

Premier projet data

Analyse descriptive des symptômes et diagnostics COVID-19

Ce projet constitue ma première réalisation pratique en analyse de données. L'objectif était de maîtriser le cycle de vie complet de la donnée, depuis l'extraction jusqu'à la visualisation, en respectant un cahier des charges fonctionnel strict.

Le jeu de données est issu de la plateforme Kaggle : "COVID-19 Patient Symptoms and Diagnosis Dataset". Il recense les données démographiques et cliniques de patients répartis sur plusieurs zones géographiques.

Source :

- [Lien vers le dataset Kaggle](#)
- Lien vers le cahier des charges : [!\[\]\(30a147af384f9f71632c2ff17bc706c8_img.jpg\) Cahier des charges](#)
- Lien vers google sheet :
[!\[\]\(9b33568d5c136f08ca688ce48be37574_img.jpg\) Analyse des symptômes et diagnostics COVID-19](#)

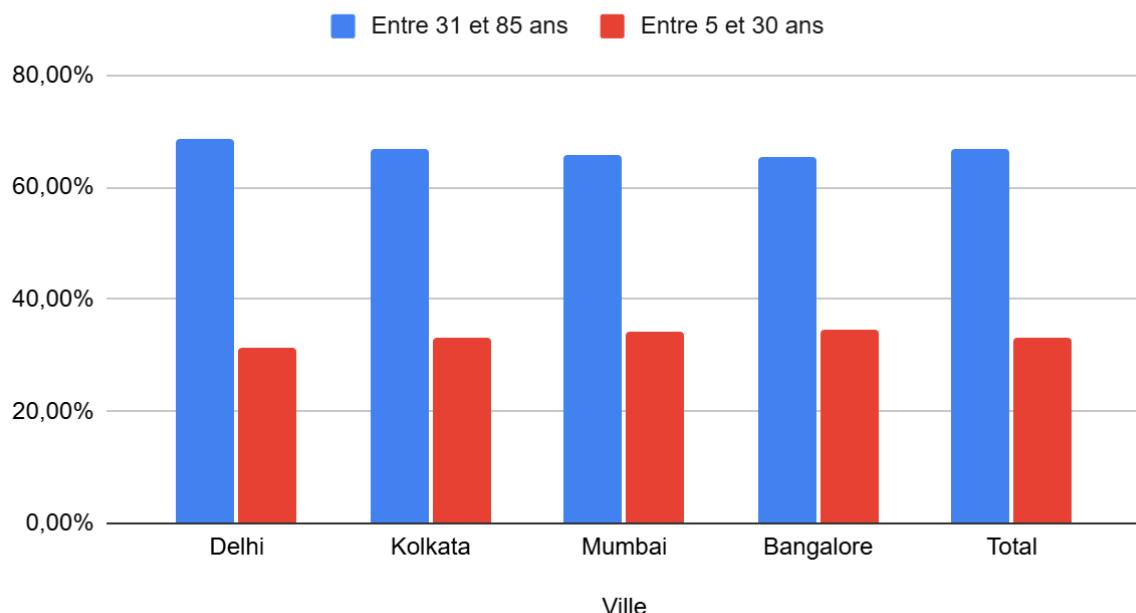
Pour mener à bien cette analyse, j'ai utilisé **Google Sheets** comme outil central :

1. **Préparation des données** : Nettoyage des entrées erronées et normalisation des unités de mesure (conversion températures F°/C°).
2. **Ingénierie des fonctionnalités** : Création de colonnes calculées pour faciliter l'analyse, en utilisant des fonctions logiques telles que **NB . SI . ENS** (COUNTIFS) pour croiser les variables (Âge vs Diagnostic / Ville vs Symptômes).
3. **Visualisation** : Conception de graphiques comparatifs (histogrammes) permettant d'identifier rapidement les corrélations entre l'âge des patients et la prévalence des symptômes.

Ce projet m'a permis de valider mes compétences fondamentales en manipulation de tableurs et en interprétation statistique descriptive.

