Doffe Gabrielle 14/12/2024

Exercise 1 report

Table of contents

Regression Dataset

A regression dataset is basically a collection of data where the goal is to predict a number, or a continuous value based on certain inputs. For example, if we have data about the size of houses and their prices, we could use that to predict the price of a house based on its size.

For this course i used the “VGChartz Games Sales Dataset”. It’s a collection of five video game websites: VGChartz, gamrFeed, gamrReview, gamrTV, and gamrConnect. VGChartz sits at the center of the network and is a video game sales tracking website, providing weekly sales figures of console software and hardware by region. The site was launched in June 2005 and is owned by Brett Walton. VGChartz provides tools for worldwide data analysis and regular reviews of the data it provides.

Ydata-profiling

Overview

Brought to you by YData

Overview | Alerts 2 | Reproduction

Dataset statistics

Number of variables	10
Number of observations	37715
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	2.9 MiB
Average record size in memory	80.0 B

Variable types

Text	3
Numeric	5
DateTime	1
Categorical	1

With this first overview, we can see that the dataset is well-structured, with no missing values or duplicate rows, making it ready for analysis. It contains a mix of text, numeric, date, and categorical variables.

Overview

Brought to you by YData

Overview | Alerts 2 | Reproduction

Alerts

`total_shipped` is highly skewed (γ1 = 52.53992885)	Skewed
`total_sales` has 912 (2.4%) zeros	Zeros

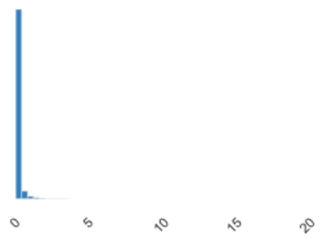
The dataset has two issues to be aware of. First, the total_shipped column is highly skewed, meaning most games have low shipment numbers, but a few have very high values, which can make analysis harder. Second, the total_sales column has 912 rows (2.4%) with zero sales. These zeros might mean the games had no sales or that the data is missing. We have to check what these zeros represent and decide whether to keep, remove, or adjust them for the analysis.

total_sales

total_sales
Real number (ℝ)

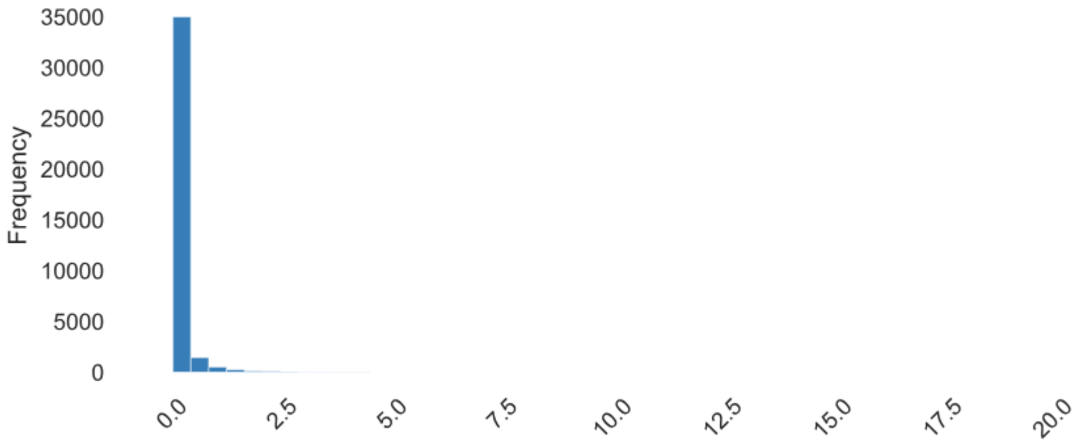
ZEROS

Distinct	426	Minimum	0
Distinct (%)	1.1%	Maximum	20.32
Missing	0	Zeros	912
Missing (%)	0.0%	Zeros (%)	2.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.34056842	Memory size	294.8 KiB



More details

Statistics | Histogram | Common values | Extreme values



Histogram with fixed size bins (bins=50)

The ‘total_sales’ column shows significant skewness in the data. Out of 37,715 entries, 426 unique sales values are observed, which is just 1.1% of the total, indicating that many games share the same sales figures. The sales range from 0 to 20.32, with an average value of only 0.34, meaning most games sold very little. Additionally, 2.4% of the rows (912 games) have zero sales, suggesting either no sales were recorded for these games, or the data might be incomplete for those entries.

