MSE | MASTER OF SCIENCE IN ENGINEERING

Projet

Année 2023-2024

New York Taxi Trip Data

Big Data Analytics [MA_BDA]

Tanya Kemm Pereira

Calvin Uebelhart

Raphaël Besson

Hes·so

Haute Ecole Spécialisée de Suisse occidentale

Fachhochschule Westschweiz

University of Applied Sciences and Arts Western Switzerland

7 juin 2024

SOMMAIRE

1. Introduction
2. Données
3. Choix des questions
4. Preprocessing
5. Réponses aux questions
6. Conclusion

INTRODUCTION

- Analyse de données des trajets des taxis jaunes de New York
- Entre 2013 et 2023, période de 10 ans
- Représente un dataset (.parquet) de 14 GB avec plus d'un milliard de trajets
- Extrait du site officiel = bonne fiabilité

DONNÉES

- **Tpep_pickup/dropoff_datetime** : date et heure de mise en service et arrêt du compteur
- **Passenger_count** : nombre de passagers dans le véhicule
- **Trip_distance** : distance parcourue, en miles, indiquée par le taximètre
- **PULocationID/DOLocationID** : zone TLC Taxi où le taximètre a été activé/arrêté
- **Payment_type** : code numérique indiquant le mode de paiement du passager
- **Tip_amout** : montant du pourboire
- **Total_amout** : montant total facturé aux passagers

DONNÉES

- Les **autres features** ont été **exclues** car :
  - Proportion de valeurs manquantes importantes
  - Données non pertinentes pour les questions choisies

VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	DOLocationID
2	2013-01-01 00:39:00	2013-01-01 00:55:00	3.0	3.86	1.0	NULL	238	116
2	2013-01-01 00:12:00	2013-01-01 00:16:00	5.0	0.0	1.0	NULL	264	264
2	2013-01-01 00:02:00	2013-01-01 00:03:00	3.0	0.0	1.0	NULL	264	264
2	2013-01-01 00:38:00	2013-01-01 00:38:00	2.0	0.0	1.0	NULL	264	264
2	2013-01-01 00:03:00	2013-01-01 00:04:00	4.0	0.0	1.0	NULL	264	264
2	2013-01-01 00:03:00	2013-01-01 00:04:00	3.0	0.0	1.0	NULL	146	146
2	2013-01-01 00:05:00	2013-01-01 00:09:00	4.0	0.0	1.0	NULL	146	146
2	2013-01-01 00:06:00	2013-01-01 00:09:00	4.0	0.0	1.0	NULL	146	146
2	2013-01-01 00:14:00	2013-01-01 00:16:00	3.0	0.0	1.0	NULL	146	146
2	2013-01-01 00:04:00	2013-01-01 00:06:00	5.0	0.0	1.0	NULL	130	130

payment_type	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improvement_surcharge	total_amount	congestion_surcharge	airport_fee
2	15.0	0.5	0.5	0.0	0.0	0.0	16.0	NULL	NULL
1	3.5	0.5	0.5	0.12	0.0	0.0	4.62	NULL	NULL
1	2.5	0.5	0.5	0.25	0.0	0.0	3.75	NULL	NULL
2	2.5	0.5	0.5	0.0	0.0	0.0	3.5	NULL	NULL
1	3.0	0.5	0.5	0.07	0.0	0.0	4.07	NULL	NULL
1	2.5	0.5	0.5	0.25	0.0	0.0	3.75	NULL	NULL
1	4.5	0.5	0.5	1.25	0.0	0.0	6.75	NULL	NULL
1	4.0	0.5	0.5	0.15	0.0	0.0	5.15	NULL	NULL
1	3.0	0.5	0.5	0.05	0.0	0.0	4.05	NULL	NULL
1	3.0	0.5	0.5	0.45	0.0	0.0	4.45	NULL	NULL

CHOIX DES QUESTIONS

- **Comment la pandémie de COVID-19 a-t-elle affecté l'industrie du taxi ?**
  - En utilisant des analyses statistiques
- **Est-possible de prédire les pourboires ?**
  - En utilisant un modèle de machine learning

PREPROCESSING

- **tpep_pickup_datetime/tpep_dropoff_datetime** :
  garder les années entre 2013 et 2023, en éliminant les années aberrantes.

- **Passenger_count** :
  Supprimer les lignes avec des valeurs manquantes ou inférieurs à 1

summary	passenger_count
count	1028161845
mean	1.637341715398902
stddev	1.2789692250398288
min	1.0
max	9.0

pickup_year	count
2001	27
2002	487
2003	50
2004	1
2008	760
2009	1280
2010	1
2011	4
2012	1
2013	171816340
2014	165447580
2015	146039232
2016	131131805
2017	113500386
2018	102870524
2019	84597309
2020	24649266
2021	30903983
2022	39655622
2023	38310128

pickup_year	count
2013	170506769
2014	164328908
2015	144938331
2016	130107651
2017	112365818
2018	100903485
2019	81546552
2020	22910378
2021	28199888
2022	36772447
2023	35581618

PREPROCESSING

- **Trip_distance** :
  Conservé les lignes supérieures à 0.

```
+-------+------------------+
|summary|     trip_distance|
+-------+------------------+
|  count|        1028161845|
|   mean|3.0015497211631086|
| stddev| 3.745036451274469|
|    min|              0.01|
|    max|            791.54|
+-------+------------------+
```