1. Analyse de la répartition par âge

Répartition des cas par tranche d'âge

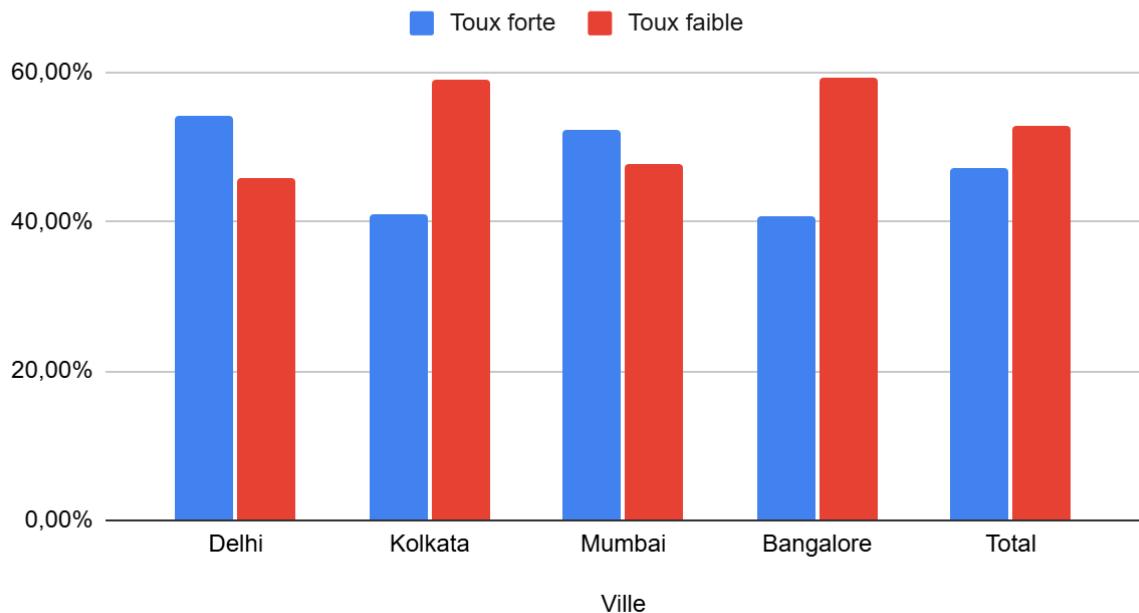


Observation :

On constate une tendance identique dans toutes les villes : la tranche d'âge des **31-85 ans** est systématiquement la plus touchée, représentant environ **65% à 70%** des cas positifs. À l'inverse, les 5-30 ans ne représentent qu'un tiers des contaminations (env. 30-35%).

2. Analyse de la toux comme symptôme

Prévalence de la toux forte chez les cas positifs



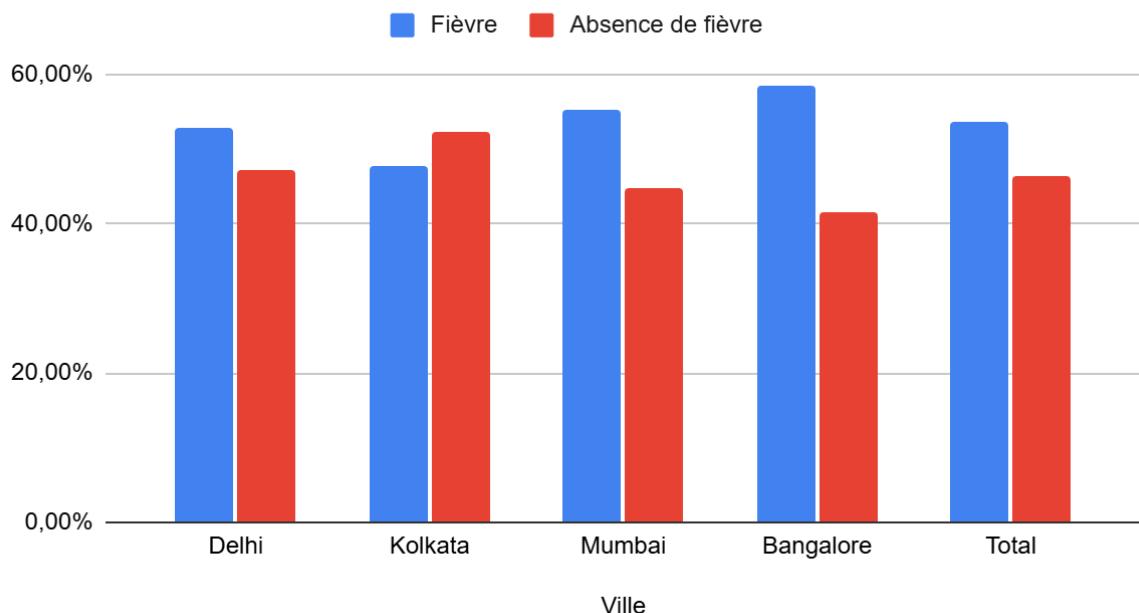
Observation :

Il y a une forte disparité géographique.

- À Delhi et Mumbai, la toux forte est majoritaire chez les patients positifs (barre bleue dominante).
- À Kolkata et Bangalore, c'est l'inverse : la majorité des cas positifs présentent une toux faible ou absente (barre rouge dominante, proche de 60%).
- Au total, la répartition est presque équilibrée (environ 48% forte / 52% faible).

3. Analyse de la fièvre comme symptôme

Prévalence de la fièvre chez les cas positif



Observation :

La fièvre est un indicateur plus constant que la toux.

- Dans 3 villes sur 4 (Delhi, Mumbai, Bangalore), la fièvre est présente chez la majorité des malades.
- Bangalore se distingue avec le taux de fièvre le plus élevé (proche de 60%).
- Kolkata est une exception notable où l'absence de fièvre est légèrement majoritaire

J'en conclus donc que l'analyse croisée des 482 cas démontre que si l'âge (31-85 ans) est un facteur de risque constant, la symptomatologie (fièvre, toux) varie fortement selon les zones géographiques.

Cela suggère l'importance d'adapter les protocoles de détection aux spécificités locales plutôt qu'à un modèle unique.