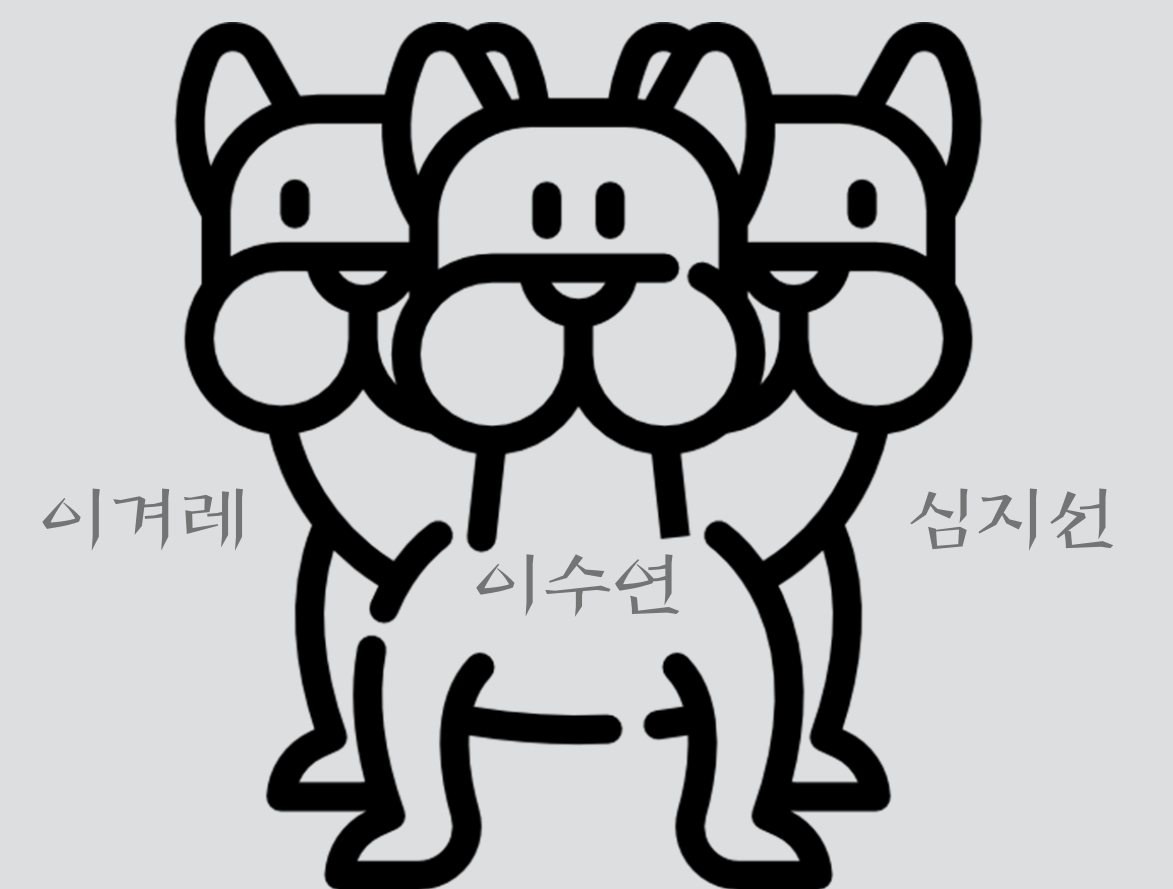


Cerberus





Mushroom Classification

데이터에 기반하여 특정 버섯이 독버섯인지
아닌지를 예측하고 실제 결과와 대비해본다

1. Introduction

- (Week 1) Objectives

- ① Kaggle의 Dataset을 가져와 자료의 Basic Description 분석 및 시각화
- ② Github에 Commit하는 과정을 통한 기본 사용방법 익히기

