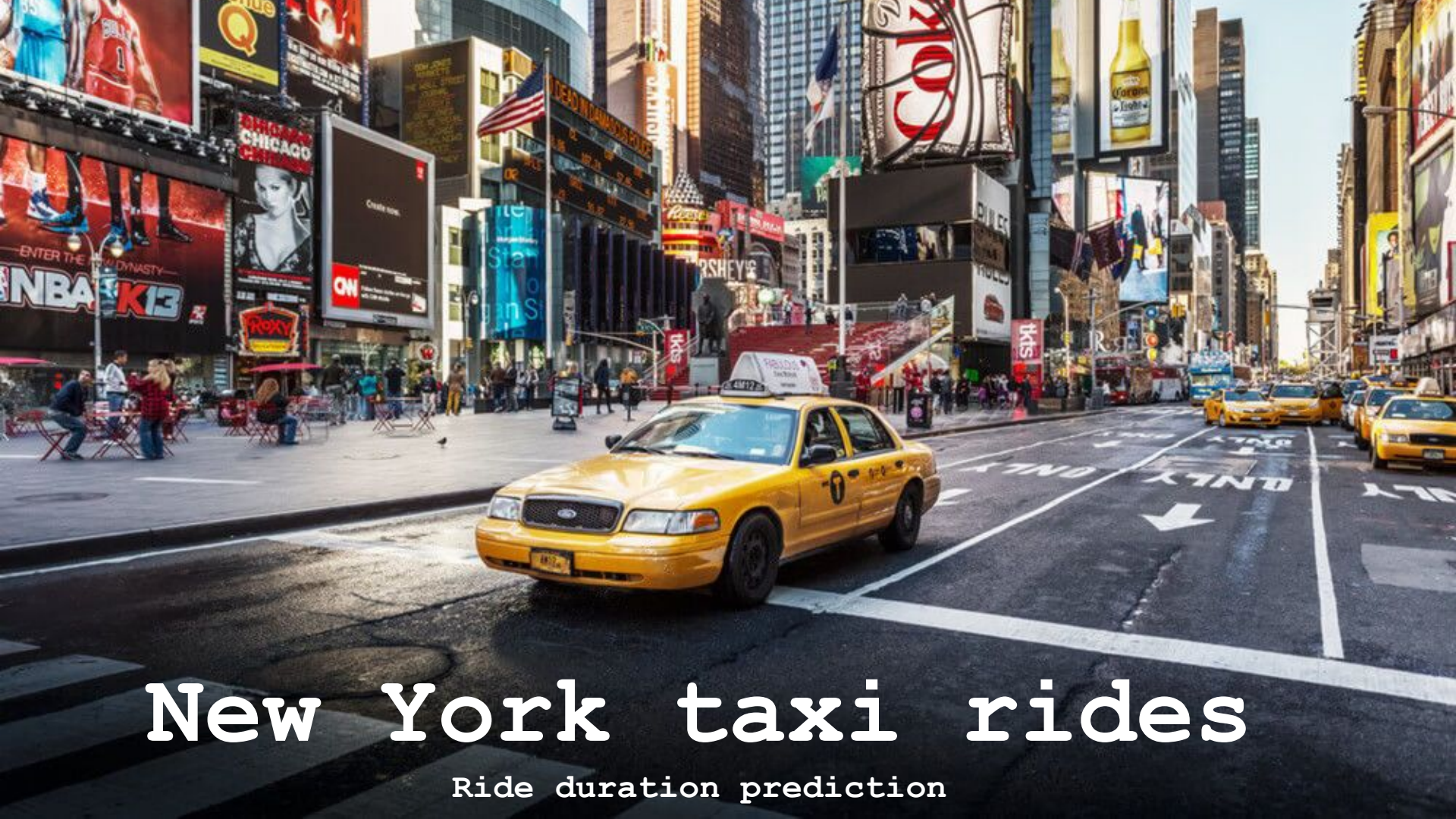


New York taxi rides

Ride duration prediction

Target and Datasets description

“Predict taxi ride Fare in NYC”

Datasets description:

