SMART CHOICE ANALYTICS

Project Charter

NOMBRE DEL PROYECTO	MENTOR	PATROCINADOR
SMART CHOICE ANALYTICS	EMILIO SANTANDER	HENRY

	EXPECTED START DATE	EXPECTED COMPLETION
Proyecto Final Data Science Henry PT-07	29/02/2024	07/06/2024

Equipo del Proyecto

Rol	Name	e-mail	Name-Abr	Git Hub
Machine Learning	Diego Osorio	dosoriofc@gmail.com	DO	dosoriofc
Data Engineer	David Marimón	david.neko26@gmail.com	DM	DaAnMaGi
Data Engineer	Salomon Orozco	Salomonorozcojaramillo@gmail.com	SO	SaloLL
Data Analytics	Marcela Correal	mcorreal@gmail.com	MC	MarceCorreal
Data Analytics	Juan Garate	garatejb@gmail.com	JG	Batxa

Links asociados

- Documentos internos
  - Repositorio del proyecto https://github.com/Batxa/DS_ProjectFinal.git
  - Follow-up de tareas: https://trello.com/b/sH9ofad9/tpfinal
  - Cronograma: https://app.teamgantt.com/projects/gantt?ids=3939144
- Documentos externos (Soy Henry)
  - Presentación y estructura del proyecto: https://docs.google.com/presentation/d/17WLH3D5cgEbNA3XIKYKMEi20Q9pbo6kOXmD74nRhQaI/edit#slide=id.g1f3119b515c_0_841
  - Enunciado del proyeto (README): https://github.com/soyHenry/PF_DS/blob/FULL-TIME/README.md
  - Diccionario de datasets originales" https://docs.google.com/document/d/1ASLMGAgrviicATaP1UJlflpmBCXtuSTHQGWdQMN6_2I/edit
  - Detalles de estructura, hitos y entregables: https://docs.google.com/document/d/e/2PACX-1vRtTsN_N3Z0DTLbh_-Xw2OxhOWeV5jmTlSRUNzTBpWM9mTnxsT03674UheR4f0hfULc2v4_sW3IgDTv/pub

- Rúbrica de evaluación: https://docs.google.com/document/d/1tBuh1LSCmvQB5Wd7-Cj4jj_o5zLk8vtBQFtDhF8oeSY/edit#heading=h.okiecvgw2s1c

Metodología de trabajo

Se utilizará la metodología Scrum-Agile, en la cual se aplicarán los siguientes conceptos:

- Iteraciones cortas (Sprints):
  - El trabajo se divide en 3 sprints de 2 semanas de duración cada uno.
  - Al final de cada sprint se debe obtener uno o más entregables.
- Reuniones diarias (Daily Stand Ups):
  - Reuniones diarias de 45 minutos donde el equipo comparte actualizaciones sobre lo que han hecho desde la última reunión, lo que planean hacer hoy y cualquier obstáculo que estén enfrentando.
  - El resultado de cada reunión será la priorización de tareas y la indicación de los próximos pasos de cada integrante
- Colaboración con el cliente:
  - Implica trabajar en estrecha colaboración con el cliente (mentor) bajo una comunicación abierta y continua, en función de comprender sus necesidades, expectativas y cualquier cambio requerido.
  - El mayor beneficio es que se obtiene un feedback temprano y regular, garantizando que el producto final satisfaga las expectativas.
- Demostraciones y retrospectivas:
  - Al final de cada sprint, se lleva a cabo una demostración del trabajo completado y una retrospectiva para identificar qué se hizo bien, qué se podría mejorar y cómo.
- Entrega continua:
  - Se busca la entrega de software en incrementos pequeños y frecuentes, en lugar de esperar hasta el final del proyecto para entregarlo todo de una vez.
- Equipos multifuncionales y autoorganizados:
  - Los equipos Agile son multifuncionales, lo que significa que incluyen todas las habilidades necesarias para completar el trabajo, y son autoorganizados, lo que les permite tomar decisiones y gestionar su propio trabajo.