The histogram confirms this skewed distribution, with most of the sales clustered near 0 and very few games reaching higher sales values. This shows that only a small portion of games are commercially successful, while most have minimal sales.

Autoviz

- **Key Visualizations**: Features like ‘critic_score’, ‘user_score’, and ‘total_sales’ showed distinct relationships with one another in the pair plots. Higher scores are typically associated with increased sales, demonstrating the importance of scores for a game's commercial success.

- **Outliers**: The variables ‘total_shipped’ and ‘total_sales’ contained significant outliers.

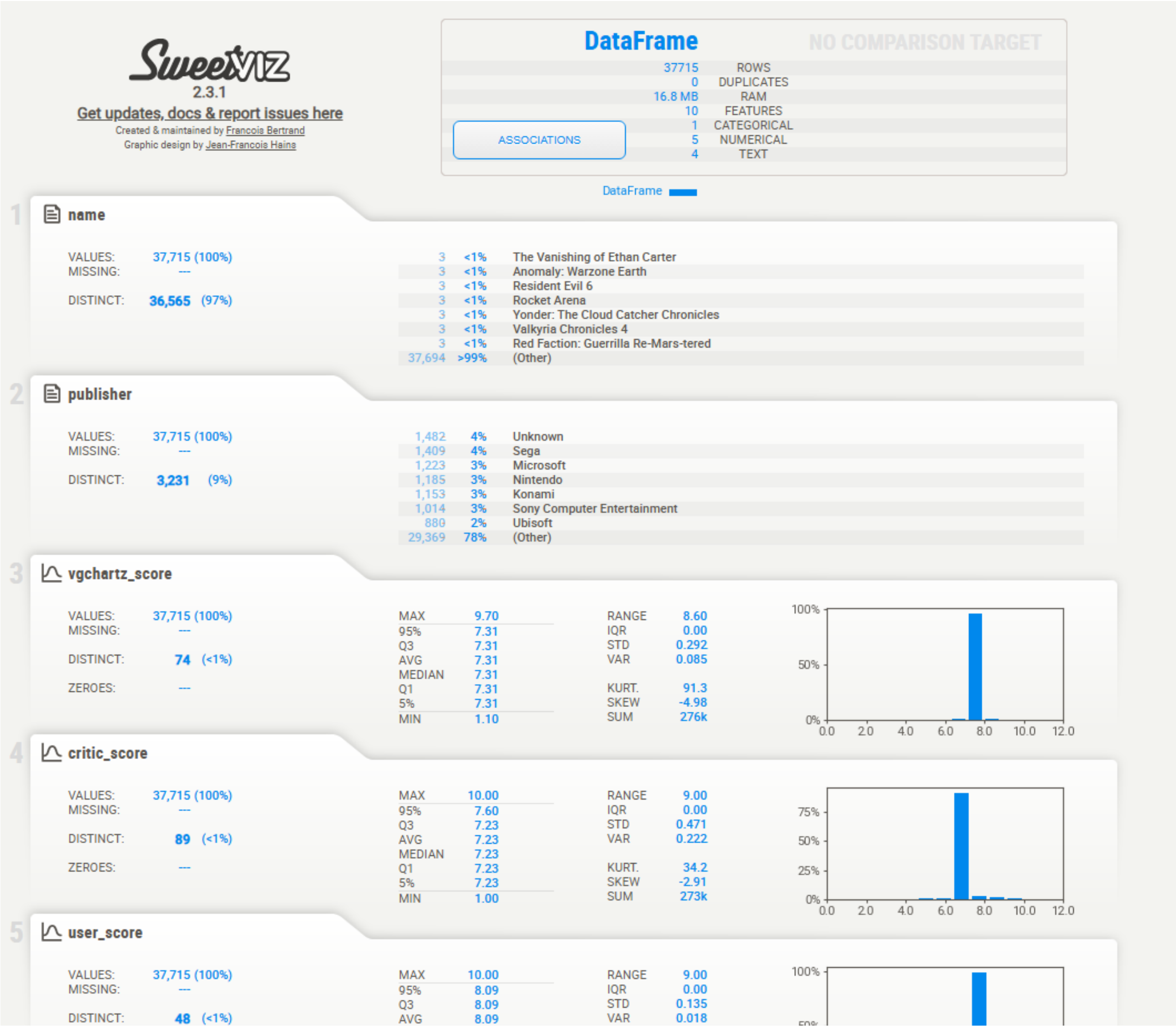
- **Feature Relationships**: Especially between ‘vgchartz_score’ and ‘critic_score’, strong correlations were found, suggesting possible redundancy.

