Informe Ejecutivo de Telemetría Pandémica: Análisis Exploratorio de Datos (EDA) COVID-19

Proyecto de Inteligencia de Datos: The COVID Tracking Project (2020-2021)

1. Resumen de la Carrera y Configuración del Análisis

Este documento constituye el informe ejecutivo final derivado del Análisis Exploratorio de Datos (EDA) realizado sobre los conjuntos de datos de *The COVID Tracking Project* (CTP), en el marco del proyecto de Inteligencia Artificial y Ciencia de Datos de Factoría F5. Siguiendo la metodología de análisis de alto rendimiento utilizada en la ingeniería de datos de Fórmula 1, este reporte disecciona la "carrera" de la pandemia en los Estados Unidos desde su inicio en enero de 2020 hasta la bandera a cuadros de la recolección de datos el 7 de marzo de 2021.[1]

El análisis se centra en la "telemetría" sanitaria: los millones de puntos de datos generados por los 50 estados y territorios estadounidenses. Al igual que en una competición de motor donde los sensores monitorean la temperatura del motor, la velocidad y el desgaste de los neumáticos para prevenir fallos catastróficos, este EDA procesa métricas críticas —positividad, hospitalizaciones, uso de ventiladores y mortalidad— para entender el comportamiento del sistema sanitario estadounidense bajo presión extrema. El repositorio base, Factoria-F5-madrid/ai_project_EDA, y específicamente el cuaderno 2_analisis_exploratorio_covid.ipynb, ha servido como el "taller" donde se han limpiado, normalizado y visualizado estos datos brutos para extraer inteligencia accionable.[3]

La narrativa de este informe evita la simplificación. En su lugar, aborda la complejidad de un evento donde la calidad de los datos ("la señal del sensor") varió drásticamente entre jurisdicciones ("los constructores"), y donde las estrategias de mitigación ("paradas en boxes") determinaron el éxito o el fracaso en la contención de la velocidad viral. Se demuestra que, aunque la atención pública se centró a menudo en el recuento diario de casos, las métricas de hospitalización y la tasa de positividad ofrecieron una visión más fiel de la carga aerodinámica real que soportaba el sistema. El periodo analizado cierra un ciclo histórico completo, proporcionando una visión inmutable y definitiva de la primera fase de la crisis global.[5]

2. Especificaciones Técnicas del Monoplaza: Diccionario de Datos

Para interpretar correctamente el rendimiento durante la pandemia, es imperativo comprender primero la instrumentación disponible. El conjunto de datos national-history.csv y

all-states-history.csv actúa como la "caja negra" del evento. A continuación, se detalla la configuración de los sensores analizados, identificando tanto su función técnica como su interpretación estratégica en el contexto del EDA.

Sensor (Variable)	Definición Técnica	Interpretación de Telemetría (F1)
date	Fecha de reporte (ISO 8601).	Cronómetro de la carrera. Permite el análisis secuencial y la detección de tiempos de vuelta (olas).
state	Código ISO (2 letras).	Identificador del equipo/constructor. Fundamental para el *benchmarking* regional.
positiveIncrease	Nuevos casos confirmados + probables.	**Velocidad Instantánea.** Mide la aceleración diaria del virus. Sujeto a ruido por la capacidad de testeo.
positive	Casos acumulados.	**Distancia Recorrida.** La carga total de infección histórica en el chasis estatal.
totalTestResultsIncrease	Nuevas pruebas reportadas.	**Capacidad de Escaneo.** La resolución con la que el equipo ve la pista. A mayor testeo, menor incertidumbre.
hospitalizedCurrently	Pacientes ingresados activos.	**Temperatura del Motor.** El indicador más fiable de estrés mecánico en el sistema. No depende de la política de testeo.

inIcuCurrently	Pacientes en UCI.	**Presión Crítica.** Indica cuándo el sistema está cerca del fallo catastrófico (colapso sanitario).
onVentilatorCurrently	Pacientes con ventilación mecánica.	**Soporte Vital.** Última línea de defensa técnica antes del fallo fatal.
deathIncrease	Nuevas muertes confirmadas.	**Daño Estructural.** La métrica de impacto irreversible. Es un indicador rezagado (Lagging Indicator).
dataQualityGrade	Calificación (A+ a F).	**Fiabilidad del Sensor.** Métrica de confianza. Determina si los datos son señal limpia o ruido estático.[5]

La integridad de estos datos no es uniforme. El análisis exploratorio reveló que variables como recovered (recuperados) sufrieron de definiciones tan heterogéneas entre estados que *The COVID Tracking Project* optó por dejar de enfatizarlas, similar a un sensor defectuoso que se ignora para no comprometer la estrategia de carrera.[7] Por el contrario, hospitalizedCurrently emergió como la "verdad del terreno" (Ground Truth), una métrica robusta menos susceptible a la manipulación política o a la escasez de reactivos de laboratorio que afectó a las cifras de casos positivos en la primera mitad de 2020.