Generalidades del Proyecto

Cliente	El proyecto está dirigido a propietarios y directivos de restaurantes y bares de Estados Unidos, interesados en obtener información clave sobre las tendencias e intereses de los clientes del sector, con el propósito de que ésta sirva como input para sus proyectos y estrategias que mejoren el desempeño de su marca.
Alcance	Los datos que involucra este proyecto se limitan a los restaurantes y bares ubicados en el territorio de los Estados Unidos, de todas las categorías desde el año 2010 al año 2021

	Los mismos se disponibilizarán en un Data WareHouse que podrá ser consultada en cualquier momento por el cliente. Los reviews son proporcionados por la plataforma Yelp, y los encontrados en el sistema Google Maps.
Objetivo 1	Desarrollar una herramienta de gestión de la información (Dashboard) que ayude a los directivos de restaurantes a mejorar la elaboración de estrategias o campañas de marketing y los servicios ofrecidos a sus clientes a través de las siguientes funcionalidades: a. Segmentación de clientes: Hacer uso de técnicas de clustering para segmentar a los clientes en grupos homogéneos, en función de sus patrones de consumo y comportamiento, contribuyendo a personalizar ofertas y estrategias de marketing. b. Análisis de sentimientos en reseñas: Aplicar técnicas de procesamiento de lenguaje natural (NLP) y análisis de sentimientos para comprender las opiniones de los usuarios sobre el negocio, extrayendo información valiosa sobre los intereses y gustos de los clientes. c. Detección de tendencias en la industria: Emplear técnicas de análisis de datos y series temporales para identificar tendencias emergentes en la industria de restaurantes, bares y negocios a nivel local y nacional, contribuyendo a la adaptación oportuna del negocio a los cambios en la demanda del mercado. La herramienta permitirá al usuario ejecutar filtros de acuerdo a establecimientos, lugares, categorías, reviews y calificaciones, lo que permitirá sacar conclusiones de valor para el desarrollo de estrategias.
Objetivo 2	Presentar un análisis del mercado de restaurantes y bares estadounidense basados en las reseñas de clientes de las plataformas de Yelp y Google Maps, identificando e ilustrando las características más relevantes del mercado, mediante el análisis de las relaciones entre las variables que generan un mayor impacto en las evaluaciones de los establecimientos. El estudio, basado en la utilización de la herramienta, incluirá elementos tales como: a. Características de los establecimientos tales como: Categoría, Atributos/Misceláneos, Nivel de precios, reviews, Rating/Stars asignados b. Cantidad de comentarios positivos/neutros/negativos (obtenido del análisis de sentimiento) c. Características de los usuarios tales como: número de reseñas suministradas, número de años registrado en Yelp, número de amigos en Yelp d. Ranking de Restaurantes y bares por Comentarios positivos (obtenido del análisis de sentimiento, con número de reseñas como segundo criterio de ordenamiento) e. Ranking de Restaurantes por Comentarios Negativos (obtenido del análisis de sentimiento, con número de reseñas como segundo criterio de ordenamiento)

Entregables Finales	Herramienta de Análisis de Datos: Dashboard que brinda la posibilidad de filtrar por cada uno de los segmentos definidos, y así mismo, visualizar dinámicamente el movimiento que los indicadores presentan, de acuerdo con los intereses del estudio particular.
	Análisis del sector gastronómico en base a los datos que se analizan con la herramienta a entregar.
	Sistema de Clasificación desarrollado con herramientas de *Machine Learning*, que permita el reconocimiento y segmentación de grupos de clientes. Los cuales serán entregados de la siguiente manera: **Sprint 1:** 1. Charter del proyecto (incluye fundamentos del proyecto, stack tecnológico y flujo de trabajo, Alcance, objetivo, entregables). 2. Exploración de los datos (EDA). **Sprint 2:** 3. Reporte de Data Analytics y producto Machine Learning. **Sprint 3:** 4. Dashboard final. 5. Producto final de Machine Learning.
Sprint 1 **EDA** (Exploratory Data Analysis)	Para la realización del EDA, se realiza en forma previa la fase de ETL (Extraction, transformation and loading of data) en forma separada, debido a la gran cantidad de información. La estructura y contenido del EDA es el siguiente: * Importación de librerías * Carga de datos * Pre-procesamiento de datos: gestión de tipos de datos, valores duplicados y nulos, gestión de características * Análisis de datos: incluye series de tiempo, distribuciones y formatos diversos para analizar la evolución a través del tiempo de características * Conclusiones y resumen de hallazgos
Sprint 1 **KPIs**	Se incluirán en el Dashboard 3 indicadores, diseñados para apoyar el diseño de las estrategias de mejora de los restaurante y bares. Los mismos estarán graficados dinámicamente en el Dashboard, herramienta que permitirá iterar gráficamente suposiciones de comportamiento. Los 3 KPIs son los siguientes:
KPI 1	Aumentar en un (10%) Número de Calificaciones con valores mayores a (3.5), en el último trimestre
KPI 2	Aumentar en un (10%) Número de Reseñas consideradas como Positivas en el último trimestre
KPI 3	Aumentar en un 10% las ventas del último trimestre