- **Distribution Patterns**: The ‘total_sales’ distribution's skewness was brought to light by the visualizations, and this could influence regression analysis. This realization implies that to improve model performance, transformations might be required to normalize this feature.

	Data Type	Missing Values%	Unique Values%	Minimum Value	Maximum Value
name	object	0.000000	96		
publisher	object	0.000000	8		
vgchartz_score	float64	0.000000	NA	1.100000	9.700000
critic_score	float64	0.000000	NA	1.000000	10.000000
user_score	float64	0.000000	NA	1.000000	10.000000
total_shipped	float64	0.000000	NA	0.000000	496.400000
total_sales	float64	0.000000	NA	0.000000	20.320000
release_date	object	0.000000	19		
genre	object	0.000000	0		
img_url	object	0.000000	90		

The dataset includes both textual and numerical data, and it is complete with no missing values. Publishers are restricted to a tiny set of eight unique values, but most game names and picture URLs are unique. Normalized ratings are shown by scores like ‘vgchartz_score’, ‘critic_score’, and ‘user_score’, which range from 1 to 10. There is a noticeable bias towards lower values in both ‘total_shipped’ (0 to 496.4) and ‘total_sales’ (0 to 20.32), which exhibit broad ranges. While genre seems to have no variance, rendering it possibly irrelevant, the ‘release_date’ column only contains 19 unique values, suggesting aggregated date categories. Although the dataset is generally well-structured, handling skewed data and determining column significance may be necessary.

Sweetviz



After analysing the dataset with Sweetviz, I found it to be well-structured and completely free of missing values. The columns are diverse, including numerical data like ‘total_sales’ and ‘vgchartz_score’, as well as textual data like ‘name’ and ‘publisher’. What stood out to me was the strong skewness in ‘total_sales’ and ‘total_shipped’, where most values are very low, but a few games have extremely high numbers, which could create bias during analysis. Some columns, like ‘genre’, seem less useful due to lack of variety or being constant. Overall, the dataset is clean and rich in information, but the distribution of values needs to be addressed for more balanced insights.

