Project Idea 2 - France COVID-19 Key Figures

Group 4.

2025-09-30

1 Link to the dataset

Kaggle - Coronavirusdataset France (file: `chiffres-cles.csv`)
Actual URL: https://www.kaggle.com/datasets/mclikmb4/coronavirusdataset-france?select=chiffres-cles.csv Google drive URL: https://drive.google.com/file/d/1rXHdGEDWFAMaitmkNSgehAt_e2FaC_PZ/view?usp=sharing

2 Introduction to the dataset

This dataset provides daily COVID-19 surveillance indicators for France at multiple geographic granularities (country, region, department, overseas collectivities). Each record includes a calendar date, a location code, and a location name, enabling comparisons across space and time. Indicators cover hospitalized patients, ICU occupancy, cumulative deaths, cumulative recoveries, and daily flows of new admissions (hospital and ICU). Source/provenance fields support auditability. The structure suits descriptive analyses and visualizations, with optional regional comparisons to highlight spatial heterogeneity. These indicators and their definitions are documented on the Kaggle dataset page (mclikmb4, 2020-2021).

3 Dataset justification

Relevance: Directly biomedical/public-health, reflecting real-world hospital and ICU loads during COVID-19.

Size/structure: The file far exceeds the minimum requirements (well over 100 rows and more than 20 columns) and includes both categorical (granularity, location IDs, sources) and continuous (counts) variables.

Accessibility/ethics: Publicly accessible aggregated, de-identified counts suitable for academic use.

Analytical potential: Enables trend estimation, wave identification, geographic comparison, and lead-lag analysis between admissions (“flow”) and occupancy (“stock”).

Ethical use. The dataset consists of aggregated, de-identified counts without PII; no patient-level identifiers are present, aligning with course requirements for ethical, public data.

4 Variables description

Key columns:
date (daily), granularity (country, region, department), location_code (location code), location_name (location name).

Indicators:
- hospitalized - current hospitalized patients
- icu_patients - current ICU patients
- deaths - cumulative deaths
- recovered - cumulative recoveries
- new_hospitalizations - new daily hospital admissions
- new_icu_admissions - new daily ICU admissions

Additional fields:
confirmed_cases and tested may be present with different levels of completeness.
Note: Due to several missing/invalid values (NaN/Inf), the tested column is largely unusable for analysis and is excluded from primary summaries and plots.

Source metadata:
source_name, source_url, source_archive, source_type.

Table 1: Row counts by geographic granularity

granularity	n
department	40715
region	7708
country	817
overseas_collectivity	131
world	83

Table 2: Summary statistics for key numeric indicators

variable	n	mean	sd	median	min	max
confirmed_cases	3081	121010.685	508142.429	27.0	0	3560764
deaths	47928	920.086	4150.452	135.0	0	70574
hospitalized	46826	578.225	2597.057	91.0	0	33497
icu_patients	46743	80.489	387.667	10.0	0	7148
new_hospitalizations	46095	32.664	166.648	4.0	0	4281
new_icu_admissions	46095	5.421	28.033	0.0	0	771
recovered	46712	3949.800	17835.138	645.5	0	299624
tested	0	NaN	NA	NA	Inf	-Inf

5 Research question(s)

1. **National waves:** How did France's national hospitalization and ICU occupancy evolve across early pandemic waves (2020-2021)?

2. **Flow-stock timing:** Do peaks in new hospital admissions precede peaks in current hospitalizations, and by roughly how many days?

6 Data cleanup and processing plan

- **Parsing and types:** Ensure the `date` field is properly parsed as a date variable and convert indicator fields into numeric types for consistency.
- **Subsetting:** For national trends, include only rows classified as country with `location_code` = "FRA". For geographic comparisons, restrict the dataset to rows where `granularity` is region.
- **Missingness:** Quantify missing values for each column and handle them transparently by applying listwise deletion for plotted series (no imputation).
- **Duplicates:** Identify and remove duplicate entries defined by the combination of `date` and `location_code`.
- **Provenance:** Retain all source metadata fields, and include them in the appendix when relevant for transparency.

7 Descriptive statistics (figures in Appendix)

France's national indicators exhibit multi-wave patterns during 2020-2021. Hospital occupancy and ICU burden rise and fall in tandem with case surges, while cumulative deaths increase monotonically. The timing relationship between new admissions (flow) and current occupancy (stock) suggests admissions lead occupancy by several days. For visuals supporting these statements, see Appendix Figures A1-A3. Tables above summarize structure and central tendencies.

Across all rows, the median current hospitalizations was 91, with an IQR of 25-285; ICU occupancy had a much lower median, which is expected since ICU is a subset of the total hospital (median 10), consistent with ICU being a subset of total hospital burden.

