Simple outlier detection

Your task to compute and visualize outliers in a time series, where you need to prepare and aggregate the underlying raw data upfront.

Language, libraries

Python, preferred data analysis library, preferred visualization library

Data

The underlying data is derived from a former Kaggle competition, that was about predicting "Click-Through-Rates" (*CTR*). *CTR* is defined as *clicks* divided by *impressions* and it measures how often advertisements are clicked relative to how often they are shown.
The data is available here: https://www.kaggle.com/c/avazu-ctr-prediction

CTR over time

The data set includes a lot of dimensional fields with (categorical) information about the environment (device, location, etc) but we are only concentrating on CTR and time series relevant fields, such as

- *click*: 0/1 for non-click/click
- *hour*: format is `YYMMDDHH`, so 14091123 means 23:00 on Sept. 11, 2014 UTC.

The first task is to aggregate data by "*hour*", calculate *CTR* and plot the resulting time series.

Outlier Detection

Second, build a simple outlier detection algorithm based on a "moving average". A data point is identified as an outlier, if it is more than 1.5 standard deviations apart from its calculated moving average (for simplicity's sake, we will assume a Gaussian distribution here).

The outcome of this task is a plot, that highlights all found outliers.

GitHub

Please create a GitHub repository and upload your code. Also create a README so that we can easily clone and run your project.