PhikMatrix

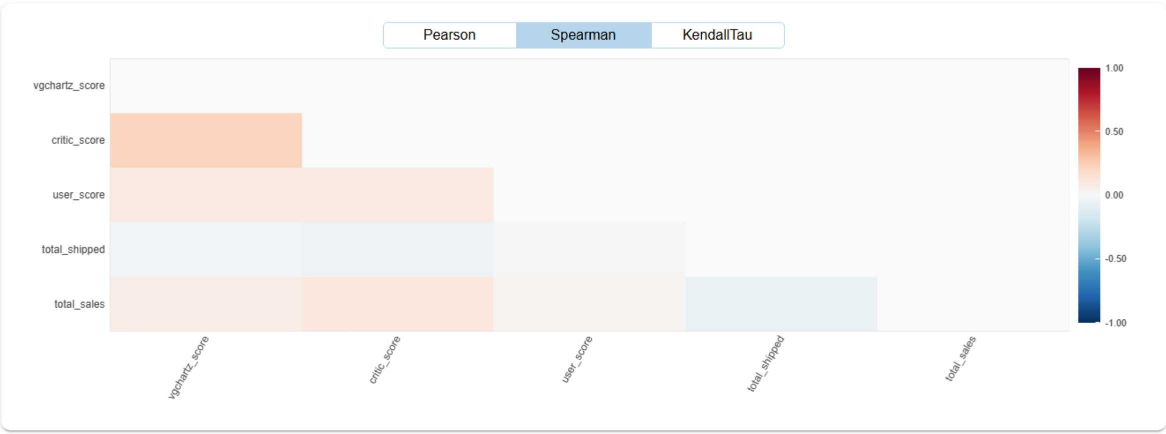
	name	publisher	vgchartz_score	critic_score	user_score	total_shipped	total_sales	release_date	genre	img_url
name	1.0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
publisher	1.0	1.000000	0.000000	0.000000	0.852382	0.000000	0.000000	0.993680	0.000000	0.000000
vgchartz_score	1.0	0.000000	1.000000	0.570668	0.561636	0.680172	0.279752	0.000000	0.416811	1.000000
critic_score	1.0	0.000000	0.570668	1.000000	0.156133	0.409855	0.352788	0.957563	0.013445	0.908168
user_score	1.0	0.852382	0.561636	0.156133	1.000000	0.939956	0.000000	0.000000	0.000000	1.000000
total_shipped	1.0	0.000000	0.680172	0.409855	0.939956	1.000000	0.000000	1.000000	0.000000	1.000000
total_sales	1.0	0.000000	0.279752	0.352788	0.000000	0.000000	1.000000	0.000000	0.236887	0.999922
release_date	1.0	0.993680	0.000000	0.957563	0.000000	1.000000	0.000000	1.000000	0.843402	0.970776
genre	1.0	0.000000	0.416811	0.013445	0.000000	0.000000	0.236887	0.843402	1.000000	0.977786
img_url	1.0	0.000000	1.000000	0.908168	1.000000	1.000000	0.999922	0.970776	0.977786	1.000000

Games with higher user ratings typically have more shipments, according to the strong correlation (0.94) between the ‘total_shipped’ statistic and the ‘user_score’. Likewise, ‘critic_score’ and ‘release_date’" have a strong correlation (0.96), pointing to a possible connection between critic assessments and the release date. It's interesting to note that ‘total_sales’ and ‘total_shipped’ have a moderate correlation (0.68), suggesting that shipments sometimes but not always accurately reflect sales. Very weak correlations between the genre and publisher columns and most variables suggest non-linear relationships or limited influence. Since the diagonal numbers indicate self-correlation, they are all 1, as would be predicted.

While category variables like genre might not be very important, the matrix generally indicates meaningful associations for numerical variables like ‘total_shipped’, ‘critic_score’, and ‘total_sales’. Selecting which features are crucial for additional research or modelling might be aided by this matrix.

DataPrep

Correlations



The Dataprep correlation heatmap shows how different variables in the dataset are related. For example, there is a moderate positive correlation between ‘vgchartz_score’ and ‘critic_score’, meaning games with higher critic scores often get better VGChartz ratings. Similarly, ‘user_score’ and ‘total_shipped’ are moderately correlated, suggesting games rated higher by users tend to have more units shipped. However, the correlation between ‘total_sales’ and ‘total_shipped’ is weaker, indicating other factors influence sales. Overall, while some relationships are clear, none are very strong, showing that sales and shipments likely depend on multiple factors.

Classification Dataset

Ydata-Profiling

A classification dataset is a collection of data used to train and test machine learning models for classification tasks. In these tasks, the goal is to predict a specific category or class label for each instance in the dataset based on its features.

Overview

Brought to you by YData

Overview Alerts 24 Reproduction

Dataset statistics

Number of variables	23
Number of observations	8124
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.4 MiB
Average record size in memory	184.0 B

Variable types

Categorical	22
Boolean	1

This dataset contains 23 variables and 8,124 observations, with no missing or duplicate values, meaning it's complete and well-structured. Most variables are categorical, representing features like characteristics or categories, while 1 variable is Boolean, likely the target (‘edible’ or ‘poisonous’ for mushrooms). The dataset is small (1.4 MiB), making it efficient to process. Overall, it's a clean and ready-to-use dataset for classification tasks, where the goal is to predict a specific category based on the input features.