8 Planned statistical methods

- **Lagged cross-correlation** between `new_hospitalizations` (flow) and `hospitalized` (stock) to estimate lead time from admissions to occupancy.
- **Regional comparison** of ICU vs hospital burden by wave period (medians, IQRs).
- **Simple time-series decomposition** on national hospitalizations to separate trend/seasonal/residual components (if applicable).

9 Limitations

Several fields like `tested` and early `confirmed_cases` have bad coverage over time, and indicators are hospital-centric rather than community-representative. Counts are aggregated and de-identified, so patient-level cannot be controlled. Because the dataset mixes granularities (national, regional, departmental), comparing across levels requires careful subsetting (`granularity == "country"` for national trends). These constraints limit causal interpretation, so we have to focus more on descriptive trends and clearly labeled comparisons.

10 Appendix

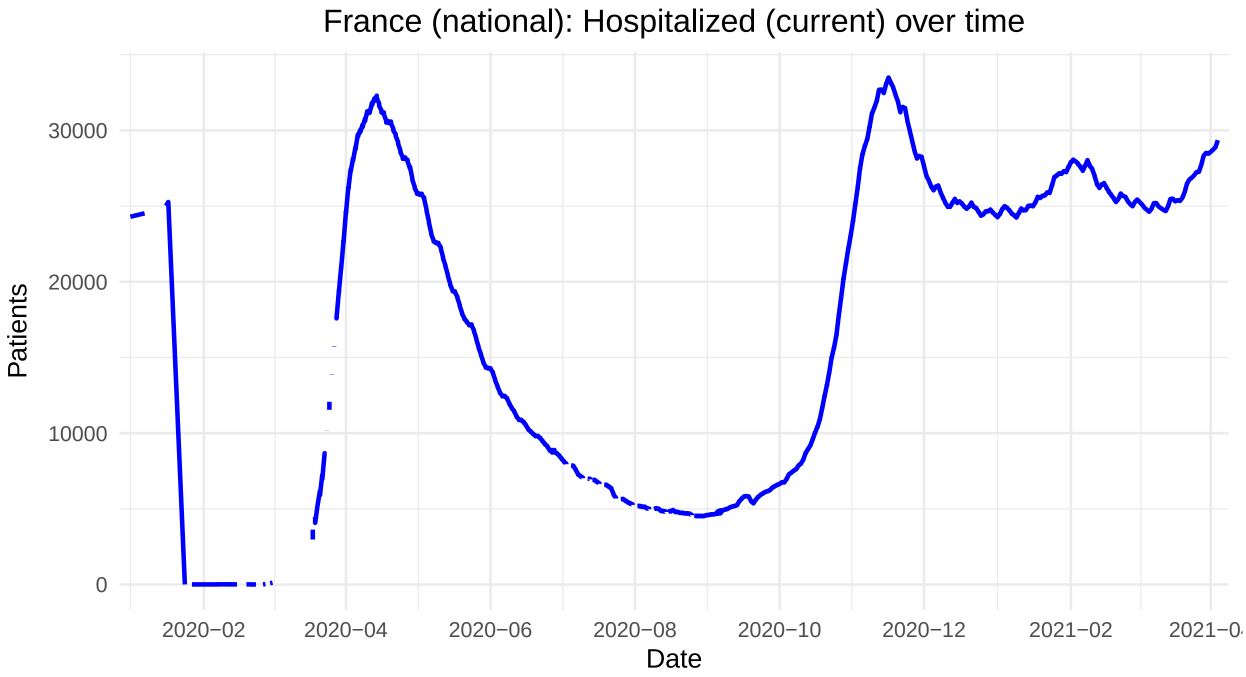

Figure 1: France (national): Hospitalized (current) over time.

