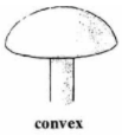
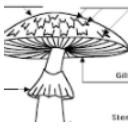
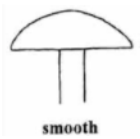
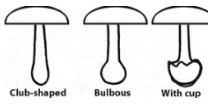
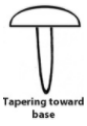
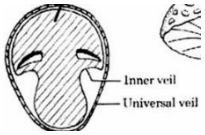
Python Final Project

Developer: **Opal Peltzman** – 208521385

Introduction

Attributes Information –

ID	Column	Description
0	Classes	Whether the mushroom is edible or poisonous. (**e** = edible), (**p** = poisonous)
1	Cap shape	Whether the mushroom cup shape is knobbed or sunken. (**k** = knobbed ), (**s** = sunken ) (**b**= bell ), (**c** = conical ) (**x** = convex ), (**f** = flat )
2	Cap surface	Whether the mushroom cup surface is fibrous, grooves, scaly or smooth. (**f** = fibrous ), (**g** = grooves ), (**y** = scaly ), (**s** = smooth )
3	Cap color	Whether the mushroom cup color is brown, buff, cinnamon, gray, green, pink, purple, red, white or yellow. (**n** = brown), (**b** = buff/pale ), (**c** = cinnamon ), (**g** = gray) , (**r** = green) , (**p** = pink) , (**u** = purple) , (**e** = red) , (**w** = white) , (**y** = yellow)
5	Odor	Measured mushroom odor level, units undefined.

6	Gill attachment	Represents the gill attachment to the mushroom stalk. (**a** = attached  or ) , (**d** = descending ) ,  (**f** = free -> not attached) , (**n** = notched -> smoothly notched and running briefly down stem )
7	Gill spacing	Gill = wide and thin sheet-like plates radiating from stem.  (**-1**= close ), (**0**= crowded ) , (**1**= distant )
10	Stalk shape	The shape of the stem. (**e** = enlarging ), (**t** = tapering )
16	Veil type	 In the CSV mushroom file there is <u>no Veil type column</u> (**u =** universal), (**p** = partial)
17	Veil color	(**n** = brown), (**o** = orange) , (**w** = white) , (**y** = yellow)
18	Ring number	Number of ring on the mushroom stalk. (**n** = none), (**o** = one) , (**t** = two)
21	Population	Measure of density of mushroom clusters
22	Latitude	Latitude of sample
23	Longitude	Longitude of sample

Feature analysis -

ID	Column	Categorical	Numerical	Nominal	Ordinal
0	Classes	V	-	V	-
1	Cap shape	V	-	V	-
2	Cap surface	V	-	V	-
3	Cap color	V	-	V	-
5	Odor	-	V	-	
6	Gill attachment	V	-	-	V
7	Gill spacing	V	-	-	V
10	Stalk shape	V	-	V	-
16	Veil type	V	-	V	-
17	Veil color	V	-	V	-
18	Ring number	V	-	-	V
21	Population	-	V	-	
22	Latitude	-	V	-	
23	Longitude	-	V	-	

Data size – number of data rows=8124 multiply the number of columns=13, total of **105612**.

(Excluding column ‘Veil type’, that contains null for every row in data set)

Initial data analysis

Feature analysis -

ID	Column	Categorical	Nominal	Ordinal	Explanation
0	Classes	V	V	-	There is no hierarchy between the two different attributes.
1	Cap shape	V	V	-	There is no hierarchy between the cap shapes.
2	Cap surface	V	V	-	There is no hierarchy between the cap surfaces.
3	Cap color	V	V	-	There is no hierarchy between the cap colors.
5	Odor	-	-		-
6	Gill attachment	V	-	V	**There are levels of the attachment to the mushroom stalk. We can refer those levels as an order of the strength of the attachment.** **The order - a. free, b. notched, c. attached, d. descending**
7	Gill spacing	V	-	V	**There are levels of the gill's density.** **The crowded type is more compact than close, crowded and close are more compact then distant.** **The order – a. crowded, b. close, c. distant**
10	Stalk shape	V	V	-	There is no hierarchy between the stalk shapes.
16	Veil type	V	V	-	There is no hierarchy between the Veil types.
17	Veil color	V	V	-	There is no hierarchy between the veil colors.
18	Ring number	V	-	V	**There is a numerical order of the possible counts of rings on the mushroom stalk.** **The order – a. 0, b. 1, c. 2**
21	Population	-	-		-
22	Latitude	-	-		-
23	Longitude	-	-		-

Feature descriptions -

	classes	cap_shape	cap_surface	cap_color	odor	gill_attachment
count	8124	8124	8124	8124	8124	7374
unique	2	6	4	10		2
top	e	X	y	n		f
freq	4208	3656	3244	2284		7184
mean					2.568453964	
std					4.739983902	
min					-13.02	
25%					-0.63	
50%					2.5	
75%					5.7325	
max					19.45	

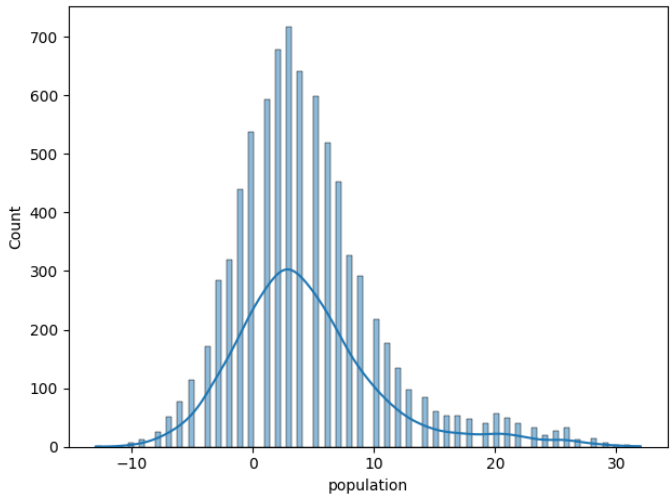
	gill_spacing	stalk_shape	veil_color	ring-number	population	latitude	longitude
count	8034	8124	8124	8038	8124	8124	8124
unique		2	4	3			
top		t	w	o			
freq		4608	7924	7408			
mean	-0.838063231				4.462210734	56.96074098	-98.2174377
std	0.368415725				6.199476561	5.301214152	1.018065831
min	-1				-13	10	-102.092822
25%	-1				0	57.38854412	-98.90204172
50%	-1				4	57.55979291	-98.21861757
75%	-1				7	57.71956383	-97.5282973
max	0				32	58.439199	-94.75088618

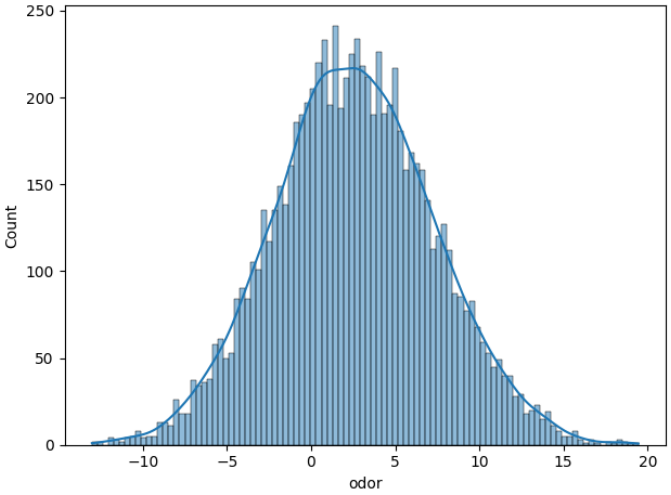
Dealing with missing data - with code at final_project.py file.