- **Payment_type** :
  Assuré que les valeurs soient comprises entre 1 et 6.
  (1 = carte de crédit, 2 = espèces, 3 = sans frais, 4 = litige, 5 = inconnu, 6 = voyage annulé)

- **Tip_amount** :
  Garanti que les valeurs soient égales ou supérieures à 0.

```
+-------+------------------+
|summary|        tip_amount|
+-------+------------------+
|  count|        1028161845|
|   mean|1.8242027050914496|
| stddev| 2.571036397465161|
|    min|               0.0|
|    max|             391.0|
+-------+------------------+
```

PREPROCESSING

- **Total_amount** :
  Garder les valeurs supérieures à 0.

```
+-------+------------------+
|summary|      total_amount|
+-------+------------------+
|  count|        1028161845|
|   mean|16.738070693639692|
| stddev|13.939369524643432|
|    min|              0.01|
|    max|            399.99|
+-------+------------------+
```

- **PULocationID/DOLocationID** :
  Doit correspondre à des valeurs entre 1 et 265, qui sont les zones valides de New York

OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough
1	0.116357	0.000782	Newark Airport	1	EWR
2	0.433470	0.004866	Jamaica Bay	2	Queens
3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx
4	0.043567	0.000112	Alphabet City	4	Manhattan
5	0.092146	0.000498	Arden Heights	5	Staten Island

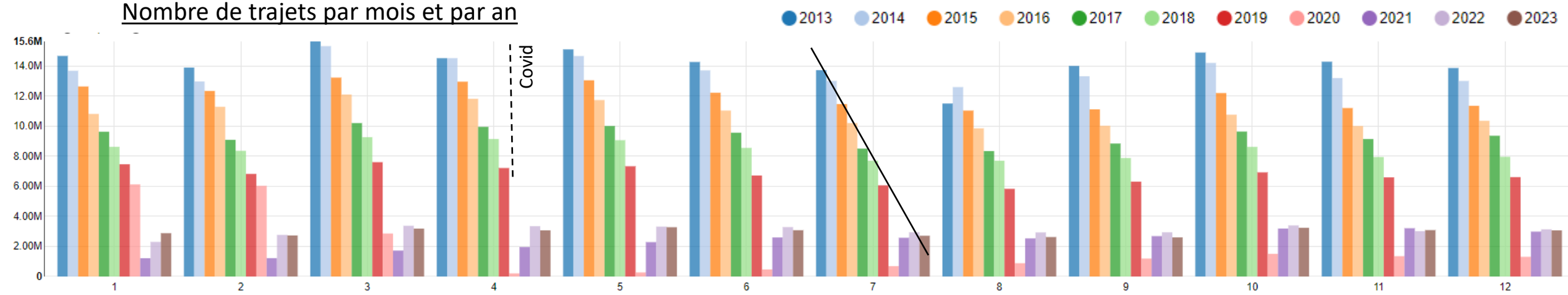
```
+-------+------------------+
|summary|      PULocationID|
+-------+------------------+
|  count|        1028161845|
|   mean|162.71745904556496|
| stddev| 66.71198648049561|
|    min|                 1|
|    max|               265|
+-------+------------------+
```

```
+-------+------------------+
|summary|      DOLocationID|
+-------+------------------+
|  count|        1028161845|
|   mean|160.83411356798598|
| stddev| 70.41217223597155|
|    min|                 1|
|    max|               265|
+-------+------------------+
```

COMMENT LA PANDÉMIE DE COVID-19 A-T-ELLE AFFECTÉ L'INDUSTRIE DU TAXI ?

- **Analyses statistiques** :
  - **Diminution** du nombre de trajets chaque année, concurrence avec les **VTC** comme Uber depuis 2011
  - Forte diminution depuis la pandémie du **Covid-19** en **mars 2020** (confinement)
  - 2022 suit la tendance observée avant le Covid et 2023 est similaire à 2023 (stabilisation)

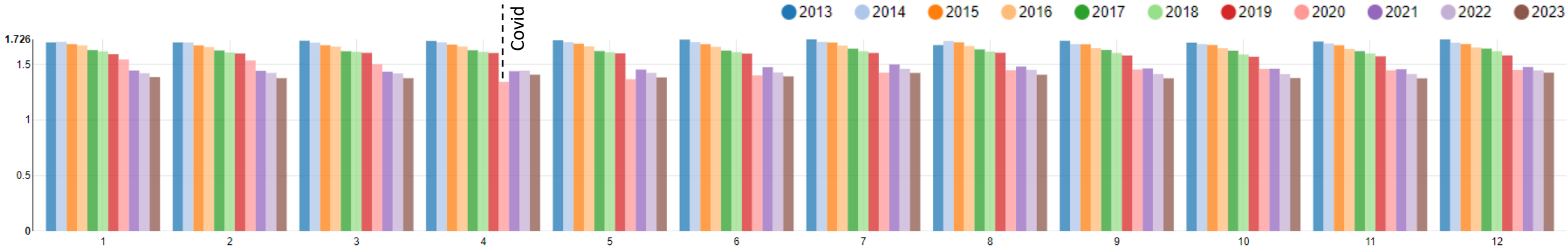
Nombre de trajets par mois et par an



COMMENT LA PANDÉMIE DE COVID-19 A-T-ELLE AFFECTÉ L'INDUSTRIE DU TAXI ?

