Montreal Road Collisions

http://donnees.ville.montreal.qc.ca/dataset/collisions-routieres

Contents

Introduction

The purpose of this project is to use acquired techniques.

In our project we will get use of:

- Docker images and stacks
- Volumes, containers
- Spark Clusters
- GCP
- HDFS, MondoDB and Parquet
- Pyspark, RDD.

Project Structure

Data:

- accidents.csv
- accidents_new.csv
- municipalities.csv
- final.csv

Code:

- data_cleaning.ipynb
- data_EDA.ipynb
- csv-to-parquet.py
- Write_features_MongoDB.ipynb
- Prediction_Subset.ipynb
- Prediction_and_model_export.ipynb

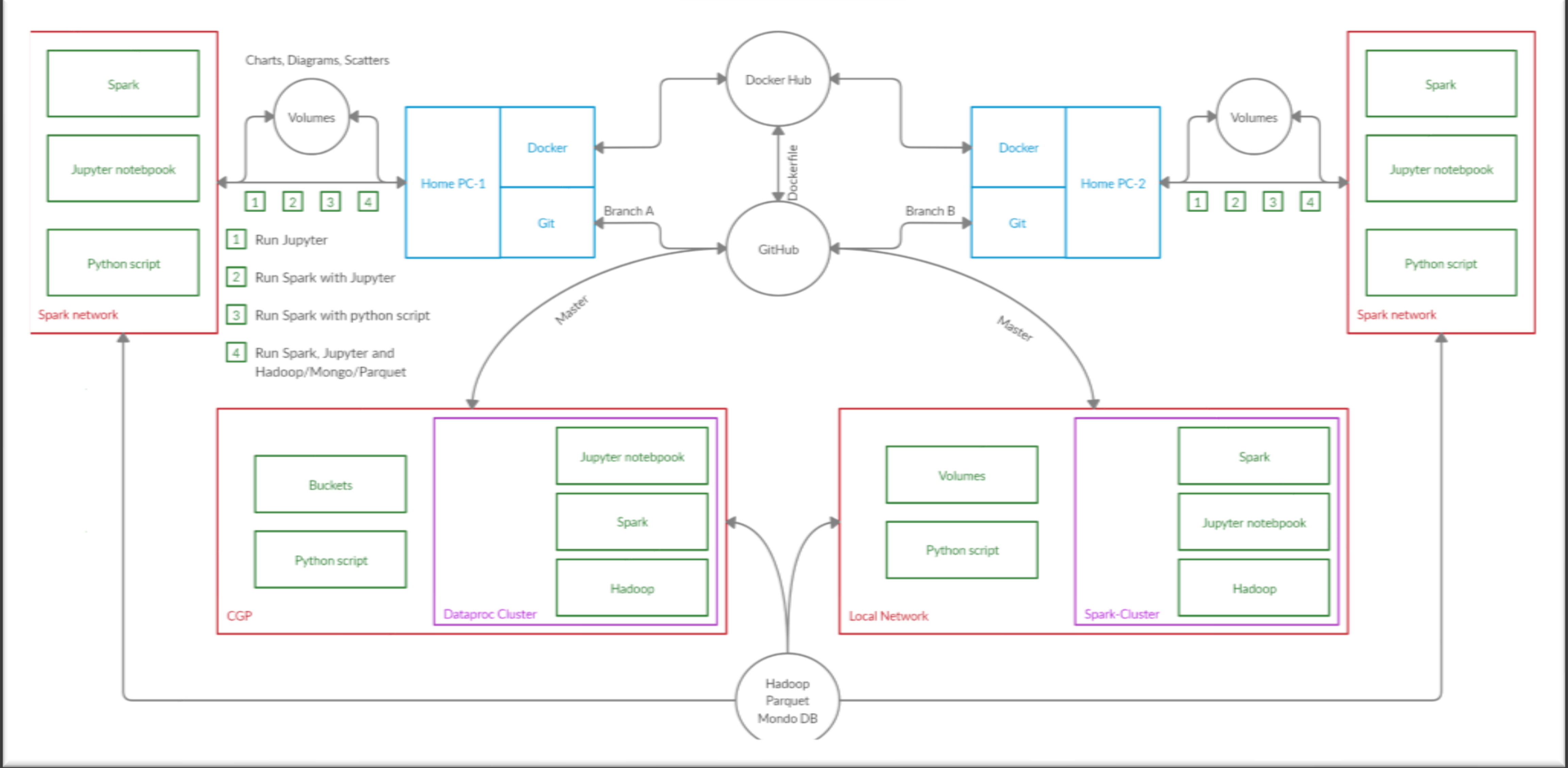
Packaging and Deploying:

