Informe OLAP

Sistemas de la Información 2024-2025

22/01/2025

Óscar Lestón Casais

Contenido

Índice de Figuras

1. Resumen

Este proyecto académico ofrece una visión integral de la Fórmula 1, analizando datos históricos desde 1950 hasta la mitad de la temporada 2024. A través del diseño de herramientas OLAP, el objetivo es facilitar el análisis estratégico, operativo y comercial. Las fuentes de datos incluyen un extenso *dataset* de *Kaggle* y APIs complementarias para información de clima y comunicaciones de radio.

El sistema permite explorar estadísticas de rendimiento, analizar factores como clima, estrategias en boxes y audiencias televisivas. Además, se han generado tableros interactivos para visualizar resultados, meteorología y comunicaciones de radio que han sido capturadas en tiempo real, destacando patrones clave y tendencias en la Fórmula 1. Las transformaciones ETL estructuran los datos en un modelo optimizado para consultas avanzadas, abarcando información desde tiempos de vuelta y resultados de carreras hasta detalles de circuitos y condiciones climáticas.

Este análisis representa una herramienta poderosa para optimizar decisiones y entender la dinámica competitiva y comercial de la Fórmula 1.

2. Introducción

El presente proyecto académico se centra en la industria de la Fórmula 1, un deporte de alta competencia donde cada segundo cuenta, tanto dentro como fuera de la pista. Este proyecto busca proporcionar una visión integral y multidimensional de los datos relacionados con las carreras, los equipos, los pilotos y diversos factores externos, con el fin de facilitar el análisis estratégico, operativo y comercial. Mediante la implementación de un esquema en estrella y un entorno analítico, se pretende ayudar a resolver desafíos clave en la competencia y en el negocio de la Fórmula 1.

El sistema permitirá a los usuarios explorar y analizar grandes volúmenes de datos históricos desde 1950 hasta mediados de la temporada de 2024 (de momento), abarcando estadísticas de rendimiento de equipos y pilotos junto con el impacto de variables externas como las condiciones meteorológicas y la audiencia televisiva. Esto se traduce en la posibilidad de optimizar estrategias en pista mediante el análisis detallado de patrones de éxito y áreas de mejora, evaluar el impacto de factores como los tiempos en boxes y el clima en los resultados, y comprender las dinámicas comerciales del deporte, como las tendencias de audiencia y su relación con las decisiones estratégicas.

Como visión de este proyecto se aspira a constituir una herramienta que transforme datos complejos en decisiones fundamentadas, potenciando el rendimiento y el impacto comercial de la Fórmula 1 en un contexto dinámico y exigente.

3. Fuentes de datos y modelado.

La fuente de información primaria utilizada fue el *dataset* de *Kaggle* "*Formula 1 World Championship History (1950-2024)"* [1] con los siguientes ficheros .CSV acompañados de la descripción que aporta su autor y la dimensionalidad de los datos de cada uno:

1. **Track_Information.csv:** Contiene detalles sobre los circuitos donde se han llevado a cabo carreras, incluyendo su ubicación, longitud y características únicas. (77x8)
2. **Team_Details.csv:** Proporciona información sobre los constructores, incluyendo su historia, logros y rendimiento a lo largo de diferentes temporadas. (212x4)
3. **Constructor_Performance.csv:** Detalla el rendimiento de los constructores en carreras individuales, mostrando cómo han evolucionado los equipos a lo largo de los años. (12415x5)
4. **Constructor_Rankings.csv:** Ofrece las clasificaciones anuales de los constructores, destacando las dinámicas competitivas dentro del deporte. (13516x6)
5. **Driver_Details.csv:** Incluye información completa sobre los pilotos, como sus datos personales, estadísticas de carrera y logros. (1150x8)
6. **Driver_Rankings.csv:** Muestra las clasificaciones anuales de los pilotos, destacando quién lideró las tablas y qué tan reñidas fueron las competencias por el campeonato. (34608x6)
7. **Race_Schedule.csv:** Enumera todas las carreras realizadas desde 1950 hasta 2024, junto con detalles como la fecha, ubicación y nombre de la carrera. (1125x17)
8. **Race_Results.csv:** Proporciona resultados detallados de cada carrera, incluyendo posiciones finales, puntos obtenidos y otros métricas clave. (26686x16)
9. **Lap_Timings.csv:** Contiene datos sobre los tiempos de vuelta registrados por los pilotos durante las carreras, ofreciendo información sobre la consistencia de su desempeño. (575880x5)

