|  |
| --- |
| **Project | Cement sales** |

For this project, we use a varied dataset: mainly product sales data from an industrial company. As our objective was to analyse the impact of logistics costs on discounts. For this we used the data to determine the distances to customers from the shipping points.

**Sources**

Our initial idea was to get distance information through the google maps API, this access worked initially. Eventually we had to accept less precise distances.

**Data Cleaning process**

The cleaning process consisted of identifying outliers and missing values between records. We also made sure that the fields were formatted in the way we expected to be able to make the comparison we were hoping to make. Our biggest challenge is certainly to be able to calculate distances using the postcodes.

**Data Analysis**

We have two sources of data, on the one hand sales data and on the other hand geographic information data. We obtained the distances to the customers and we will analyse the quantities and distances in relation to the final prices (discounts). For this we have to take a segment that is homogeneous, i.e. that has similar characteristics in all operations.

**Hiphotesys test**

Our hypothesis is that discounts increase as delivery distances increase. Once we had the clean data we performed 3 hypothesis tests.