- Dockerfile
- jupyter_local-compose.yml
- jupyter_bind-compose.yml
- jupyter-compose.ym
- spark-compose.yml
- spark_bind-compose.yml
- spark_bind_hdfs-compose.yml
- spark_hdfs-compose.yml
- spark_mongo-compose.yml

Prediction Script:

- pred.py
- model.py

Packaging and deployment schema



Techniques used

- Jupyter
- Jupyter + Spark
- Jupyter + Spark + Volumes
- Spark + Parquet + Hadoop
- Jupyter + Spark + MondoDB
- GCP Dataproc cluster
- Running script on a Spark cluster
- Running script on a Spark cluster from an image

Data Cleaning

- Reading the dataset and checking its structure.
- Choosing some columns and renaming them.
- Adjusting the types.
- Reading municipalities dataset and merging it with collisions.
- Exploring collisions in each municipality and doing some other basic explorations to better understand the data.
- Dealing with nulls: removing all the rows with unknown categories anyway.
- Writing the "clean" data to another csv file.

EDA

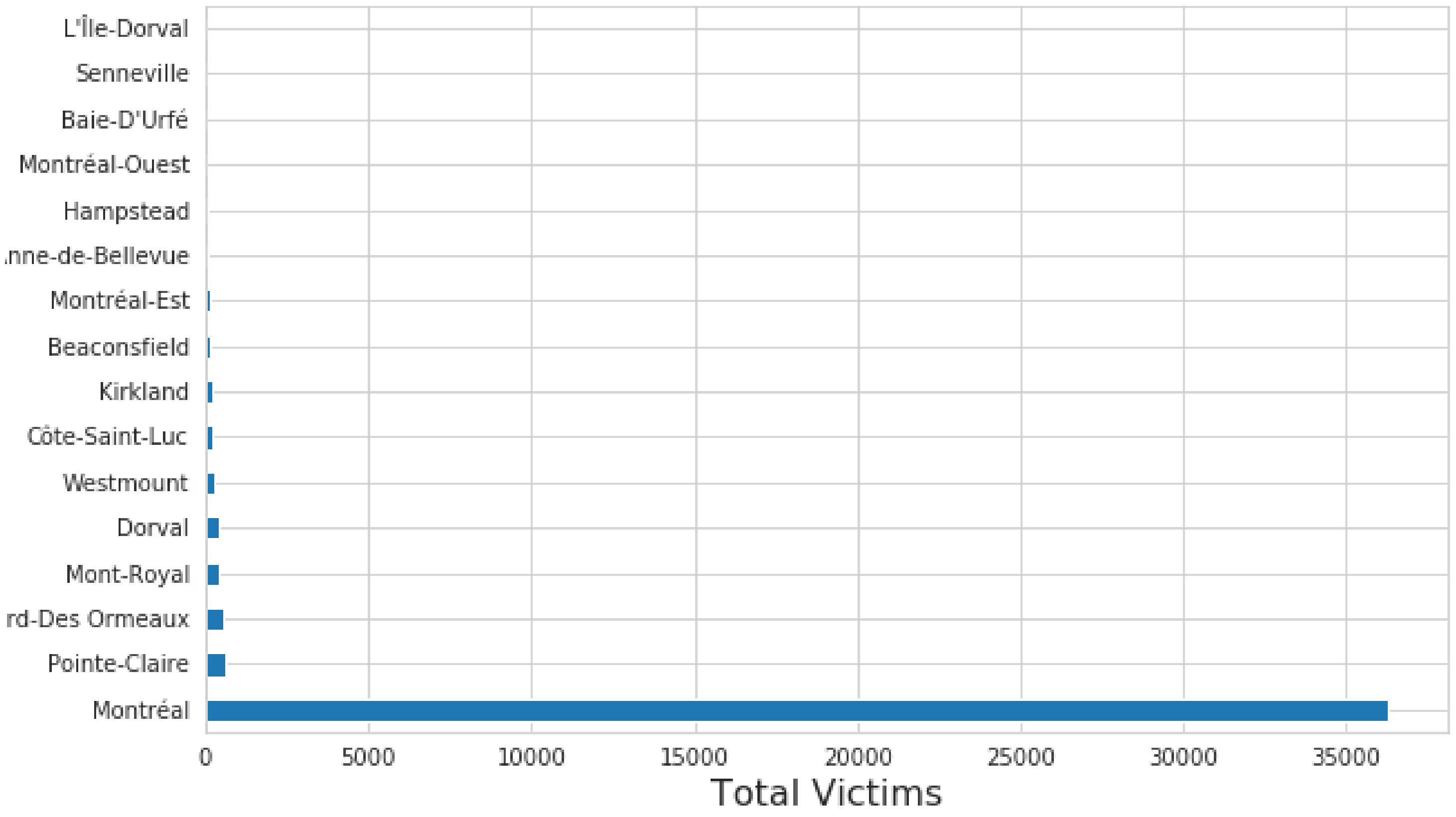
Understand the data through numerical and visual methods, by applying various data perspectives.
The goal is to find the features that will allow us to better understand our data and answer our questions.