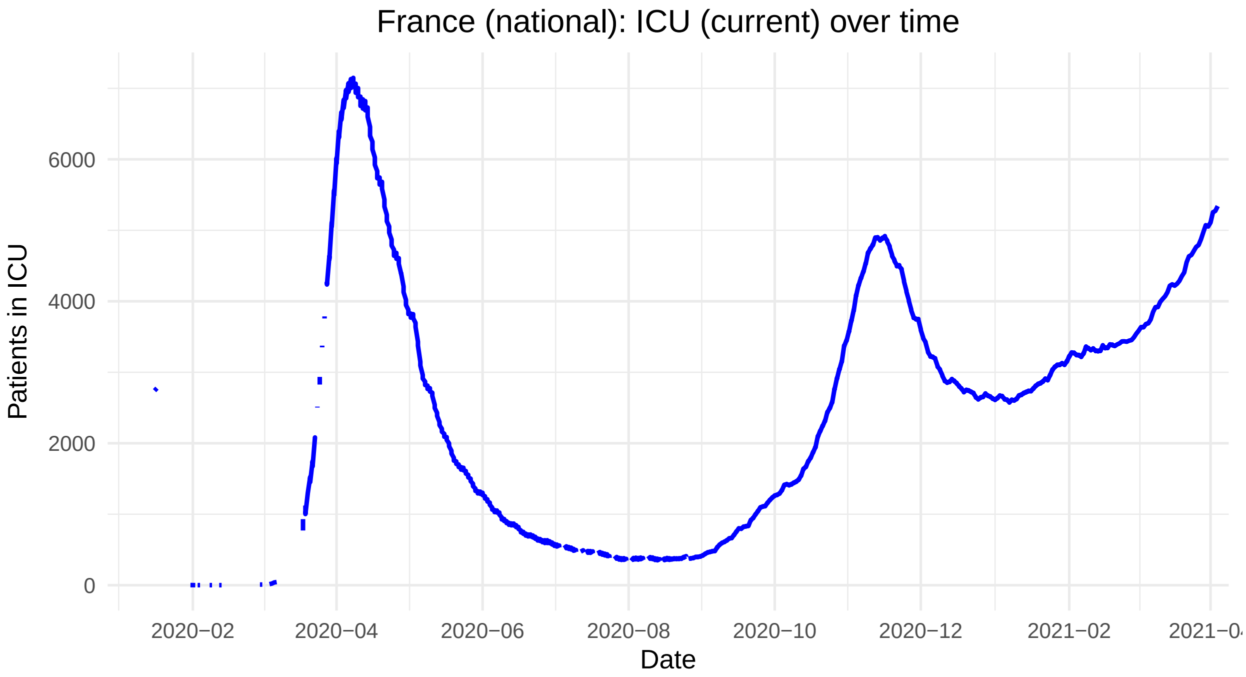

Figure 2: France (national): ICU (current) over time.

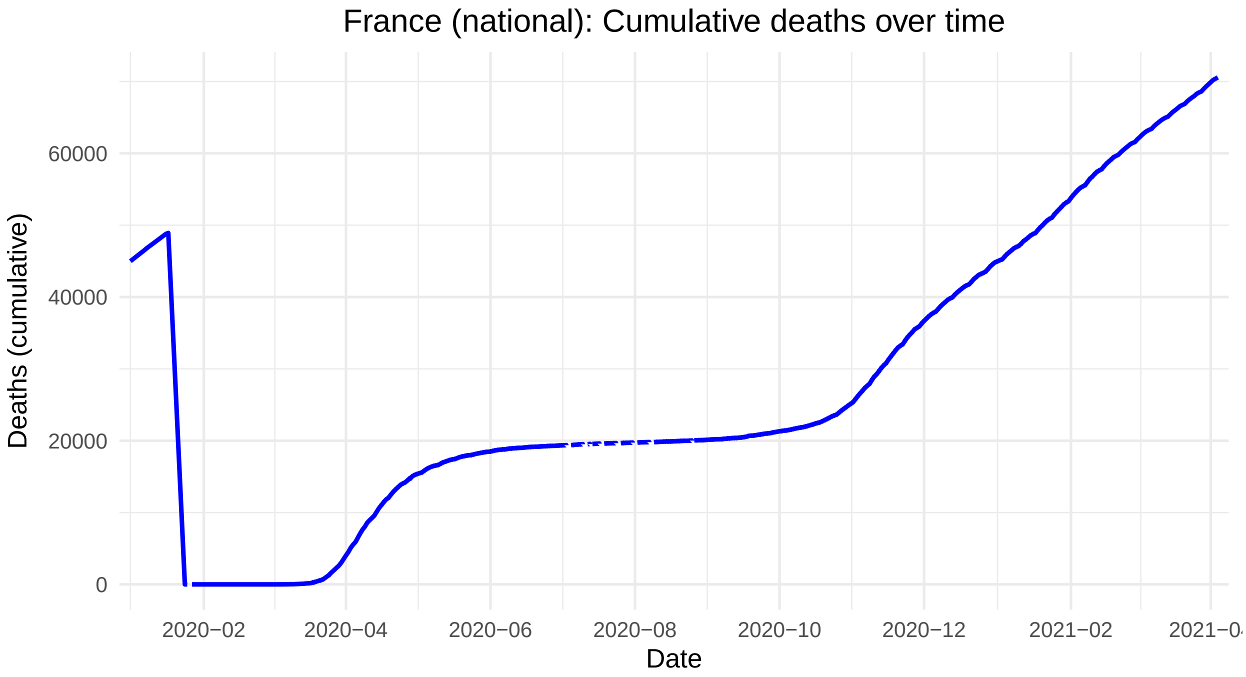

Figure 3: France (national): Cumulative deaths over time.

11 References

mclikmb4, (2021, April 4), Coronavirus-dataset France, Kaggle, https://www.kaggle.com/datasets/mclikmb4/coronavirusdataset-france?select=chiffres-cles.csv