10. **Pit_Stop_Records.csv:** Ofrece información sobre las paradas en boxes realizadas durante las carreras, incluyendo tiempos y estrategias, que a menudo influyen en el resultado de la carrera. (10987x6)
11. **Qualifying_Results.csv:** Detalla los resultados de las sesiones de clasificación, que determinan la parrilla de salida para cada carrera. (10040x9)
12. **Sprint_Race_Results.csv:** Incluye datos sobre las carreras sprint, carreras más cortas introducidas para decidir las posiciones de salida para la carrera principal. (300x14)
13. **Season_Summaries.csv:** Resume cada temporada, incluyendo el número de carreras, campeones y momentos clave.[1]
14. **Race_Status.csv:** Contiene códigos y descripciones relacionadas con el estado de los autos durante una carrera, como si un auto terminó, se retiró o fue descalificado. (139x2)

Cada fichero .CSV se corresponde de forma directa con una tabla en la base de datos como veremos en el diagrama estrella de la Figura 1 representadas por las tablas con el nombre en morado. Existen contadas excepciones como *Season_summaries.csv* o alguna columna de *race_schedule.csv* que son enlaces a la Wikipedia y se descartan porque no se consideran de valor.

Como fuentes secundarias se usaron:

- API *openf1*, para la obtención de URLs a ficheros mp3 de las comunicaciones radio entre piloto y boxes. Dará lugar a una dimensión en el diagrama estrella *radio_coms* (3497x4) [2]
- API *open-meteo* en su versión histórica para obtener información meteorológica por horas de los días de carrera. Dará lugar a la dimensión *race_weather* (27000x12) [3]. Contiene temperatura, humedad, lluvia, presión, cobertura de nubes, viento (velocidad, dirección y ráfagas) y radiación solar para cada hora del día de carrera.
- *Liberty Media Group digital summary*, proporciona datos de espectadores de televisión por año de fórmula 1 entre el 2008 y el 2021. Da lugar a la dimensión *tv_viewership* (14x2).

Todas ellas representadas en naranja en Figura 1.

[1] Se descarta, solo vincula un año al enlace de la Wikipedia de la temporada y no se considera útil.