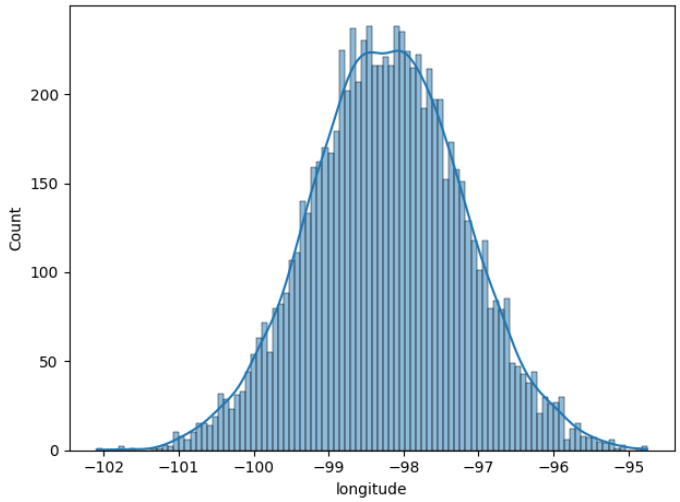
Save fixed data to CSV file - with code at final_project.py file. (File name = ‘fixed_data’)

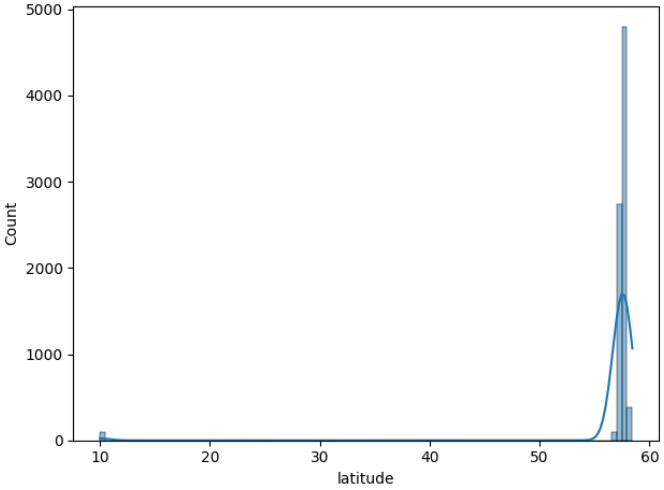
Exploratory data analysis

Histograms -

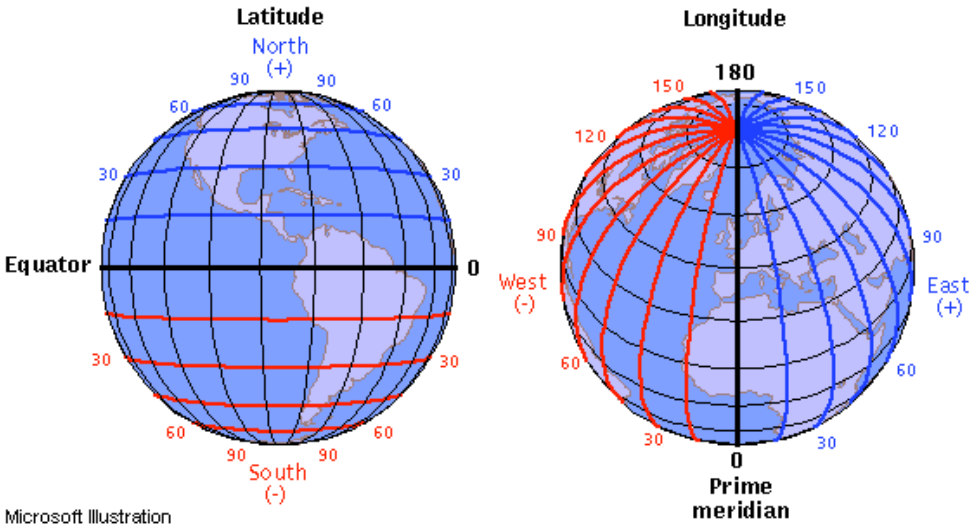








Understanding **Latitude** and **Longitude**:

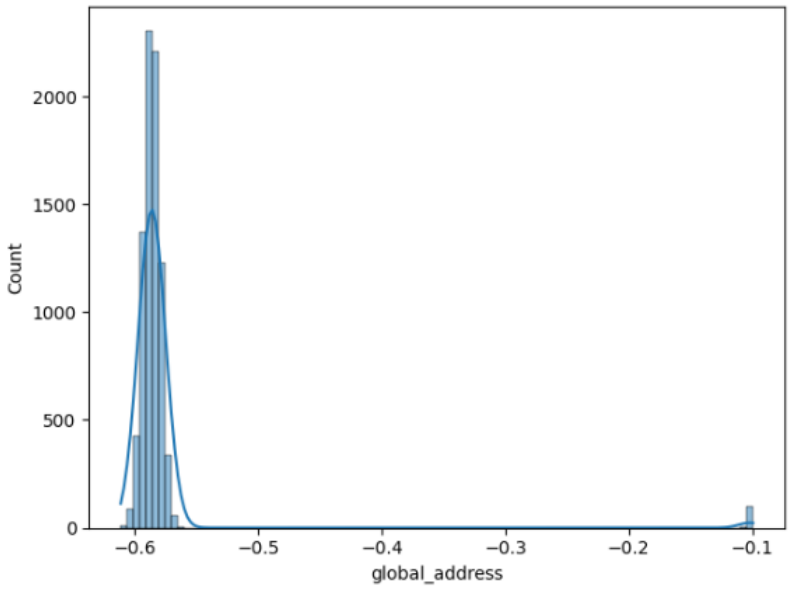


Credit: Illinois State University

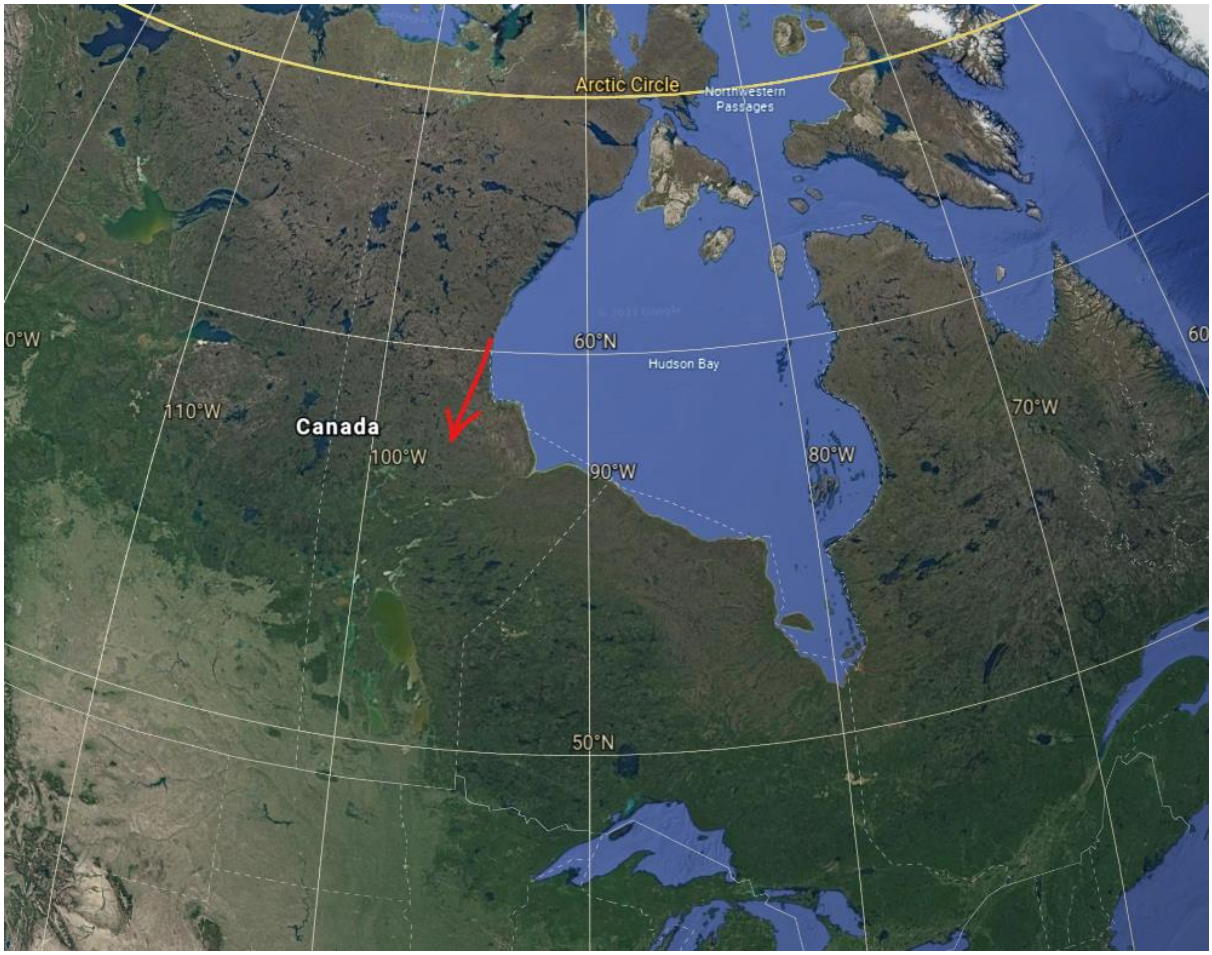
I created a new feature based on Latitude and Longitude.

The feature will be a global address that is given as Latitude divided by Longitude,

This will give the "intersection" where the place is located.

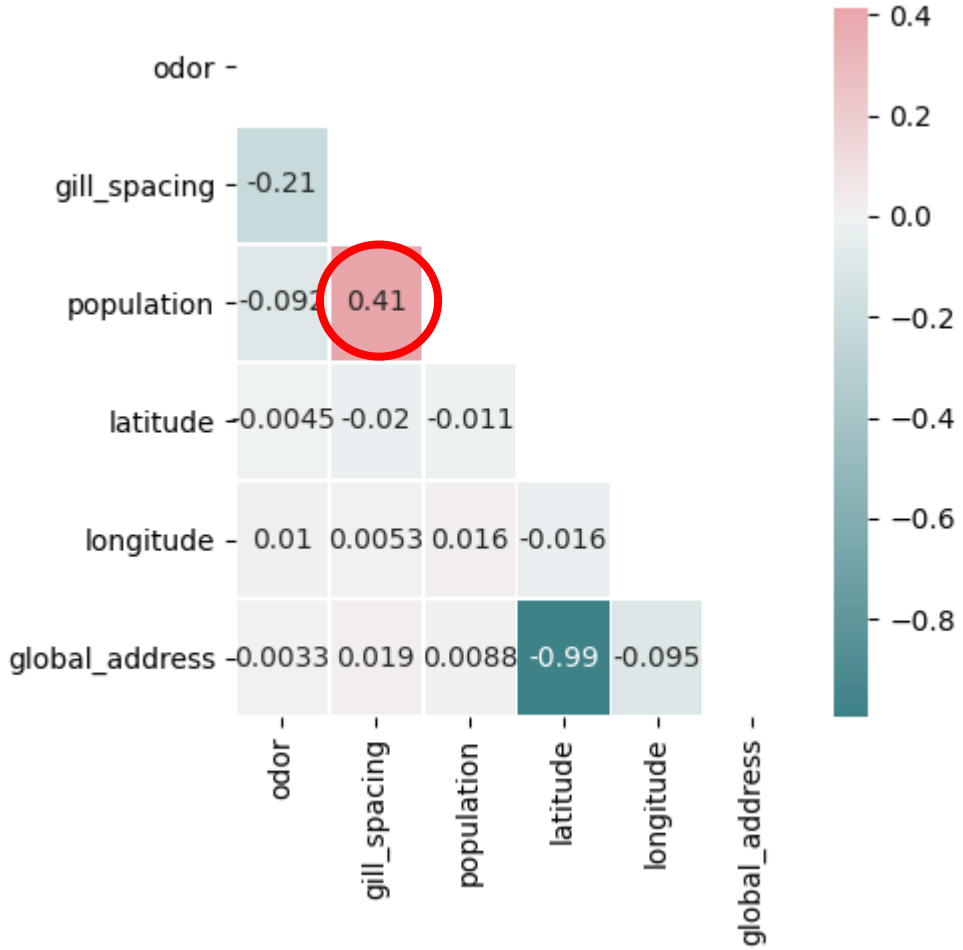


According to the new feature, we can see that most of our mushroom data is from Canada -

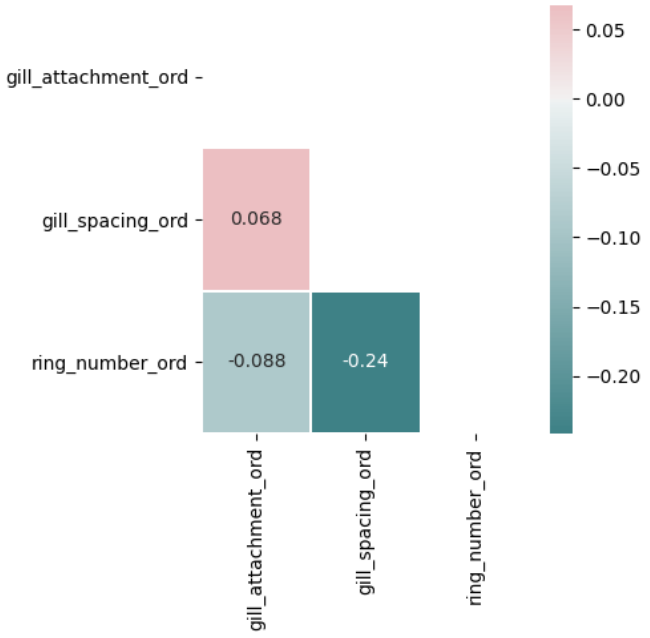

Arctic Circle
Northwestern Passages
0°W
60°N
Hudson Bay
110°W
Canada
100°W
90°W
80°W
70°W
50°N