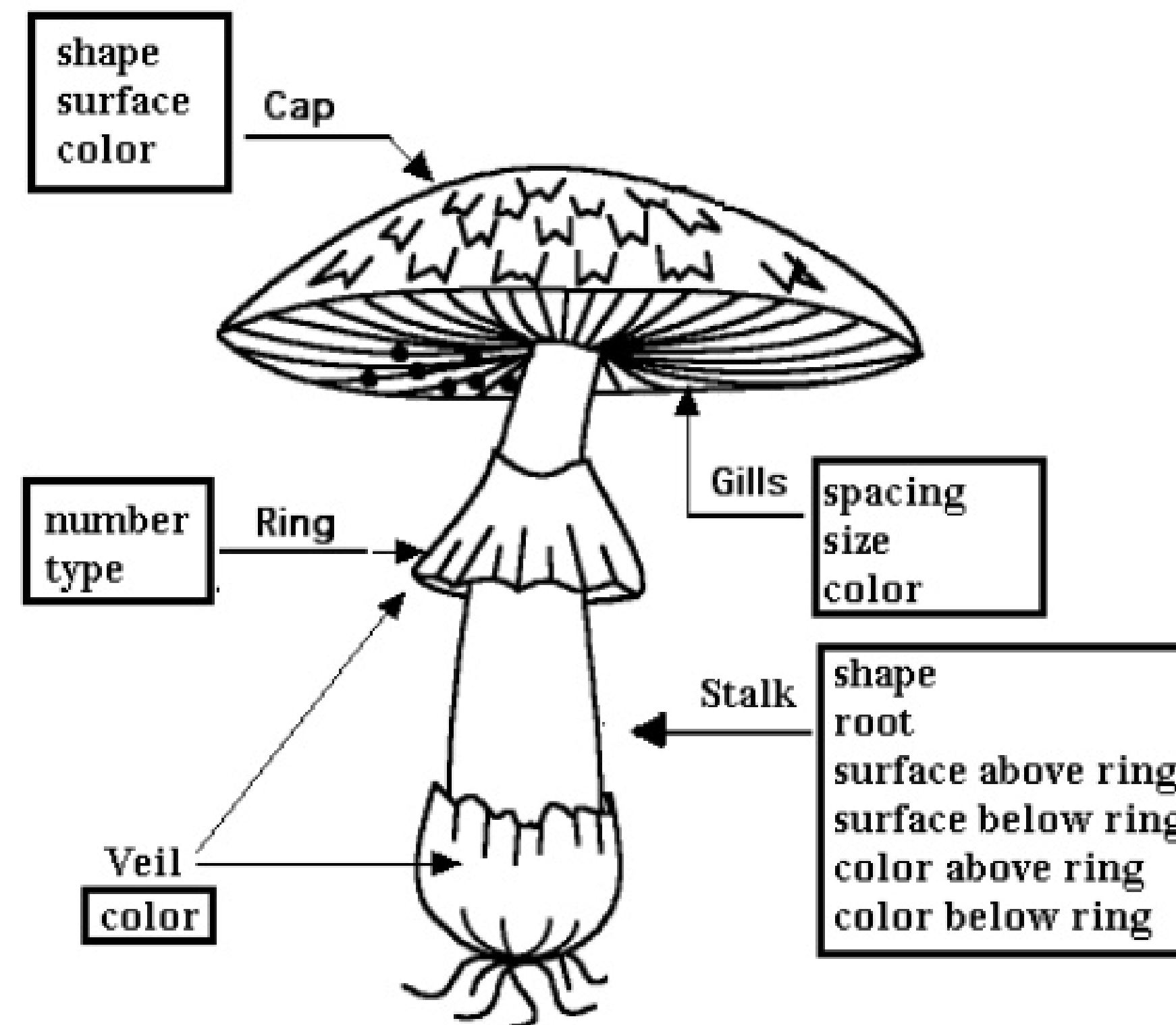
1. Introduction

- Domain Knowledge

15 Attributes

5 Sections:

1. Cap
2. Ring
3. Veil
4. Gills
5. Stalk




2. Data Used

'Mushroom Classification'

- UCI에서 Machine Learning 교육을 위해 제공한 자료
- 가상의 버섯 8000여개를 23개의 feature로 labeling한 text data

Mushroom Classification Dataset ▶



Reviewed Dataset

Mushroom Classification

Safe to eat or deadly poison?

