Test 1

In the lambda function, I define a customed exception called `SmallerException` which is caught in the Step Function when the lambda function fails to run. Please replace the **Resource** field to the arn of the lambda function when testing the Step Function.

Test 2

I only test the 2nd script on AWS Athena. Some issues are observed during my testing.

1. When I use `split` function to get the file name, Athena complains array out of boundary. I end up to use `split_part` function.
2. It returns false when I run `strpos(useragent, 'Mozilla') = 1` for those records using browser in Athena. I test the same in Google Cloud BigQuery, it returns true.
3. I can't find the timezone format '`%z`', so I use `concat` to concatenate the datetime and timezone.

Test 3

It is assumed that the given S3 bucket has been set up to trigger the lambda function when a file is uploaded.

This lambda function is fired when one or more XML files are uploaded to the given S3 bucket. It will perform,

1. Download each file to tmp folder
2. Parse each XML file, retrieve the required fields and write to a JSON file
3. Upload the JSON file to the processed folder

Test 4

This residual versus fitted value plot shows residual values increase as the size of the fitted value increases.The residuals are getting larger and larger as the value of the response increases. Ideally plotting residuals versus the value of a fitted response should produce a distribution of points scattered randomly about 0, regardless of the size of the fitted value.

The plot implies R-squared, the proportion of the variance in the dependent variable (response) that is predictable from the independent variables (regressors), is pretty low. This model tends to be underfitting, which is not good for business use as a rough approximation. One possible solution is to transform the response, by modeling its logarithm or square root, etc..