This was done by the usage of:

- Bar and Pie Charts
- Heatmap
- Aggregates

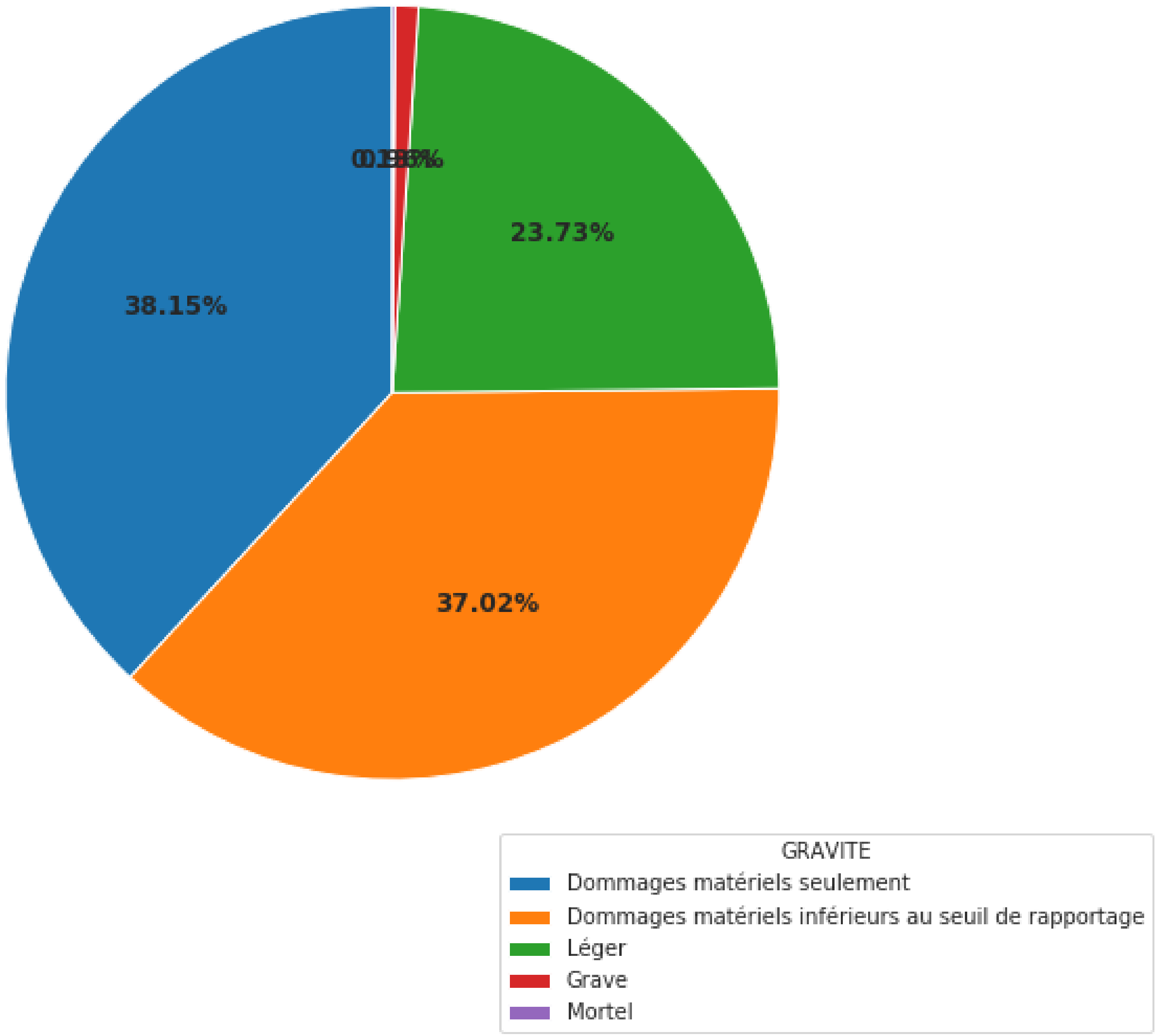
Bar Chart

- Montreal has the largest number of victims.
