Difference between Anomaly and Outliers

Kenan Sooklall

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Outlier

A legitimate data point that is far away from the mean or median in a distribution.

Anomaly

An illegitimate data point that's generated by a different process than whatever generated the rest of the data.

Let's take a hypothetical example A recent data breach has happened at a big bank and a customer called stating he account might have

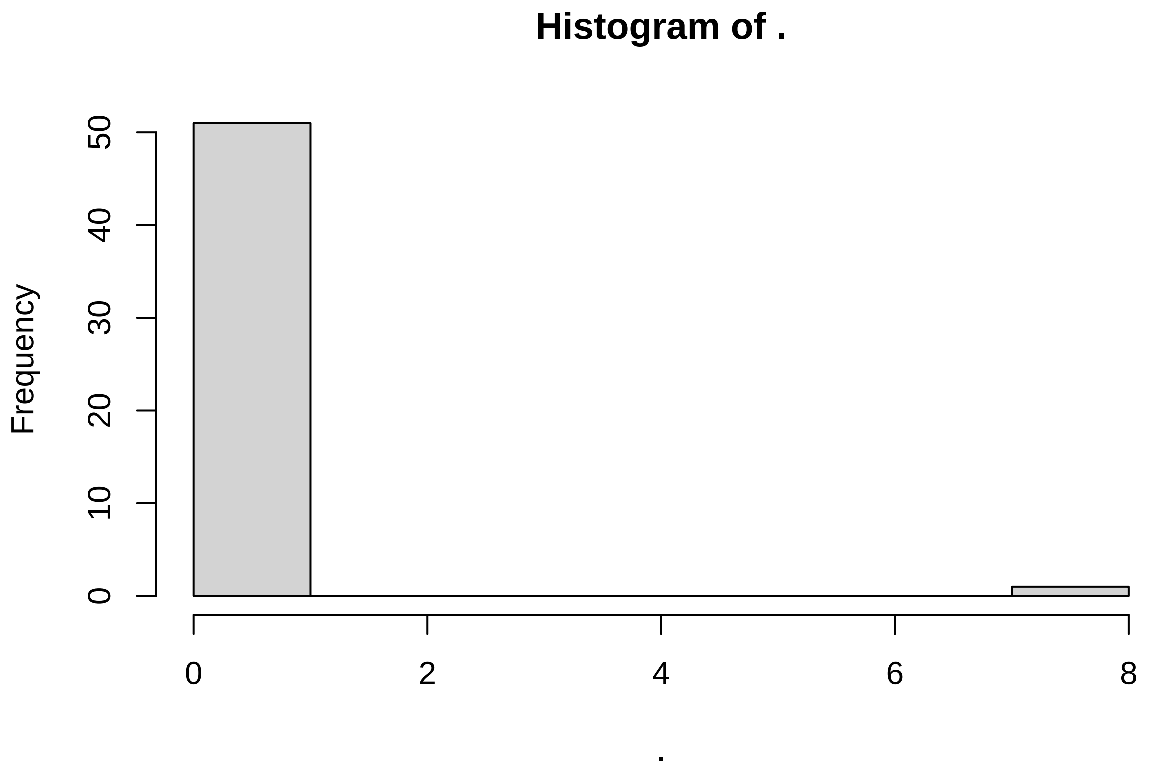
```
##        item cost
## 1     bagle 2.50
## 2    coffee 1.50
## 3     apple 0.65
## 4     water 2.50
## 5  sandwich 5.50
## 6     bagle 2.50
## 7    coffee 1.50
## 8     apple 0.65
## 9     water 2.50
## 10 sandwich 5.50
## 11    bagle 2.50
## 12   coffee 1.50
## 13    apple 0.65
## 14    water 2.50
## 15 sandwich 5.50
## 16    bagle 2.50
## 17   coffee 1.50
## 18    apple 0.65
## 19    water 2.50
## 20 sandwich 5.50
## 21    bagle 2.50
## 22   coffee 1.50
## 23    apple 0.65
## 24    water 2.50
## 25 sandwich 5.50
## 26    bagle 2.50
## 27   coffee 1.50
## 28    apple 0.65
## 29    water 2.50
## 30 sandwich 5.50
```

The Outlier

Outliers are easy to detect because they can be explicitly defined.

- All values outside 3σ
- Values outside $1.5 \pm IQR$

Here we first try values outside 1 std, first scale the data then plot the histogram