Overview | Alerts 24 | Reproduction

Alerts

`veil-type` has constant value "p"	Constant
`bruises` is highly overall correlated with `class` and 8 other fields	High correlation
`cap-color` is highly overall correlated with `ring-type` and 1 other fields	High correlation
`class` is highly overall correlated with `bruises` and 9 other fields	High correlation
`gill-attachment` is highly overall correlated with `gill-color` and 4 other fields	High correlation
`gill-color` is highly overall correlated with `bruises` and 6 other fields	High correlation
`gill-size` is highly overall correlated with `class` and 6 other fields	High correlation
`gill-spacing` is highly overall correlated with `habitat` and 2 other fields	High correlation
`habitat` is highly overall correlated with `bruises` and 3 other fields	High correlation
`odor` is highly overall correlated with `bruises` and 7 other fields	High correlation
`population` is highly overall correlated with `gill-size` and 3 other fields	High correlation
`ring-number` is highly overall correlated with `odor` and 3 other fields	High correlation
`ring-type` is highly overall correlated with `bruises` and 11 other fields	High correlation
`spore-print-color` is highly overall correlated with `bruises` and 5 other fields	High correlation
`stalk-color-above-ring` is highly overall correlated with `class` and 7 other fields	High correlation
`stalk-color-below-ring` is highly overall correlated with `class` and 6 other fields	High correlation
`stalk-root` is highly overall correlated with `bruises` and 8 other fields	High correlation
`stalk-shape` is highly overall correlated with `cap-color` and 5 other fields	High correlation
`stalk-surface-above-ring` is highly overall correlated with `bruises` and 2 other fields	High correlation

The 24 alerts highlight issues in the dataset, such as one column ('veil-type') having the same value for all rows, making it useless for analysis and worth removing. Most other alerts point out high correlations between certain features ('bruises' and 'class' or 'odor'), meaning these features provide similar information. This can cause redundancy in the dataset, making it more complex to work with. To fix this, we can just remove unnecessary or redundant features and keep only the ones that are most useful for predicting the targe). This will make our model simpler and more efficient.

`gill-attachment` is highly imbalanced (82.7%)	Imbalance
`veil-color` is highly imbalanced (90.2%)	Imbalance
`ring-number` is highly imbalanced (73.5%)	Imbalance

This part of the alerts shows that three features ('gill-attachment', 'veil-color', and 'ring-number') are highly imbalanced, meaning one category in these features dominates the others. For example, in 'veil-color', 90.2% of the data belongs to a single category. Imbalances like this can reduce the usefulness of these features for making predictions, as the model might not learn enough about the less common categories.