- L’île -Dorval has the minimum number of victims.



Pie Chart

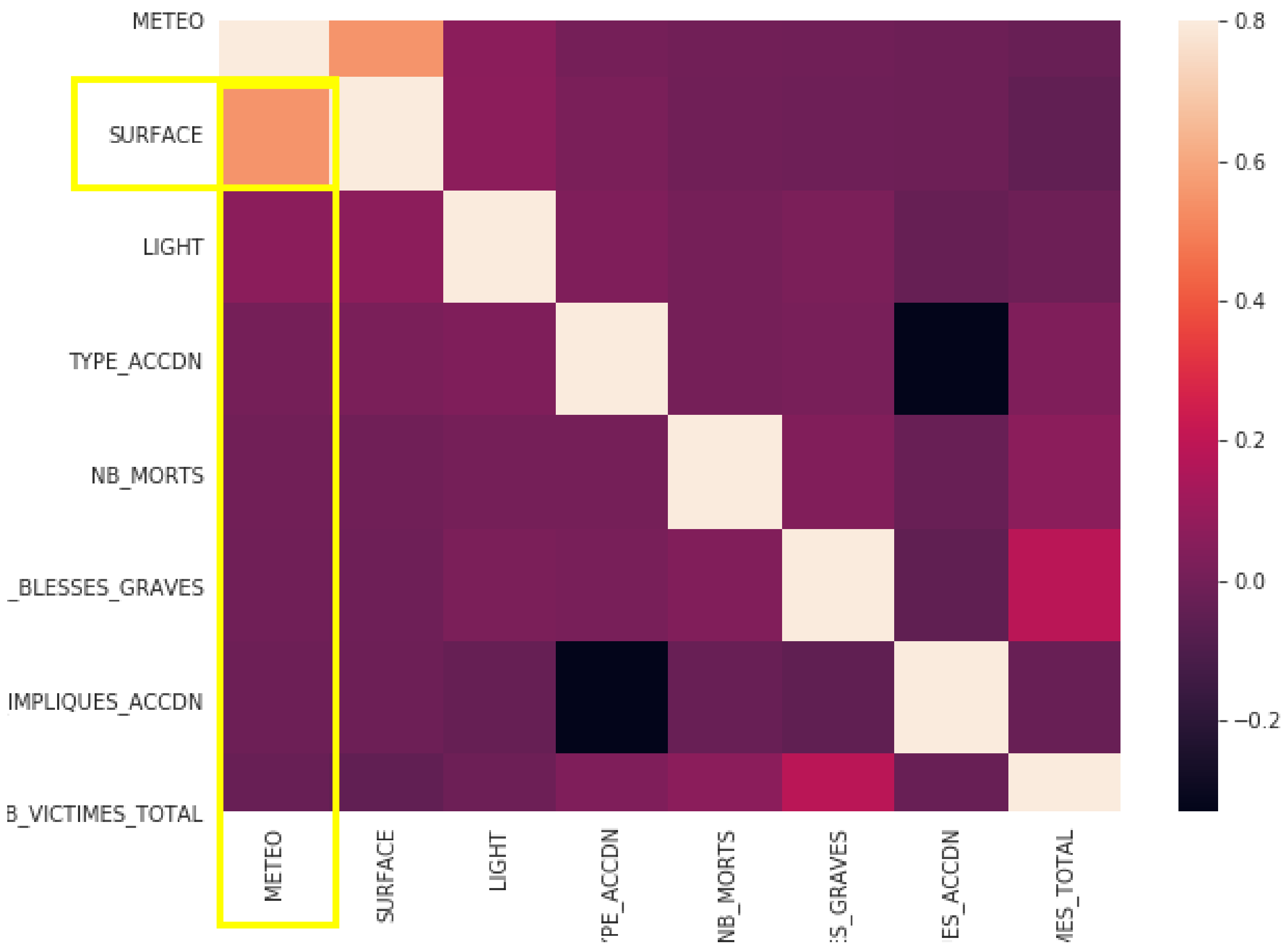
Material Damages are the most common.