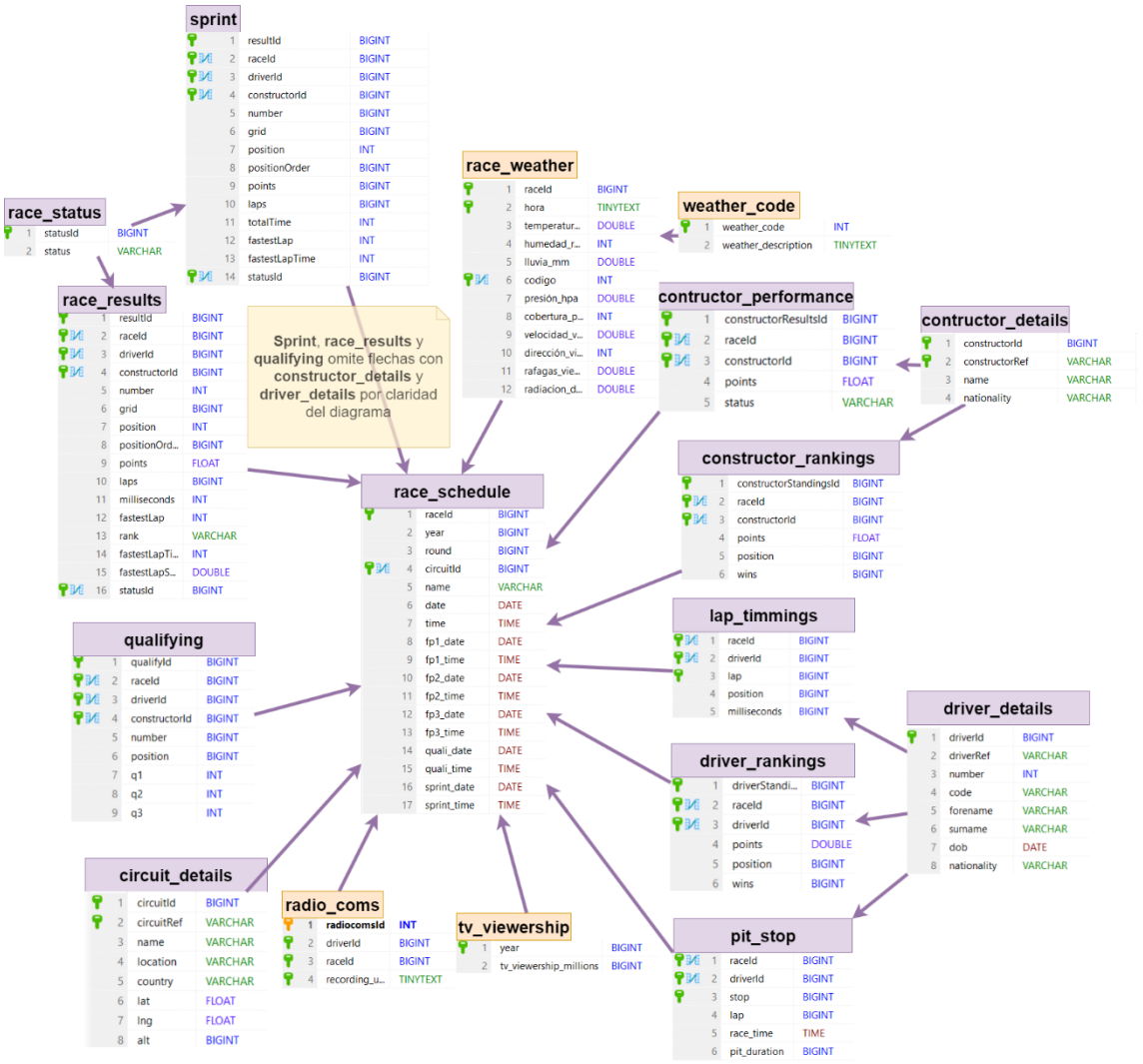

Figura 1: Diagrama estrella

4. Descripción procesos ETL

Las transformaciones de las que consta el proceso ETL están estructuradas de forma que hay una transformación para cada tabla de la base de datos.

La lógica empleada para las trasformaciones que se refieren al *dataset* que se usa como fuente de información primaria y para la fuente secundaria de *tv_viewership* (ya que también parte de un .CSV) siguen el mismo esquema (N entradas → N salidas) (Figura 2):

1. Extraer filas del fichero CSV.
2. Modificar el formato si es necesario para que sea compatible con le base de datos.[2]
3. Guardar en la tabla que corresponda.

[2] Este paso es especialmente necesario en caso de fechas u horas, pero no es necesario para todas las transformaciones, como por ejemplo para *pit_stop* que no fue necesario (Figura 3). Por otro lado, desconozco si al haber utilizado MariaDB este formateo es menos automático que si usase MySQL.

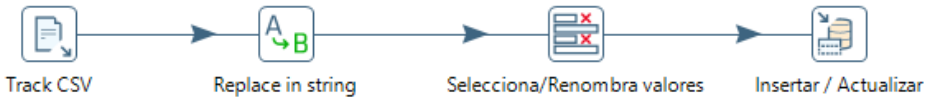

Figura 2: Lectura de CSV con adaptación de formato

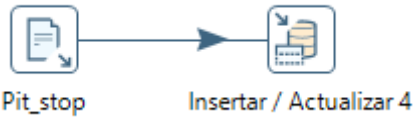

Figura 3: Lectura de CSV sin adaptación de formato

Una vez guardados estos datos en el *data warehouse* podemos seguir con las transformaciones de las fuentes secundarias ya que necesitaremos hacer uso de las fuentes primarias para relacionar los datos que vamos a obtener de las APIs.