=> Set : about 1,45M records and 11 attributes

Records of rides in New York in 2016

source:

Codebook ½ : taxi dataset

id - a unique identifier for each trip
vendor_id - a code indicating the provider associated with the trip record
pickup_datetime - date and time when the meter was engaged
dropoff_datetime - date and time when the meter was disengaged
passenger_count - the number of passengers in the vehicle (driver entered value)
pickup_longitude - the longitude where the meter was engaged
pickup_latitude - the latitude where the meter was engaged
dropoff_longitude - the longitude where the meter was disengaged
dropoff_latitude - the latitude where the meter was disengaged
store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
trip_duration - duration of the trip in seconds
source:https://www.kaggle.com/c/nyc-taxi-trip-duration/data

Codebook 2/2 : weather

Weather data collected from the National Weather Service.
It contains the first six months of 2016, for a weather station in central park.

It contains for each day :
minimum temperature
maximum temperature
average temperature
precipitation
snow fall,
current snow depth.
The temperature is measured in Fahrenheit and the depth is measured in inches.
T means that there is a trace of precipitation.

Source: *http://w2.weather.gov/climate/xmacis.php?wfo=okx.*

.

Steps

I - Data Cleaning and Manipulation

II - EDA

III- Modeling

- 1) Unsupervised learning
- 2) Supervised Learning

IV - First conclusions

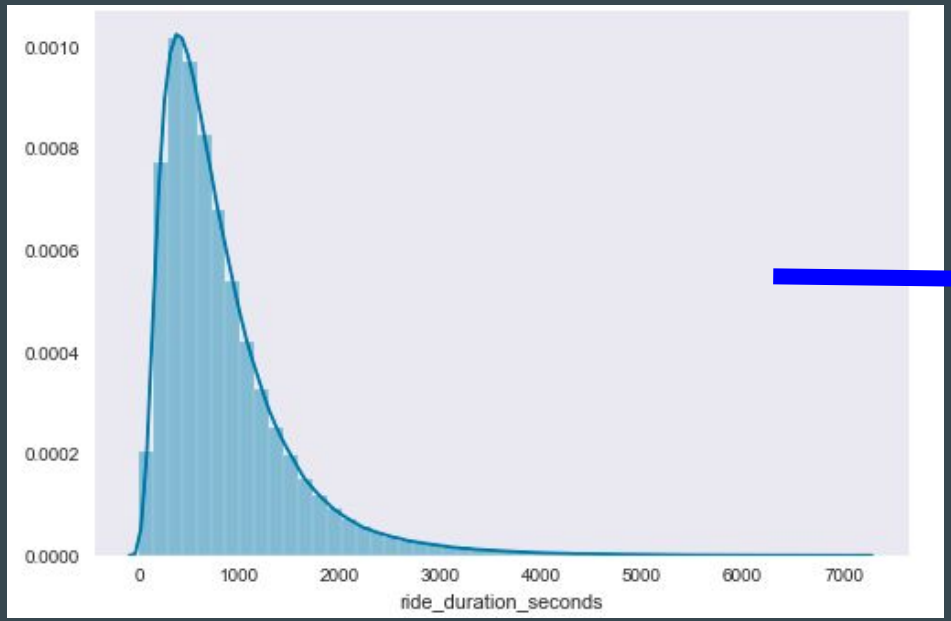
Data Cleaning

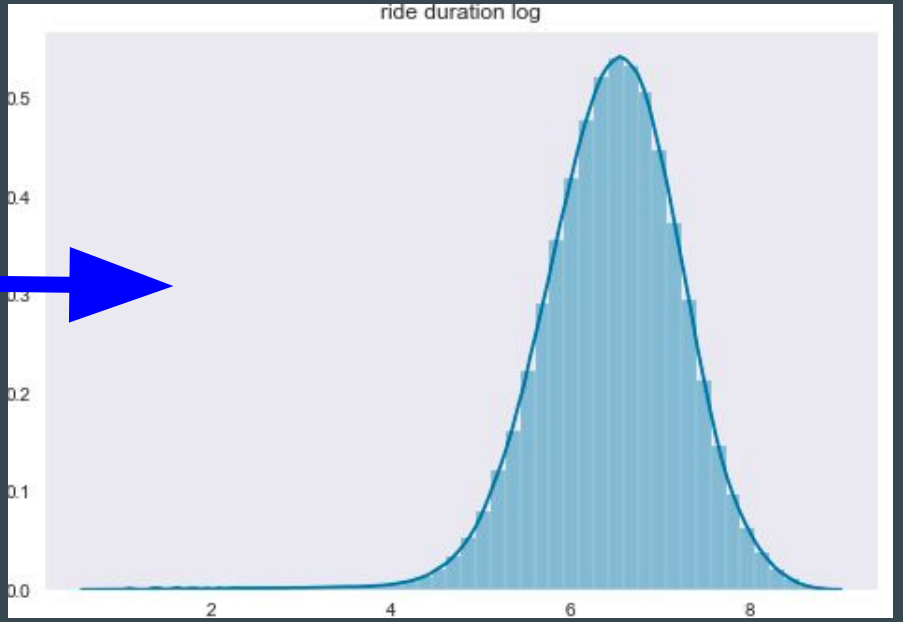
- drop duplicates
- drop useless columns
- drop inconsistent data ex: ride duration above 2hours, number passenger null

Data Manipulation

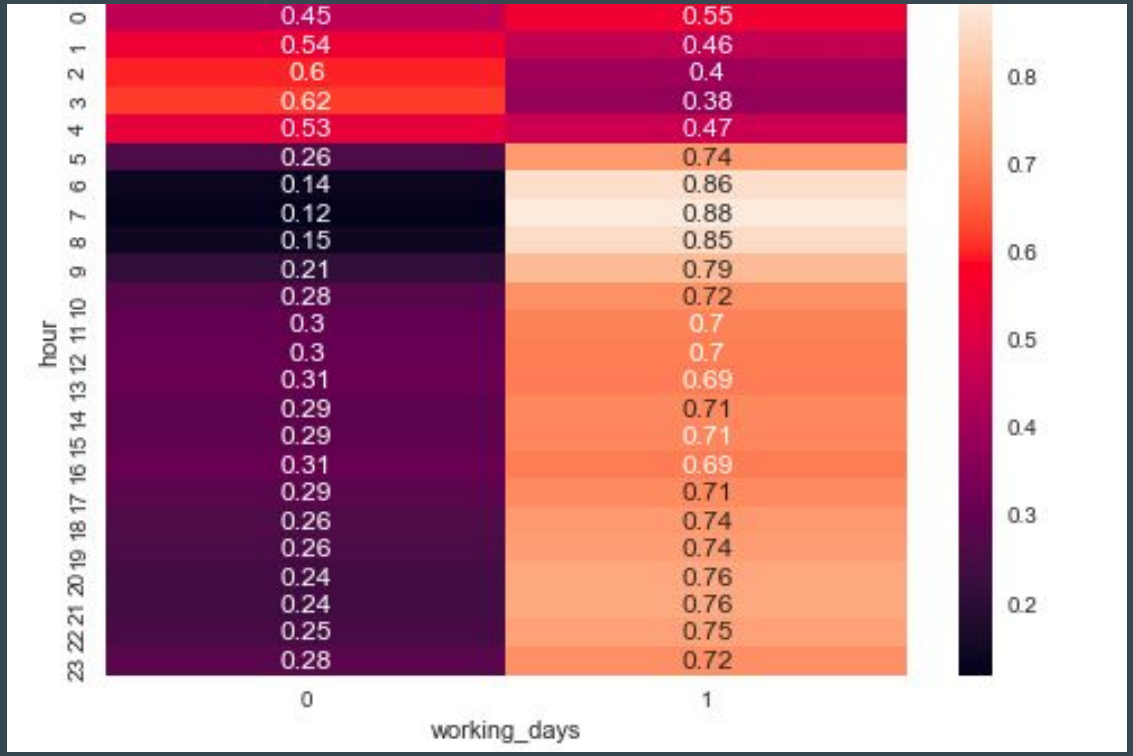
- Convert to datetime format
- Split date in Month, day, weekend...
- Merge dataset
- convert units
- calculate an approximation of the distance