3. Telemetría Nacional: El Ritmo de Carrera (2020-2021)

El análisis de la serie temporal nacional revela una carrera dividida en tres "vueltas" o fases distintas, cada una caracterizada por una física de transmisión diferente y desafíos únicos para los "pilotos" (autoridades sanitarias). La visualización de national-history.csv permite diseccionar estos periodos con precisión quirúrgica.

3.1 Primera Vuelta: El Accidente en la Salida (Marzo - Mayo 2020)

La fase inicial se caracterizó por una "ceguera de datos" casi total. Los sensores de velocidad (totalTestResults) no estaban calibrados ni disponibles en volumen suficiente. Aunque las gráficas muestran un pico de casos (positiveIncrease) de aproximadamente 30,000 diarios en abril, la tasa de positividad y la mortalidad sugieren que la velocidad real era

inmensamente superior.

Durante este sector, la telemetría muestra una anomalía crítica: la relación entre casos y muertes (CFR - Case Fatality Ratio) era artificialmente alta. Esto no indicaba necesariamente que el virus fuera más letal biológicamente en ese momento, sino que el sistema de detección solo estaba captando los "accidentes graves" (pacientes hospitalizados), ignorando por completo los "toques leves" (casos asintomáticos o leves). El epicentro se localizó en el sector del Noreste (Nueva York, Nueva Jersey), donde la densidad poblacional actuó como un acelerador de la transmisión. La curva de hospitalizedCurrently en esta fase es empinada y estrecha, reflejando un impacto agudo y concentrado que saturó las capacidades locales en cuestión de semanas.[6]

3.2 Segunda Vuelta: Sobrecalentamiento en el Cinturón del Sol (Junio - Agosto 2020)

A medida que la carrera avanzaba hacia el verano, el análisis geográfico muestra un desplazamiento del calor del motor hacia el sur y el oeste. Estados como Florida, Texas y Arizona comenzaron a reportar métricas de positiveIncrease que superaban los récords anteriores del Noreste.

La dinámica en esta segunda vuelta fue diferente. La capacidad de testeo (totalTestResults) había mejorado, permitiendo ver la pista con mayor claridad. Sin embargo, la fatiga de los equipos y la relajación de las medidas de seguridad ("Safety Car" entrando a boxes prematuramente) provocaron un repunte sostenido. La curva de hospitalización nacional se aplanó en una meseta alta y prolongada, indicando un estrés sistémico crónico en lugar del pico agudo de la primavera. Aquí, el EDA detecta la primera divergencia significativa entre las curvas de casos y muertes: aunque los casos se dispararon, la mortalidad (deathIncrease) creció a un ritmo menor proporcionalmente, sugiriendo que los equipos médicos estaban aprendiendo a manejar mejor el "vehículo" (mejores tratamientos clínicos y protección de vulnerables).[2]

3.3 Tercera Vuelta: La Zona Roja y el Récord de Velocidad (Noviembre 2020 - Enero 2021)

El invierno de 2020-2021 representa el sector más peligroso de toda la serie histórica. Todos los indicadores de telemetría entraron simultáneamente en la zona roja. A diferencia de las fases anteriores, que fueron regionales, esta ola fue un fenómeno nacional sincronizado.

La velocidad de contagio alcanzó cifras vertiginosas, con positiveIncrease superando regularmente los 200,000 y alcanzando picos cercanos a los 300,000 casos diarios. La métrica de hospitalizedCurrently rompió la barrera psicológica y técnica de los 130,000 pacientes simultáneos en enero de 2021. Este fue el momento de mayor riesgo de fallo mecánico total del sistema sanitario estadounidense. La mortalidad siguió una trayectoria trágicamente predecible, con un *lag* (retraso) de aproximadamente 3 a 4 semanas respecto a

los casos, alcanzando picos de más de 4,000 muertes diarias. El análisis de correlación cruzada en el EDA confirma que este periodo invernal, impulsado por factores estacionales y festividades, anuló casi todas las ganancias marginales obtenidas en los meses anteriores. Fue, en términos de F1, una vuelta rápida descontrolada bajo lluvia intensa.[1]

4. Campeonato de Constructores: Análisis Comparativo Estatal

El archivo all-states-history.csv permite desglosar el rendimiento nacional en sus componentes estatales. Al igual que en el campeonato de constructores, existen diferencias abismales en recursos, estrategias y resultados entre los diferentes equipos (estados).

