

Data Science projects

Hamza Imloul

February 3, 2022

1 Introduction

These projects are developed using R/Python. To view tutorials, user guides, and further documentation, please visit my [portfolio](#).

2 Machine learning

2.1 NLP: Sentiment Analysis using NLTK

Run a lexicon-based sentiment analysis (eg. NLTK Vader Sentiment Analyser) on the textual data, then report and discuss the results. Does the lexicon sentiment score associate with the venue ratings provided by the users?

Pre-process the text review data and create a new column in the data frame which will hold the cleaned review data.

Removed punctuations and digits characters, lemmatization, Stopwords, Stemming, and synonyms.

Built a supervised learning model for text analysis.

Split the dataset into a train and test-set.

Structured the data set of textual strings.

Method 1 : **Bag of Words**.

Method 2 : **TF-IDF** (term frequency–inverse document frequency).



Figure 1: Reported waste in the transportation process.

2.2 Machine Learning Analysis with Venue Review Data in Calgary

Tags: Geospatial Analytics

For this second task, we would like you to analyse a dataset that contains review data of different venues in the city of Calgary, Canada. With the help of several machine learning techniques that we have learnt in the course, you will be tasked to distill insights from this social media dataset.

Two of its notable features are the geocoding of every reviewed venues and the availability of a considerable amount of text data in it, which lend to its ability to be processed using spatial and text analysis techniques respectively.

As a prelude to the analysis prompts below, have a brief think about some of these questions:

What can we discover about the venue review data?

Are there any spatial patterns that can be extracted from the data?

Can we build a machine learning model that predicts review rating for unseen data points using the text of the reviews?

2.3 Anomaly detection tool exposing the wasting time in the process

dddd

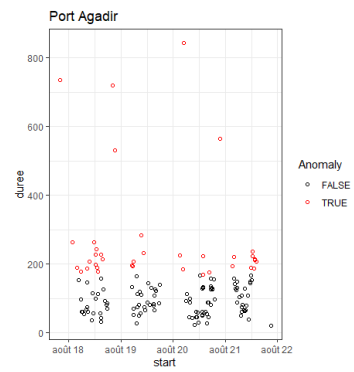


Figure 2: Reported waste in the transportation process.

2.4 Driver Behavior clustering

In this project, we tried to create a statistical model to cluster driver behavior based on CAN Bus sensors data.

We will use Hierarchical clustering to identify and group the drivers based on their behavior and driving style. This identification of drivers can be used for improvements.

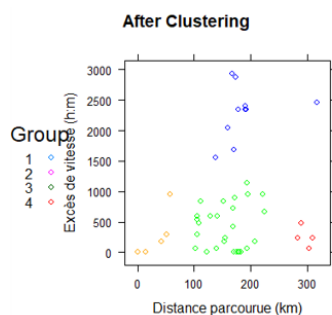


Figure 3: Reported waste in the transportation process.

3 Operations research

3.1 Decision support tool for the renewal of equipment.

Saved operational and maintenance costs.

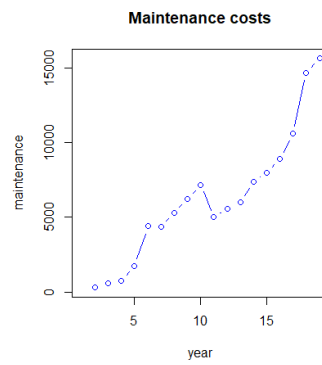


Figure 4: Reported waste in the transportation process.

3.2 Data pre-processing tool, of geofencing events dataset using R

Data preparation, is the most important part of the data analysis process, increasing value from data. Importing data set, understand the business problem, and select variables in the output.

Using Rest API calls to request data from servers of both platforms, it is the best approach since the data received is raw, in a JSON format, and not consuming storage resources. But requires a specific knowledge of HTML request and Restful APIs structure.

Helpfully, a useful method to manage and keep the work organized is to work with geofencing alerts since the two platforms send them by e-mail. And by that way, the mailbox can be considered a centralized fleet management platform specifically for the pickup, travel and delivery operations.

The inconvenient of this technic is that every geofencing alert mail is incorporating a descriptive text and image, what consume storage and wireless network data. In addition, data retrieving demands an appropriate program to connect with mailbox provider server, and the use of regular expressions to extract the needed information from every mail. But for now, it still the most affordable technic, since it provides us the needed and real-time updated information.

