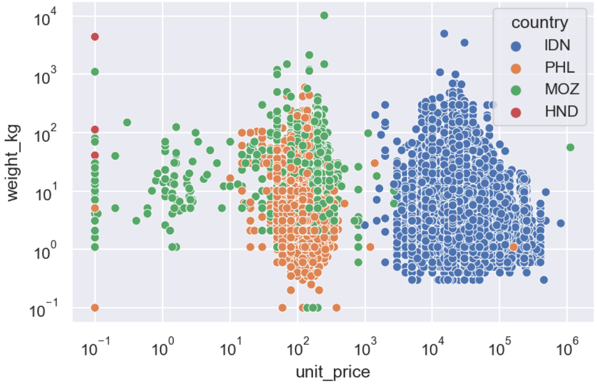
Outlier Detection Methodology

1 A first look at the data

Central to this project is developing a method of flagging potential outliers. This is not an uncommon problem in statistics. In our catch data, we are mainly concerned with these variables that the user inputs:

- unit price
- weight (in kg and lbs)
- count

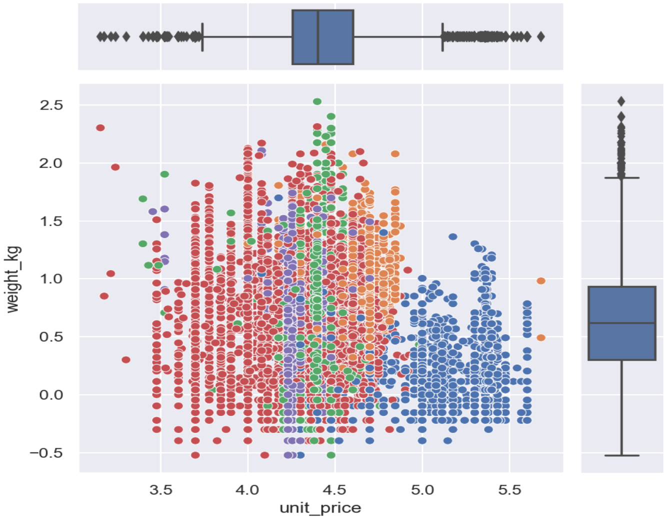
We can think of the data in terms of price and quantity. Unit price is the only variable that we can call price, while the two weight and one count variables can be considered quantities. Thus the data is treated like points on a 2D grid, as shown below.



Note that in all plots, the axes are in log-scale, and a value of 10^{-1} means the original value was 0.

As we can see, there are distinct clusters of data based on the country. Thus it would be a good idea to check for potential outliers relative to a country. But this is not granular enough. Let's zoom in on the Indonesia cluster on the right, and look specifically at the samples that come from the five most popular fish. Each color represents a distinct fish.

We can once again see distinct clusters, now between different fish species. Thus, we can tell that, for example, the two orange points on the far right are probably not supposed to be there; all the other orange points are near the middle.

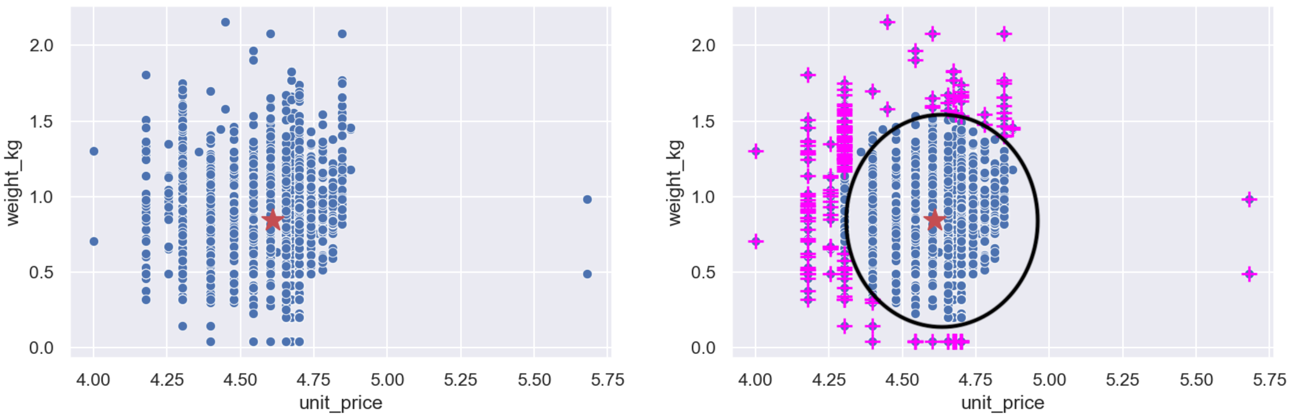


2 Mahalanobis method

Once we’ve focused on a particular fish species from a particular country, how do we flag points? In statistical theory, there are many tests for deciding if a data sample might be an outlier. One of the most popular, and simple, methods is to look at a box plot like the ones above. Points outside of the whiskers get flagged as outliers. But this really only works with 1-dimensional data; our data is 2-dimensional. For our bivariate data, I looked into the following techniques:

- Bagplot (computationally complicated and would have to write from scratch)
- Fitting to probability distributions (data did not fit well)
- Mahalanobis distance (simple to set up, works OK for most bivariate data distributions)

