NYC Taxi Analytics

Dashboard Interactif

Analyse des trajectoires de taxis à New York City

basée sur 200 000 courses aléatoires

Anas JEBALI

Rania CHIRANE

Open Data et Web des Données

Année universitaire 2024–2025

Source des données : NYC Taxi & Limousine Commission

Table des matières

1. Introduction

Chaque jour, des milliers de taxis jaunes sillonnent les rues de New York, tissant un réseau invisible de déplacements à travers les cinq boroughs. Ces trajets, enregistrés par la *Taxi & Limousine Commission*, constituent une mine d'or pour comprendre les dynamiques urbaines de la ville qui ne dort jamais.

Ce projet propose une exploration visuelle et analytique de ces données à travers un dashboard interactif. L'objectif n'est pas simplement d'afficher des chiffres, mais de raconter l'histoire de la mobilité new-yorkaise : où vont les gens, quand, et pourquoi certaines zones génèrent plus de revenus que d'autres.

Périmètre de l'étude

- **Volume** : 200 000 trajectoires échantillonnées aléatoirement
- **Source** : NYC TLC Trip Record Data
- **Granularité** : Course individuelle avec pickup/dropoff, tarif, pourboire

2. Méthodologie et Pipeline de Travail

Afin de structurer notre démarche, nous avons suivi un pipeline de traitement de données inspiré des pratiques professionnelles en data science. L'objectif était de transformer un dataset brut de courses de taxis new-yorkais en un dashboard complet mêlant analyse, visualisation et intelligence artificielle.

2.1. Compréhension du Jeu de Données

Nous avons débuté par l'exploration des données fournies par la Taxi & Limousine Commission (TLC). Cette étape a permis d'identifier les variables essentielles (horodatages, distances, montants, pourboires) ainsi que les limites du dataset :

- impossibilité de localiser les trajets directement (pas de coordonnées GPS),
- nécessité de relier les `PULocationID` et `DOLocationID` à de véritables zones géographiques.

Cette contrainte nous a amenés à créer un module externe contenant les coordonnées des 263 zones TLC.

2.2. Prétraitement et Création de Variables

Cette phase visait à enrichir les données afin de faciliter l'analyse :

- conversion des dates et extraction des composantes temporelles (heure, jour, mois),
- création de la durée de course,
- création de la vitesse moyenne,
- calcul de la distance géographique (formule de Haversine),
- création de ratios d'intérêt (prix par mile),
- nettoyage des valeurs aberrantes.

2.3. Exploration Analytique et Visualisations

Plusieurs visualisations exploratoires ont été produites pour comprendre les patterns :

- analyse des zones les plus actives,
- analyse temporelle (heures, jours, tendances),
- graphe réseau des flux entre zones (NetworkX),

— cartes 2D et 3D (Plotly & Folium).

Ces analyses ont orienté la construction du dashboard.

2.4. Modélisation par Apprentissage Automatique

Deux approches ont été développées :

— **Prédiction du pourboire** avec un modèle Random Forest,

— **Segmentation des clients** avec K-Means et visualisation PCA en 3D.

L'objectif était d'offrir une aide à la décision et une compréhension plus fine des comportements des usagers.

2.5. Intégration dans Streamlit

Enfin, toutes les analyses ont été intégrées dans une interface interactive :

— pages et onglets thématiques,

— interactions en temps réel,

— optimisation avec le cache,

— design sombre inspiré des outils financiers.

3. Journal de Bord

Ce journal retrace les grandes phases d'avancement du projet, depuis le choix du dataset jusqu'à la conception finale du dashboard interactif.

3.1. Début du Projet : Exploration et Préparation

Au lancement du projet, un premier dataset avait été considéré, mais son manque de richesse a motivé un changement vers les données de taxis new-yorkais (NYC TLC). Nous avons chargé et exploré un échantillon de 200 000 courses afin de conserver un volume représentatif tout en assurant des performances acceptables dans Streamlit.

Les premières actions majeures ont été :

— comprendre la structure des données TLC,

— identifier les variables pertinentes,

— détecter la nécessité d'ajouter les coordonnées des zones,

— construire un module Python dédié aux zones TLC.

3.2. Milieu du Projet : Analyse, Création des Modèles et Visualisations

Une fois les données préparées, la phase centrale du projet a consisté à :

— créer des visualisations exploratoires (barplots, heatmaps, flux entre zones),

— construire le graphe réseau des déplacements,

— implémenter les premières cartes (2D et 3D),

— développer le modèle de prédiction de pourboire via Random Forest,

— réaliser une segmentation des clients à l'aide de K-Means et d'une réduction dimensionnelle PCA.

Cette étape a permis de poser les fondations du futur dashboard.

3.3. Fin du Projet : Intégration, Optimisation et Design

La dernière phase a été consacrée à l'intégration de toutes les composantes dans une application Streamlit cohérente :

— création des différentes pages et sections du dashboard,
— harmonisation visuelle (thème sombre, typographies, animations),
— optimisation des temps de chargement via le cache Streamlit,
— réalisation des tests utilisateurs,
— rédaction du rapport et préparation de la présentation finale.

4. Architecture du Dashboard

Le dashboard a été conçu avec **Streamlit**, un framework Python qui permet de créer des applications web interactives sans écrire de JavaScript. L'interface adopte un thème sombre professionnel, inspiré des outils d'analyse financière.

4.1. Stack Technique

Composant	Rôle
Streamlit	Framework principal, gestion de l'état et du cache
Plotly	Graphiques interactifs (3D, heatmaps, barres)
NetworkX	Modélisation du réseau de zones
D3.js	Visualisation du graphe de connexions
Scikit-learn	Modèles de prédiction (Random Forest, K-Means)
Pandas	Manipulation et agrégation des données

TABLE 1 – Technologies utilisées

4.2. Organisation de l'Interface

L'interface se divise en sections repliables, permettant à l'utilisateur de naviguer sans être submergé d'informations. Chaque section s'ouvre avec un effet d'agrandissement du titre, offrant une expérience fluide.

