Analyse de la mobilité urbaine par lignes et modes de transport – Chicago vs Philadelphie

---

Analyse Exploratoire des données (EDA)

Objectif de l'EDA

Cette phase d'analyse exploratoire vise à comprendre la structure, la qualité et les caractéristiques statistiques des jeux de données de fréquentation des transports publics pour deux villes : **Philadelphie** et **Chicago**. L'EDA permet de valider la cohérence des données avant les étapes de nettoyage, de transformation et d'analyse comparative.

Sources de données

Philadelphie

Deux fichiers au format CSV ont été utilisés :

- **Average_Daily_Ridership_By_Route – City of Philadelphia** : fréquentation moyenne journalière par ligne.
- **Average_Daily_Ridership_By_Mode – City of Philadelphia** : fréquentation moyenne journalière par mode de transport.

Chicago

Les données de Chicago proviennent de deux sources :

- Un fichier **Excel** pour la fréquentation journalière par mode (Bus / Rail).
- Un fichier **CSV (issu d'une conversion RDF)** pour la fréquentation journalière par ligne.

Philadelphie – Analyse Exploratoire

Données par ligne (By Route)

Dimensions du jeu de données

- Nombre d’observations : 10 994
- Nombre de variables : 6

Variables

- Calendar_Year (int) : année calendaire
- Calendar_Month (int) : mois calendaire
- Route (object) : identifiant de la ligne
- Average_Daily_Ridership (int) : fréquentation moyenne journalière
- Source (object) : source de collecte (APC)
- ObjectId (int) : identifiant technique

Statistiques descriptives clés

- Période couverte : 2019 – 2025
- Fréquentation moyenne journalière : ~3 385 passagers
- Forte dispersion (écart-type ~7 455), indiquant une hétérogénéité importante entre les lignes

Qualité des données

- Valeurs manquantes : aucune
- Doublons : aucun détecté

Données par mode (By Mode)

Dimensions du jeu de données

- Nombre d’observations : 492
- Nombre de variables : 6

Variables

- Calendar_Year, Calendar_Month
- Mode : type de transport (Bus, Heavy Rail, Regional Rail, etc.)
- Average_Daily_Ridership
- Source, ObjectId

Statistiques descriptives clés

- Fréquentation moyenne journalière : ~106 416 passagers
- Valeurs élevées pour les modes lourds (Heavy Rail, Regional Rail)

Qualité des données

- Aucune valeur manquante
- Aucun doublon

Chicago – Analyse Exploratoire

Données par mode (By Mode – Excel)

Dimensions du jeu de données

- Nombre d'observations : 9 100
- Nombre de variables : 5

Variables

- service_date (datetime) : date du service
- day_type : type de jour (W = weekday, U = weekend/holiday)
- bus : fréquentation bus
- rail_boardings : fréquentation rail
- total_rides : fréquentation totale

Statistiques descriptives clés

- Période couverte : 2001 – 2025
- Fréquentation moyenne journalière totale : ~1,21 million de trajets
- Variabilité significative selon les périodes et le type de jour

Qualité des données

- Aucune valeur manquante
- Aucun doublon

Données par ligne (By Route – CSV)

Dimensions du jeu de données

- Nombre d'observations : 1 092 438
- Nombre de variables : 4

Variables

- route (object) : identifiant de la ligne
- date (object) : date du service

- day_type (object) : type de jour
- rides (int) : nombre de trajets

Statistiques descriptives clés

- Fréquentation moyenne journalière par ligne : ~5 834 trajets
- Valeur maximale observée : 45 177 trajets

Qualité des données

- Aucune valeur manquante
- Aucun doublon
- Avertissement de type (DtypeWarning) sur la colonne route, nécessitant une standardisation ultérieure

L'analyse exploratoire met en évidence :

Cette phase valide la faisabilité des étapes suivantes : nettoyage léger, standardisation des formats temporels et analyses comparatives inter-villes.