Heatmap

METEO (weather) has strong correlation with SURFACE.

Surface: State of the running surface during the accident (wet, dry, ice,..)



Prediction Subject

Predict ***number of collisions*** in Montreal based on:

- Lighting conditions
- Day of the week
- Month
- Weather conditions

Restructuring Data

<u>Original dataset:</u>

one row = one collision

<u>Needed:</u>

- one row= one day with the number of collision per day
- divide each day into day and night in terms of lighting.

Original Dataset

collision - oriented structure

ID	DATE	WEEK_DAY	LIGHT	METEO	STREET
SPVM _ 2012 _ 5093	2012/01/01	DI	3	12	COUSINEAU
SPVM _ 2012 _ 5924	2012/01/01	DI	3	12	BD ST JACQUES
SPVM _ 2012 _ 6851	2012/01/01	DI	3	11	SHERBROOKE O
SPVM _ 2012 _ 647	2012/01/01	DI	3	12	DE SALABERRY
SPVM _ 2012 _ 2412	2012/01/01	DI	3	12	BILLERON
SPVM _ 2012 _ 19368	2012/01/01	DI	1	12	JARRY
SPVM _ 2012 _ 9840	2012/01/01	DI	1	12	FACE PTE S CH DE LA
SPVM _ 2012 _ 20821	2012/01/01	DI	3	12	ST ZOTIQUE
SPVM _ 2012 _ 20822	2012/01/01	DI	3	11	DE LA ROCHE
SPVM _ 2012 _ 11843	2012/01/01	DI	3	12	ST LAURENT
SPVM _ 2012 _ 22656	2012/01/01	DI	3	14	RIVARD
SPVM _ 2012 _ 23880	2012/01/01	DI	3	11	CHARTRAND
SPVM _ 2012 _ 11844	2012/01/01	DI	3	12	SHERBROOKE
SPVM _ 2012 _ 23881	2012/01/01	DI	3	14	HENRI BOURASSA E
SPVM _ 2012 _ 25227	2012/01/01	DI	3	12	COUTURE
SPVM _ 2012 _ 11845	2012/01/01	DI	3	12	ST PIERRE
SPVM _ 2012 _ 25230	2012/01/01	DI	2	11	LACORDAIRE
SPVM _ 2012 _ 2413	2012/01/01	DI	3	11	BONIN
SPVM _ 2012 _ 25232	2012/01/01	DI	3	17	DELAGE
SPVM _ 2012 _ 12628	2012/01/01	DI	1	99	DE MAISONNEUVE E
SPVM _ 2012 _ 26596	2012/01/01	DI	3	11	IBERVILLE
SPVM _ 2012 _ 26597	2012/01/01	DI	1	12	D'IBERVILLE
SPVM _ 2012 _ 12884	2012/01/01	DI	3	11	DELORIMIER
SPVM _ 2012 _ 26598	2012/01/01	DI	3	11	15E AV MTL
SPVM _ 2012 _ 29113	2012/01/01	DI	3	11	MICHEL BOUVIER
SPVM _ 2012 _ 12885	2012/01/01	DI	3	11	WOLFE
SPVM _ 2012 _ 29114	2012/01/01	DI	1	13	LANGELIER