The number exceeds the speed limit.

Rate of call responsiveness.

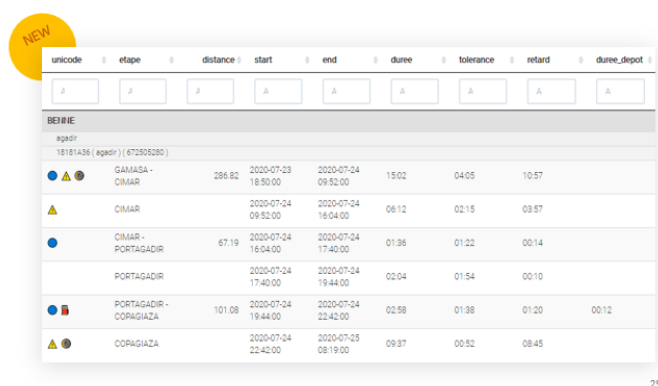
The number of times when the driver started after the planned starting time in a group.

The number of days off in a month.

Exceeding driving time limit (9 hours).

Brutal acceleration.

Brutal braking.



unicode	etape	distance	start	end	duree	tolerance	retard	duree_depot
18181436 (agadir) (67205280)								
● ▲ ●	GAMASA - CIMAR	286.82	2020-07-23 18:50:00	2020-07-24 09:52:00	15:02	04:05	10:57	
▲	CIMAR		2020-07-24 09:52:00	2020-07-24 16:04:00	06:12	02:15	03:57	
●	CIMAR - PORTAGADIR	67.19	2020-07-24 16:04:00	2020-07-24 17:40:00	01:36	01:22	00:14	
	PORTAGADIR		2020-07-24 17:40:00	2020-07-24 19:44:00	02:04	01:54	00:10	
● ●	PORTAGADIR - CORAGIAZA	101.08	2020-07-24 19:44:00	2020-07-24 22:42:00	02:58	01:38	01:20	00:12
▲ ●	CORAGIAZA		2020-07-24 22:42:00	2020-07-25 08:19:00	09:37	00:52	08:45	

Figure 5: Reported waste in the transportation process.

3.3 Fleet assignement using R

Logistics touches every part of a business and even goes beyond to include suppliers, carriers and customers. Needless to say, the logistics are complicated with many interrelated components. This high degree of complexity is difficult to manage without the use of models that faithfully represent the processes and their interactions. Reduce costs and downtime. In this Article, we will apply Hungarian Algorithm, which is an optimization algorithm.

Modelling the problem After applying design research, some If you are not familiar with trasnportation problems, you can check this good website.

Algorithm We use the Hungarian algorithm, Assign n missions to n available trucks E_{rv} set of available real vehicles.

E_{dv} = set of dummy vehicles.

E_s = set of sources of missions.

E_d = set of destinations of missions.

Cost matrix

$$c_{vp_{ij}} = cf_v(d_{pi} + d_{ij}) + ca_{ij} - p_{ij} \quad (1)$$

where

$c_{vp_{ij}}$ = the cost to assign the truck v located in p to the task i to j.

cf_v = cost per km of fuel consumption for the truck v.

ca_d = cost per km of the resource (driver) allocation.

d_{ij} = distance traveled to perform task from the source i to the destination j.

d_{pi} = distance traveled to perform task from the truck position p to the source i.

b_{ij} = bonus of driver to performs the mission i to j.

p_{ij} = profit of transportation from the source i to the destination j.

the vp_{ij}^{th} element is the cost of assigning the resource v located in p, to the task: source i to destination j. the matrix is 4-dimensionnal. We research to add penalty cost related to in-site time in sources and destinations, to avoid missions with a large cycle time. + Probability to blockage InSite due to infraction (wheels state. . .).

3.4 Modelled the parcels' routing time.

Problem: calculate the delay $J+i$ of delivery of a package in order to obtain the number of engaged packages to be delivered within $J+i$.

Inputs: datetime-depot, abb-depot, abb-destination, id-package.

data sets : intra-trips, inter-trips, Delays-matrix, Available-Trucks-capacity, packages-details.