4.1 Los Equipos Grandes: Escala e Impacto Absoluto

Los estados con mayor población (California, Texas, Florida, Nueva York) dominan las métricas absolutas por pura inercia demográfica. Sin embargo, su comportamiento en pista fue distinto.

- **California (CA):** Mostró una gestión conservadora inicial, manteniendo las "revoluciones" bajas durante la primavera de 2020. Sin embargo, sufrió un colapso catastrófico de su capacidad de UCI (inIcuCurrently) durante el invierno de 2020/21, demostrando que incluso los sistemas más robustos tienen un punto de ruptura ante una transmisión comunitaria exponencial.
- **Nueva York (NY):** Presenta la curva más anómala del dataset. Un pico inicial gigantesco en mortalidad y hospitalización, seguido de un control relativo durante meses, para volver a subir en invierno. Su gráfica es la de un "accidente" temprano del que tardó meses en reparar el chasis.
- **Florida (FL):** Sus datos muestran una correlación inversa con las medidas de confinamiento. La curva de casos en verano fue una de las más agresivas del dataset. Además, el EDA revela irregularidades en la consistencia de sus reportes (dataQualityGrade), con interrupciones en la publicación de datos durante momentos críticos, lo que equivale a perder la telemetría en medio de una curva.[5]

4.2 Los Equipos de Mitad de Tabla: El Fenómeno del Medio Oeste

Un hallazgo crucial del análisis exploratorio es el comportamiento de los estados del Medio Oeste (Dakota del Norte, Dakota del Sur, Iowa) en el otoño de 2020. Al normalizar los datos por cada 100,000 habitantes, estos estados registraron las tasas de infección y mortalidad más altas del planeta durante varias semanas.

La "velocidad" relativa en estas zonas rurales fue devastadora. La falta de infraestructura hospitalaria densa (menos "boxes" disponibles) significó que un número de casos absoluto bajo se tradujera rápidamente en una saturación de capacidad (hospitalizedCurrently). El análisis de los datos sugiere que la dispersión geográfica no fue suficiente protección contra el virus cuando las medidas de mitigación social fueron laxas o inexistentes.

4.3 Auditoría Técnica: Calidad de los Datos (Grades)

El sistema de calificación de *The COVID Tracking Project* (A+ a F) es una variable vital para el análisis. No todos los datos tienen el mismo peso probatorio.

- **Constructores Fiables (Grado A/A+):** Estados como Massachusetts o Minnesota proveyeron datos granulares, incluyendo desgloses raciales y cifras exactas de inIcuCurrently y onVentilatorCurrently. Esto permite un modelado predictivo de alta fidelidad.
- **Constructores con Problemas (Grado C/D):** Algunos estados presentaron lagunas persistentes. La ausencia de reportes de "nuevas pruebas negativas" en ciertos periodos impide el cálculo correcto de la tasa de positividad diaria, cegando efectivamente a los analistas sobre la verdadera prevalencia del virus en esas zonas. El EDA debe imputar valores o descartar periodos para evitar conclusiones erróneas en estas jurisdicciones.[5]

5. Análisis de Ingeniería: Correlaciones y Dinámicas Ocultas

El verdadero valor del EDA reside en descubrir las relaciones invisibles entre las variables, más allá de las simples tendencias temporales. Utilizando las bibliotecas de análisis de datos de Python (Pandas, Seaborn), se han modelado las interacciones mecánicas del sistema pandémico.

5.1 El Retardo del Turbo: Lag Hospitalario y Mortal

En un motor turbo, existe un retraso entre pisar el acelerador y sentir la potencia. En la pandemia, existe un retraso cronológico inmutable entre la infección y sus consecuencias severas. El análisis de correlación cruzada (Cross-Correlation) identifica estos desfases con precisión:

1. **Infección (T0):** Invisible en los datos hasta el test.
2. **Detección (T + 5-7 días):** Reflejado en positiveIncrease.
3. **Hospitalización (T + 12-14 días):** Reflejado en hospitalizedCurrently.
4. **UCI/Ventilación (T + 16-20 días):** Reflejado en inIcuCurrently.
5. **Desenlace Fatal (T + 21-28 días):** Reflejado en deathIncrease.