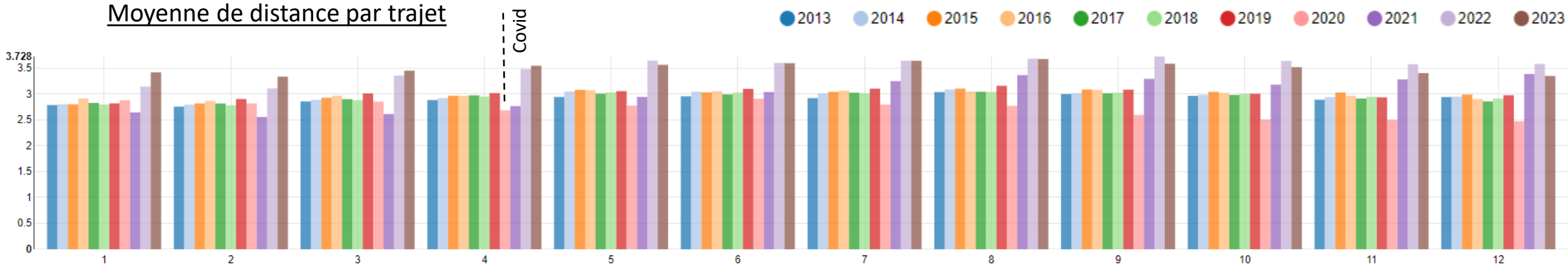
- **Analyses statistiques** :
  - **Diminution** progressive du nombre de passagers par trajet, de 1.7 à 1.3
  - Baisse plus prononcée depuis la pandémie en **mars 2020**
  - En 2023, la moyenne est similaire voir plus bas que 2020 = effets de la pandémie sur les habitudes de transport

Moyenne du nombre de passagers par trajet



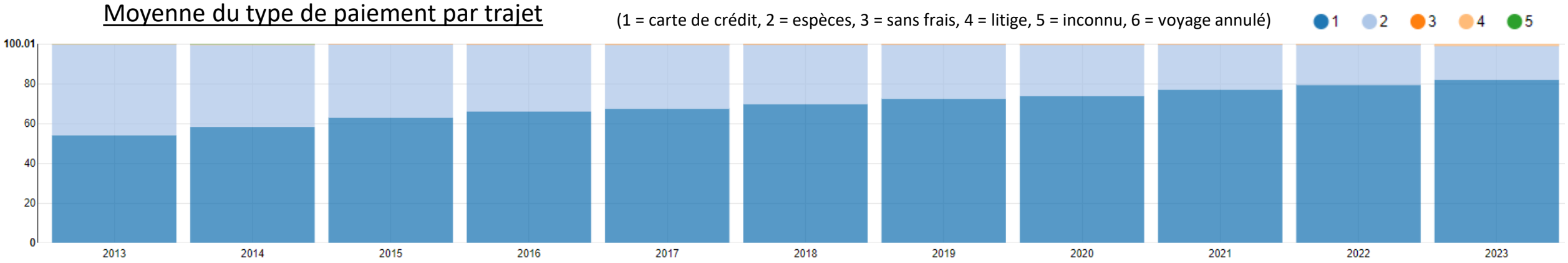
COMMENT LA PANDÉMIE DE COVID-19 A-T-ELLE AFFECTÉ L'INDUSTRIE DU TAXI ?

- **Analyses statistiques** :
  - Entre 2013 et 2019, la moyenne est restée **stable**
  - Diminution légèrement plus marquée depuis la pandémie en **mars 2020**
  - En 2022 et 2023, les moyennes sont plus élevées, indiquant un changement de comportement de mobilité



COMMENT LA PANDÉMIE DE COVID-19 A-T-ELLE AFFECTÉ L'INDUSTRIE DU TAXI ?

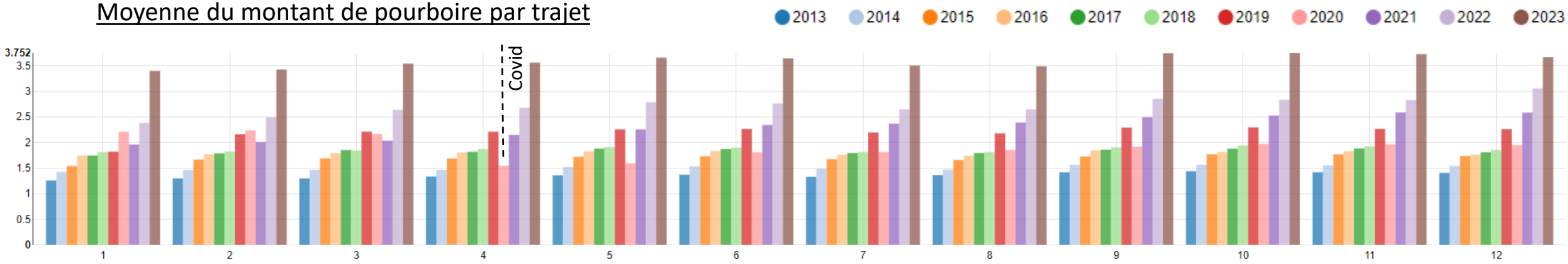
- **Analyses statistiques** :
  - Augmentation continue de l'utilisation de la carte de crédit par rapport à l'argent liquide au fil du temps
  - La pandémie en **2020** a confirmé cette tendance
  - En 2022 et 2023, il n'y a pas eu de changement significatif de cette tendance



COMMENT LA PANDÉMIE DE COVID-19 A-T-ELLE AFFECTÉ L'INDUSTRIE DU TAXI ?