Correlation heat map -

Numerical correlation heat map –

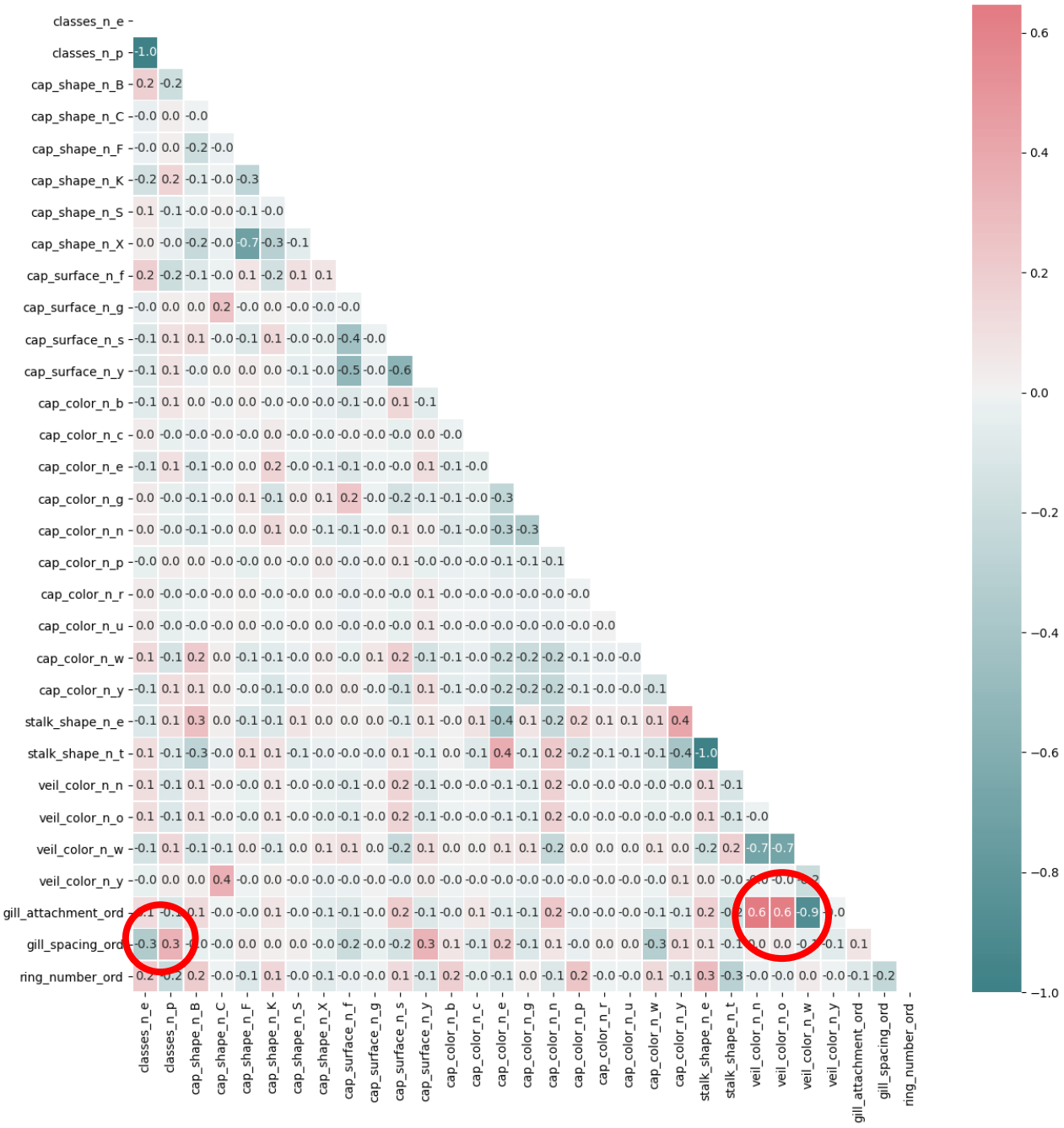


Ordinal features Correlation heat map –

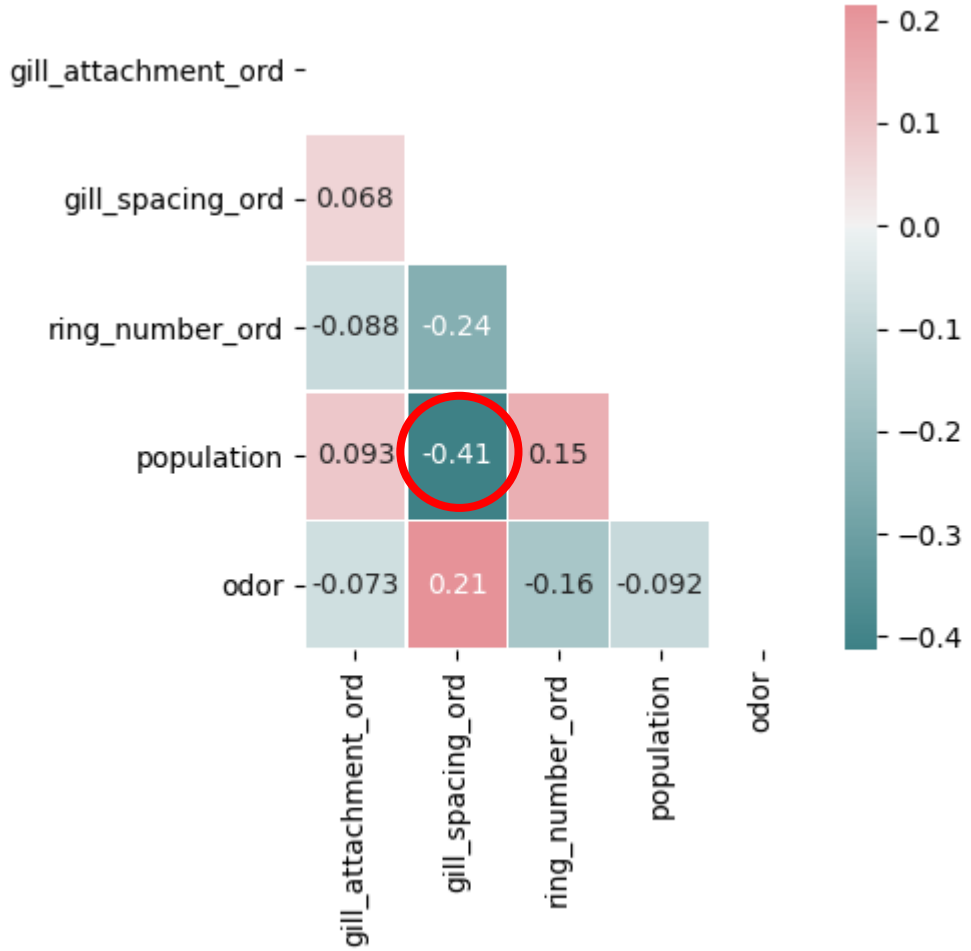


Nominal with **numerical** features Correlation heat map –

Ordinal with **nominal** features Correlation heat map –



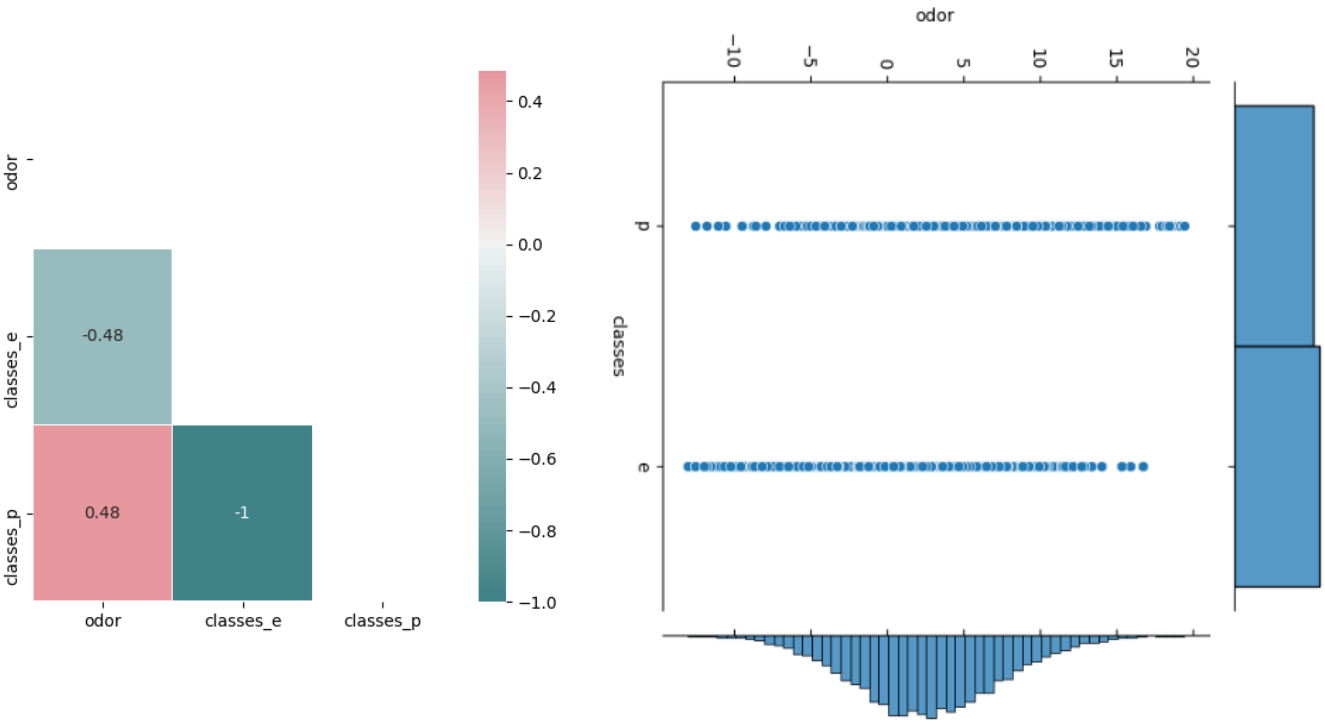
Ordinal with **numerical** features Correlation heat map –



Features -

<u>Classes</u> feature –

According to the nominal with numerical features correlation map, we saw correlation=0.5 between **classes** to **odor** and another (smaller) correlation=0.2 between **population** to **classes**.



Classes pivot table -

pivot table with numerical features -

classes	e	p