Exploratory Data Analysis : distribution shape of ride duration

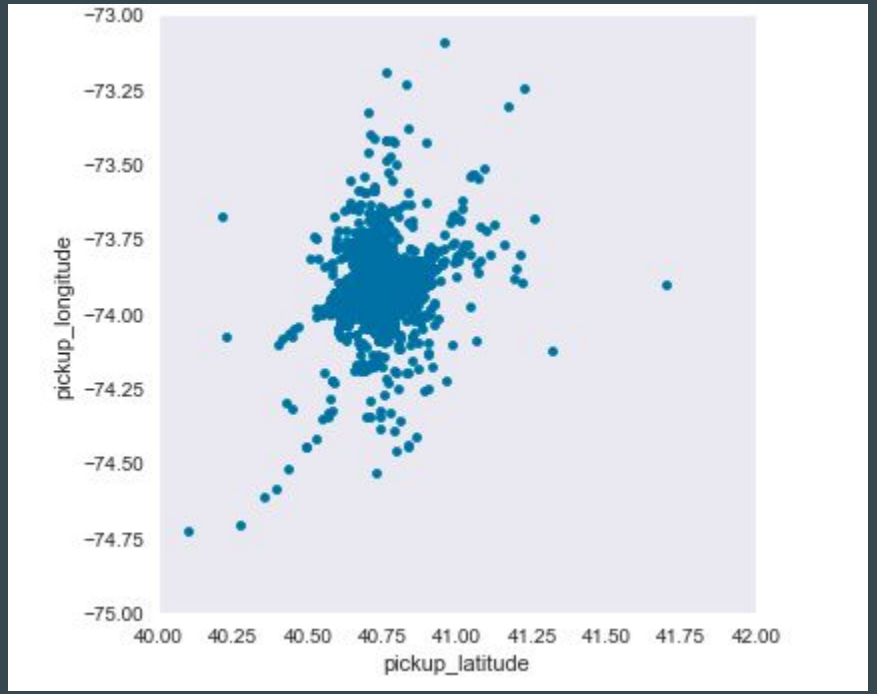


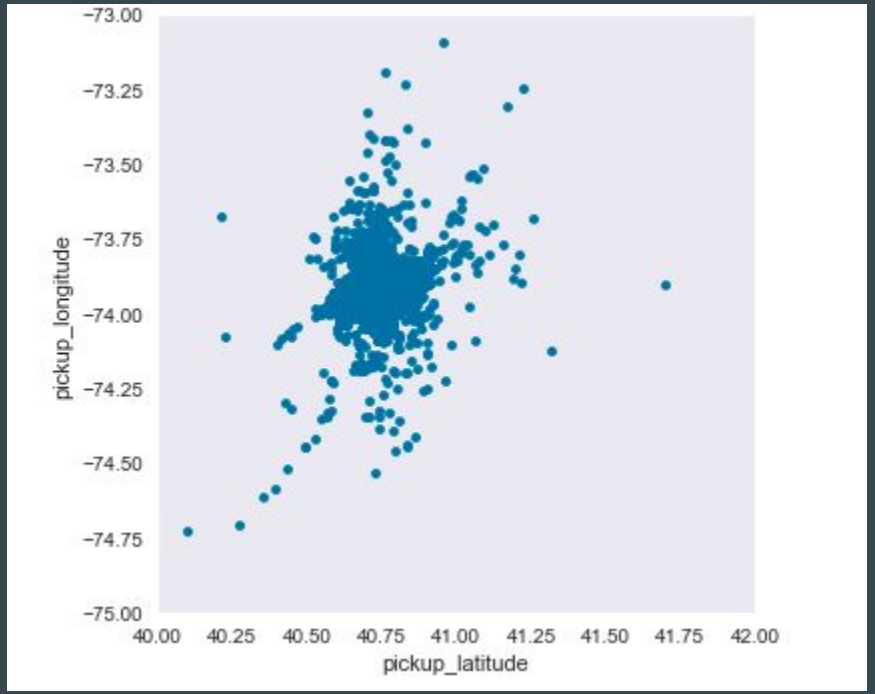


How rides are splitted depending on the hour



Pickup points an dropoff points distribution





Weather data : provides a few information but not enough

Modeling : Unsupervised learning 1/2

Target : "*To cluster the insignificant data in order to save information without too much attributes during the prediction*"