- **Analyses statistiques** :
  - Entre 2013 et 2018, la moyenne augmente et se stabilise
  - Diminution plus marquée pendant la pandémie en 2020, distance sociale, masques
  - En 2022 et surtout 2023, les moyennes sont plus élevées. Elles sont influencées par des **distances** plus longues et un possible **changement** de la manière dont les clients perçoivent et récompensent les services.

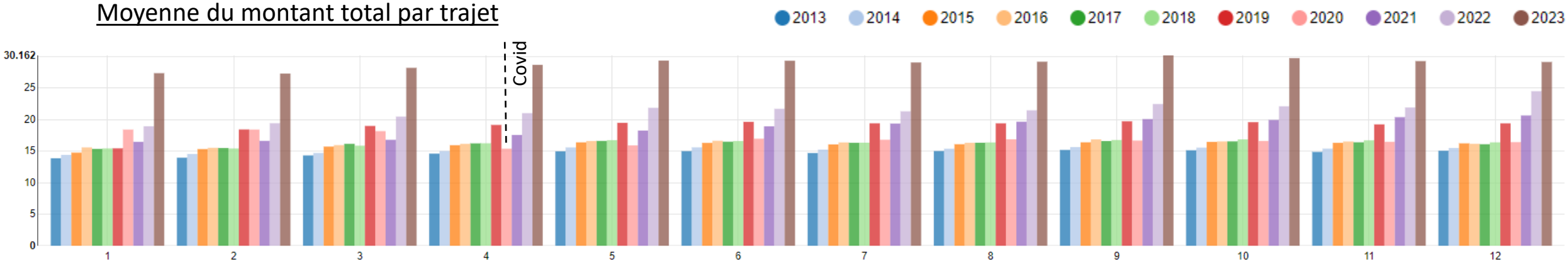
Moyenne du montant de pourboire par trajet



COMMENT LA PANDÉMIE DE COVID-19 A-T-ELLE AFFECTÉ L'INDUSTRIE DU TAXI ?

- **Analyses statistiques** :
  - Entre 2013 et 2018, la moyenne augmente et se stabilise
  - Diminution légèrement plus marquée pendant la pandémie en **2020** par rapport à 2019 mais similaire à 2018
  - 2022 et surtout 2023 montrent des moyennes plus élevées = à cause de l'augmentation des **surcharges** et l'**inflation** des prix