Autoviz

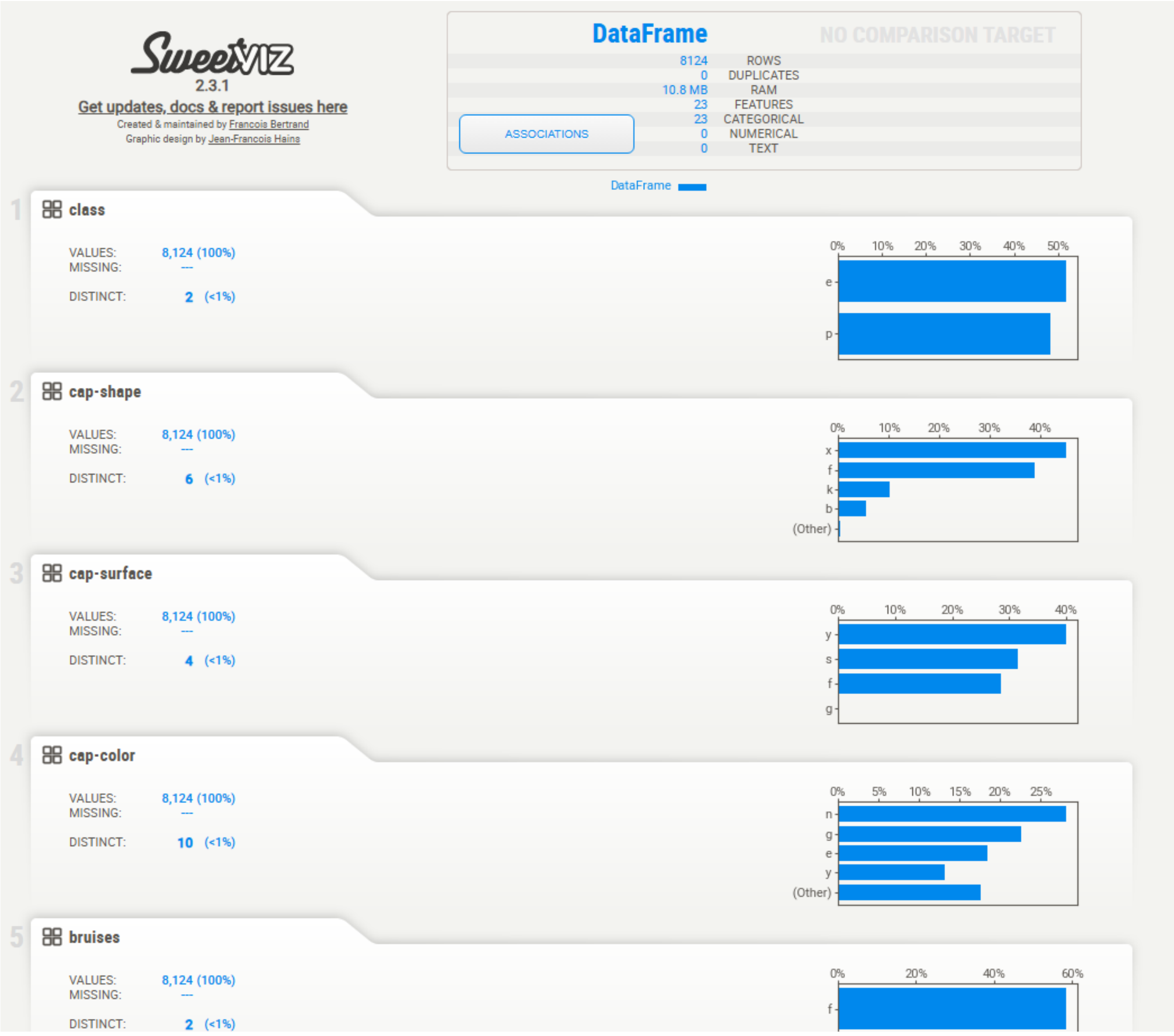
	Data Type	Missing Values%	Unique Values%	Minimum Value	Maximum Value
class	object	0.000000	0		
cap-shape	object	0.000000	0		
cap-surface	object	0.000000	0		
cap-color	object	0.000000	0		
bruises	object	0.000000	0		
odor	object	0.000000	0		
gill-attachment	object	0.000000	0		
gill-spacing	object	0.000000	0		
gill-size	object	0.000000	0		
gill-color	object	0.000000	0		
stalk-shape	object	0.000000	0		
stalk-root	object	0.000000	0		
stalk-surface-above-ring	object	0.000000	0		
stalk-surface-below-ring	object	0.000000	0		
stalk-color-above-ring	object	0.000000	0		
stalk-color-below-ring	object	0.000000	0		
veil-type	object	0.000000	0		
veil-color	object	0.000000	0		
ring-number	object	0.000000	0		
ring-type	object	0.000000	0		
spore-print-color	object	0.000000	0		
population	object	0.000000	0		
habitat	object	0.000000	0		

Th dataset appears to be clean and well-structured with the following observations:

- Data Types: All the features are of type object, which indicates they are categorical variables. This makes sense for a classification dataset like this one, where features represent descriptive categories or characteristics.
- Missing Values: There are no missing values in the dataset. Every feature has 0% missing data
- Unique Values: The percentage of unique values for each feature is 0%, suggesting that the dataset contains categorical variables with a predefined and limited number of categories.
- No Numerical Data: There are no numerical features, meaning this dataset entirely relies on categorical features to predict the target variable (class), making it suitable for algorithms that handle categorical data well, like decision trees or ensemble methods.

The dataset is clean, complete, and primarily made up of categorical variables. It is ready for preprocessing steps like encoding (one-hot encoding or label encoding) before being used for machine learning models.