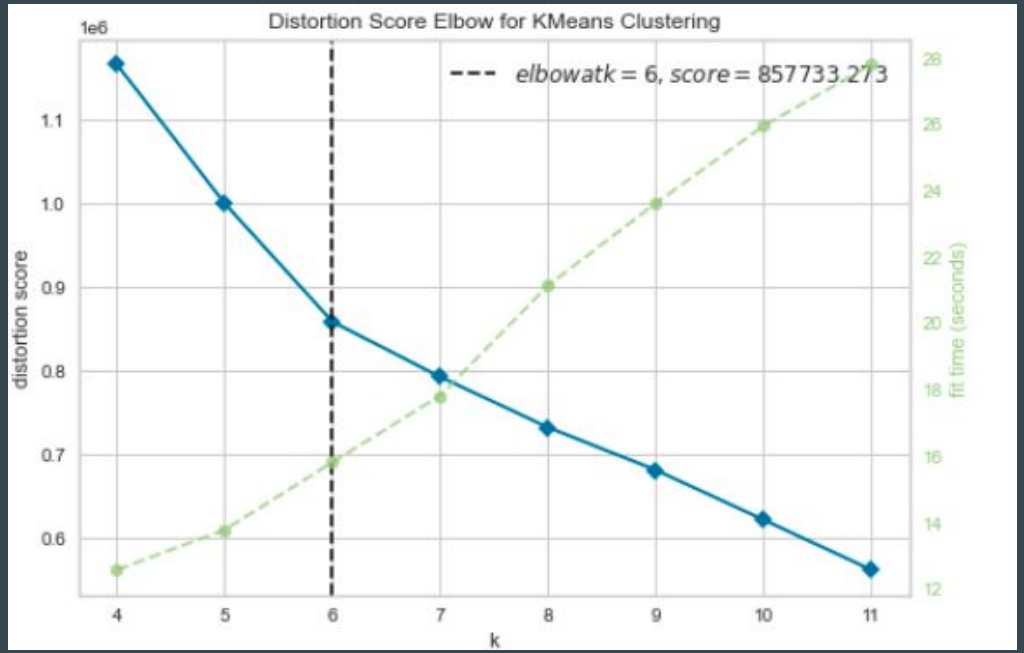
Step 1: Feature selection (manual)