UCI ML UCI Machine Learning • last updated 2 years ago







Data Overview **Kernels** Discussion Activity

Download (30 KB) [New Kernel](#)

388


2. Data Used

- 버섯 하나 당 23개의 Feature로 이루어져 있음
- 각 Feature는 알파벳 글자 하나로 Labeling 되어져 있음

mushrooms.csv (365.24 KB)					
20 of 23 columns					
	 class	 cap-shape	 cap-surface	 cap-color	 bruises
	edible=e, poisonous=p	bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s	fibrous=f, grooves=g, scaly=y, smooth=s	brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y	bruises=t, no=f
	<div>e52%</div> <div>p48%</div> <div>Other (0)0%</div>	<div>x45%</div> <div>f39%</div> <div>Other (4)16%</div>	<div>y40%</div> <div>s31%</div> <div>Other (2)29%</div>	<div>n28%</div> <div>g23%</div> <div>Other (8)49%</div>	<div>f58%</div> <div>t42%</div> <div>Other (0)0%</div>

2. Data Used

- 'class'라는 feature는 해당 버섯이 식용버섯인지 독버섯인지를 나타냄
- 우리의 목표는 버섯의 특징을 입력 받아 식용인지 아닌지 판단하는 **Binary Classifier**를 **신경망**으로 구현하는 것
- 따라서 'class'를 신경망의 Y값으로 잡아 **Supervised Learning**을 시킬 예정

A class 
edible=e, poisonous=p

e	52%
p	48%
Other (0)	0%

2. Data Used

- 보통 대략 70%의 Dataset을 신경망의 Training에, 30%를 Test에 활용함
- 따라서 총 8124개의 버섯 데이터 중 **6124개를 Training**에, **2000개를 Test**에 사용할 예정

3. Methods

- 언어 : Python3
- 환경 : Jupyter Notebook
- 라이브러리
 - Pandas, Matplotlib : 데이터 가공 및 시각화
 - Tensorflow, Numpy : 신경망 구현



4. Results

(1) Dataset Format Analysis

- **shape** : 8124 X 23 꼴의 2차원 Matrix 형식 -> 총 8124개의 버섯이 23개의 feature들로 labeling 되어있음

```
In [10]: mushrooms.shape
```

```
Out [10]: (8124, 23)
```

- **head()** : head는 가장 앞의 5개 row를 제시, 각 데이터가 23개의 feature들로 구분됨을 알 수 있음

In [9]: mushrooms.head()

Out [9]:

	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color	population
0	p	x	s	n	t	p	f	c	n	k	...	s	w	w	p	w	o	p	k	s
1	e	x	s	y	t	a	f	c	b	k	...	s	w	w	p	w	o	p	n	n
2	e	b	s	w	t	l	f	c	b	n	...	s	w	w	p	w	o	p	n	n
3	p	x	y	w	t	p	f	c	n	n	...	s	w	w	p	w	o	p	k	s
4	e	x	s	g	f	n	f	w	b	k	...	s	w	w	p	w	o	e	n	a

5 rows x 23 columns

- **describe()** : 각 feature들이 몇 종류의 값으로 구분 되는지와 그 횟수를 알 수 있음

In [6]: mushrooms.describe()