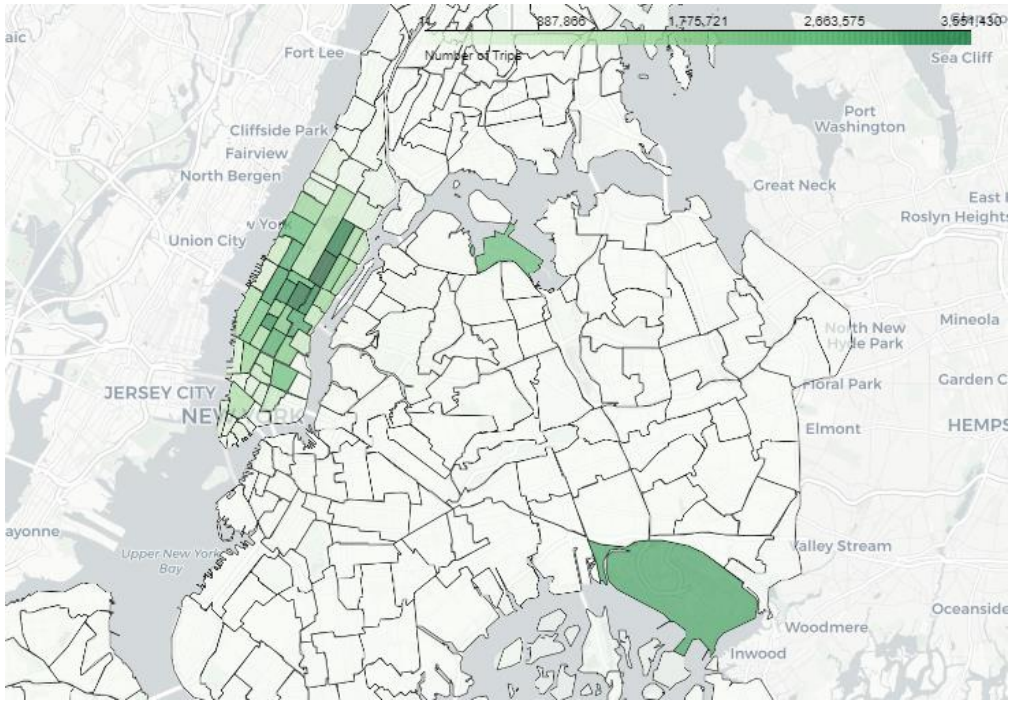
Moyenne du montant total par trajet



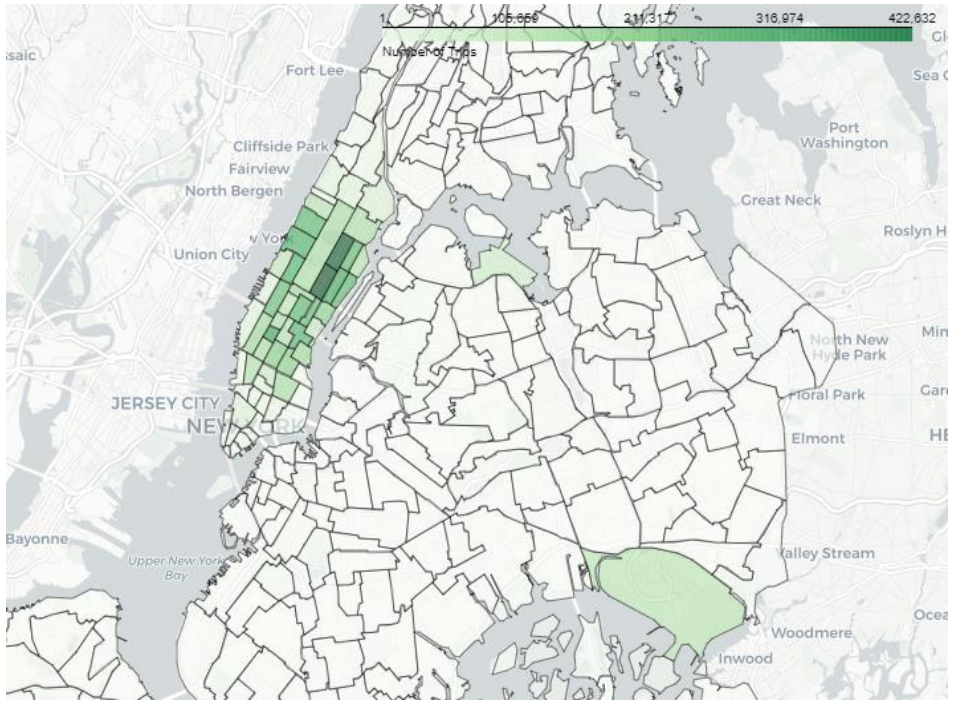
COMMENT LA PANDÉMIE DE COVID-19 A-T-ELLE AFFECTÉ L'INDUSTRIE DU TAXI ?

Nombre de trajets par année selon la localisation de départ :

PULocationID en 2019



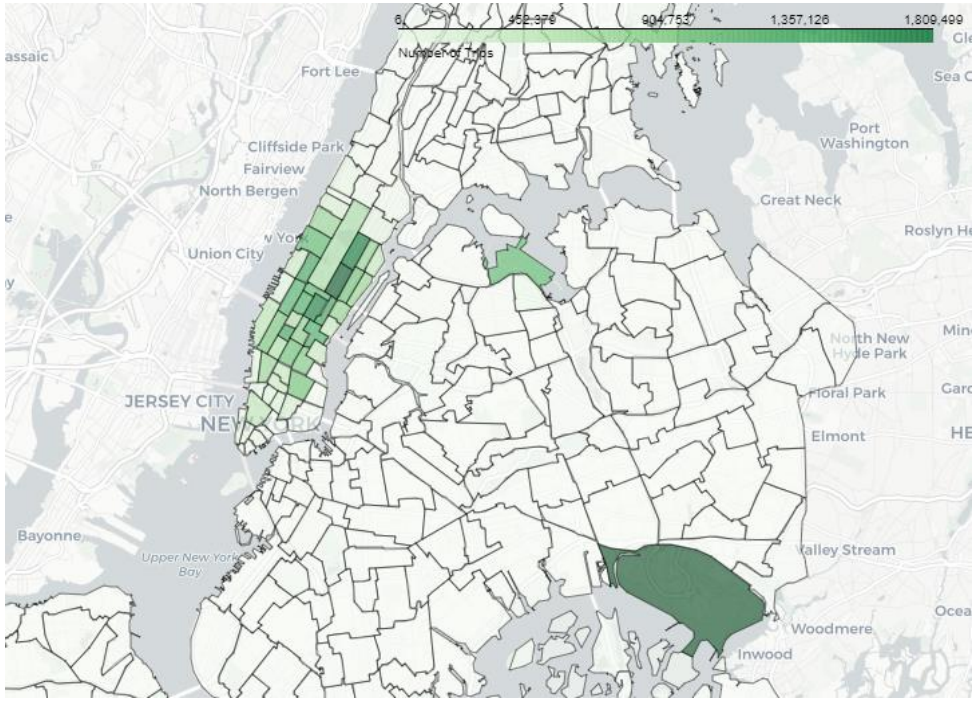
PULocationID en 2020 (pendant la pandémie)