SPVM _ 2012 _ 31121	2012/01/01	DI	3	12	BEACONSFIELD BD
SPVM _ 2012 _ 6852	2012/01/01	DI	1	11	BELMONT
SPVM _ 2012 _ 15336	2012/01/01	DI	3	14	VAN HORNE
SPVM _ 2012 _ 15337	2012/01/01	DI	3	14	EDOUARD MONTPETIT
SPVM _ 2012 _ 649	2012/01/01	DI	3	14	SOMMERSET
SPVM _ 2012 _ 2414	2012/01/01	DI	3	14	COTE VERTU
SPVM _ 2012 _ 16957	2012/01/01	DI	3	11	PL DES COOPERATIVES
SPVM _ 2012 _ 8609	2012/01/01	DI	3	14	BRIAND
SPVM _ 2012 _ 16959	2012/01/01	DI	3	14	ST LAURENT
SPVM _ 2012 _ 18257	2012/01/01	DI	3	12	CHARLAND
SPVM _ 2012 _ 2416	2012/01/01	DI	3	12	ROBERTSON VSL
SPVM _ 2012 _ 8614	2012/01/01	DI	3	17	DE L EGLISE
SPVM _ 2012 _ 18258	2012/01/01	DI	1	12	42E RUE
SPVM _ 2012 _ 23882	2012/01/02	LU	3	11	LACORDAIRE
SPVM _ 2012 _ 15341	2012/01/02	LU	1	12	VAN HORNE
SPVM _ 2012 _ 23883	2012/01/02	LU	1	11	LEGER
SPVM _ 2012 _ 29116	2012/01/02	LU	3	11	ST JUST
SPVM _ 2012 _ 2417	2012/01/02	LU	3	11	HENRI BOURASSA
SPVM _ 2012 _ 19369	2012/01/02	LU	3	11	BERRI
SPVM _ 2012 _ 20823	2012/01/02	LU	3	12	ROSEMONT
SPVM _ 2012 _ 25234	2012/01/02	LU	1	11	FRADETTE
SPVM _ 2012 _ 29130	2012/01/02	LU	3	11	BEAUBIEN
SPVM _ 2012 _ 30398	2012/01/02	LU	3	11	VOIE SERV A 40 O
SPVM _ 2012 _ 31123	2012/01/02	LU	1	11	JEAN YVES
SPVM _ 2012 _ 1125	2012/01/02	LU	1	11	BD ST JEAN STAT
SPVM _ 2012 _ 7466	2012/01/02	LU	1	12	LASALLE
SPVM _ 2012 _ 15338	2012/01/02	LU	3	12	LEGARE
SPVM _ 2012 _ 10376	2012/01/02	LU	1	11	STE CATHERINE O
SPVM _ 2012 _ 21804	2012/01/02	LU	1	11	AV LAURIER E
SPVM _ 2012 _ 26599	2012/01/02	LU	3	11	BELANGER
SPVM _ 2012 _ 31122	2012/01/02	LU	1	11	CALAIS
SPVM _ 2012 _ 5094	2012/01/02	LU	1	12	MARCHE CENTRAL
SPVM _ 2012 _ 7467	2012/01/02	LU	1	11	ST PATRICK
SPVM _ 2012 _ 7470	2012/01/02	LU	3	12	3E AV
SPVM _ 2012 _ 16960	2012/01/02	LU	3	14	PAPINEAU
SPVM _ 2012 _ 12995	2012/01/02	LU	1	11	ONTARIO
SPVM _ 2012 _ 15339	2012/01/02	LU	3	14	DES JOCKEYS
SPVM _ 2012 _ 7	2012/01/03	MA	1	11	GOUIN O
SPVM _ 2012 _ 1128	2012/01/03	MA	3	11	COTE DE LIESSE

