Column Boxplots

Boxplots for each Data Column:

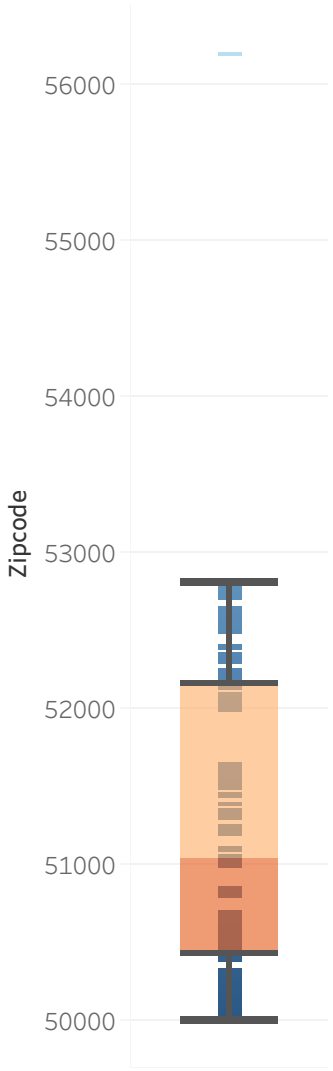
Note: Red lines indicate extreme outlier tails that have been omitted. These indicate a very skewed distribution of points.

Column Boxplots

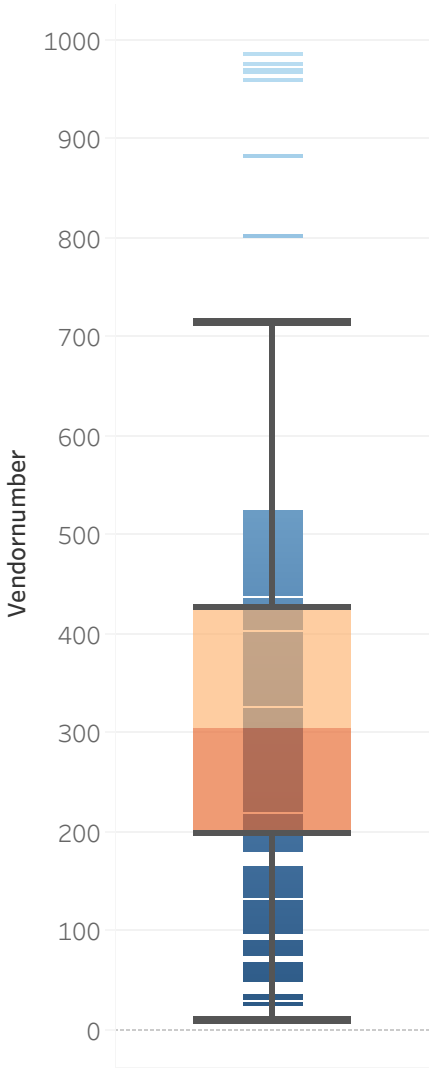
StateData

These measures are bounded, and stay in a predictable range and scale.