Phase de Nettoyage et Standardisation des Données

Objectifs du nettoyage

La phase de nettoyage vise à :

- Harmoniser les périodes temporelles entre les villes (à partir de 2019).
- Uniformiser la **structure des fichiers** (colonnes, formats, granularité).
- Rendre les données comparables entre **Philadelphie** et **Chicago**.
- Produire des jeux de données propres et prêts pour l'analyse et la visualisation (Power BI).

Nettoyage des données – Chicago

Données par mode (Excel)

Étapes appliquées :

- Suppression des données antérieures à 2019 afin d'assurer la cohérence temporelle.
- Suppression de la variable total_rides, redondante avec les données par mode.
- Transformation des colonnes bus et rail_boardings via un pivot (melt) pour obtenir une structure longue.

- Extraction des variables temporelles Calendar_Year et Calendar_Month.
- Agrégation mensuelle par moyenne afin de passer d'une granularité journalière à mensuelle.

Structure finale :

- Calendar_Year
- Calendar_Month
- Mode (bus / rail_boardings)
- Average_Monthly_Ridership

Données par ligne (CSV)

Étapes appliquées :

- Conversion de la colonne date en format datetime.
- Suppression des données antérieures à 2019.
- Extraction de l'année et du mois.
- Agrégation mensuelle par ligne (moyenne des trajets journaliers).
- Conservation des colonnes essentielles uniquement.

Structure finale :

- Calendar_Year
- Calendar_Month
- route
- Rides

Le volume initial (>1 million de lignes) a été réduit tout en conservant l'information analytique pertinente.

Nettoyage des données – Philadelphie

Données par mode (CSV)

Étapes appliquées :

- Filtrage des modes afin de conserver uniquement Bus et les modes contenant Rail.
- Regroupement des modes ferroviaires (Heavy Rail, Regional Rail) sous une catégorie unique : rail_boardings.
- Agrégation mensuelle par moyenne afin de conserver une valeur représentative.
- Renommage des variables pour assurer la cohérence inter-villes.

Structure finale :

- Calendar_Year
- Calendar_Month
- Mode
- Average_Monthly_Ridership

Données par ligne (CSV)

Étapes appliquées :

- Suppression des colonnes techniques non analytiques (Source, ObjectId).
- Renommage de la variable de fréquentation pour uniformisation (Rides).

Structure finale :

- Calendar_Year
- Calendar_Month
- Route
- Rides

Sauvegarde des jeux de données nettoyés

Les jeux de données nettoyés ont été exportés au format CSV dans un répertoire dédié afin de garantir la traçabilité et la reproductibilité du projet :

- PHI_route.csv
- PHI_mode.csv
- CHI_route.csv
- CHI_mode.csv

Ces fichiers constituent la **base finale** utilisée pour les analyses comparatives et la visualisation sous Power BI.

Cette étape prépare le terrain pour la phase suivante : **analyse exploratoire avancée et visualisation des tendances de fréquentation**.

Modélisation des données & Power BI (Schéma en étoile)

Objectif de la modélisation

La phase de modélisation vise à transformer les données nettoyées en un **schéma en étoile (Star Schema)** optimisé pour l'analyse dans Power BI. Ce choix garantit :

- de bonnes performances de calcul (DAX),
- une lisibilité claire du modèle,
- une séparation nette entre faits et dimensions,
- une comparabilité directe entre villes, modes et routes.

Le modèle final est structuré autour de **2 tables de faits** (ridership par mode et par route) et **4 dimensions** (Date, Ville, Mode, Route).

Tables de faits

FACT_MODE – Ridership mensuel par mode

Rôle : table centrale contenant les indicateurs de fréquentation mensuelle par mode de transport.

Granularité :

- 1 ligne = 1 mode × 1 mois × 1 ville

Colonnes :

- Id_Date : clé étrangère vers DIM_DATE
- Id_City : clé étrangère vers DIM_CITY
- Id_Mode : clé étrangère vers DIM_MODE
- Average_Monthly_Ridership : mesure principale (fréquentation moyenne mensuelle)

Source :

- Fusion des données CHI_mode et PHI_mode

Transformations clés :

- Harmonisation des noms de modes
- Jointures avec les dimensions Date, Ville et Mode
- Suppression des colonnes textuelles remplacées par des identifiants

---

FACT_ROUTE – Ridership mensuel par route

Rôle : table de faits dédiée à l'analyse de la fréquentation par ligne (route).