Step 2 : Defining the number of clusters with the K-Elbow method

Step 3 : Running PCA

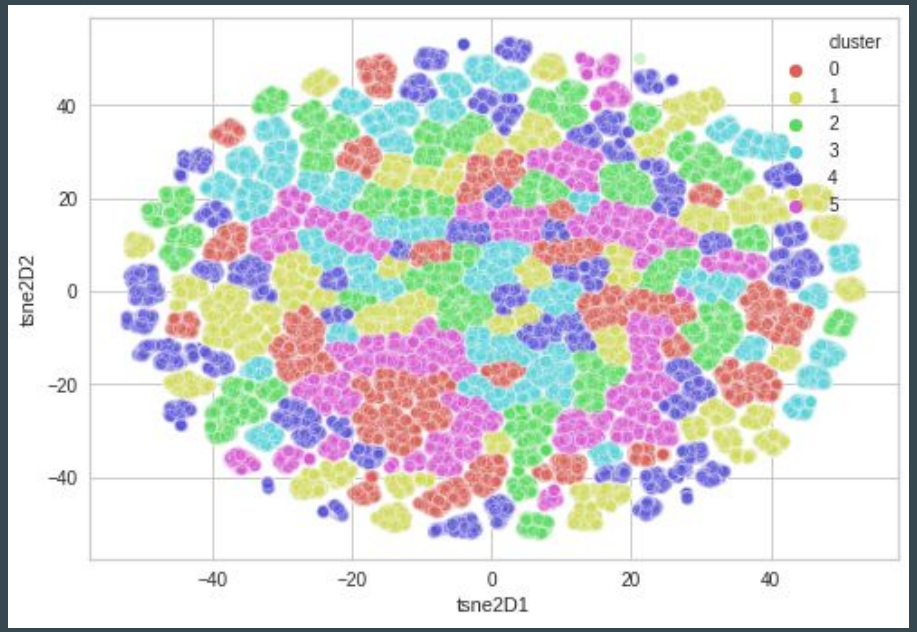
Step 3: Applying KMean model

Step 4 : Cluster Visualization with TSNE

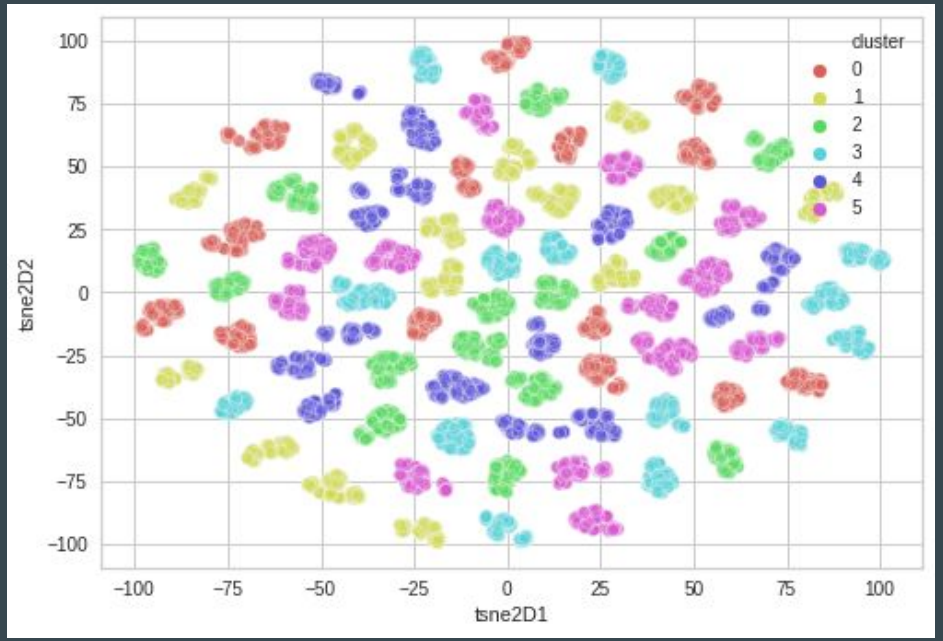


Modeling : Unsupervised learning 2/2

For 0.1 % of the dataset



For 0.01 % of the dataset



Modeling : Supervised learning

	hour	new_passenger_count	weekday	distance_KM	cluster
0	17	1.0	Monday	1.498521	2
1	0	1.0	Sunday	1.805507	0
2	11	1.0	Tuesday	6.385098	4
3	19	1.0	Wednesday	1.485498	5
4	13	1.0	Saturday	1.188588	2

New dataset for the prediction of ride duration

Model tested:

Linear regression with a R-squared = 0.464

OLS with constant

Model still running: *Generalized Linear Model with a Tweedie distribution* (gamma

OLS results

OLS Regression Results

Dep. Variable:	ride_duration_seconds	**R-squared:**	0.538
Model:	OLS	**Adj. R-squared:**	0.538
Method:	Least Squares	**F-statistic:**	8.039e+04
Date:	Mon, 03 Aug 2020	**Prob (F-statistic):**	0.00
Time:	16:33:21	**Log-Likelihood:**	-7.2666e+06
No. Observations:	966962	**AIC:**	1.453e+07
Df Residuals:	966947	**BIC:**	1.453e+07
Df Model:	14		
Covariance Type:	nonrobust		

Assumption check done.

No multicolinearity

Normality assumption not respected

Potential issue with linearity and autocorrelation

First conclusion

Number of clusters may not be relevant

Other regression models to test in order to improve the R-squared