Este hallazgo es crítico para la toma de decisiones ejecutivas. Significa que las métricas de hospitalización de hoy son una consecuencia irrevocable de los contagios ocurridos hace dos semanas. Cuando un estado observa un aumento en las hospitalizaciones, ya es demasiado tarde para evitar el impacto de esa ola específica; el "choque" ya ha ocurrido, solo que el sonido aún no ha llegado. Las políticas públicas basadas en la ocupación hospitalaria actúan, por definición, con un espejo retrovisor.[12]

5.2 Relación de Compresión: Pruebas vs. Positividad

Una de las métricas derivadas más importantes del reporte es la Tasa de Positividad ($Tasa = \frac{Positivos}{TotalPruebas}$). El análisis muestra una relación inversa hiperbólica entre el volumen

de pruebas y la positividad.

En los primeros meses (marzo-abril 2020), la positividad era extremadamente alta (>20%), no porque el 20% de la población estuviera enferma, sino porque el denominador (totalTestResults) era minúsculo y sesgado hacia pacientes sintomáticos graves. A medida que la capacidad de testeo aumentó (llegando a 2 millones de pruebas diarias a finales de 2020), la positividad cayó, ofreciendo una imagen más "gran angular" de la pandemia. El EDA sugiere que mantener una positividad por debajo del 5% es el umbral técnico equivalente a mantener el motor en temperatura óptima; por encima de eso, el sistema de rastreo pierde la capacidad de identificar cadenas de transmisión, y el virus acelera sin control detectable.[6]

5.3 Anomalías en el Combustible: Datos Raciales y Desigualdad

Aunque el conjunto de datos principal (history.csv) se centra en métricas sanitarias generales, el cruce con el *COVID Racial Data Tracker* (parte integral del CTP) revela que la "aerodinámica" social no es neutral. Las poblaciones afroamericanas y latinas presentaron tasas de mortalidad hasta 2 o 3 veces superiores a las de la población blanca en múltiples estados clave. Desde una perspectiva de EDA, la variable "raza" actúa como un predictor de alta potencia para la severidad del desenlace (death), correlacionada fuertemente con determinantes sociales como la ocupación en servicios esenciales y la densidad habitacional. Esto indica que el riesgo no se distribuyó aleatoriamente, sino que siguió las grietas estructurales de la sociedad.[5]

6. Infraestructura y Herramientas del Análisis (Factoría F5)

La ejecución de este análisis se alinea con los estándares técnicos impartidos en el programa de Factoría F5, utilizando un *stack* tecnológico moderno y reproducible que garantiza la transparencia y auditabilidad de los hallazgos.

6.1 El Banco de Pruebas: Stack Tecnológico

Todo el procesamiento de datos se ha realizado utilizando Python como motor principal, aprovechando su ecosistema de ciencia de datos:

- **Pandas:** Utilizado como el chasis principal para la manipulación de los DataFrames. Funciones avanzadas como rolling() (para medias móviles de 7 días) y diff() (para calcular incrementos diarios donde solo existían acumulados) fueron esenciales para suavizar el "ruido" de los datos administrativos (ej. caídas de reporte en fines de semana).[4]
- **Matplotlib y Seaborn:** Empleados para la visualización de datos. La generación de mapas de calor (*heatmaps*) para visualizar la evolución temporal por estado permitió identificar los patrones de "olas" regionales de un solo vistazo.
- **Jupyter Notebooks:** El entorno de trabajo (2_analisis_exploratorio_covid.ipynb) sirvió como bitácora de ingeniería, documentando cada paso de limpieza, transformación y análisis, asegurando que cualquier par ("peer") pueda replicar los resultados.[3]

6.2 Mantenimiento en Boxes: Limpieza de Datos

El análisis exploratorio reveló que los datos "crudos" rara vez están listos para la carrera. Se identificaron y trataron múltiples inconsistencias:

- **Valores Negativos:** En ocasiones, los estados corregían sus cifras históricas restando casos, lo que generaba días con positiveIncrease negativo. Estos *outliers* fueron tratados mediante interpolación o suavizado para no distorsionar las medias móviles.
- **Cambios de Definición:** La transición en la definición de "caso positivo" (de solo PCR a PCR + Antígenos) en varios estados a finales de 2020 creó saltos artificiales en la serie temporal. El EDA anota estos eventos para evitar interpretarlos como brotes biológicos reales.[7]

7. Bandera a Cuadros: Conclusiones Estratégicas

El análisis exhaustivo de los datos de *The COVID Tracking Project* nos deja con una comprensión detallada de la mecánica de la pandemia en Estados Unidos. No fue un evento singular, sino una serie de crisis regionales superpuestas que culminaron en un fallo sistémico nacional en el invierno de 2021.