Esquema de Trabajo

El proyecto consta de un total de 3 bloques o sprints con las siguientes características:

- Sprint 1: Puesta en marcha
  - Duración: 2 semanas
  - Actividad principal: Contextualización de situación y propuesta de solución

- Sprint 2: Data Engineering
  - Duración: 2 semanas
  - Actividad principal: Implementación de estructura de datos

- Sprint 3: Data Analytics y Machine Learning
  - Duración: 2 semanas
  - Actividad principal: Desarrollo final de la herramienta de análisis de la información y el producto de *Machine Learning*.

Stack Tecnológico	
Lenguaje Principal de programación	Python
Gestor Base de datos	SQL
Repositorio para alojar código	Git Hub
Herramienta de seguimiento del Proyecto	Trello
Plataforma de lanzamiento	Visual Code Studio
Bibliotecas de Python	Pandas, Numpy, LTK,SCikitl Learn
Servicio de Nube	GCP - Fabric

Datasets a utilizar

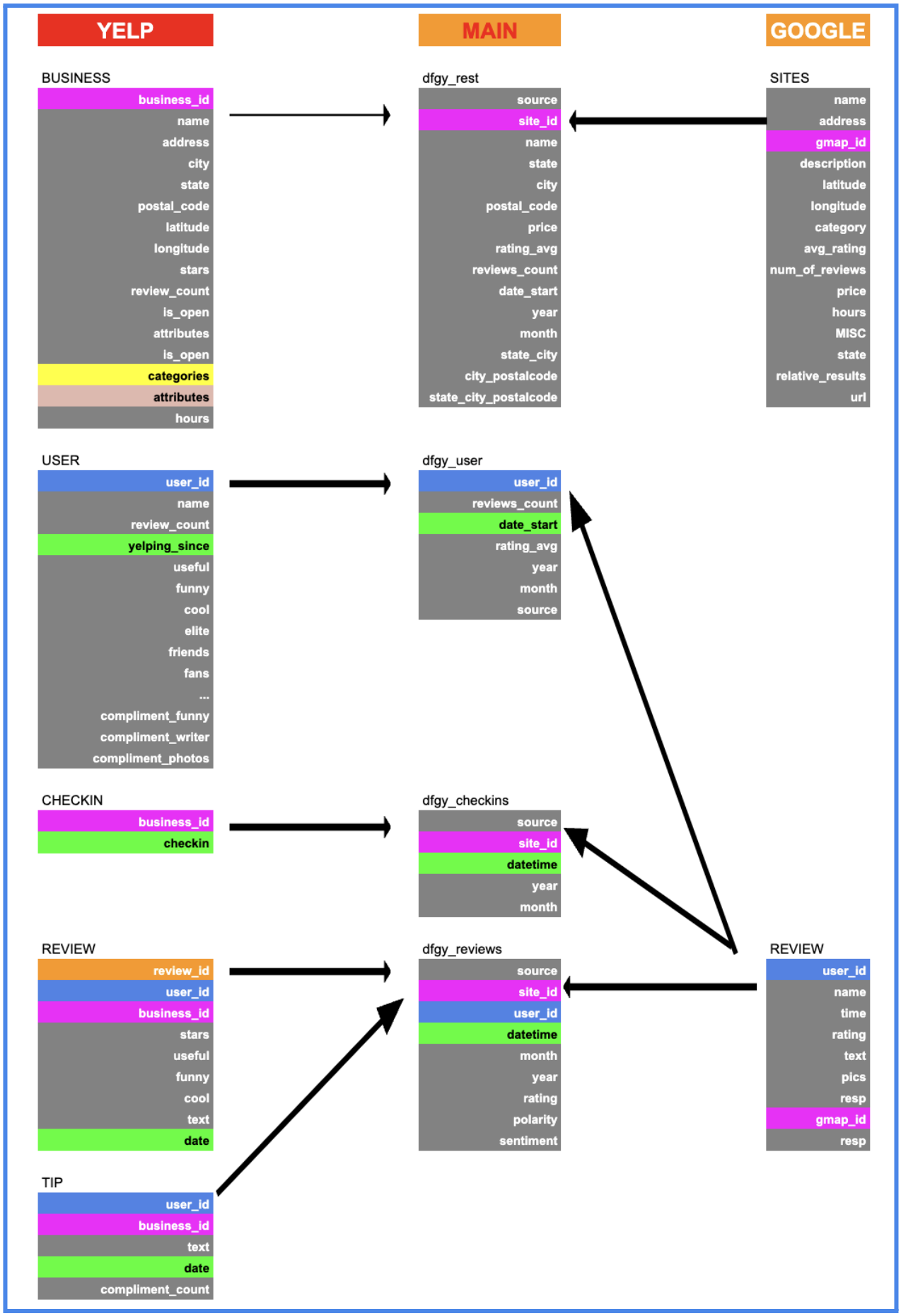
- Fuentes de datos obligatorios
  - Datasets de YELP
    - Business: contienen la información del comercio
    - Checkin: contiene registros de check-in o visitas de clientes
    - User: contiene la data propia del usuario
    - Tip: son consejos o sugerencias rápidas por parte del usuario
    - Review: contiene las reseñas completas
  - Datasets de GOOGLE
    - Sites: contiene información de los sitios
    - Reviews: contiene reseñas de usuarios
- Fuentes de datos adicionales
  - Códigos postales de USA
  - Datos demográficos