Para la obtención de las comunicaciones de radio se usa la API *open-meteo*. Se realizan peticiones a la URL:

*https://api.openf1.org/v1/team_radio?driver_number=**number**&date=**formattedDate***

Donde number y formattedData se corresponden al número de piloto (*driver_details*) y un día de carrera (*race_schedule*) (no existen múltiples carreras en un día). En el "script formatear fecha y espera" la espera es necesaria para que la api no deniegue la petición. Al recibir un 200 OK trasformamos el JSON con la respuesta en las filas adaptadas para la tabla *radio_coms*. El numero de elementos obtenidos depende de las entradas que devuelva la API pudiendo ser una relación de 1: N donde N puede ser 0.

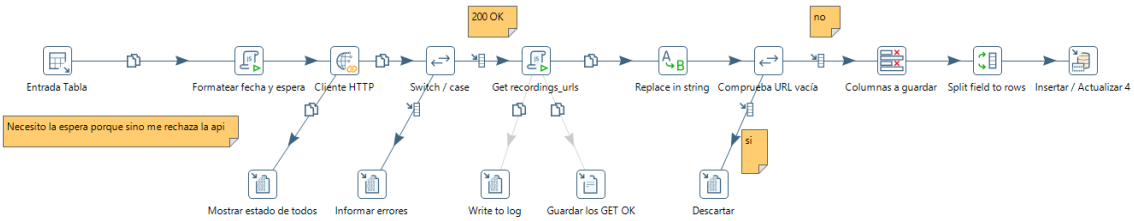

Figura 4: Transformación para registros de radio

El resumen de los pasos fue el de la Figura 5.

#	Nombre paso	Numero Copia	Le◆do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Entrada Tabla	0	0	6321	6321	0	0	0	0	Finalizado	0.1s	70,233	-
2	Formatear fecha y espera	0	6321	6321	0	0	0	0	0	Finalizado	53mn 29s	2	-
3	Cliente HTTP	0	6321	12642	0	0	0	0	0	Finalizado	53mn 29s	4	-
4	Switch / case	0	6321	6321	0	0	0	0	0	Finalizado	53mn 29s	2	-
5	Get recordings_urls	0	6319	6319	0	0	0	0	0	Finalizado	53mn 29s	2	-
6	Informar errores	0	2	2	0	0	0	0	0	Finalizado	53mn 29s	0	-
7	Replace in string	0	6319	6319	0	0	0	0	0	Finalizado	53mn 29s	2	-
8	Comprueba URL vacía	0	6319	6319	0	0	0	0	0	Finalizado	53mn 29s	2	-
9	Mostrar estado de todos	0	6321	6321	0	0	0	0	0	Finalizado	53mn 29s	2	-
10	Columnas a guardar	0	604	604	0	0	0	0	0	Finalizado	53mn 29s	0	-
11	Split field to rows	0	604	4164	0	0	0	0	0	Finalizado	53mn 29s	1	-
12	Insertar / Actualizar 4	0	4164	4164	4164	3796	0	0	0	Finalizado	53mn 29s	1	-
13	Descartar	0	5715	5715	0	0	0	0	0	Finalizado	53mn 29s	2	-

Figura 5: Resultado API radio

A nivel anecdótico los dos errores que informa son errores internos de la API al hacer una solicitud con parámetros correctos. Figura 6.

url	formatte...	HTTP_RESPONSE	HTTP_RESPONSE_CODE
https://api.openf1.org/v1/team_radio?driver_number=63&date=2021-05-02	2021-05-02	An error occurred <pre>Traceback (most recent call last): File "/usr/local/lib/python3.10/site-packages/http...	500
https://api.openf1.org/v1/team_radio?driver_number=77&date=2021-05-02	2021-05-02	An error occurred <pre>Traceback (most recent call last): File "/usr/local/lib/python3.10/site-packages/http...	500