gill_spacing	-0.716968	-0.971910
global_address	-0.579921	-0.580082
latitude	56.946471	56.976075
longitude	-98.205595	-98.230163
odor	0.351692	4.950511
population	5.820342	3.002809

pivot table with nominal features -

classes	e	p
cap_colors	1	1
cap_shapes	1	1
cap_surfaces	3	2
stalk_shape_bin	0	0
veil_colors	1	1

pivot table with ordinal features -

classes	e	p
gill_attachment_ord	1	1
gill_spacing_ord	2	2
ring_number_ord	2	2

We can see that when the **odor** is > 15, the chances for the mushroom to be poisonous are higher.

The mean **odor** for poisonous mushrooms is higher than edible mushrooms.

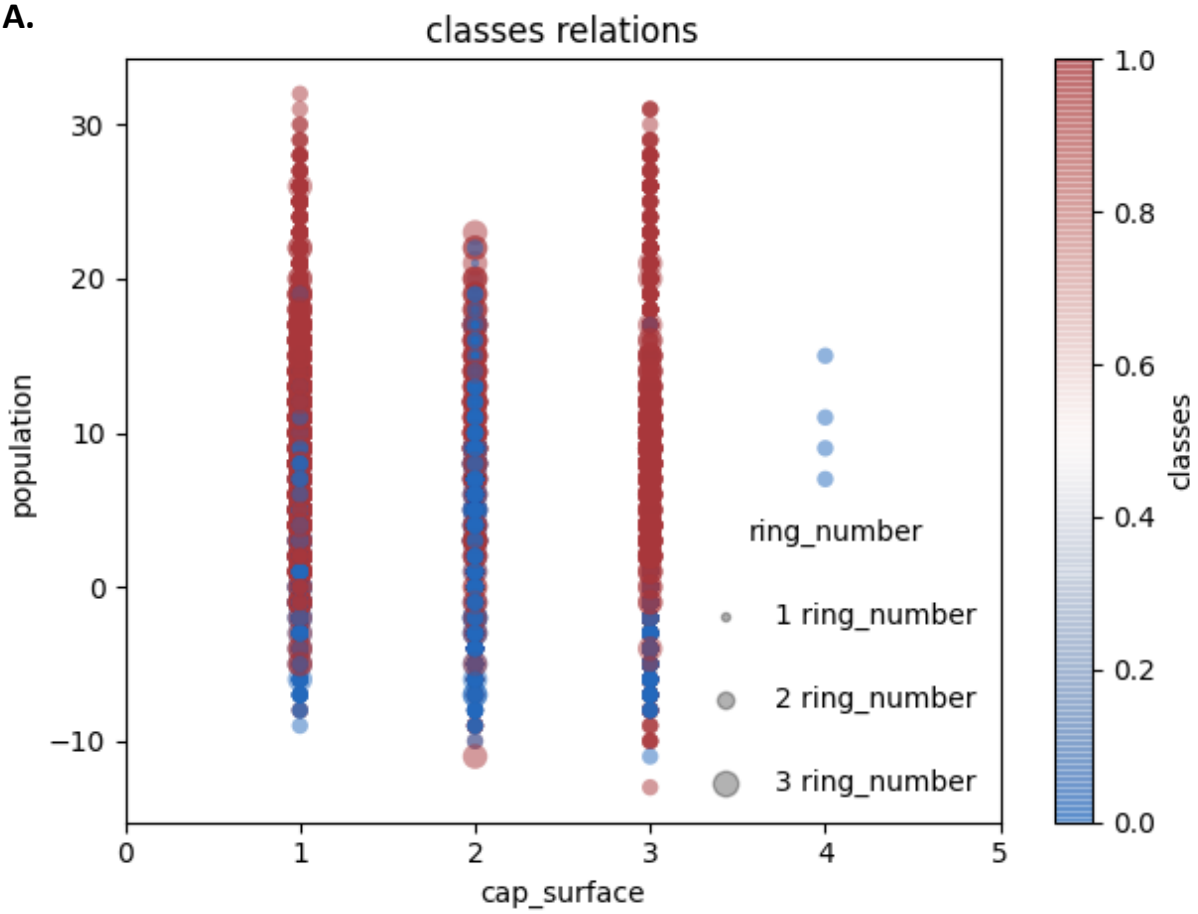
In addition, the mean **density** of mushroom clusters is higher when the mushrooms are edible.