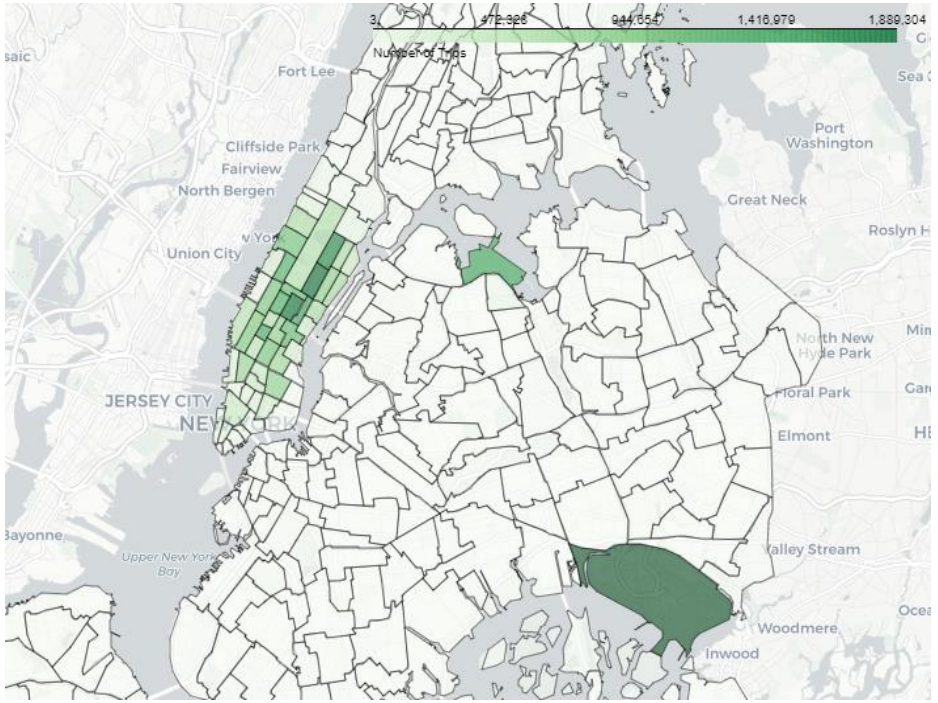
COMMENT LA PANDÉMIE DE COVID-19 A-T-ELLE AFFECTÉ L'INDUSTRIE DU TAXI ?

Nombre de trajets par année selon la localisation de départ :