Figura 6: Errores 500 radio

Para comprobar que no pusiésemos valores repetidos se realiza la consulta de la consulta de la Figura 7.

```
1 SELECT recording_url_item, COUNT(*) AS repeticiones
2 FROM radio_coms
3 GROUP BY recording_url_item
4 HAVING COUNT(*) > 1;
5
```

radio_coms (0r × 2c)

#	recording_url_item	repeticiones

Figura 7: Comprobación repetidos radio

Para la obtención de los registros climáticos se usa la siguiente consulta:

*https://archive-api.open-meteo.com/v1/archive?latitude=**lat**&longitude=**lng***
*&hourly=**hourly**&start_date=**formattedDate**&end_date= **formattedDate**&format=**format***

Los parámetros *lat* y *lng* son latitud y longitud obtenidos de *circuit_details*. Los parámetros *format* y *hourly* son comunes se insertan a través del producto cartesiano (Figura 8). El formato puede ser CSV o JSON en este caso me conviene el JSON porque usando *Pentaho Spoon* me resultó más sencillo gestionar la salida de cliente HTTP tipo JSON que una CSV. *Hourly* es la receta de parámetros que nos interesa que devuelva la API para cada hora del día seleccionado (lluvia, humedad…).

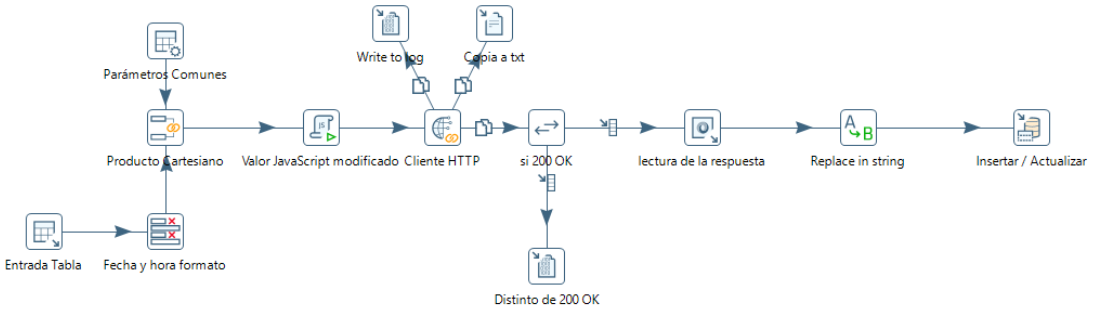

Figura 8: Transformación para registros climáticos

En esta transformación el número de entradas será el numero de filas de *race_schedule* y el número de salidas será 24 filas por cada entrada 1: 24 como máximo dependiendo de la disponibilidad de la API. 1125x24=27000. Para este caso el resumen se ve en Figura 9. A nivel anecdótico al hacer el producto cartesiano, con el que concateno los valores fijos *hourly* y *format* a las columnas extraídas de la base de datos, me dejé una segunda fila con valores a en blanco. Esa es la razón por la que en esta captura en el paso de producto cartesiano se pasa de 505 a 1006. La API devuelve un código distinto al 200 para esas peticiones así que en el paso de lectura de la respuesta la entrada vuelve a tener la cardinalidad correcta.

#	Nombre paso	Numero Copia	Le♦do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Parámetros Comunes	0	0	2	0	0	0	0	0	Finalizado	0.0s	2.000	-
2	Entrada Tabla	0	0	503	503	0	0	0	0	Finalizado	0.0s	100.600	-
3	Fecha y hora formato	0	503	503	0	0	0	0	0	Finalizado	0.0s	55.889	-
4	Producto Cartesiano	0	505	1006	0	0	0	0	0	Finalizado	0.4s	2.712	-
5	Valor JavaScript modificado	0	1006	1006	0	0	0	0	0	Finalizado	4mn 21s	4	-
6	Cliente HTTP	0	1006	3018	0	0	0	0	0	Finalizado	4mn 21s	12	-
7	Write to log	0	1006	1006	0	0	0	0	0	Finalizado	4mn 21s	4	-
8	si 200 OK	0	1006	1006	0	0	0	0	0	Finalizado	4mn 21s	4	-
9	lectura de la respuesta	0	503	12072	12072	0	0	0	0	Finalizado	4mn 21s	46	-
10	Replace in string	0	12072	12072	0	0	0	0	0	Finalizado	4mn 21s	46	-
11	Copia a txt	0	1006	1006	0	1006	0	0	0	Finalizado	4mn 21s	4	-
12	Distinto de 200 OK	0	503	503	0	0	0	0	0	Finalizado	4mn 21s	2	-
13	Insertar / Actualizar	0	12072	12072	12072	12024	0	0	0	Finalizado	4mn 22s	46	-