    - Datos de crecimiento económico
    - https://data.gov/

Alcance de la información a utilizar:

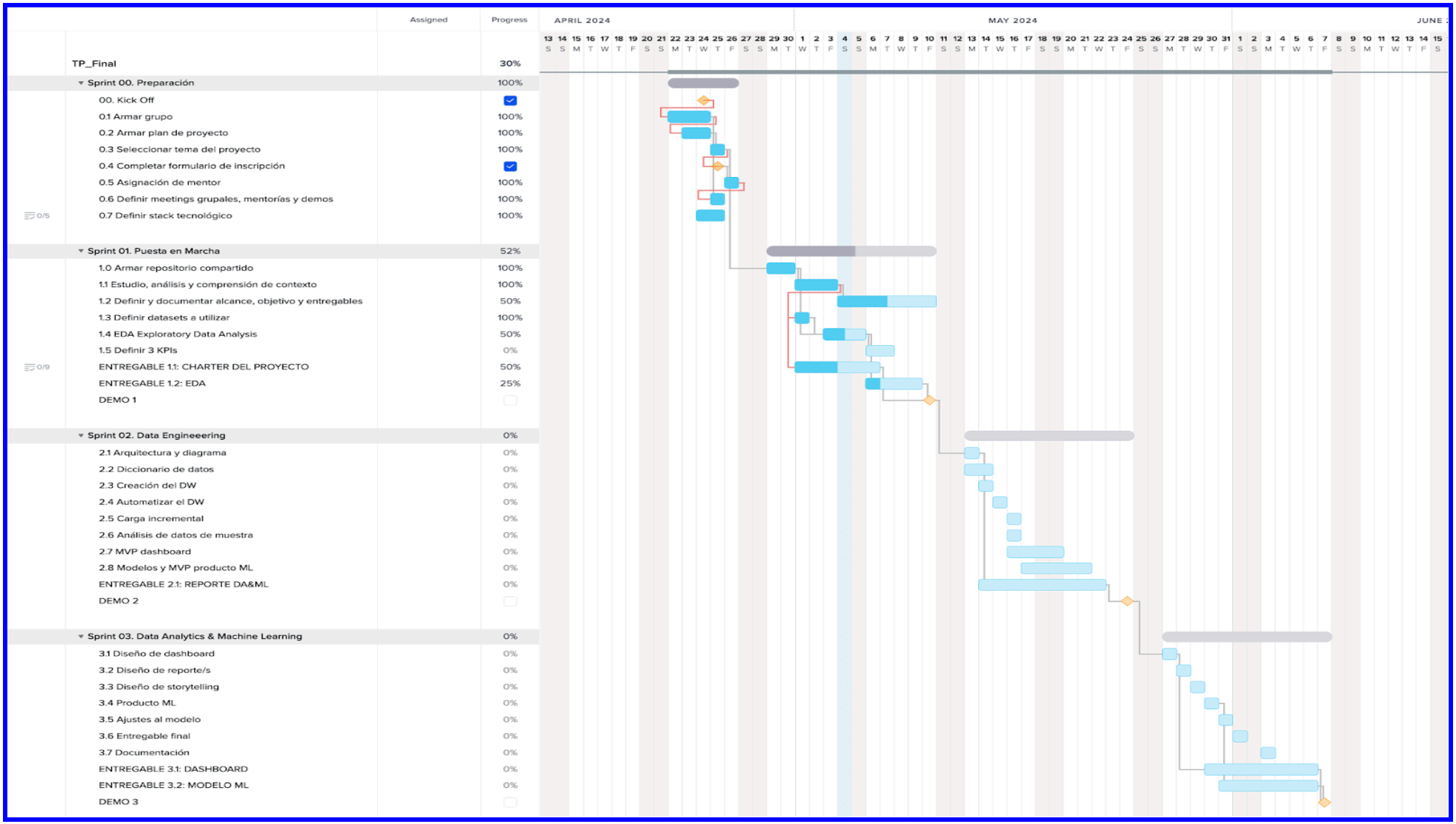
- Alcance de rubros: restaurantes y bares
- Alcance geográfico: Estados Unidos
- Alcance temporal: 2010 en adelante, la información llega hasta el 2021
- Sesgos posibles: la información contiene datos del año 2020 siendo el covid 19 un factor de alto impacto. A efectos del EDA se incluirá esta información, y luego se definirá si se utiliza o no para las proyecciones del producto final

Flujo y estructura de datos:



:

Plan y cronograma de trabajo (Esquema Gantt)



	Assigned	Progress
TP_Final		30%
Sprint 00. Preparación		100%
00. Kick Off		☑
0.1 Armar grupo		100%
0.2 Armar plan de proyecto		100%
0.3 Seleccionar tema del proyecto		100%
0.4 Completar formulario de inscripción		☑
0.5 Asignación de mentor		100%
0.6 Definir meetings grupales, mentorías y demos		100%
0/5 0.7 Definir stack tecnológico		100%