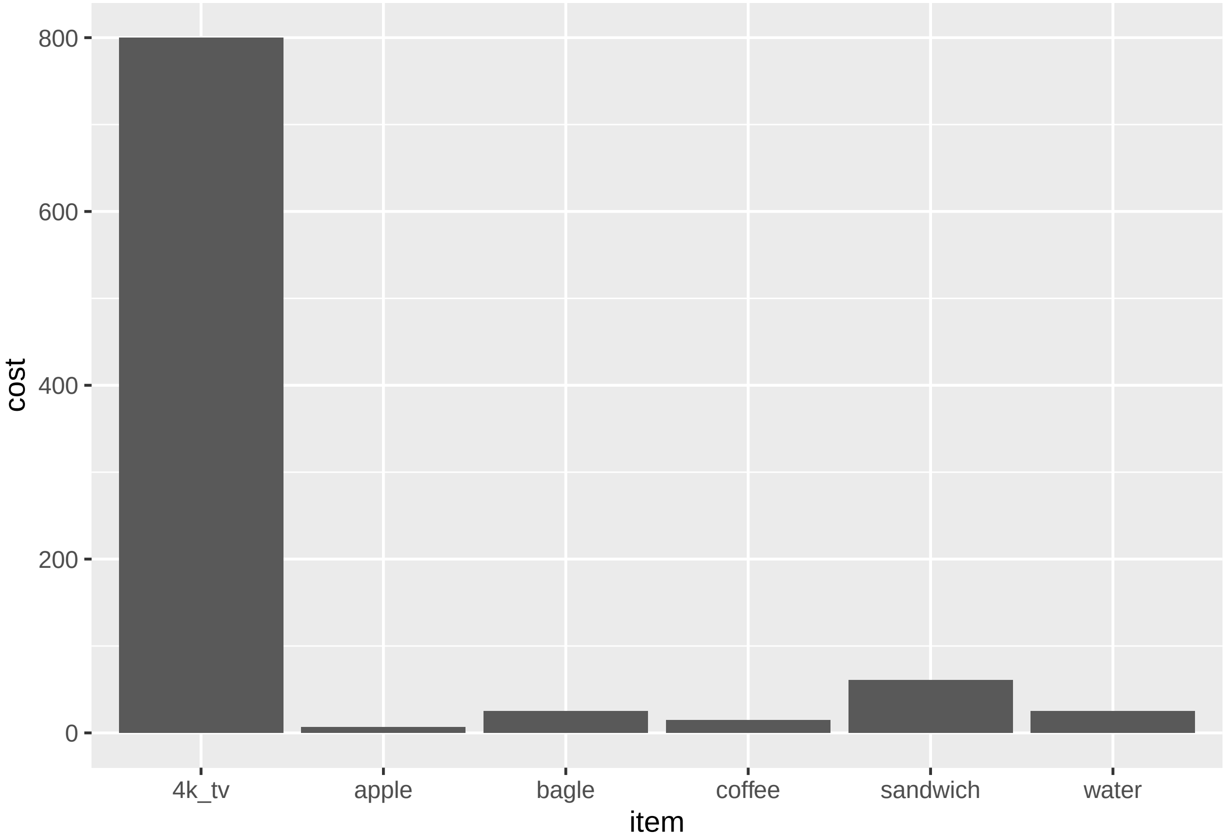


Imagine if this was a normal distribution that was right skewed the outlier is very obvious. Filter for values greater than 1.

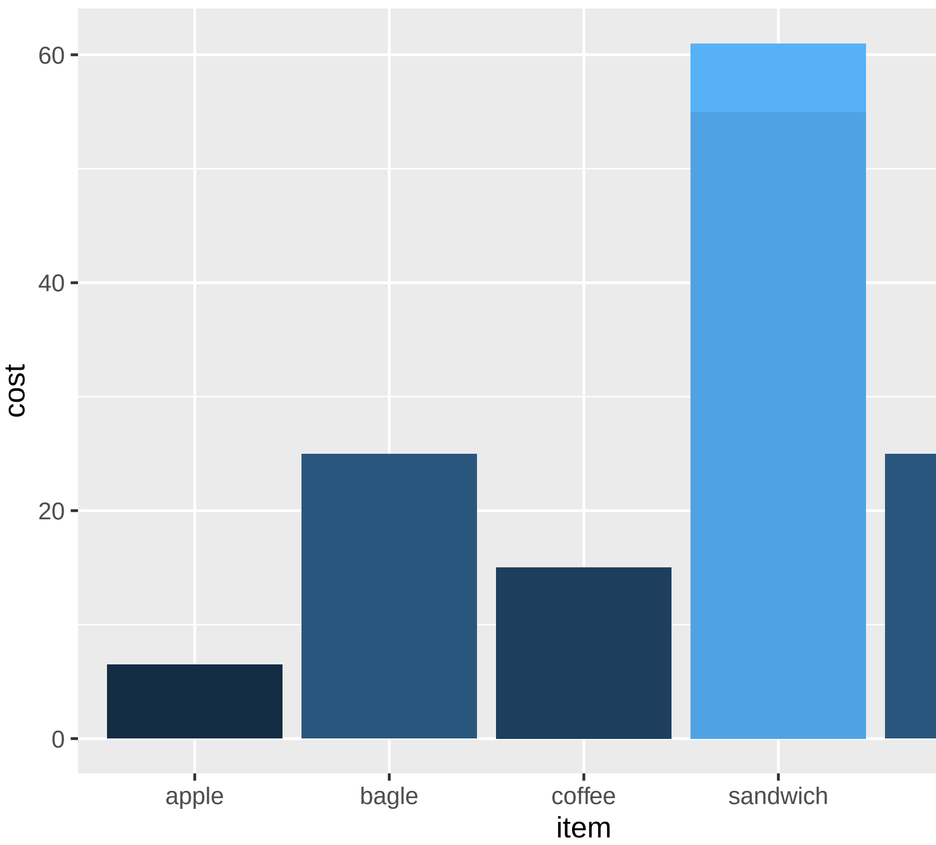
```
##    item   cost  scale
## 1 4k_tv 799.99 7.0716
```

The Anomaly

But how would you find the Anomaly or even if it exists



Everything looks normal and everything else is accounted for.