Sections du Dashboard

1. **Réseau de Taxis** — Graphe interactif des connexions entre zones
2. **Analyse Économique** — Patterns temporels et zones rentables
3. **Carte 3D** — Visualisation géospatiale des métriques
4. **Prédictions ML** — Intelligence artificielle appliquée

5. Visualisation du Réseau

La première visualisation représente NYC comme un **graphe orienté** où chaque nœud est une zone géographique et chaque arête symbolise un flux de trajets entre deux zones.

5.1. Construction du Graphe

Les données brutes sont agrégées par paire (origine, destination). Seules les routes dépassant un seuil configurable de trajets sont conservées, évitant ainsi le bruit visuel des connexions anecdotiques.

— **Nœuds** : Zones de pickup/dropoff (identifiées par LocationID)
— **Arêtes** : Nombre de trajets entre deux zones
— **Taille des nœuds** : Proportionnelle au degré (nombre de connexions)

5.2. Interactivité

Grâce à D3.js, l'utilisateur peut :

— Glisser les nœuds pour réorganiser le graphe
— Zoomer sur une région spécifique
— Survoler un nœud pour voir ses statistiques
— Filtrer par arrondissement d'origine ou de destination

Les couleurs distinguent les zones selon leur rôle : vert pour les départs, bleu pour les arrivées, rose pour les zones standard.

6. Analyse Économique Spatio-Temporelle

Cette section répond à une question simple mais cruciale : *quand et où faut-il travailler pour maximiser ses revenus ?*

6.1. Rentabilité Horaire

Le premier graphique trace le revenu moyen par course en fonction de l'heure de la journée. On observe généralement deux pics correspondant aux heures de pointe (matin et soir), avec une vallée en milieu de journée.

6.2. Patterns Hebdomadaires

Une heatmap croise les jours de la semaine avec les heures, révélant des motifs récurrents. Le vendredi soir et le samedi présentent souvent des revenus supérieurs, tandis que le dimanche matin reste calme.

Indicateur clé : Efficacité

L'efficacité est définie comme le ratio entre le revenu moyen et la distance moyenne parcourue. Une zone avec des courses courtes mais bien payées sera plus « efficace » qu'une zone nécessitant de longs trajets.

$$Efficacit = \frac{Revenumoyen}{Distancemoyenne}$$

6.3. Top Zones

Un classement horizontal affiche les six zones générant le plus de revenus. La couleur des barres encode l'efficacité : plus c'est vert, plus la zone est rentable par kilomètre parcouru.

7. Carte 3D Interactive

Pour offrir une perspective géographique, une visualisation 3D positionne chaque zone selon ses coordonnées GPS, avec une hauteur proportionnelle à la métrique choisie (revenu total, nombre de trajets, distance moyenne ou efficacité).

L'utilisateur peut :

— Sélectionner la métrique à afficher

— Ajuster le nombre de zones visibles
— Changer la palette de couleurs (Plasma, Viridis, Turbo, Hot)
— Faire pivoter la vue en 3D

Cette représentation permet d'identifier immédiatement les « hotspots » de l'activité taxi à Manhattan et dans les aéroports.

8. Intelligence Artificielle

La dernière section exploite le machine learning pour aller au-delà de l'analyse descriptive.

8.1. Prédiction de Pourboire

Un modèle **Random Forest** (100 arbres) est entraîné pour prédire le pourboire d'une course en fonction de plusieurs variables :

— Distance du trajet
— Tarif de la course
— Heure de prise en charge
— Jour de la semaine
— Zone de départ
— Zone d'arrivée
— Nombre de passagers
— Mode de paiement

Le modèle atteint un score R^2 d'environ 70%, indiquant une capacité raisonnable à expliquer la variance des pourboires. Sans surprise, le mode de paiement (carte vs. cash) ressort comme la variable la plus influente — les paiements par carte enregistrent systématiquement les pourboires, contrairement au cash.

8.2. Segmentation Clients

Un algorithme **K-Means** (3 clusters) segmente les courses en trois profils :

Segment	Caractéristiques
Économique	Trajets courts, tarifs bas, pourboires modestes
Standard	Trajets moyens, comportement « typique »
Premium	Longues distances, tarifs élevés, pourboires généreux

Table 2 – Segmentation des courses

Une visualisation 3D (réduction PCA) permet d'observer la séparation des clusters dans l'espace des features.

9. Conclusion

Ce projet illustre comment des données ouvertes, combinées à des outils de visualisation modernes, peuvent révéler les rythmes cachés d'une métropole. Le dashboard offre une lecture multi-échelle : du flux global entre arrondissements jusqu'à la prédiction individuelle d'un pourboire.

Plusieurs pistes d'amélioration restent envisageables : intégration de données météorologiques, prise en compte des événements spéciaux (concerts, matchs), ou encore déploiement d'un modèle de prédiction de la demande en temps réel.

Au-delà de l'exercice technique, ce travail rappelle que derrière chaque point de données se trouve un trajet réel — un New-Yorkais pressé, un touriste émerveillé, ou simplement quelqu'un rentrant chez lui après une longue journée.

The city that never sleeps, analyzed.

Références

— NYC Taxi & Limousine Commission — https://www.nyc.gov/site/tlc/about/data.page

— Documentation Streamlit — https://docs.streamlit.io

— Documentation Plotly — https://plotly.com/python

— D3.js — https://d3js.org