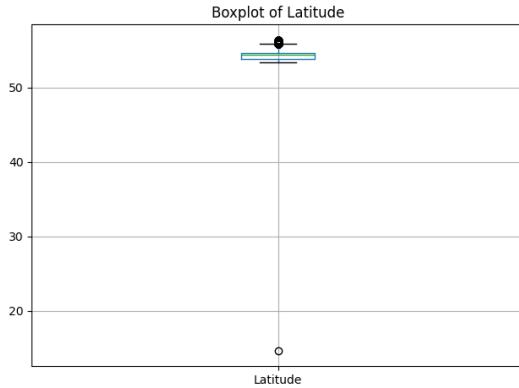


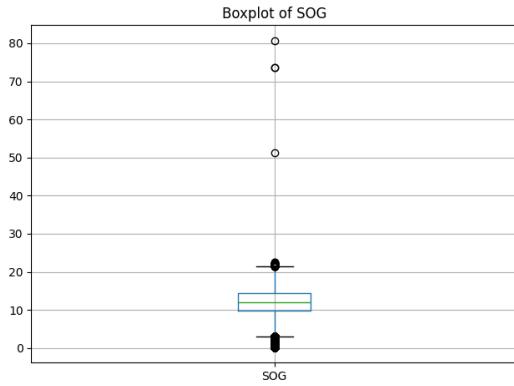
Data analysis

This report focuses only on outliers, missing values and other things needing to be cleaned – distribution is analyzed in profile.html.

1. 116 213 duplicates
2. Breadth and length: 10924 zeros
3. Latitude: one outlier

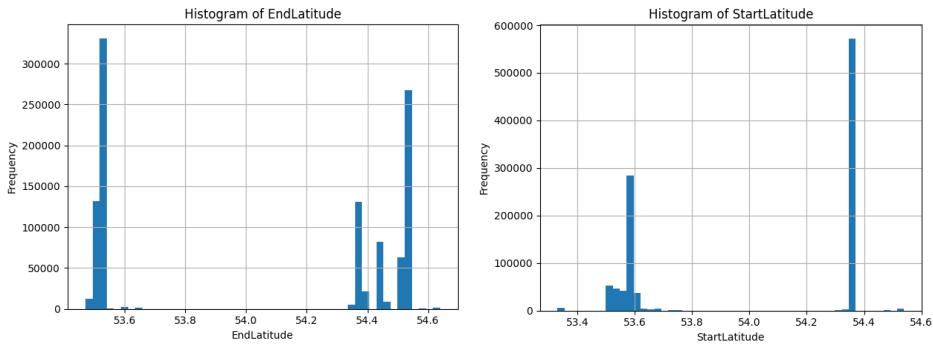


4. SOG: 3 examples of superspeed ships

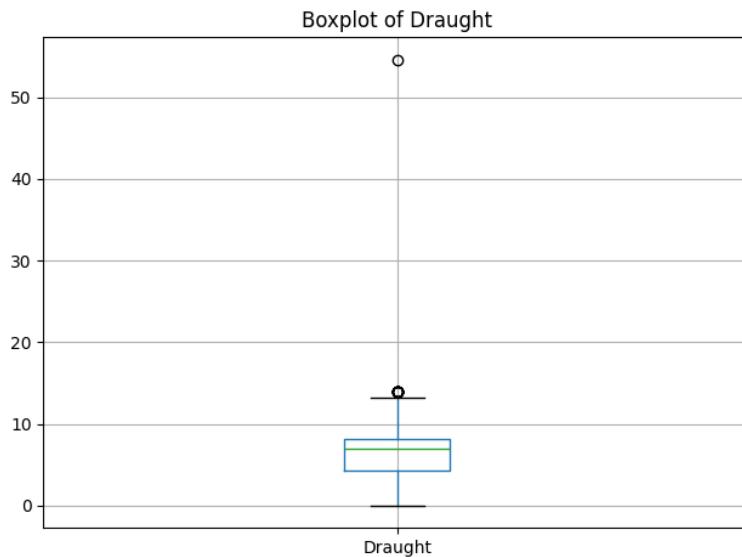


5. Why do starting and ending latitudes and longitudes differ?

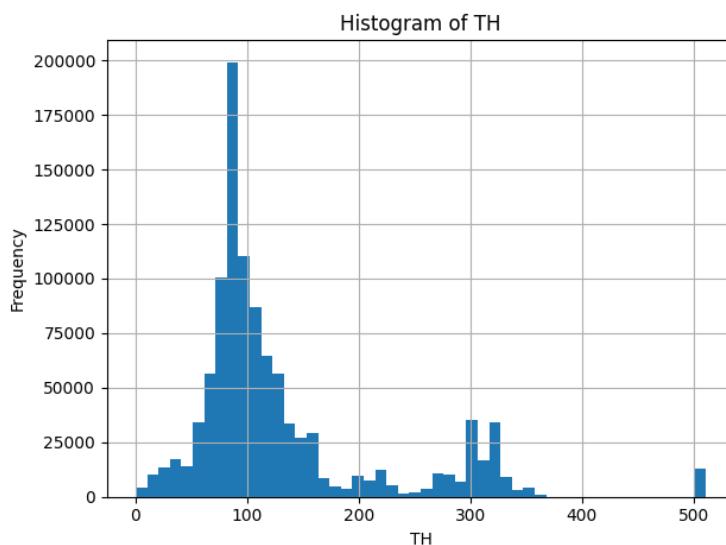
Because they start in different docks (at least let's suppose so)



6. Draught: sometimes a ship becomes Titanic (or submarine). 16 270 missing values – to be removed. 0 and 0,1 to be removed



7. TH: should only have values 0-360 but have 500. Over 360 need to be removed



How do you read box plots? The green line is a median. Blue lines are first and third quartile. The black lines are called whiskers, and to calculate them we need also IQR: distance between blue lines. So, the bottom whisker is $Q1 - 1,5 \cdot IQR$, and top whisker is $Q3 + 1,5 \cdot IQR$. The dots are outliers.

How to interpret box plots: clear outliers are errors and need to be removed. Not so clear outliers are possible and will be checked later, but error-threshold will be over them.

Coordinates: more than 5 away from median are errors. (not anomalies, but errors)

Values constant for a trip change during the trip.

Length and Breadth: 0 are errors.

Draught: 0 and 0,1 are errors

SOG: over 40 are errors.

COG and TH: valid values are between 0 and 360.

ANOMALIES:

Long stop in the middle of the sea (more than certain distance from harbor (start and end destinations))

Frequent or drastical course change (Over X degrees on this Latitude/Longitude in comparison to other examples/trips)

Frequent or drastical speed change (Over X knots, more than certain distance from harbor)

Frequent or drastical draught change (Over certain distance)

Latitude and longitude that are more than certain distance from usual coords on the same trip route.

Starting and ending point slightly differ from usual values of startPort and endPort.