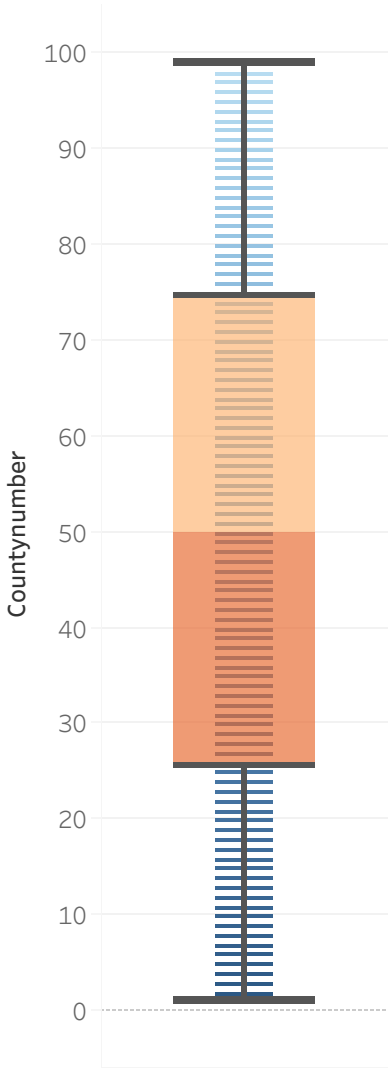
Zipcode



VendorNumber



CountyNumber

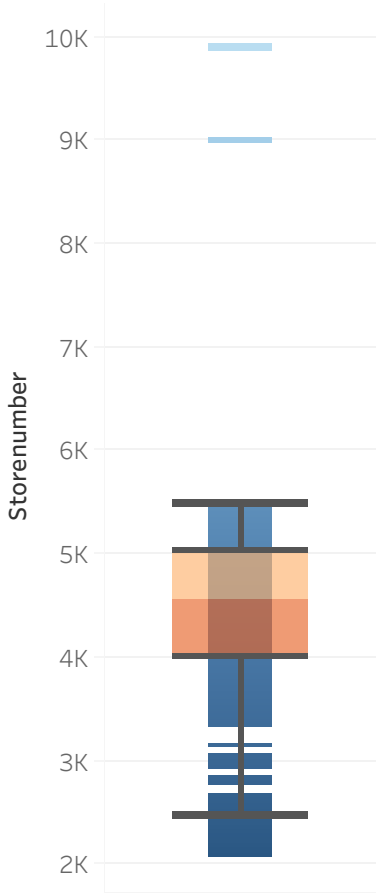


Column Boxplots

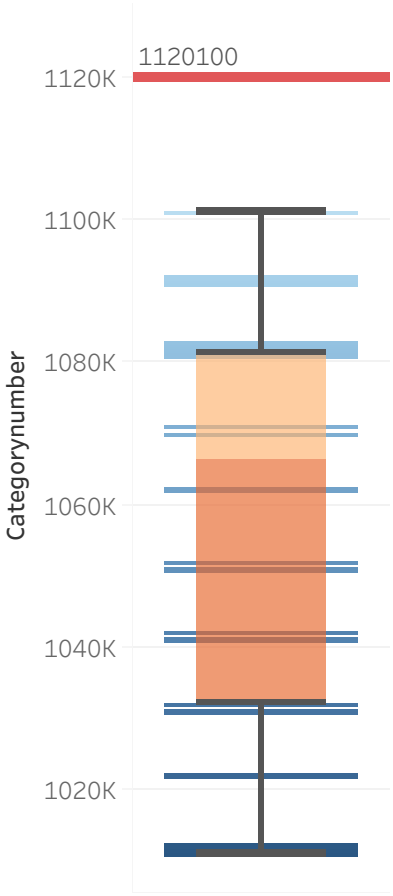
StoreItem

Data anomolies noted in all three measures:(1) **Store Number:** Massive jumps and gaps, not continually alloted, or many stores closed. (2) **Catagory Numbers:** Live in discrete zones, have large jumps to accomodate subcatagories (?). (3) **ItemNumber:** Huge jumps in two sets of numbers. New and old numbering system? **Q:** Which products are in the two different catagories? Which items are in the two different item regions?