Figura 9: Resultado Tranformación tiempo

Por otro lado, esta captura es de una parte de la transformación ya que a mitad de ejecución la API me cortó el acceso. Como las entradas estaban ordenadas, esta captura es del resultado de la ejecución del resto de peticiones. Además de eso, una de las variables que me interesa del tiempo es un código WMO [4] que resume el tiempo en esa hora. Estos códigos tienen asociadas descripciones proporcionadas por la organización mundial de meteorología. Para vincular esas descripciones generé la tabla *weather_code* con esa relación. Figura 10Figura 11

```
codigo;descripción
0;'Cielo despejado'
1;'Principalmente despejado'
2;'Parcialmente nublado'
3;'Nublado'
4;'Cubierto'
5;'Cielo con polvo,arena o humo'
6;'Neblina'
```

Figura 10: Código descripción tiempo CSV

#	weather_code	weather_description
1	0	Cielo despejado
2	1	Principalmente despejado
3	2	Parcialmente nublado
4	3	Nublado
5	4	Cubierto

Figura 11: Tabla Código descripción tiempo

Con esto finaliza el recorrido por todas las transformaciones del ETL.

5. Análisis de resultados

El presente análisis se centra en diversos aspectos clave del panorama histórico y actual de la Fórmula 1, utilizando tableros visuales interactivos que permiten explorar datos relevantes. A través de mapas, gráficos y *dashboards*, se busca responder preguntas sobre la ubicación de circuitos, la evolución del número de carreras, los pilotos más exitosos, los resultados de los grandes premios y las comunicaciones entre pilotos y equipos.

¿Dónde se ubican los circuitos de f1?

Se crea un mapa del mundo con las localizaciones de cada circuito usando sus coordenadas (en Figura 12 se puede ver la parte de Europa, Oriente Medio y norte de África).