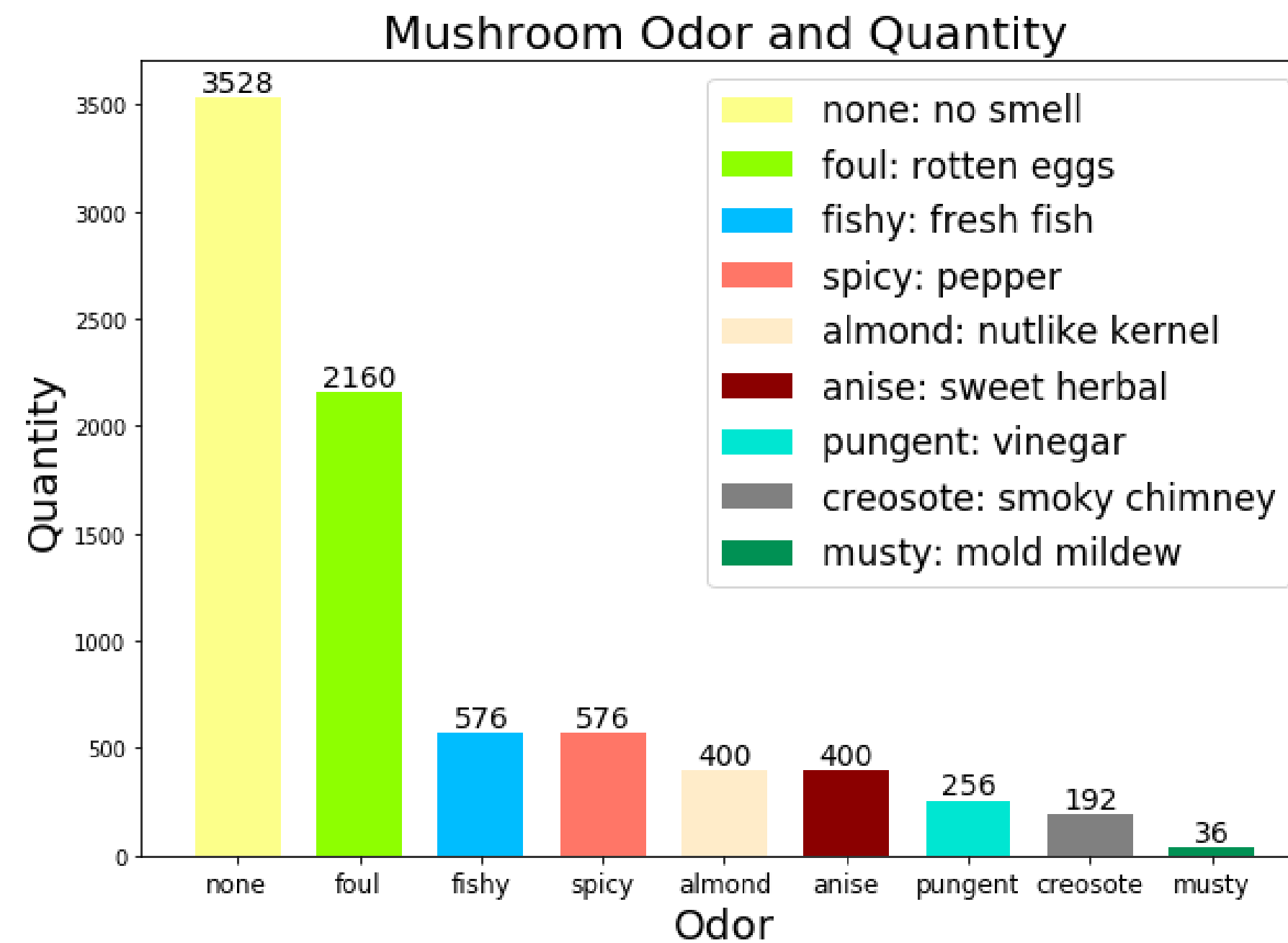