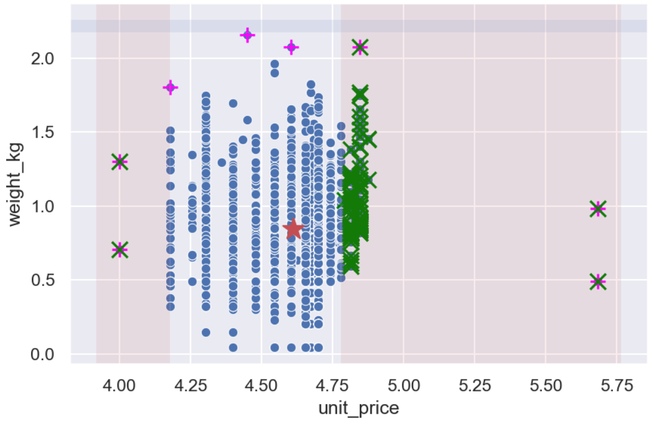
Ultimately, the Mahalanobis distance was used for its simplicity. What this technique will do is find an ellipse that fits the data well, expands the ellipse, and flags any samples outside of it. We can look at how it works with samples from Indonesia with fish species *Scomberomorus commerson*:



The red star denotes the centroid of the data, where the ellipse is centered. The magenta crosses are outside of the ellipse (the ellipse drawn is not exact). In practice, the ellipse is larger so that only a handful of points are picked up with this test.

3 Floors and ceilings

This method on its own is not perfect, as even with a large ellipse, some points will be flagged that really shouldn’t be. Thus we employ an additional filter: most fish records have specific floors and ceilings on the weight/count and price of catches. So we will flag any samples that are both (1) outside the ellipse and (2) exceed specified thresholds. For weight values, we only use the ceiling and not the floor, since the data caps itself below at 0, presumably from a check for negative values in the app.

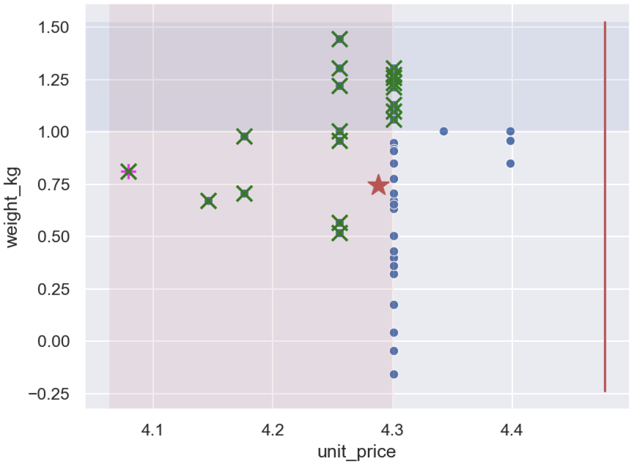


The red region exceeds the thresholds on price and the blue region exceeds the threshold on weight. The samples that are in either region are marked with a green cross.

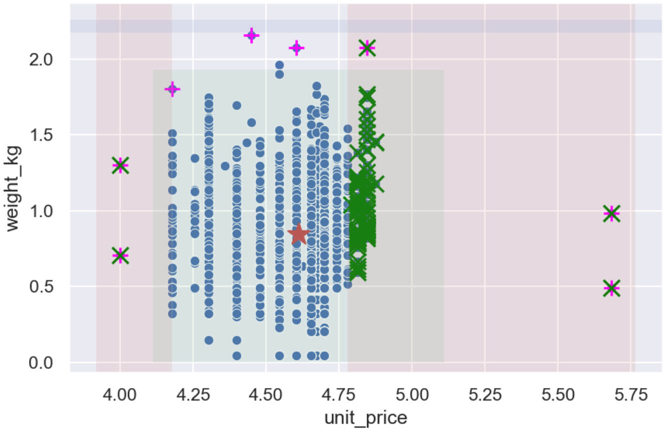
Some samples do not have thresholds on record, so we make them: In the x-direction, the thresholds are 1.5 units away from the centroid, and in the y-direction, the threshold is 2 units above. The reasoning is that these values are large enough to make you think "Ok, maybe there's an extra zero here."

4 The safe box

Even with these two methods, some points will be flagged that are "too close" to be a concern, maybe because the thresholds are too close, as in the following plot: (the fish is *Lutjanus bohar*)



Notice how the unit_price of every point fits roughly in 4.1-4.4. This is not a very large range, corresponding to roughly 12,500-25,000, and yet one point was flagged. This motivates one final check: is the sample over 0.5 units away in the x-direction, or 1 unit away in the y-direction, from the centroid? This creates a "safe box" of samples of least concern (in green):



6 Conclusion

The last plot gives the full picture: potential outliers are any points that are outside of

(1) the Mahalanobis ellipse (marked by magenta crosses)
(2) the floors/ceilings (marked by green crosses)
(3) the safe box

Thus for *Scomberomorus commerson*, we have 5 potential outliers. In particular, the two on the far right were the distant orange points from the first plot.

Out of the 43,464 records that were able to be processed (some were dropped due to missing data), this detection method flagged 205 points, or 0.5% of the data.

There are certainly shortcomings to this method. For some fish, there are many samples that deviate from what may be considered the main cluster. This could signal a change in the market for this fish, as sometimes there are too many samples to be a coincidence. What happens then is that all these points, maybe 20 of them, get flagged. Thus one improvement in the methodology is to detect if there is a cluster of points outside of the central cluster. These may then be specifically not flagged if it is known they're not true anomalies.

There's another problem with these clusters. Some fish have multiple distinct price clusters at significant scale; one cluster may have a price about 10, another about 100, and another about 10,000. This is prominent in two contexts: many of Mozambique's fish and fish that are denoted by "Mixed group" or similar. It seems like these may have to be handled carefully on a case by case basis.

Another improvement that could be made is replacing the Mahalanobis part. Initial testing for how well probability distributions could fit the data were done under the assumption the data was continuous. After looking at many plots, it's clear that the price on some fish are more discrete than continuous. Thus it may be worth looking into fitting discrete probability distributions to the price data.