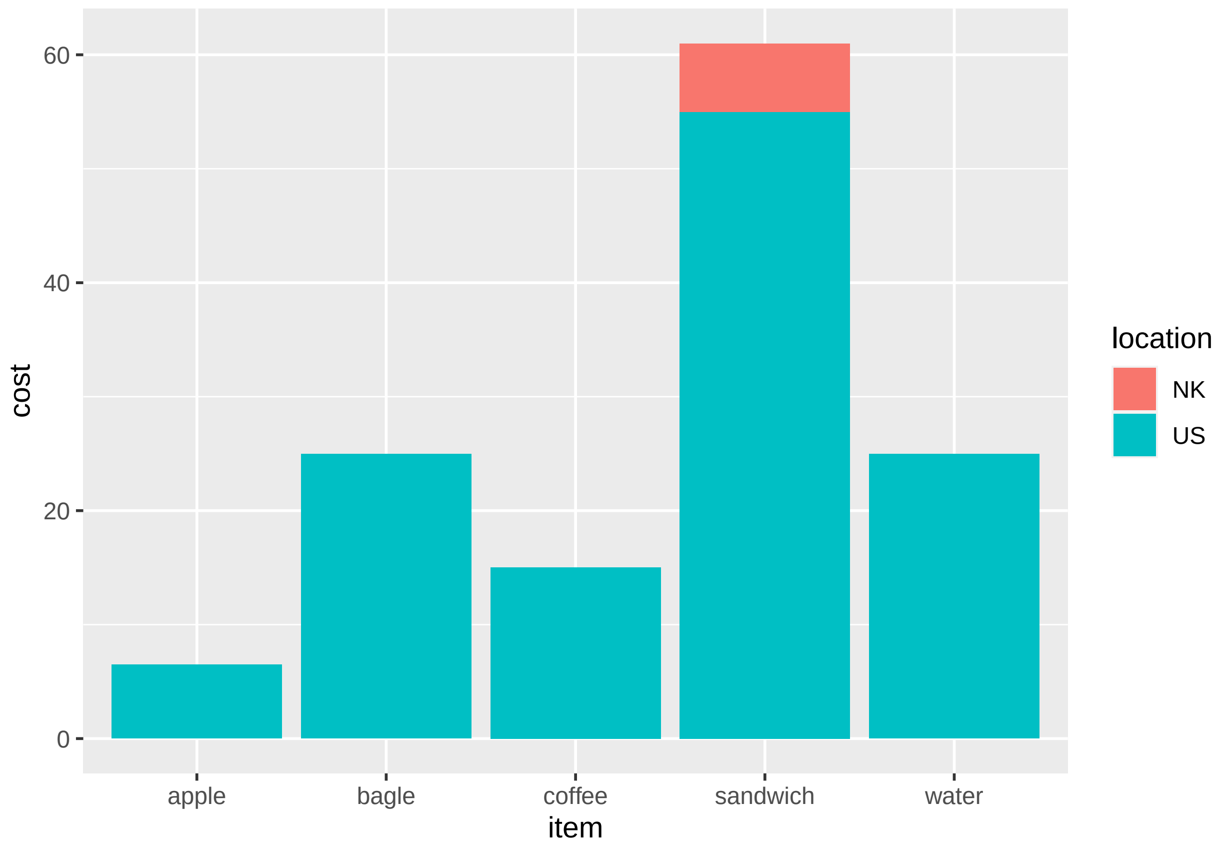


Try plotting it color change based on prices

The light blue at the top of the sandwich bar stand out, filter the data on item==sandwich

```
##        item cost     scale
## 1  sandwich  5.5 0.1124184
## 2  sandwich  5.5 0.1124184
## 3  sandwich  5.5 0.1124184
## 4  sandwich  5.5 0.1124184
## 5  sandwich  5.5 0.1124184
## 6  sandwich  5.5 0.1124184
## 7  sandwich  5.5 0.1124184
## 8  sandwich  5.5 0.1124184
## 9  sandwich  5.5 0.1124184
## 10 sandwich  5.5 0.1124184
## 11 sandwich  6.0 0.1078973
```

Make a request to the accounting department to get location data and replot



BAM!!! The anomaly found, the $6 cost wasn't done by the customer

Read world problems and cases

Sample dataset, creditcardfraud This dataset presents transactions that occurred in two days, where we have *492* frauds out of *284,807* transactions. The dataset is highly unbalanced, the positive class (frauds) account for **0.172%** of all transactions.

[Fradulent_cases] that went to trial (https://www.fincen.gov/resources/law-enforcement/case-examples?field_tags_investigation_target_id=678)