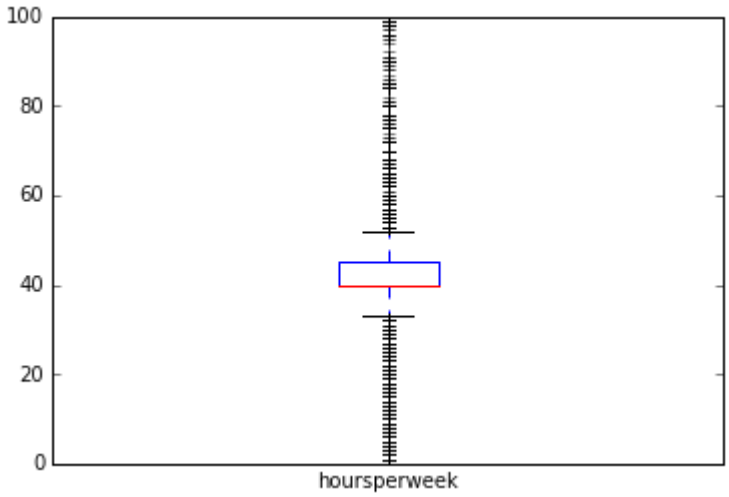
Outliers

This lesson explains what are outliers, why they happen and how to remove them.

WE'LL COVER THE FOLLOWING

What is an outlier?

Another area of cleaning can be dealing with outliers. First off, how do you define an outlier? This can require domain knowledge as well as other information, but a simple way to start is by taking a look at box plots:



Box Plot of Hours Per Week

The above plot was calculated with this command:

```
bbox = train_df['hoursperweek'].plot(kind="box")
```

Detection of an outlier

Here, anything outside the “whiskers” could be considered an outlier. As a refresher, the “whiskers” are the lines sticking out from the box and are 1.5 times the interquartile range. The interquartile range is the distance between the 25th and 75th percentiles.

Here is some code that will calculate the necessary cutoff numbers for all your number variables in a dataframe:

```
q_df = train_df.quantile([.25, .75])
q_df.loc['iqr'] = q_df.loc[0.75] - q_df.loc[0.25]
q_df.loc['whisker_length'] = 1.5 * q_df.loc['iqr']
q_df.loc['max_whisker'] = q_df.loc['whisker_length'] + q_df.loc[0.75]
q_df.loc['min_whisker'] = q_df.loc[0.25] - q_df.loc['whisker_length']
q_df
```

This code assumes your data are in a dataframe called train_df. The first line calculates the values of the 25th and 75th percentiles. It then takes the difference between the two to get the interquartile range (iqr). We get the length of the whiskers by multiplying the iqr by 1.5 and then calculate the min and max whisker value by subtracting and adding the whisker length from the 25th and 75th percentile values. Your output from this code will look something like this:

	age	fnlwgt	educationnum	capitalgain	capitalloss	hoursperweek
0.25	28.0	117827.0	9.0	0.0	0.0	40.0
0.75	48.0	237051.0	12.0	0.0	0.0	45.0
iqr	20.0	119224.0	3.0	0.0	0.0	5.0
whisker_length	30.0	178836.0	4.5	0.0	0.0	7.5
max_whisker	78.0	415887.0	16.5	0.0	0.0	52.5
min_whisker	-2.0	-61009.0	4.5	0.0	0.0	32.5

In this example, we can look more closely at the numbers for “age.” If we use the idea that any value outside of the min or max whisker values is an outlier, then any age greater than 78 or less than -2 would be considered an outlier.

Another simple way to try and identify outliers is by taking some multiple of the standard deviation of your values, usually 2 or 3 times. For example, if the standard deviation of age were 13 and the mean were 30, and you decided anything 3 times the standard deviation could be an outlier, then any age

greater than 69 (13x3 + 30) or less than -9 (30 - 13x3) would be outliers.

For age, we can also use domain knowledge and know that anything less than 0 and greater than 130 are clearly outliers.

All this being said, there are many ways to detect outliers and these are some of the simplest first attempts you could take. These methods look at the feature by itself, sometimes your outliers are multi-variate, a weight by itself might not be an outlier, but for a given height, it is. Here is a post that deals with some more advanced ways of detecting these:

https://www.kdnuggets.com/2017/01/3-methods-deal-outliers.html

Removal of an outlier

Once you think you have found some outliers, though, what should you do?

Like missing data, you need to ask yourself, why might these outliers be here? If they are actually part of your data, then likely you will want to keep them. By removing them, you would skew your training away from the truth. If they are erroneous, for example, ages over 500, you can feel safe dropping those rows if you have enough data or perhaps input a better value. Or even better, try to discover how these bad values came to be and fix them. It could be that your data pipeline broke, but the raw data is still good.

Lastly, it is always good to consider the data generation process. Do your data come from a sample? If so, maybe it isn't very representative, so things that look like outliers in your sample are actually not in the population.

There are many factors to consider and these are only some. The main point is to be thoughtful of your outliers, ignoring or dropping them without thought is usually not the best idea.