3.5 Capacitated Pickup & Delivery Vehicle Routing of taxi service

Usually the template you're using will have the page margins and paper size set correctly for that use-case. For example, if you're using a journal article template provided by the journal publisher, that template will be formatted according to their requirements. In these cases, it's best not to alter the margins directly.

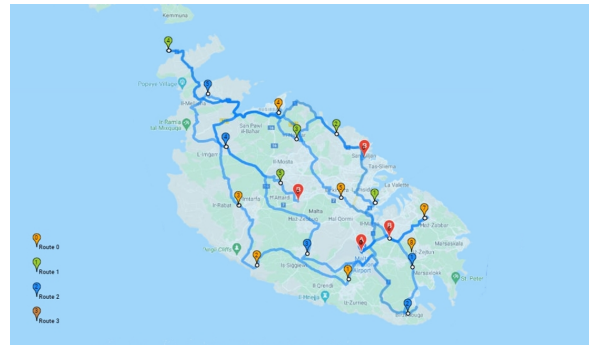


Figure 6: Optimized the VRP with multiple pickups and deliveries.

3.6 Optimized the routing of sales persons through schools in california

Real problem case.

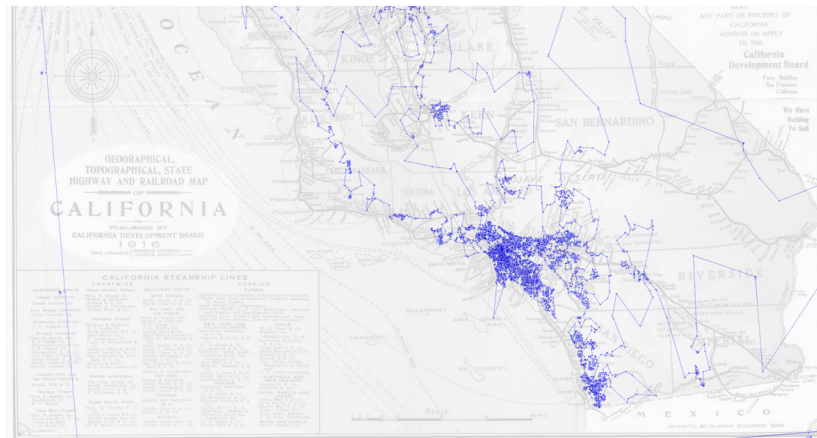


Figure 7: Optimized the routing of the sales persons.

4 Reporting

4.1 Designed a R-Shiny App to monitor the transportation traffic and mobility

We will develop a case study, which is a project I worked on for a Transportation company, managing resources of 102 trucks, and 91 drivers.

This project was initiated by the need of a real-time tracking tool for the moving resources. The result, as we see, is a spatial map with multiple layers to filter the units.

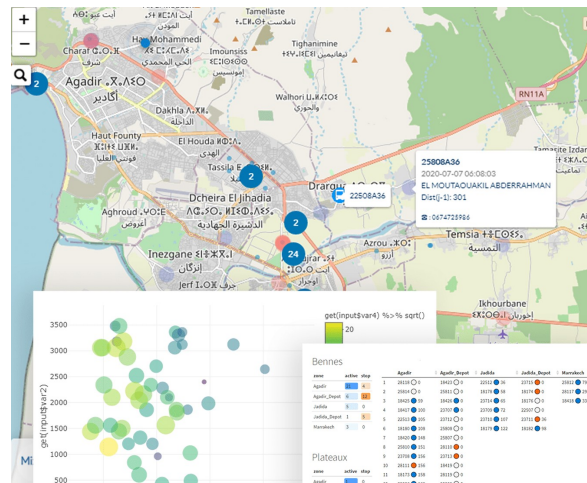


Figure 8: Tracking of vehicles status on map.

4.2 Automated Data processing for Trip metrics reporting

First you have to upload the image file from your computer using the upload link in the file-tree menu.

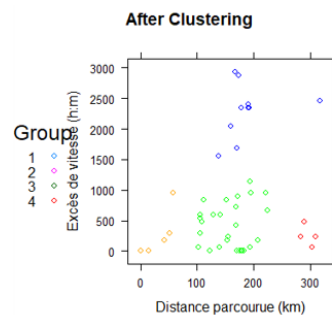


Figure 9: Reported waste in the transportation process.

