Postie Data Challenge Description and Goals

The data science take home challenge is intended to help us evaluate how you perform on a task reflective of typical data science work at Postie. The test centers on being able to access data in AWS, read it, figure out and document any problems with it, transform it, then create meaningful predictions from the data. In addition to the technical work, we are looking for inquisitiveness and an ability to try to figure out what is important with a given dataset and communicate those appropriately. Please feel free to ask questions via email to data@postie.com.

Keep in mind that this is more an exercise in critical thinking than a Kaggle competition. All of the above will be used as evaluation criteria.

Scenario:

It is July 3rd, 2017. A data analyst just started at your company on July 1st, and has done an analysis to figure out the sales for 2017-07-03 were $164065.00.

Technical Details

The system log data is located publically via http in `us-east-1` on AWS S3 in the `postie-testing-assets` bucket.

Each file contains rows with the following data:

- `timestamp` - the date of the transaction
- `website_id` - the integer value of the website the transaction took place on
- `customer_id` - the integer identifier for the customer
- `app_version` - the version of the software
- `placeholder` - a placeholder column
- `checkout_amount` - the amount in dollars of the complete transaction
- `url` - the full website url that instantiated the transaction, with product names and product count params

Analysis

1. The analyst wants you to figure out why their 2017-07-03 sales value is much lower than their previous day sales summary.
2. The new analyst would also like a report detailing other key metrics for the system. Other than the average sales value per day, are there any other metrics that they should be using? Is an average sales value the right metric to use? Explain why/why not.

3. What information can you extract from the urls? Can you infer all product prices? Is there any other information that you believe would be useful to understand what is going on?
4. Are there any interesting purchasing combinations, events, or metrics that are worth reporting and displaying? What information should the analyst know about the system that you've uncovered?
5. Can you predict a the total sales number (in dollars) for 2017-07-04? How certain is the predicted number? State explicitly what the prediction is doing, and what general steps you did to get the number you report and what assumptions you have made.
6. Is there any additional information, data or access that would make your prediction better?

Results and Formatting

Python is the preferred language of the challenge, but you may use any language you are comfortable with (e.g. R, Excel, Julia, Scala, etc.). Also any offline/non-programmatic manual file editing is great too.

Please create an analysis and report that answers the above questions. A Jupyter Notebook (please leave relevant viewable cell outputs) is preferred, but feel free to submit the report and analysis in any format (.xls/.pdf or whatever). Regardless of the format, please make sure that it contains explicit answers to the above questions, any code written (with relevant commentary), relevant graphs, and pertinent explanations. Any data produced/edited should also be submitted along with the report.

To submit you challenge, please email the analysis and results with a note to data@postie.com ; Feel free to include any metadata about how you performed the challenge in the email body.

There are two ways to deliver the code and files to us:

- tarball - please tar and gzip your project folder. Attach it to the email or include a downloadable url.
- Github - please create a publically viewable repository and link to it.

Any meaningful feedback is also appreciated!