Granularité :

- 1 ligne = 1 route × 1 mois × 1 ville

Colonnes :

- Id_Date : clé étrangère vers DIM_DATE
- Id_Route : clé étrangère vers DIM_ROUTE
- Id_City : clé étrangère vers DIM_CITY
- Rides : fréquentation moyenne mensuelle par route

Source :

- Fusion des données CHI_route et PHI_route

Transformations clés :

- Agrégation journalière → mensuelle
- Jointures avec dimensions Route, Ville et Date

Tables de dimensions

DIM_DATE – Dimension temporelle

Rôle : permettre les analyses temporelles (tendances, saisonnalité, comparaison annuelle).

Colonnes :

- Id_Date : identifiant unique (format YYYYMM)
- Year : année
- Month : numéro du mois
- Month_Name : nom du mois
- Year_Month : libellé textuel (Month - Year)

Origine :

- Générée à partir des données de routes (CHI + PHI)

DIM_CITY – Dimension géographique

Rôle : identifier la ville associée aux observations.

Colonnes :

- Id_City : identifiant de la ville (CHI / PHI)
- City : nom complet de la ville

Valeurs possibles :

- CHICAGO
- PHILADELPHIE

DIM_MODE – Dimension mode de transport

Rôle : normaliser les modes de transport pour comparaison inter-villes.

Colonnes :

- Id_Mode : identifiant unique du mode
- Mode : type de transport (bus, rail_boardings)

Remarque :

- Les différents types de rail ont été regroupés sous une seule catégorie

---

DIM_ROUTE – Dimension lignes / routes

Rôle : décrire les lignes de transport utilisées dans l'analyse par route.

Colonnes :

- Id_Route : identifiant unique de la route
- Route : code ou nom de la ligne

Origine :

- Extraction des routes distinctes depuis CHI_route et PHI_route

Relations du modèle

- DIM_DATE (1) → FACT_MODE (*)
- DIM_DATE (1) → FACT_ROUTE (*)
- DIM_CITY (1) → FACT_MODE (*)
- DIM_CITY (1) → FACT_ROUTE (*)

- DIM_MODE (1) → FACT_MODE (*)
- DIM_ROUTE (1) → FACT_ROUTE (*)

Toutes les relations sont définies en **one-to-many**, avec propagation de filtre depuis les dimensions vers les tables de faits.

Justification du choix du schéma

Ce modèle permet :

- une analyse multi-niveaux (ville, mode, route, temps),
- une réutilisation facile des dimensions dans Power BI,
- une scalabilité pour l'ajout futur d'autres villes ou modes,
- une écriture simple et performante des mesures DAX.

Cette modélisation constitue une base robuste pour les visualisations, KPI et analyses comparatives présentées dans Power BI.