Sprint 01. Puesta en Marcha		52%
1.0 Armar repositorio compartido		100%
1.1 Estudio, análisis y comprensión de contexto		100%
1.2 Definir y documentar alcance, objetivo y entregables		50%
1.3 Definir datasets a utilizar		100%
1.4 EDA Exploratory Data Analysis		50%
1.5 Definir 3 KPIs		0%
0/9 ENTREGABLE 1.1: CHARTER DEL PROYECTO		50%
ENTREGABLE 1.2: EDA		25%
DEMO 1		☐
Sprint 02. Data Engineeering		0%
2.1 Arquitectura y diagrama		0%
2.2 Diccionario de datos		0%
2.3 Creación del DW		0%
2.4 Automatizar el DW		0%
2.5 Carga incremental		0%
2.6 Análisis de datos de muestra		0%
2.7 MVP dashboard		0%
2.8 Modelos y MVP producto ML		0%
ENTREGABLE 2.1: REPORTE DA&ML		0%
DEMO 2		☐
Sprint 03. Data Analytics & Machine Learning		0%
3.1 Diseño de dashboard		0%
3.2 Diseño de reporte/s		0%
3.3 Diseño de storytelling		0%
3.4 Producto ML		0%
3.5 Ajustes al modelo		0%
3.6 Entregable final		0%
3.7 Documentación		0%
ENTREGABLE 3.1: DASHBOARD		0%
ENTREGABLE 3.2: MODELO ML		0%
DEMO 3		☐