Sweetviz



The Sweetviz report shows that the dataset is clean, with no missing values, and consists mostly of categorical features like ‘cap-color’ and ‘odor’, which describe the mushrooms. It highlights imbalances in some features and correlations between variables, meaning some features might carry redundant information. The report also likely provides insights into how these features relate to the target variable (‘edible’ or ‘poisonous’), helping identify which features are most useful for predictions.

PhikMatrix

The Phik matrix analysis evaluates the relationships between features in the dataset using the Phik correlation, which measures dependencies between categorical variable.

	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-above-ring	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-color	ring-number	ring-type	spore-print-color
class	1.000000	0.233071	0.124149	0.230056	0.752519	0.998748	0.128165	0.508024	0.758803	0.856036	...	0.776169	0.795524	0.455800	0.660890	0.055481	0.236677	0.821569	0.733773
cap-shape	0.233071	1.000000	0.207606	0.460182	0.213456	0.432120	0.093256	0.000000	0.324564	0.505914	...	0.142725	0.124807	0.206841	0.253216	0.066649	0.189694	0.322697	0.443874
cap-surface	0.124149	0.207606	1.000000	0.473019	0.126489	0.438726	0.090294	0.196297	0.223355	0.615061	...	0.171210	0.187791	0.369327	0.417745	0.288134	0.017119	0.317082	0.554377
cap-color	0.230056	0.460182	0.473019	1.000000	0.233108	0.575862	0.147887	0.526662	0.675368	0.619906	...	0.367116	0.477847	0.411750	0.497387	0.000000	0.569056	0.855786	0.553486
bruises	0.752519	0.213456	0.126489	0.233108	1.000000	0.878589	0.081533	0.431559	0.588251	0.869078	...	0.757905	0.789078	0.403711	0.636119	0.036392	0.044237	0.948824	0.638637
odor	0.998748	0.432120	0.438726	0.575862	0.878589	1.000000	0.079133	0.602253	0.930469	0.722570	...	0.744751	0.804078	0.503713	0.678057	0.000000	0.389349	0.892199	0.684594
gill-attachment	0.128165	0.093256	0.090294	0.147887	0.081533	0.079133	1.000000	0.000000	0.065724	1.000000	...	0.052368	0.073760	1.000000	1.000000	1.000000	0.000000	0.138505	0.792618
gill-spacing	0.508024	0.000000	0.196297	0.526662	0.431559	0.602253	0.000000	1.000000	0.158046	0.396656	...	0.657030	0.653183	0.345297	0.482806	0.000000	0.311877	0.399283	0.314415
gill-size	0.758803	0.324564	0.223355	0.675368	0.588251	0.930469	0.065724	0.158046	1.000000	0.928834	...	0.284602	0.201092	0.310462	0.463436	0.030070	0.274571	0.799841	0.687637
gill-color	0.856036	0.505914	0.615061	0.619906	0.869078	0.722570	1.000000	0.396656	0.928834	1.000000	...	0.649036	0.664059	0.756383	0.816473	0.959766	0.661498	0.874174	0.882052
stalk-shape	0.145232	0.296178	0.070912	0.803039	0.025189	0.802104	0.156268	0.076095	0.341422	0.802970	...	0.479365	0.500428	0.444295	0.710661	0.067180	0.442979	0.851792	0.514645
stalk-root	0.397994	0.761577	0.467074	0.733816	0.490630	0.781656	0.111734	0.471557	0.563977	0.726612	...	0.395122	0.670944	0.451756	0.506518	0.097244	0.179870	0.580866	0.713812
stalk-surface-above-ring	0.776169	0.142725	0.171210	0.367116	0.757905	0.744751	0.052368	0.657030	0.284602	0.649036	...	1.000000	0.864887	0.496058	0.694704	0.000000	0.254709	0.802785	0.588110
stalk-surface-below-ring	0.795524	0.124807	0.187791	0.477847	0.789078	0.804078	0.073760	0.653183	0.201092	0.664059	...	0.864887	1.000000	0.498278	0.710325	0.000000	0.016384	0.835002	0.588571
stalk-color-above-ring	0.455800	0.206841	0.369327	0.411750	0.403711	0.503713	1.000000	0.345297	0.310462	0.756383	...	0.496058	0.498278	1.000000	0.725691	0.765538	0.230203	0.629242	0.683496
stalk-color-below-ring	0.660890	0.253216	0.417745	0.497387	0.636119	0.678057	1.000000	0.482806	0.463436	0.816473	...	0.694704	0.710325	0.725691	1.000000	0.784775	0.551740	0.827783	0.684097
veil-color	0.055481	0.066649	0.288134	0.000000	0.036392	0.000000	1.000000	0.000000	0.030070	0.959766	...	0.000000	0.000000	0.765538	0.784775	1.000000	0.000000	0.036726	0.976086