Figure 1 : schéma en étoile

Analyse de la mobilité urbaine par lignes

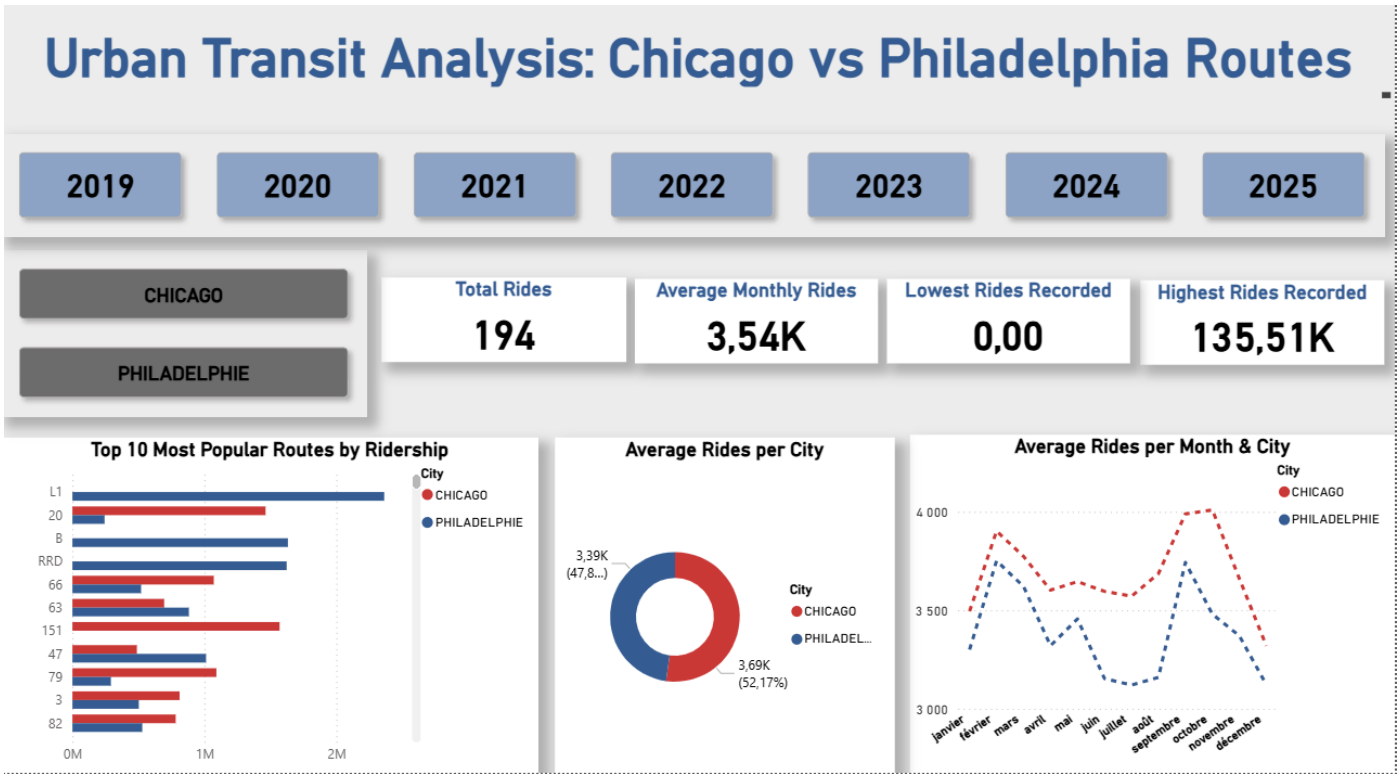

Figure 2 : Analyse de la mobilité urbaine par lignes

Ce tableau de bord propose une **analyse comparative de la fréquentation des transports urbains par lignes (routes)** entre **Chicago** et **Philadelphie**, sur la période **2019–2025**.

Filtres interactifs

- **Filtre temporel** par année (2019 à 2025)
- **Filtre géographique** par ville (Chicago / Philadelphie)

Ces filtres permettent d'analyser dynamiquement l'évolution de la fréquentation selon la période et la ville sélectionnées.

Indicateurs clés (KPI)

- **Total Rides** : nombre total de trajets enregistrés pour la sélection
- **Average Monthly Rides** : moyenne mensuelle des trajets
- **Lowest Rides Recorded** : plus faible niveau de fréquentation observé
- **Highest Rides Recorded** : niveau maximal de fréquentation enregistré

Ces indicateurs donnent une **vision synthétique du volume et de la variabilité de la demande**.

Visualisations principales

- **Top 10 des lignes les plus fréquentées**
  → Comparaison directe des lignes les plus populaires entre Chicago et Philadelphie
- **Répartition moyenne des trajets par ville (donut chart)**
  → Mise en évidence du poids relatif de chaque ville dans la fréquentation globale
- **Évolution mensuelle moyenne des trajets par ville**
  → Analyse des tendances saisonnières et comparaison des dynamiques urbaines

Insight principal :
Chicago présente globalement des volumes de fréquentation plus élevés, tandis que Philadelphie affiche une dynamique plus modérée mais régulière selon les périodes.