According to the nominal features, we can see that the most frequent **cap surface** for edible mushrooms is 3= f= **fibrous** and for poisonous is 2=y= **scaly**.

According to the ordinal feature, edible and poisonous mushrooms share the same frequency values for each ordinal features.

<u>Exploring features with different values in classes pivot tables</u> -

A.



B.

classes	ring_number_ord	
e	2	3685
	3	523
p	2	3809
	3	72
	1	35

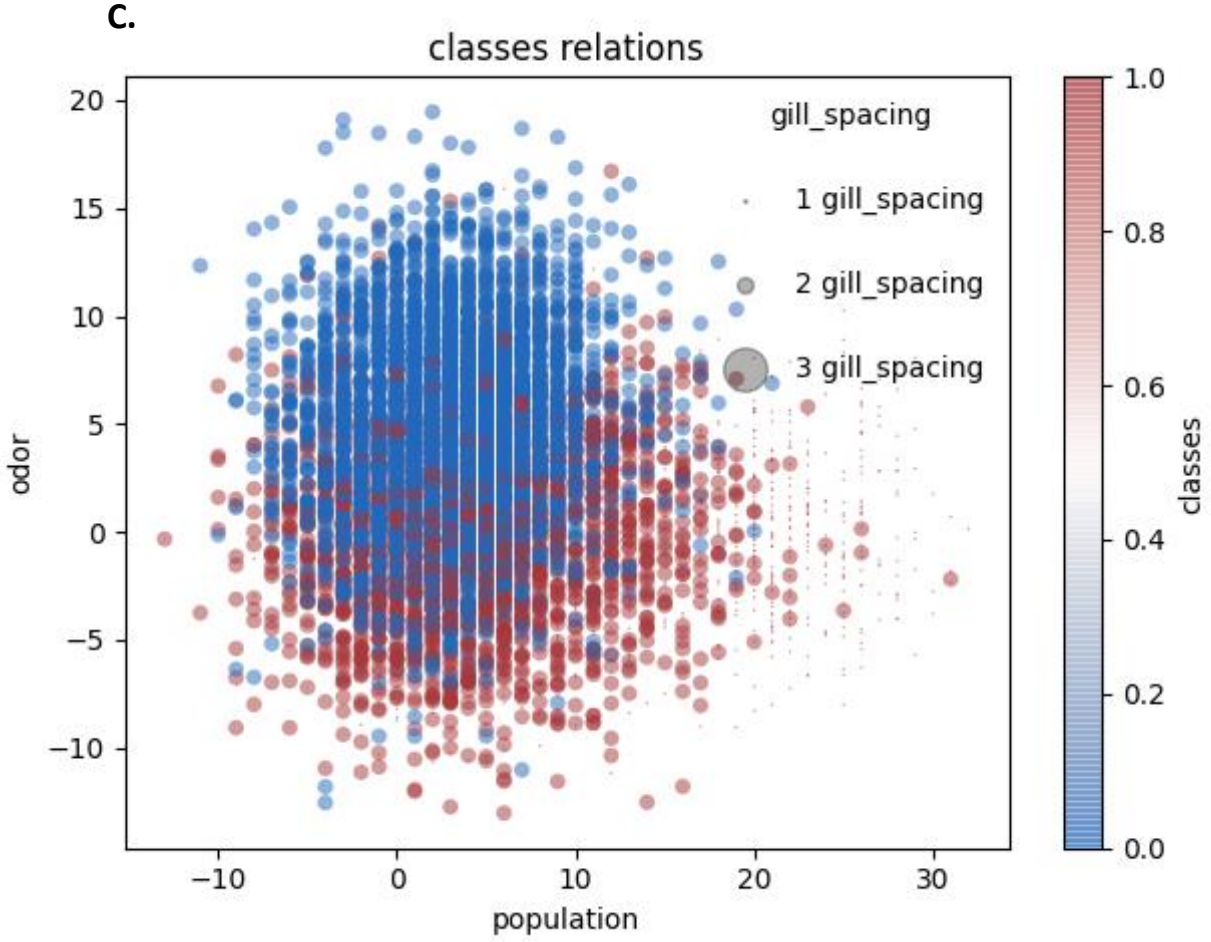
classes	cap_surfaces	
e	3	1560
	2	1504
	1	1144
p	2	1740
	1	1412
	3	760
	4	4

From A and B we can see that most poisonous mushrooms have scaly = 2= y, cap surface.

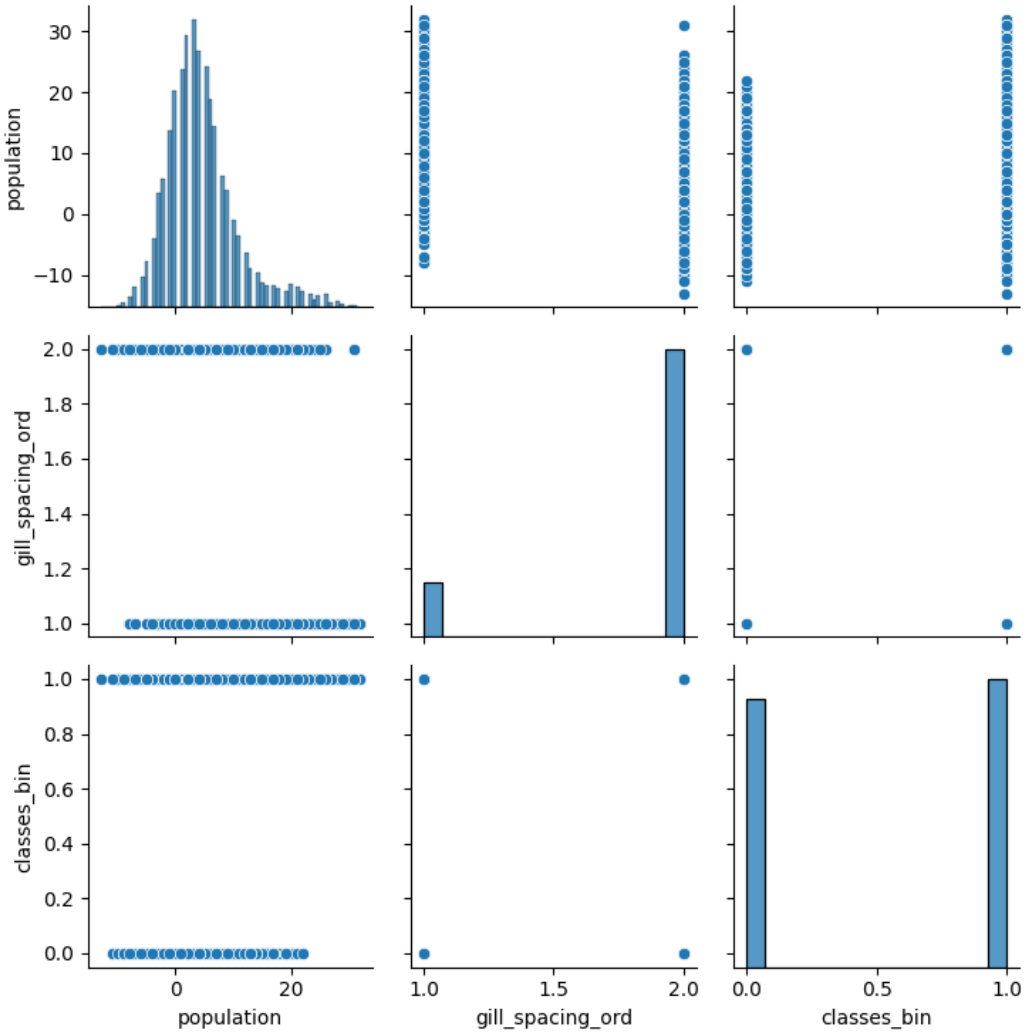
The most common cap surface for edible mushrooms is fibrous = 3= f (which is not very common for poisonous mushrooms).