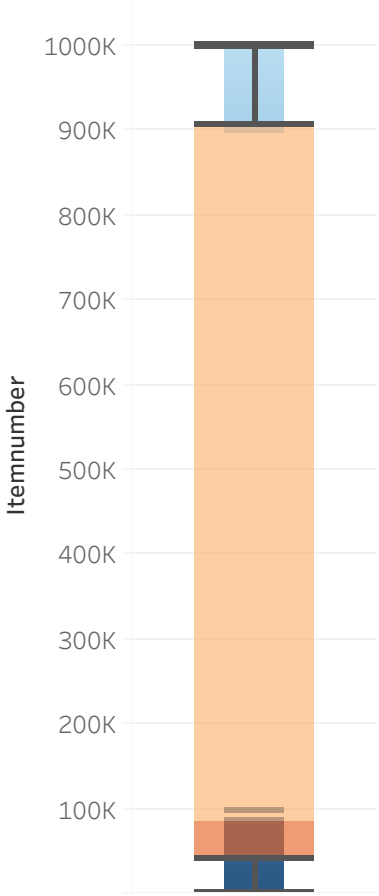
StoreNumber



CategoryNumber



ItemNumber

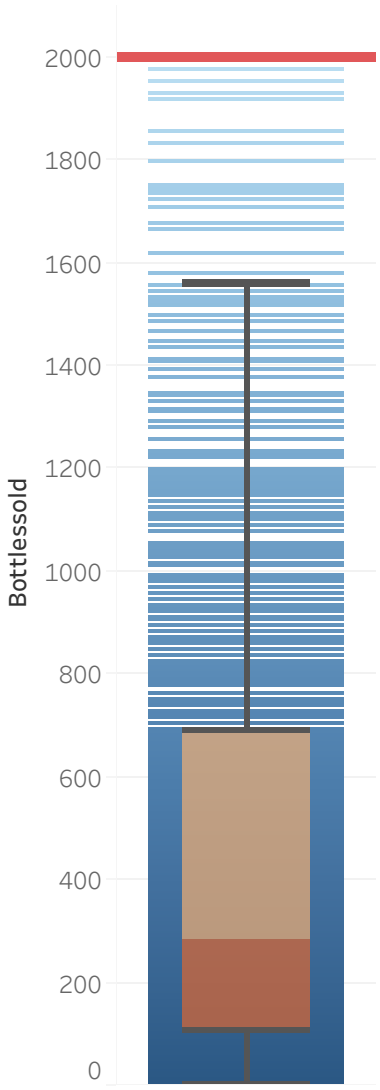


Column Boxplots

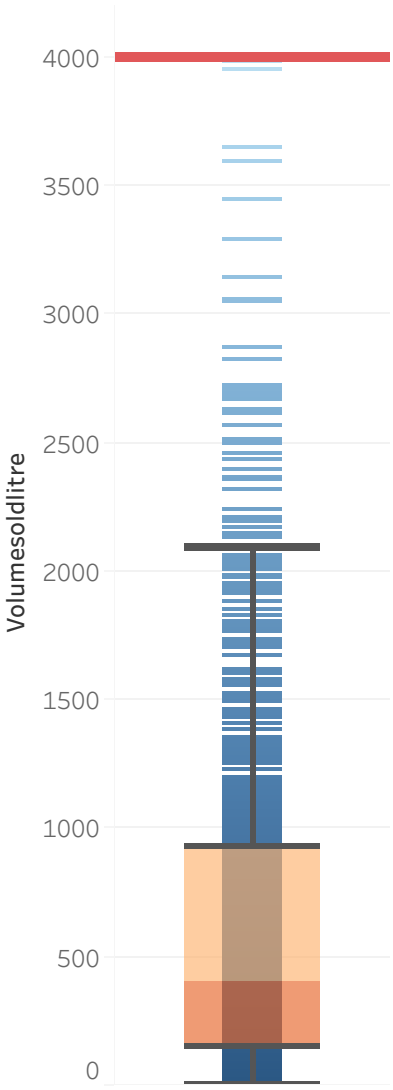
VolumeBottles

Here, all measures have have very large, dense distributions of outliers, they have been cut out to visualize the boxplot more. Q: What do these outliers represent?

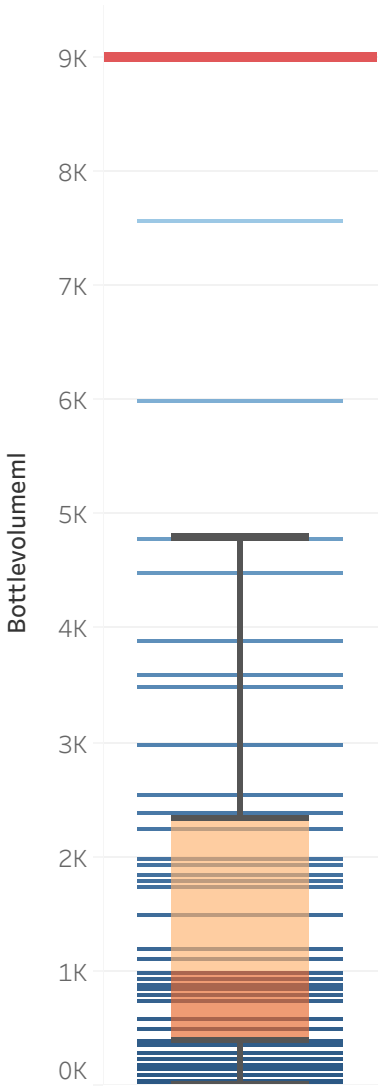
BottlesSold



VolumeSoldLitre



BottleVolumeMl



Column Boxplots

Sales

Again, we see massive outliers in the dataset. They have been clipped, as indicated by the red line, to see our boxplot better. Q: Verify data integrity by using volume, price and bottle sale data. What catagory are the very large sales in? Special promotions? Or Huge orders?