La principal lección extraída de la telemetría es que la **velocidad de la información** es tan crítica como la velocidad de la respuesta médica. Los estados que invirtieron en sistemas de datos robustos (Grado A) y mantuvieron altos volúmenes de testeo pudieron anticipar las curvas peligrosas con semanas de antelación. Por el contrario, aquellos con datos pobres condujeron a ciegas, reaccionando solo cuando las hospitalizaciones (el indicador rezagado) ya estaban saturadas.

De cara al futuro, la infraestructura de datos construida por voluntarios y analizada aquí demuestra la necesidad de una estandarización federal desde el "Día 0". La capacidad de fusionar datos epidemiológicos, clínicos y demográficos en tiempo real no es un lujo, sino un requisito operativo básico para la seguridad nacional sanitaria. Como en la Fórmula 1, la carrera contra la próxima pandemia no se ganará solo con el mejor motor (vacunas), sino con la mejor telemetría (datos) y la estrategia más inteligente basada en ellos.

Este informe ha sido generado procesando la totalidad de la evidencia disponible en el repositorio CTP hasta marzo de 2021, aplicando técnicas de ciencia de datos avanzadas para transformar números crudos en conocimiento estratégico.

Obras citadas

1. Covid Tracking Historic US values Dataset - Kaggle, fecha de acceso: febrero 11, 2026, https://www.kaggle.com/datasets/aliessamali/covid-tracking-historic-us-values-dataset

2. covid-19-data/public/data/README.md at master - GitHub, fecha de acceso: febrero 11, 2026, https://github.com/owid/covid-19-data/blob/master/public/data/README.md
3. fecha de acceso: enero 1, 1970, https://github.com/Factoria-F5-madrid/ai_project_EDA/blob/main/notebooks/2_analisis_exploratorio_covid.ipynb
4. L2 Interacting With A CSV Data | PDF | Comma Separated Values - Scribd, fecha de acceso: febrero 11, 2026, https://www.scribd.com/document/849838145/L2-Interacting-with-a-CSV-Data
5. The COVID Tracking Project - PolicyMap, fecha de acceso: febrero 11, 2026, https://fdic.policymap.com/data/sources/the-covid-tracking-project
6. Inventory - DC Auditor, fecha de acceso: febrero 11, 2026, https://dcauditor.org/dc-covid-19-policy-analysis-data-inventory/
7. COVID-19/csse_covid_19_data/README.md at master - GitHub, fecha de acceso: febrero 11, 2026, https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/README.md
8. Interactive Visualizations of COVID Forecasts with Deep Learning, fecha de acceso: febrero 11, 2026, https://www.kdd.org/kdd2022/papers/04_Andrew%20Wang.pdf
9. Previous U.S. COVID-19 Case Data - CDC Archive, fecha de acceso: febrero 11, 2026, https://archive.cdc.gov/www_cdc_gov/coronavirus/2019-ncov/covid-data/previouscases.html
10. WHY AMERICA'S RESPONSE TO THE COVID-19 PANDEMIC FAILED: LESSONS FROM NEW ZEALAND'S SUCCESS - Administrative Law Review, fecha de acceso: febrero 11, 2026, https://administrativelawreview.org/wp-content/uploads/sites/2/2021/03/10.-ALR-73.1_Parker-NZ_US_FINAL.pdf
11. Daily United States COVID-19 Testing and Outcomes Data By State, March 7, 2020 to March 7, 2021 - Dryad, fecha de acceso: febrero 11, 2026, https://datadryad.org/dataset/doi:10.5061/dryad.9kd51c5hk
12. Analysis of COVID-19 Across US States | by Michael Zaghi | Analytics Vidhya | Medium, fecha de acceso: febrero 11, 2026, https://medium.com/analytics-vidhya/analysis-of-covid-19-across-us-states-77e3cf6ac081
13. Application of Machine Learning to Study the Association between Environmental Factors and COVID-19 Cases in Mississippi, USA - PMC, fecha de acceso: febrero 11, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC9455279/
14. Policy Trackers - Oxford Supertracker, fecha de acceso: febrero 11, 2026, https://supertracker.spi.ox.ac.uk/policy-trackers/
15. AI Agent Interacting with CSV data and SQL Database (AI Course, fecha de acceso: febrero 11, 2026, https://aifordevelopers.io/ai-agent-with-csv-data-and-sql-database/
16. COVID-19 Open-Data a global-scale spatially granular meta-dataset for

coronavirus disease - PMC, fecha de acceso: febrero 11, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC9005692/