Intermediate Result

Day-oriented data with number of accidents, grouped by weather (meteo) condition.

date	light	week_day	month	meteo	count
2012-01-01-d	1	DI	1	11	2
2012-01-01-d	1	DI	1	13	1
2012-01-01-d	1	DI	1	12	4
2012-01-01-n	2	DI	1	12	12
2012-01-01-n	2	DI	1	11	10
2012-01-01-n	2	DI	1	14	8
2012-01-01-n	2	DI	1	17	2
2012-01-02-d	1	LU	1	11	9
2012-01-02-d	1	LU	1	12	3
2012-01-02-n	2	LU	1	11	7
2012-01-02-n	2	LU	1	14	2
2012-01-02-n	2	LU	1	12	3
2012-01-03-d	1	MA	1	17	1
2012-01-03-d	1	MA	1	12	1
2012-01-03-d	1	MA	1	11	22
2012-01-03-n	2	MA	1	12	1
2012-01-03-n	2	MA	1	11	15
2012-01-04-d	1	ME	1	12	6
2012-01-04-d	1	ME	1	11	18
2012-01-04-d	1	ME	1	17	4
2012-01-04-n	2	ME	1	11	2
2012-01-04-n	2	ME	1	12	3
2012-01-04-n	2	ME	1	17	10
2012-01-05-d	1	JE	1	17	2
2012-01-05-d	1	JE	1	11	30
2012-01-05-d	1	JE	1	12	8
2012-01-05-n	2	JE	1	11	10
2012-01-05-n	2	JE	1	17	5
2012-01-05-n	2	JE	1	12	4
2012-01-06-d	1	VE	1	12	10
2012-01-06-d	1	VE	1	13	1

Final Dataset

Date - oriented structure

```
df_p.show(20)
```

date	count	day	night	DI	LU	MA	ME	JE	VE	SA	1	2	3	4	5	6	7	8	9	10	11	12	11-12	13-14-15	16-17-18	19
2012-01-01-d	7	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-01-n	32	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-02-d	12	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-02-n	12	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-03-d	24	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-03-n	16	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-04-d	28	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-04-n	15	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2012-01-05-d	40	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-05-n	19	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-06-d	28	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
2012-01-06-n	12	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-07-d	29	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-07-n	16	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-08-d	20	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-08-n	15	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-09-d	26	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-09-n	16	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-10-d	33	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2012-01-10-n	14	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

Prediction and Results

- We used ***Random Forest Regressor***
- Results are not good with ***CVS = 0.6 and MAE = 6.67***
- Possibly add some features using external sources or develop new features based on ratios or dependencies of existing features.
- Strongest feature importance: ***Day, Night, Sunday, Saturday, Snow, January, December, Friday***

Predicting with specific features

- Model exported using *joblib* library

- Created a script that takes 4 arguments:
    - ➢ day/night
    - ➢ weekday
    - ➢ month
    - ➢ weather condition

- Output: number of collisions

```
nasta@LAPTOP-3QFQBQU5 MINGW64 ~/BIG_DATA/2019
$ python pred.py 'day','DI',2,'neige'

Successfully loaded features!

Showing results:
Input features: ['day', 'DI', '2', 'neige']
Predicted number of collisions: 29
```

Questions?

Thank you!