Analyse de la mobilité urbaine par mode de transport

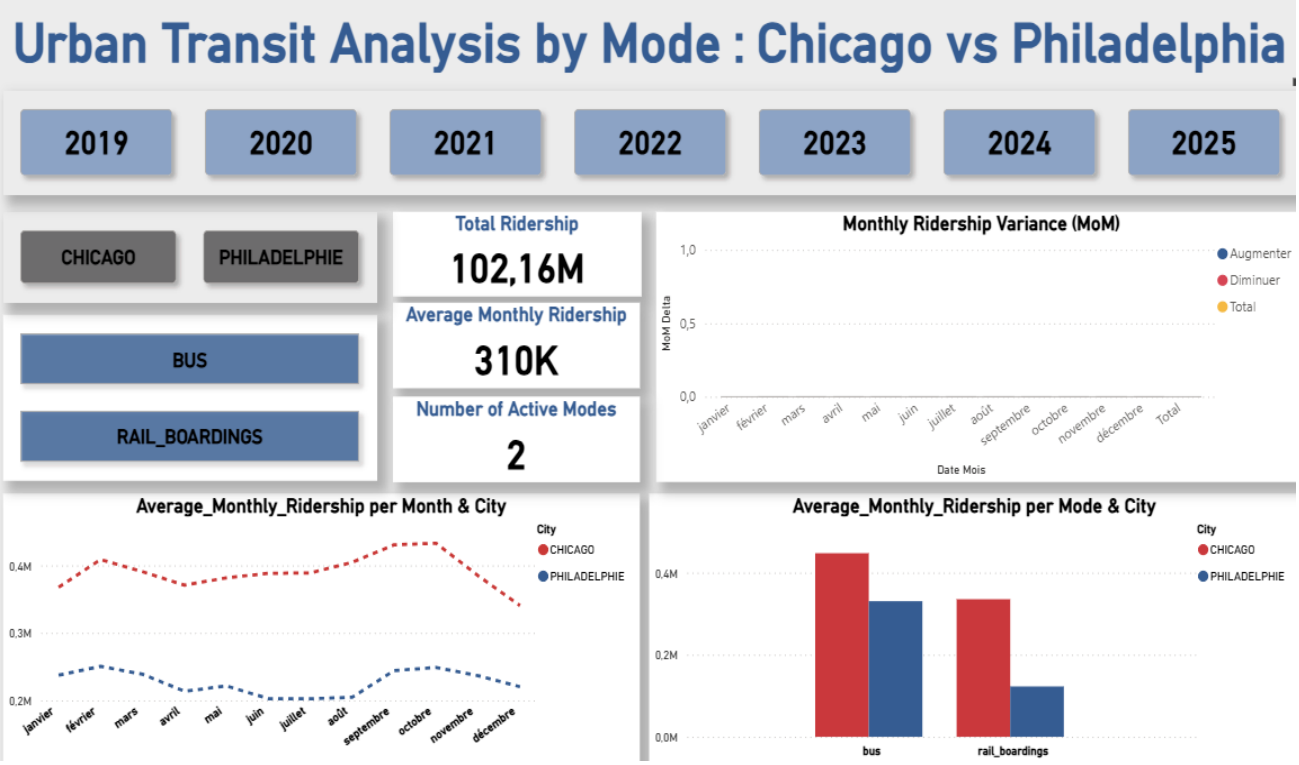

Figure 3 : Analyse de la mobilité urbaine par mode de transport

---

Ce tableau de bord se concentre sur l'**analyse des modes de transport**, principalement **Bus** et **Rail**, afin d'évaluer leur contribution à la mobilité urbaine dans les deux villes.

Filtres disponibles

- **Année**
- **Ville**
- **Mode de transport** (Bus / Rail)

Indicateurs clés

- **Total Ridership** : nombre total de passagers sur la période
- **Average Monthly Ridership** : moyenne mensuelle de passagers
- **Number of Active Modes** : nombre de modes de transport analysés

Ces KPI permettent de **quantifier l'importance globale de chaque mode**.