Many features, such as ‘odor’ "gill-color‘, and ‘ring-type‘ show moderate to strong correlations with others, indicating that they carry significant predictive information and play an important role in understanding patterns in the data. Some variables, like ‘stalk-color-above-ring‘, and ‘stalk-color-below-ring‘, exhibit particularly high correlations, suggesting a strong dependency or overlap in the information they provide. On the other hand, certain features, like ‘veil-color‘ and ‘ring-number’, have weaker correlations with others, indicating they might contribute less unique information to the dataset.

These results are useful for improving the dataset's quality and efficiency. Features with very high correlations may be redundant, we could consider removing or combining them to reduce dimensionality without losing important information.

DataPrep

Overview

Dataset Statistics

Number of Variables	23
Number of Rows	8124
Missing Cells	0
Missing Cells (%)	0.0%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	10.3 MB
Average Row Size in Memory	1.3 KB
Variable Types	Categorical: 23

Dataset Insights

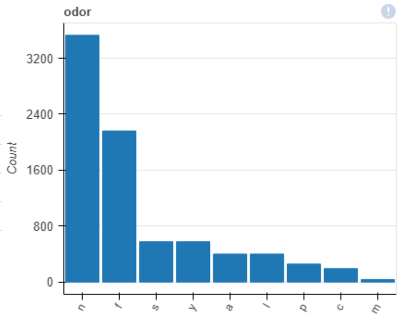
veil-type has constant value "p"	Constant
class has constant length 1	Constant Length
cap-shape has constant length 1	Constant Length
cap-surface has constant length 1	Constant Length
cap-color has constant length 1	Constant Length
bruises has constant length 1	Constant Length
odor has constant length 1	Constant Length
gill-attachment has constant length 1	Constant Length
gill-spacing has constant length 1	Constant Length
gill-size has constant length 1	Constant Length

1 2 3

From the DataPrep report, the dataset is clean, with no missing values or duplicate rows, and all 23 variables are categorical. It occupies 10.3 MB of memory and has consistent data formats, making it efficient to process. Insights like ‘veil-type’ being constant across all rows (value ‘p’) indicate that this feature adds no variability and can be safely dropped.

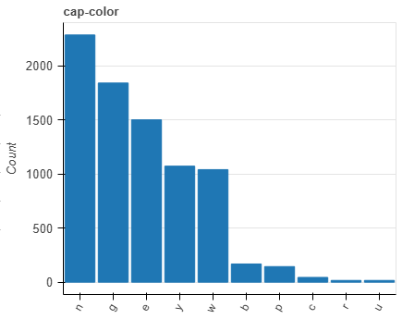
odor
categorical
Show Details

Approximate Distinct Count	9
Approximate Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory Size	536184



cap-color
categorical
Show Details

Approximate Distinct Count	10
Approximate Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory Size	536184



Detailed feature analyses, such as for ‘odor’ and ‘cap-color’, show distinct categories with varying distributions. For example, certain categories in ‘odor’ and ‘cap-color’ dominate, as seen in their bar charts, which might indicate feature imbalances. These dominant values may strongly influence the dataset's overall structure and predictions, highlighting the importance of preprocessing steps like balancing data if needed. Overall, the dataset is ready for machine learning tasks but could benefit from dimensionality reduction and encoding strategies to optimize model performance.

Sources and help

Documentation :

DataPrep : https://pypi.org/project/dataprep/

Sweetviz: https://pypi.org/project/sweetviz/

Autoviz : https://pypi.org/project/autoviz/

Phik: https://pypi.org/project/phik/

Ydata-Profiling: https://pypi.org/project/ydata-profiling/

AI:

ChatGpt to understand the different analysis and explain better.