The Cat Food Canspiracy

Outliers in Cat Food Data

Prompt

Could you describe my dataset?
Could you boxplot ‘price’?
Could you plot ‘price’ vs ‘weight’ by ‘company’?

Strategy

1. Find outliers by studying the statistics.
2. Picture the outliers and
   learn about their nature before dropping them.

Code & Results

1. Descriptive statistics

This code describes the dataset.

```
data.describe()
```

	price	weight	star_rating	total_comments	adult	chicken	wet	from_ocean	fish	
count	164.000000	164.000000	164.000000	164.000000	164.000000	164.000000	164.000000	164.000000	164.000000	164
mean	1188.853659	2.246902	4.307927	837.439024	0.689024	0.280488	0.637195	0.640244	0.292683	0
std	1320.939228	2.452212	0.449674	2031.010946	0.419174	0.450613	0.408222	0.481399	0.456388	0
min	40.000000	0.050000	2.500000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0
25%	420.000000	0.830000	4.100000	15.250000	0.500000	0.000000	0.500000	0.000000	0.000000	0
50%	669.000000	1.20 0000	4.400000	115.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0
75%	1500.000000	3.000000	4.500000	521.250000	1.000000	1.000000	1.000000	1.000000	1.000000	1
max	6550.000000	14.000000	5.000000	12982.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1

Here, you can see that the **price** changes from $40 to $6,550.

Is it okay or unusual?

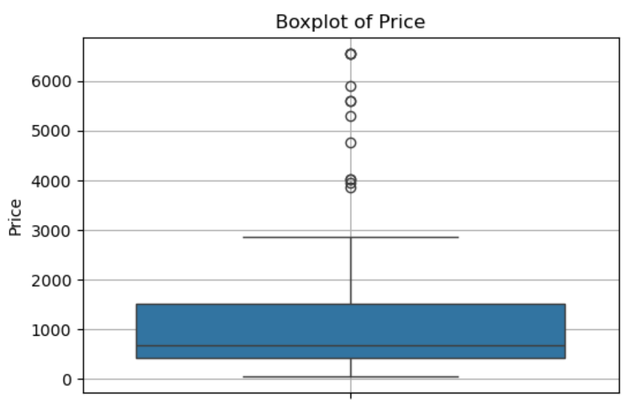
💡Tips

Don't rush to delete data that looks unusual.

2. Visualization

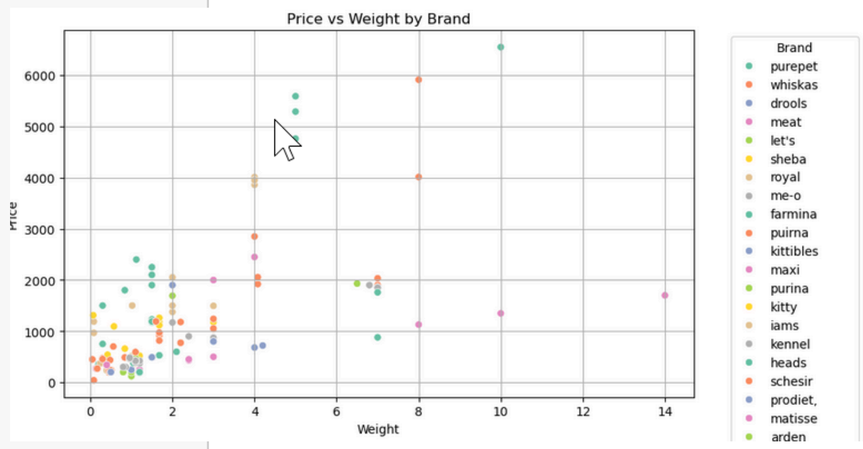
This code makes a box plot.

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(6, 4))
sns.boxplot(y=data['price'])
plt.title('Boxplot of Price')
plt.ylabel('Price')
plt.grid(True)
plt.show()
```



Some high prices are just exceptions. But what if the brand's marketing is the reason? Let's make a scatterplot of "price" versus "weight" by "company."

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))
sns.scatterplot(data=data, x='weight', y='price', hue='company', palette='Set2')
plt.title('Price vs Weight by Brand')
plt.xlabel('Weight')
plt.ylabel('Price')
plt.legend(title='Brand', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.grid(True)
plt.tight_layout()
plt.show()
```



On the graph we can see that the prices by brand are in line with the packaging. This means that these high prices are probably due to the brand's policy, not outliers. Let's keep them.

Made by: @katerynakononova / Machine learning: for humans on cats
Watch the short: https://youtube.com/shorts/_JV9BgA27VQ

Website: https://katerynakononova.github.io/meowlearning/

Dive deeper into the world of AI with *Machine Learning: for Humans on Cats* — now on Amazon!
📘👉 https://www.amazon.com/dp/B0CW9SFYXF