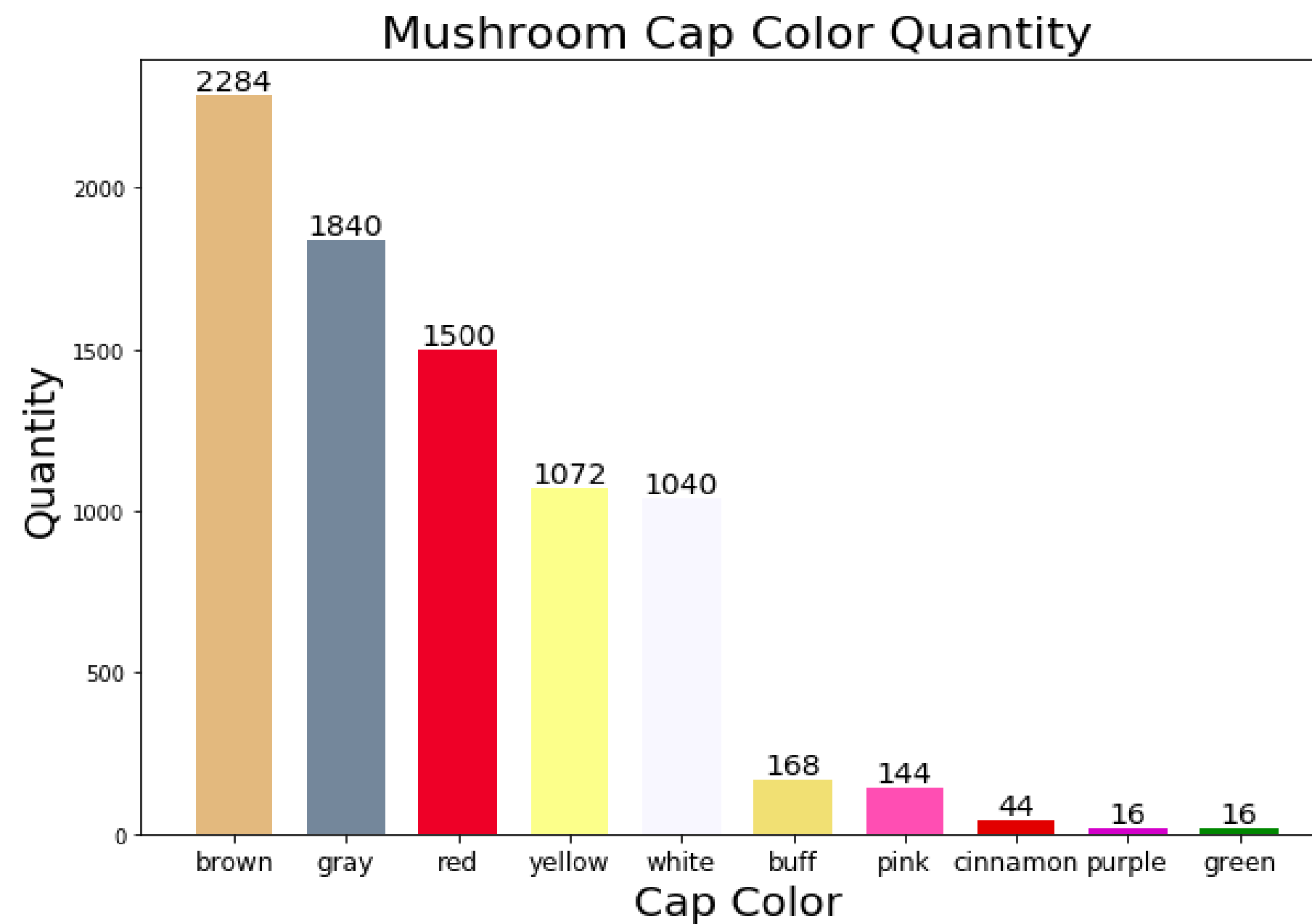
Out [6]:

