Code Challenge V2

CreditShelf Code Challenge V0.2 Java/Python Senior Developer

Welcome to CreditShelf code challenge.

In Creditshelf we strive to challenges us each day for making more of our data

We hope you find this challenge interesting, but most important, it leads the way for you to become a Senior Developer at Creditshelf.

The challenge doens't have any starter code. Please make your own environment that can be reproducible by us (use virtual environments or containers).
The objective of this exercise is to create a small tool that can show the distribution of the crashes of bicycles relative to bicycles stations.

- The challenge is compose of 5 different stages. Each stage describe one key skill that a Data Engineer in Creditshelf has.
- The challenge is not been to be full finished but of course, the more you achieve, the better.

What to Deliver?

You need to deliver a link of a github public repository with:

- Base Code
- README.md with instructions how to run it.

Stage 0

1. Extracting from external sources `https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions-Crashes/h9gi-nx95`. First step is to do some data mining in the dataset.

TODO

Create an Script that given a `BOROUGH` it give the collisions information for injured killed cycles.

Stage 1

2. Loading the data
   The second stage we are going to store the information in an external database (with a proper data model).

Recommendation:

- Try to think if you need more than one entity to store the information.

TODO

Using the script that you created before to insert the data in a data store.

Stage 2

3. Visualization
   The third stage is to visualize the data.

TODO

Create website that show information of where the crashes happend using a map
extra: filter by BOROUGH the crash accidents

Stage 3

4. Analytics I
   Save bike stations information `https://s3.amazonaws.com/tripdata/index.html` from 2019. The file names are `201901-citibike-tripdata.csv.zip` where 2019 is the year and 01 is the month.
   Aggregate the files to have only the stations used in one year. The data source has bikes rides information that contains the station name and id where the user grab the bike and the end station where he left the bike (start_station_id, end_station_id)

TODO

Create an script for extracting the information from the bike stations and persist the data in a database.

Stage 4

5. Analytics II
   Show the information of the bycicle stations in the same map and how close are they are from a crash place.

Recommendation:

- Use visualization techniques for showing how close are they from the crash station.

TODO

- Show the bicycle stations in a map.
- On select one of the stations you can see how close are they from a crash place.

Final Words

Good luck with the project and We hope this can be your first project in CreditShelf!!!