In addition, edible mushrooms more likely to have <u>two (=3)</u> rings.

C.



D.



E.

classes	gill_spacing_ord	
e	2	3017
	1	1191
p	2	3806
	1	110

From the two graphs (C&D) we can see that most of the data has gill spacing = 2 = close.

The majority of mushrooms with the population that is higher than 20 (= higher density) are edible mushrooms = 1.

The edible mushrooms appear to have more gill spacing (= 1 = crowded) than poisonous mushroom.

Moreover, the majority of poisonous mushrooms (= 0) have density that is lower than 10.

Classes' relation summary –

Important features:

Feature:	E=edible		P=poisonous
	Most frequent:		
Cap surfaces	3 = f = fibrous -> are more likely to be edible. (value 2-> 1503) (value 3-> 1560)		2 = y = scaly -> When scaly it is not for certain poisonous. (value 2-> 1740) (value 3-> 760)
Gill spacing	2,(1191 from value) 1 When value=1, it is more likely to be edible.		2, (110 from value) 1
Ring numbers	One ring, When there are <u>two</u> rings, it is more likely to be edible.		One ring
	Mean:		
odor	0.3	<	4.9
population	5.8	>	3

Classification model

Gaussian Naïve Bayes Classification –

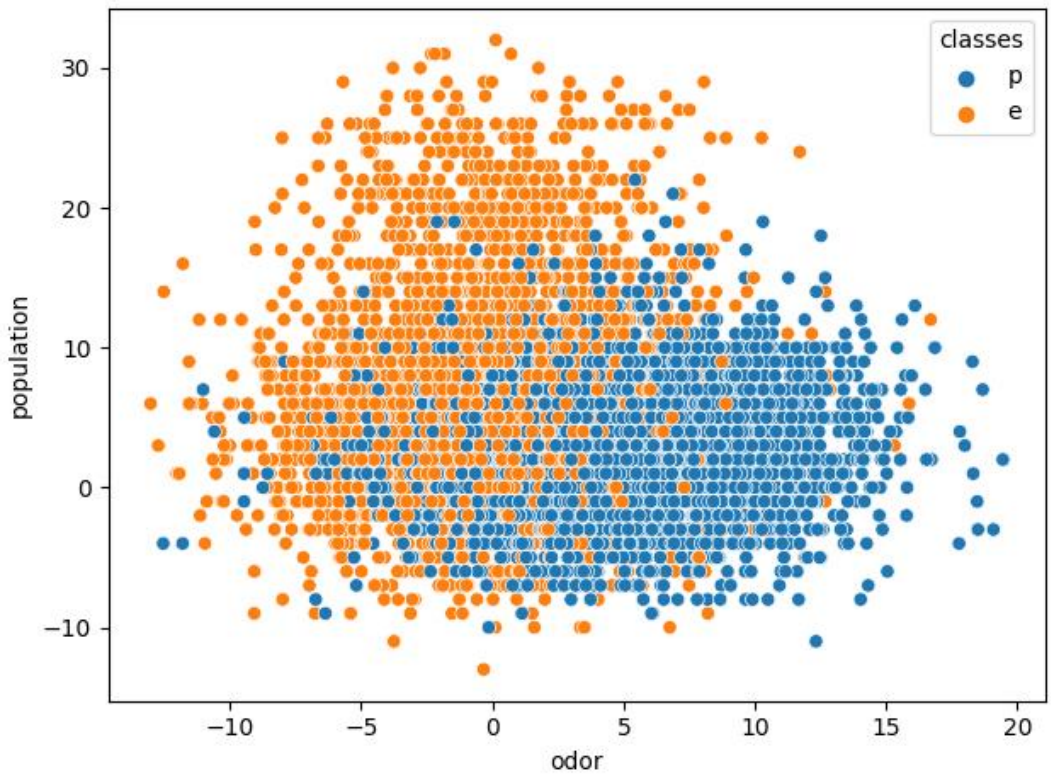
According to the exploratory data analysis, I concluded that the two best distinguishing features between poisonous mushrooms to edible mushrooms are ‘population’ and ‘odor’.

In addition, I checked the accuracy level of the two models. The model that uses all the data, had an accuracy level of 66%, compared to the model that uses only 2 features, which had an accuracy level of 74%.