Automated Data processing for Trip metrics reporting

4.3 Designed an interactive app visualize the weather data

Real problem case.

4.4 Data Analysis of the NanoString assay

Bro that's good

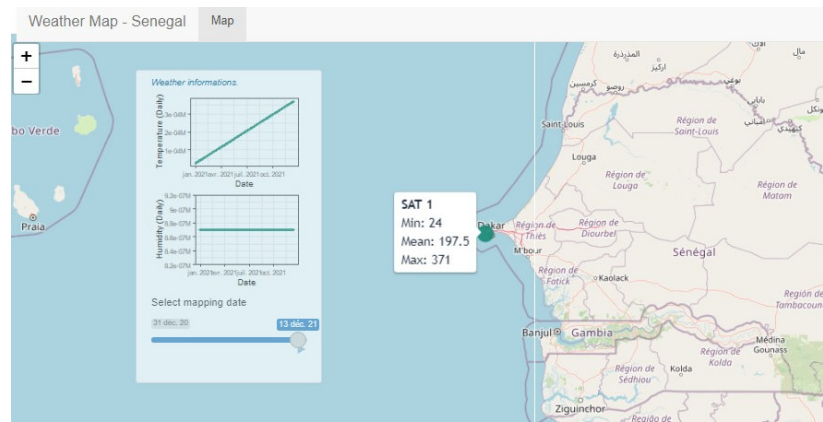


Figure 10: R Shiny app.

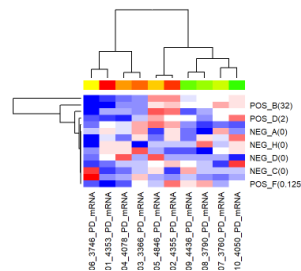


Figure 11: Reported waste in the transportation process.

4.5 Filling automation of delivery paper sheets

Real problem case.

text mining method to retrieve information from scientific journal papers on the related topics

Entering exit vouchers in a database is important since it allows the transport operations carried out to be justified. In this document, we propose a new input method, automating this task, and analyzing the advantages of this method.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	date	n_bon	societe	site	transporteur	vehicule	conducteur	cin	code	quantite	libelle	poids_vide	heure_date_e	poids_brut	heure_date_s
2	10/07/2020	3004052705	Ciments du Maroc	USINE AIT BAKA	MIXTRA	25807A36	SAID EL ABID	QA111352	130053	25700	Coke de pétrole brut MTS	17		43	10072020 114701
3	09/07/2020	3004052569	Ciments du Maroc	USINE AIT BAKA	MIXTRA	25807A36	SAID EL ABID	QA111352	130053	26020	Coke de pétrole brut MTS	17		43	09072020 223406
4	09/07/2020	3004052358	Ciments du Maroc	USINE AIT BAKA	MIXTRA	25811A36	ABDELHADI SISSIN	JTS0761	130053	26060	Coke de pétrole brut MTS	17		43	09072020 153028
5	10/07/2020	3004052668	Ciments du Maroc	USINE AIT BAKA	MIXTRA	25811A36	ABDELHADI SISSIN	JTS0761	130053	26340	Coke de pétrole brut MTS	17		44	10072020 101956

Figure 12: R Shiny app.

4.6 Designed a Power BI app for retail data

Real problem case.

Entering exit vouchers in a database is important since it allows the transport operations carried out to be justified. In this document, we propose a new input method, automating this task, and analyzing the advantages of this method.