Visualisations principales

- **Évolution mensuelle moyenne par ville**
  → Comparaison des tendances temporelles entre Chicago et Philadelphie
- **Comparaison de la fréquentation par mode et par ville**
  → Le **bus** apparaît comme le mode dominant dans les deux villes
- **Analyse des flux mensuels (MoM)**
  → Identification des cycles de croissance et quantification de l'impact des crises sectorielles, illustrant notamment la rupture brutale de fréquentation survenue en 2020.

Insight principal :
Le **bus reste le mode de transport le plus utilisé**, tandis que le **rail** joue un rôle complémentaire, plus marqué à Chicago qu'à Philadelphie.

Recommandations Stratégiques Basées sur l'Analyse

Optimisation des lignes à forte fréquentation

Constat : Les analyses par ligne montrent que certaines routes figurent de manière récurrente parmi les plus fréquentées, en particulier à Chicago. Ces lignes concentrent une part significative du ridership total.

Recommandations :

- Renforcer la fréquence de passage sur les lignes les plus sollicitées, notamment aux heures de pointe.
- Prioriser ces lignes pour la maintenance préventive et le renouvellement du matériel roulant.
- Étudier la mise en place de services express sur les axes à forte demande.

Objectif : améliorer la capacité du réseau et réduire la congestion.

Réallocation des ressources sur les lignes à faible performance

Constat : Certaines lignes présentent une fréquentation durablement faible, quel que soit le mois ou l'année analysée.

Recommandations :

- Réévaluer les horaires et la fréquence de ces lignes.
- Adapter le type de véhicule (capacité réduite) sur les segments peu fréquentés.
- Étudier une fusion ou une restructuration de certaines lignes à faible usage.

Objectif : optimiser les coûts opérationnels sans dégrader la qualité globale du service.

Ajustement de l'offre selon les tendances temporelles

Constat : Le ridership présente des variations saisonnières marquées, avec des tendances différentes entre Chicago et Philadelphie.

Recommandations :

- Mettre en place une planification dynamique des fréquences selon les mois.
- Adapter les stratégies de transport par ville, en tenant compte des spécificités locales.

Objectif : aligner l'offre de transport avec la demande réelle des usagers.

Valorisation du bus comme pilier de la mobilité urbaine

Constat : Le mode bus domine largement le ridership dans les deux villes, comparé au rail.

Recommandations :

- Maintenir les investissements prioritaires sur le réseau bus.
- Optimiser la ponctualité et la régularité via des couloirs dédiés et une meilleure gestion du trafic.
- Positionner le rail comme un mode complémentaire sur les axes structurants.

Objectif : maximiser l'impact des investissements publics.

Détection proactive des anomalies de fréquentation

Constat : Le dashboard met en évidence des baisses soudaines et des variations mensuelles négatives sur certaines périodes.

Recommandations :

- Mettre en place des alertes automatiques dans Power BI (baisse MoM significative, écart anormal par rapport à la moyenne).
- Croiser ces anomalies avec des facteurs externes (travaux, événements, incidents).

Objectif : permettre une réaction rapide et limiter l'impact sur les usagers.

Utilisation du dashboard comme outil d'aide à la décision

Constat : Le modèle en étoile et l'interactivité du dashboard facilitent l'analyse par des profils non techniques.

Recommandations :

- Utiliser le dashboard comme outil de pilotage mensuel.
- L'intégrer aux réunions décisionnelles et former les équipes métiers à son exploitation.

Objectif : renforcer une culture data-driven au sein des agences de transport.

Conclusion

Ces recommandations, basées sur l'analyse comparative du ridership à Chicago et Philadelphie, fournissent des leviers concrets pour optimiser l'offre de transport urbain. Le dashboard Power BI constitue ainsi un outil stratégique d'aide à la décision pour une gestion plus efficace, adaptative et orientée usagers du réseau de transport.