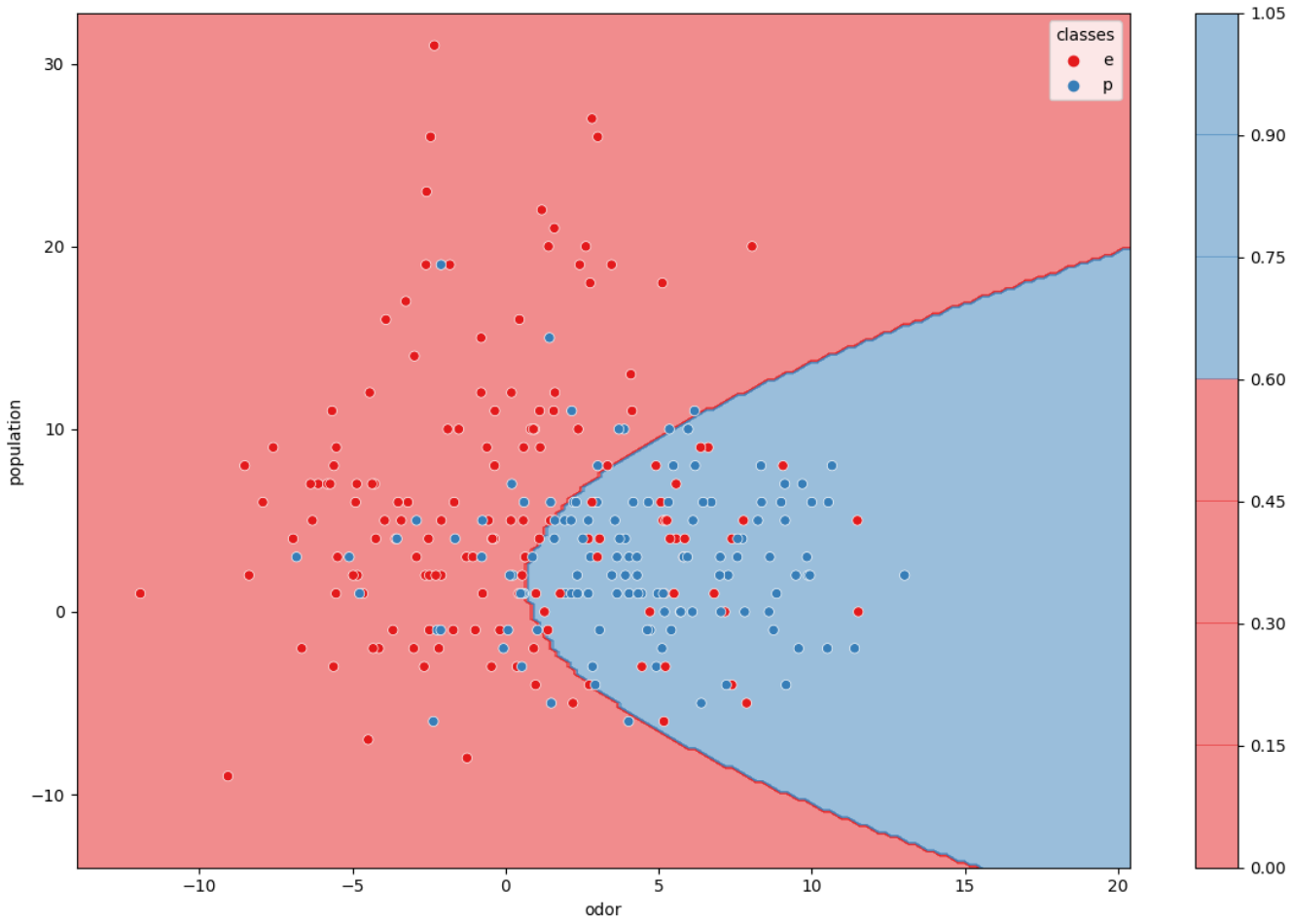
Classification report -

```
accuracy:  0.7375677006400788

classification_report:
              precision    recall  f1-score   support

           e       0.74      0.73      0.74      1020
           p       0.73      0.75      0.74      1011

    accuracy                           0.74      2031
   macro avg       0.74      0.74      0.74      2031
weighted avg       0.74      0.74      0.74      2031
```

Odor and **Population** scatterplot –



GNB classifier result –

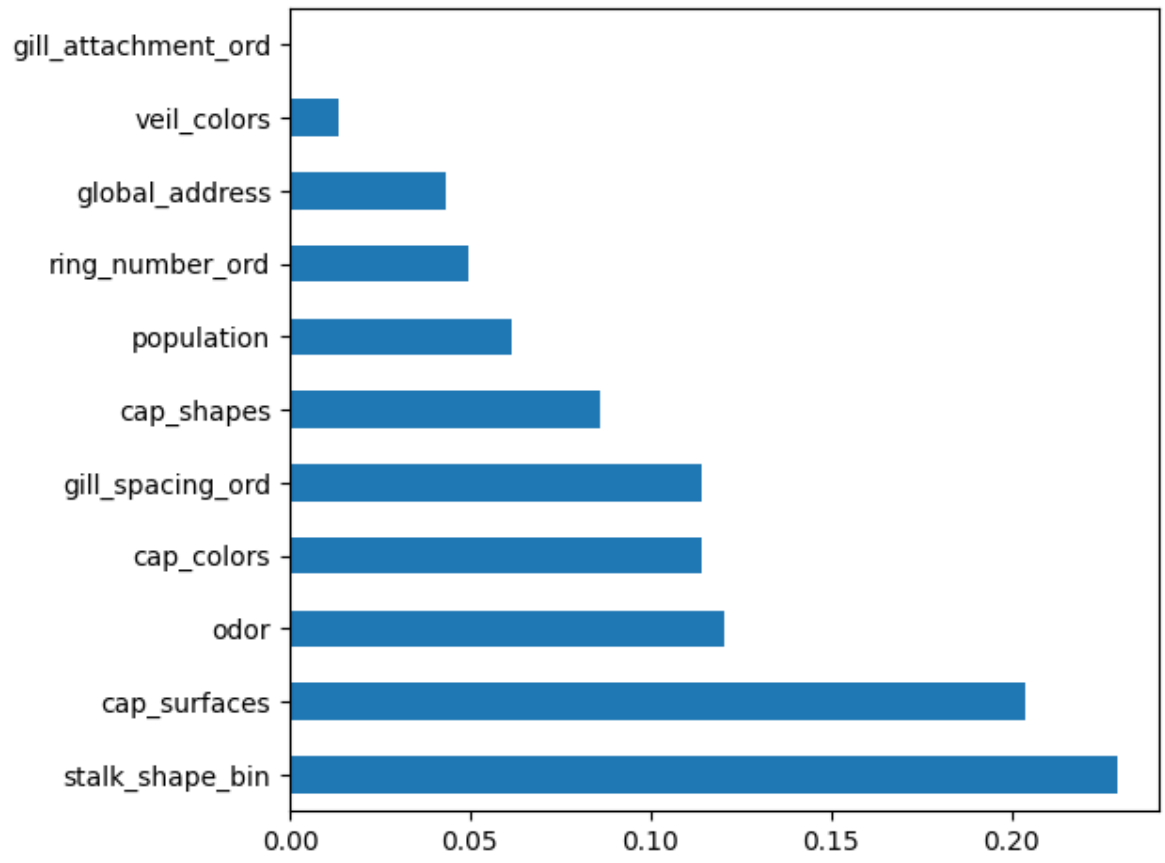


Decision Tree Classification –

a. Using all data features -

```
classification_report:
              precision    recall  f1-score   support

           e       0.88      0.88      0.88      1236
           p       0.88      0.87      0.88      1202

    accuracy                           0.88      2438
   macro avg       0.88      0.88      0.88      2438
weighted avg       0.88      0.88      0.88      2438
```

Feature importance –



b. Using most relevant features –

I choose the features: ‘stalk_shape_bin’, ‘cap_surfaces’, ‘odor’ -

```
classification_report:
              precision    recall  f1-score   support

           e       0.70      0.75      0.73      1236
           p       0.72      0.68      0.70      1202

    accuracy                           0.71      2438
   macro avg       0.71      0.71      0.71      2438
weighted avg       0.71      0.71      0.71      2438
```

From the classification report, we can see that our accuracy did not improve.

*The Tree.png is included in the final project folder.

Summary –

In the data I received, there was a missing Colum, the 'veil type' feature.

In addition, in order to represent correlation between the different features and to use them in both (Naïve Bayes and Decision tree) classification models I had to convert those features to numbers.

In the exploratory data analysis, two features distinguished best between poisonous mushrooms and edible mushrooms, odor and population.

However, neither of these features gave certainty regarding whether the mushrooms are edible or poisonous. As presented in our models accuracy (in both models the accuracy is lower than 90%).

According to the classification report, (using Naive Bayes or Decision tree), we cannot determine which accuracy features (recall or precision) is better for our data.

קישור לסרטון הגנת הפרויקט –

https://drive.google.com/drive/folders/1_217MgR51hsvAMkxsKOvO4RBDuXOw7r6?usp=sharing