PULocationID en 2022



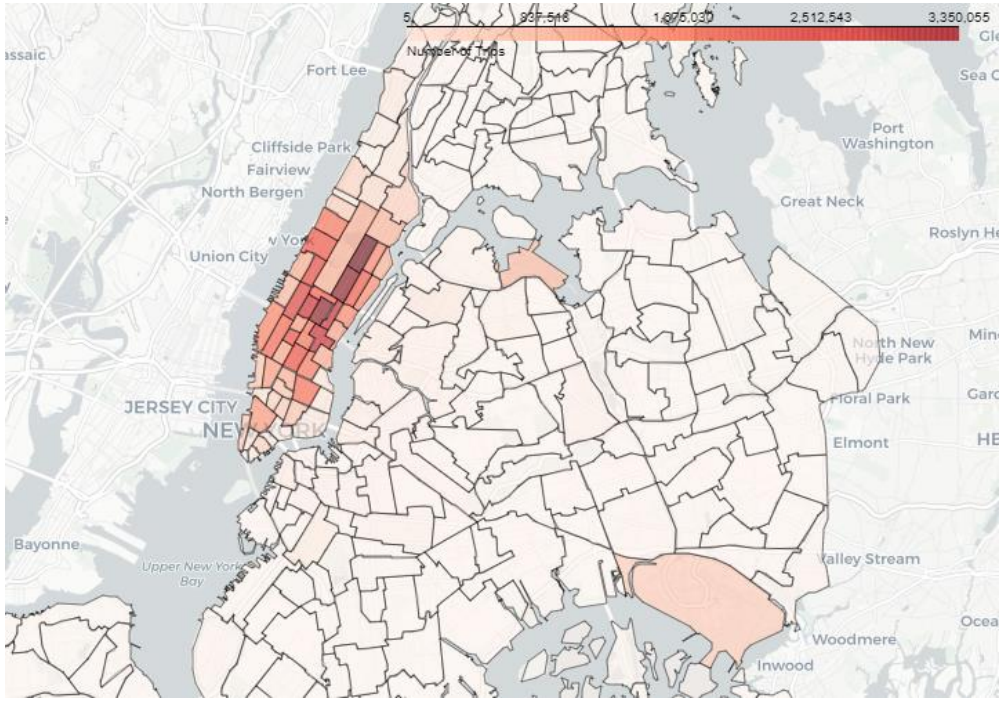
PULocationID en 2023



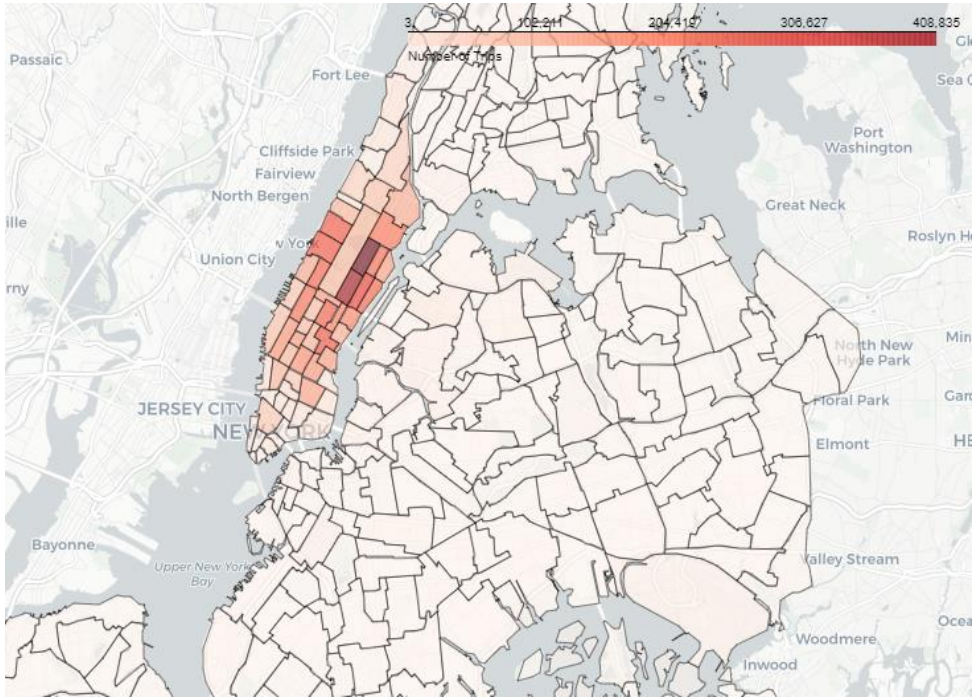
COMMENT LA PANDÉMIE DE COVID-19 A-T-ELLE AFFECTÉ L'INDUSTRIE DU TAXI ?

Nombre de trajets par année selon la localisation d'arrivée :

DOLocationID en 2019



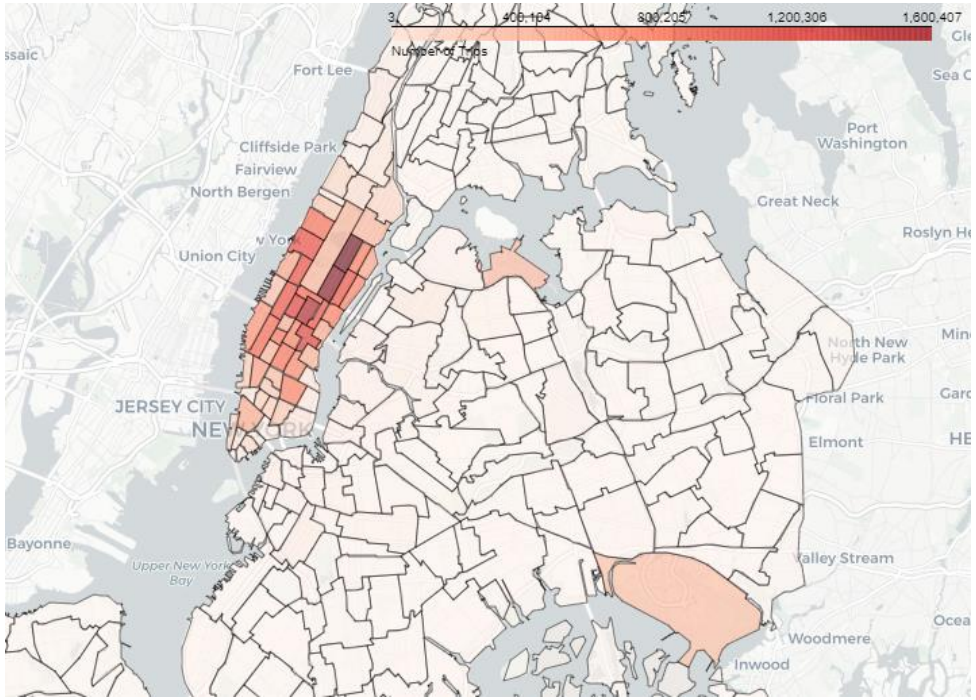
DOLocationID en 2020 (pendant la pandémie)