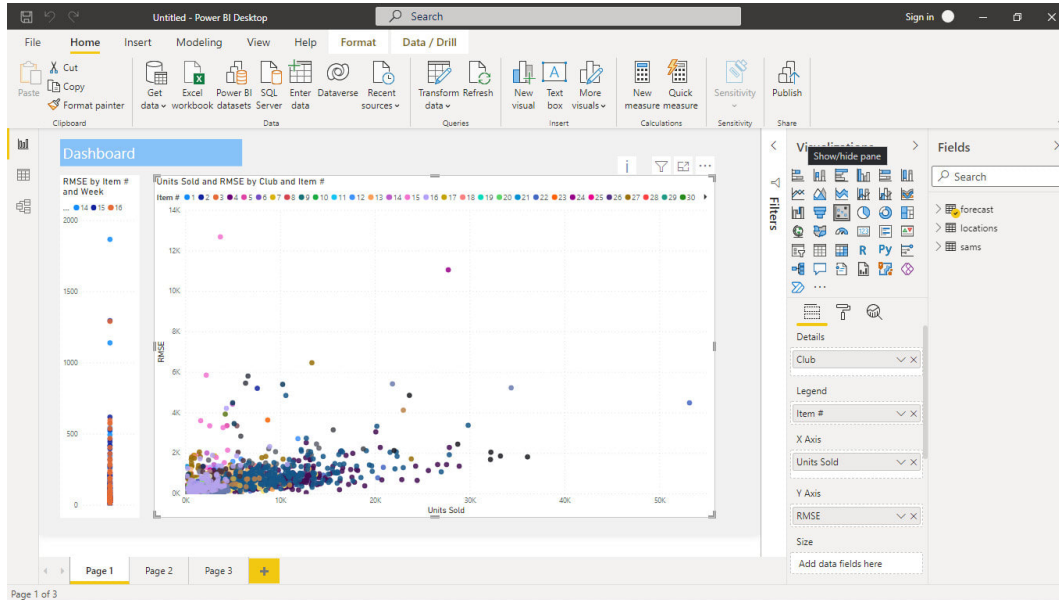


Figure 13: Power BI app.

5 Geospatial analysis

5.1 Mobility Patterns Analysis in Cambridge

Analysed the mobility patterns of users from Gowalla, a now-defunct online geo-social network from a decade ago.

On Gowalla, users were able to check in at different locations across the course of the day. The dataset that is provided to you (available on Moodle) is a subset of Gowalla users located in Cambridge, UK and, although with some personal identifiers of the users removed, you could trace the movements of particular individuals on certain days, according to their check-ins.

For two selected users, Calculated: the maximum displacement, the average displacement, and the total distance travelled on the day.

Route for User N°1.



Figure 14: Route for User N°1.

Comparative analysis of check-in frequencies and network centrality:

Urban Planning Application

to suggest some application to the urban planning constancy: We can suggest that cultural facilities could be built near the roads colored in yellow, which are close to the city center. And you can see that the centrality of the graph is close to the colleges of the University of Cambridge



Figure 15: network centrality in Cambridge

5.2 Stop & Search data analysis to predict crime outcome in London

The first task is to create four **Kernel Density Estimation** maps based on police stop and search data.

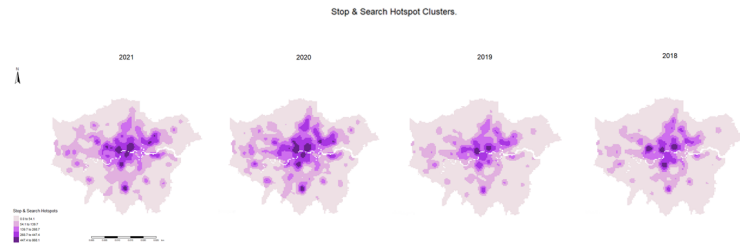


Figure 16: Kernel Density Estimation maps

5.3 Designed an app visualize the locations of a research center members

Real problem case.

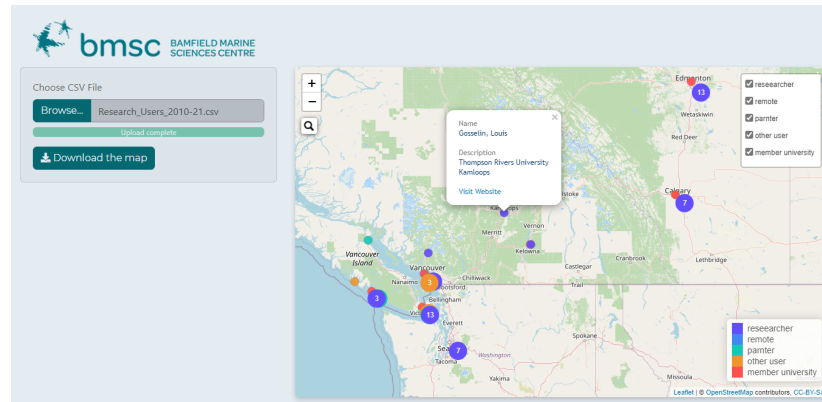


Figure 17: R Shiny app.