	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color	pop
count	8124	8124	8124	8124	8124	8124	8124	8124	8124	8124	...	8124	8124	8124	8124	8124	8124	8124	8124	
unique	2	6	4	10	2	9	2	2	2	12	...	4	9	9	1	4	3	5	9	
top	e	x	y	n	f	n	f	c	b	b	...	s	w	w	p	w	o	p	w	
freq	4208	3656	3244	2284	4748	3528	7914	6812	5612	1728	...	4936	4464	4384	8124	7924	7488	3968	2388	

4 rows x 23 columns

(2) Histogram

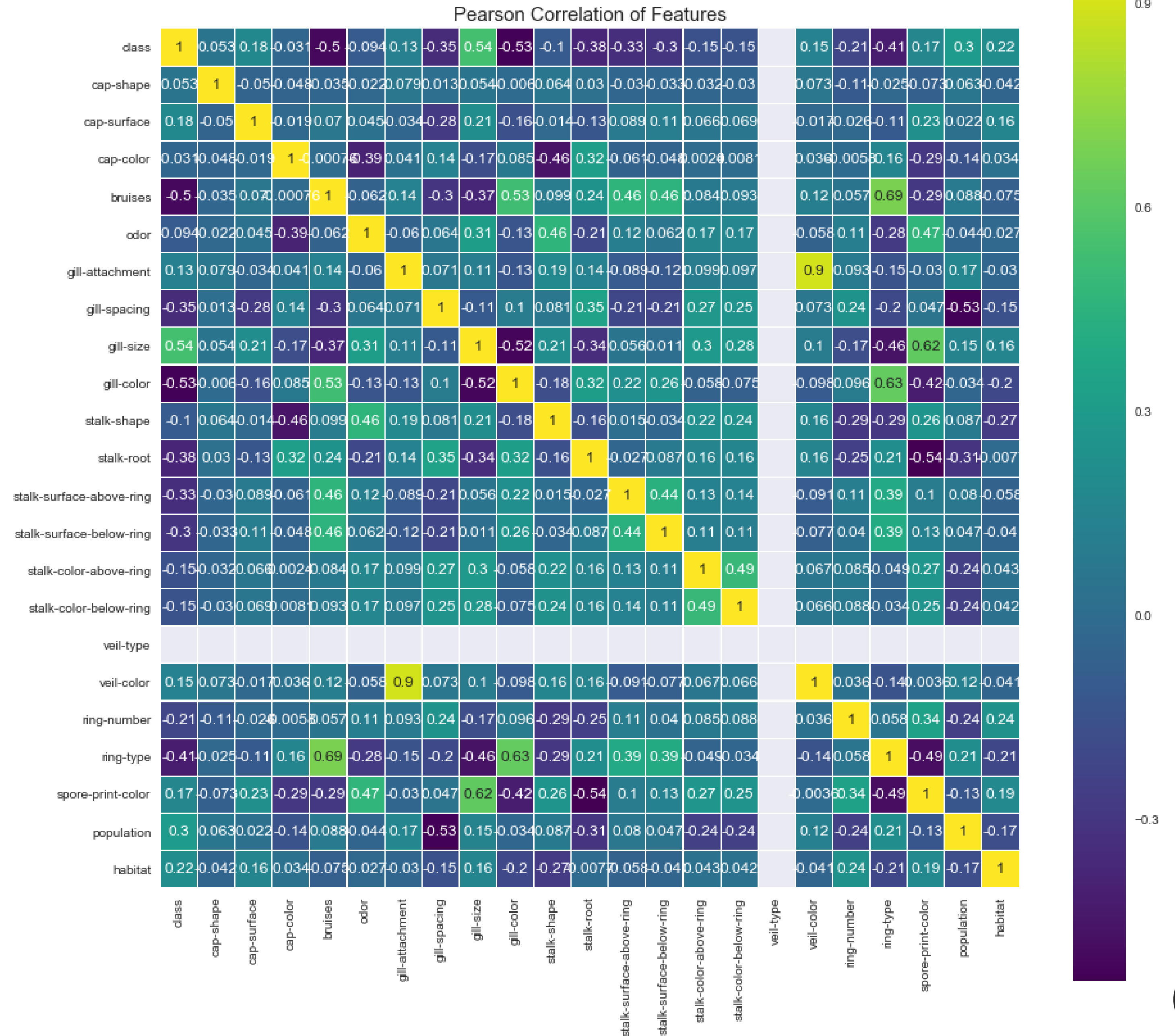
- 대표적으로 cap-color와 odor를 선택해 Histogram으로 시각화
- 해당 feature들이 class 수가 가장 많았기 때문에 시각화하는 것이 효과적이라고 판단하였음



(3) Colleration Matrix



- 각 데이터 간의 feature별 상관관계를 Matrix로 시각화
- Veil-color와 Gill-Attachment는 0.9라는 높은 상관관계를 가짐
- Veil-type은 모든 버섯이 p라는 변하지 않는 값을 갖기 때문에 상관관계를 알 수 없어 회색으로 표현됨





Conclusion

Work Done : Pandas와 Matplotlib를 이용해 버섯 데이터의 모양
과 특성을 여러 가지 방법으로 분석 + Github에 올림

To Do : 머신러닝을 통해 임의의 특성을 가진 버섯이 식용 버섯
인지 독버섯인지 가려내는 프로그램을 작성