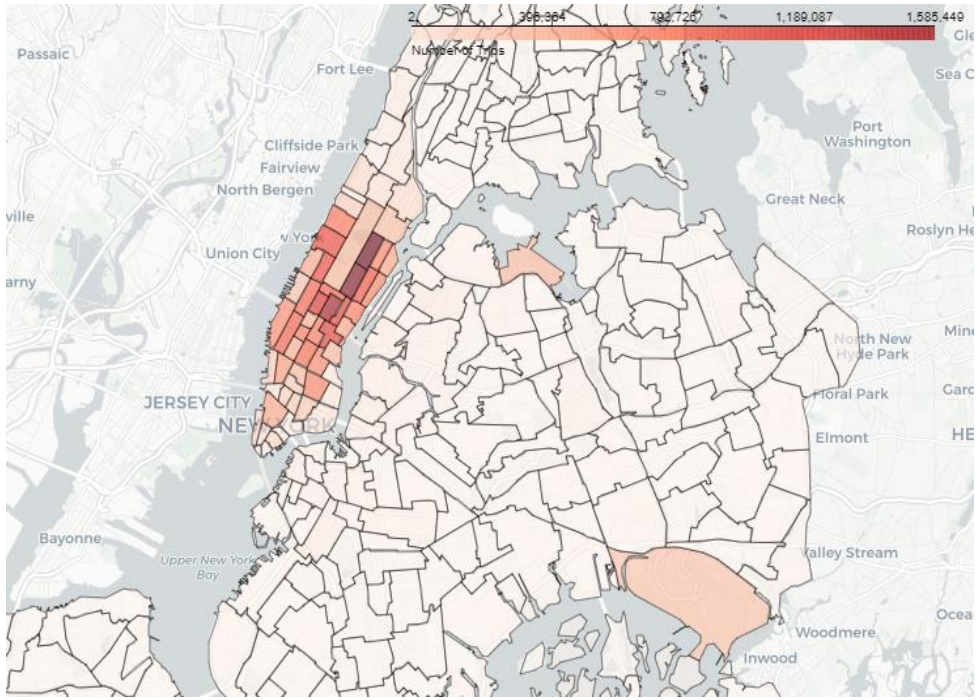
COMMENT LA PANDÉMIE DE COVID-19 A-T-ELLE AFFECTÉ L'INDUSTRIE DU TAXI ?

Nombre de trajets par année selon la localisation d'arrivée :

DOLocationID en 2022



DOLocationID en 2023



EST-POSSIBLE DE PRÉDIRE LES POURBOIRES ?

- Les features utilisées doivent pouvoir être connues par les chauffeurs de taxi avant de prendre un client
- **But: pouvoir choisir la zone optimale en fonction des conditions connues**
- Moyenne des pourboires: 1.824
- Moyenne des pourboires par zone:

PULocationID	avg_tip_amount
1	10.126331339124736
10	6.4969084882899155
93	6.209428014113195
215	6.109431763666204
132	5.8472571356932175
70	5.651203898703956
219	5.636215238746887
205	0.6422383419689114
147	0.6317241379310348
32	0.6172149095829745
159	0.6117803606134661
94	0.592773601732705
167	0.5623512849162011
254	0.532980675656889

EST-POSSIBLE DE PRÉDIRE LES POURBOIRES ?

- **Features:**
  - Zone de pickup
  - Date (année, mois, jour de la semaine, heure du jour)
  - Distance du trajet

- La distance de trajet est de loin la feature la plus corrélées au montant du pourboire (~0.5)

- Problème: la distance du trajet ne peut pas être connue à l'avance par le chauffeur de taxi

- Solution: un modèle estime la distance du trajet en fonction de la date et de la zone de pickup puis l'estimation est utilisée dans un autre modèle pour déterminer le montant du pourboire

EST-POSSIBLE DE PRÉDIRE LES POURBOIRES ?

Modèles

Algorithme	Features
Régression linéaire	zone de pickup
Régression linéaire	zone de pickup, date, distance de trajet
Random forest	date, distance de trajet
Random forest	date, zone de pickup, distance de trajet
Random forest	Date, zone de pickup, [estimation] distance de trajet

EST-POSSIBLE DE PRÉDIRE LES POURBOIRES ?

Métriques

- **MAE** = Mean Average error (erreur moyenne quelques soit la direction)
- **RMSE** = Root Mean Squared Error (plus d'importance aux grandes erreurs)

EST-POSSIBLE DE PRÉDIRE LES POURBOIRES ?

Importances des features

trip_distance 0.818
year 0.0857
pickup location 0.0594
Hour 0.0241
Weekday 0.007
Month 0.006

EST-POSSIBLE DE PRÉDIRE LES POURBOIRES ?

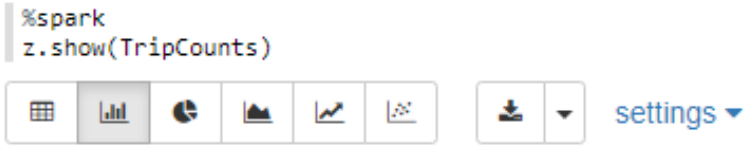
Résultats

Algorithme	Features	RMSE	MAE
Régression linéaire	zone de pickup	2.570	1.627
Régression linéaire	zone de pickup, date, distance de trajet	2.350	1.506
Random forest	date, zone de pickup	2.333	1.543
Random forest	date, zone de pickup, distance de trajet	2.128	1.399
Random forest	Date, zone de pickup, [estimation] distance de trajet	2.288	1.493

CONCLUSION

- La configuration du **cluster dataproc** sur google cloud était limitée et on n'a pas pu faire tourner le code avec notre dataset stocké dans un bucket.
    - On a donc fait tourner les notebooks Zeppelin **en local**

- **Limitation** dans le type de **visualisation** des données avec **spark**
- Transformer la questions pourboire en classification



- Question COVID: bon résultat
- Question Pourboire: résultats mitigés

SOURCES

- TLC Trip Record Data : https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- Image : https://www.fotocommunity.de/photo/yellow-cab-in-new-york-joachimj/29451183


QUESTIONS ?
TIMES SQUARE CHURCH
NYC TAXI