Data Mining report

Preprocessing part

Issues we noticed:

- each datasets has only two features that include null values `etablissement_postal` and `next_etablissement_postal`
- more than 60% of `next_etablissement_postal` is null.
- some routes of packages are not logical example:
    - a package goes from `etab_0001` to `etab_00031` and in the next time, it goes from `etab_0099` to `etab_0007` (the dataset is sorted by `date`).

Our preprocessing:

- We decided to drop null values of `etablissement_postal`, as they're just 2.7% of the whole dataset.
- We agreed on filling null values of `next_etablissement_postal` for each package/receptacle with the next value of `etablissement_postal` for that same package/receptacle.
- The remaining null values for `next_etablissement_postal` (concerning the last route for each package/receptacle, as it doesn't have any next route to be used for filling null values) are filled with the `etab_....` that appears the most frequently with the `etab_....` in `etablissement_postal` column of the same row, example:
    - if the package's last route has a null value in `next_etablissement_postal`, we can't see the next route for this package to fill the null value as it doesn't exists, instead, we iterate through the whole dataset and count for number of `etablissements` that appears in a route with each `etablissement`, and we use this information to fill remaining null values of `next_etablissement_postal`.