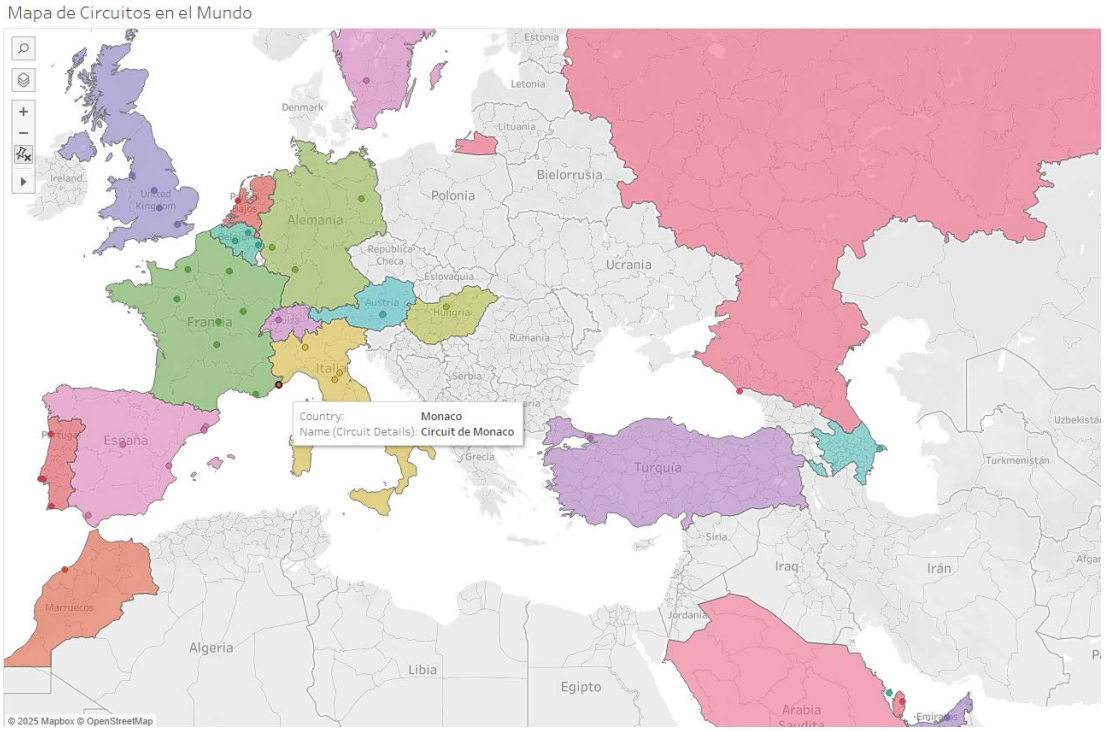

Figura 12: Mapamundi con circuitos

Puede resultar útil ya que la densidad de los circuitos se correlaciona directamente con la movilización de la región para con los eventos de fórmula 1.

¿En qué años hubo más carreras? ¿Como progresa el número de carreras a lo largo de los años?

En la Figura 13 se puede ver un diagrama de barras con el número de carreras por año. Se pasa de 7 carreras al año en 1950 a 24 en la temporada de 2024. En el rango de años de los 70 a los 80 progresivamente se llegan a duplicar las carreras por año consolidándose en 16. Especial mención al 2020 que por la pandemia del COVID se reducen el número de carreras.

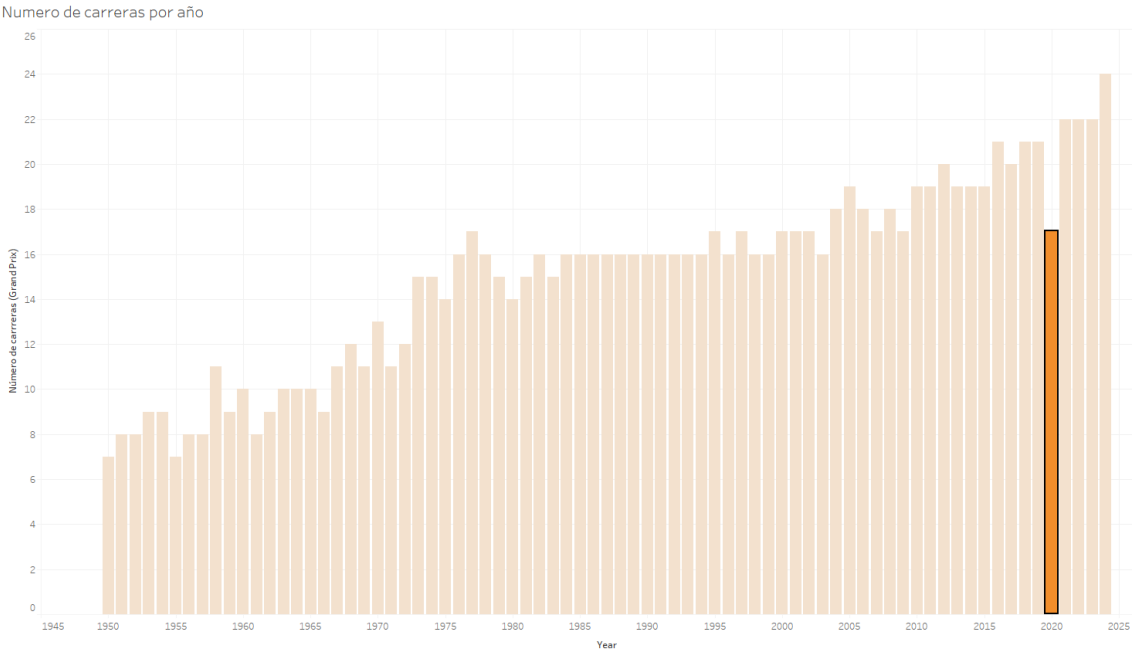

Figura 13: Diagrama de barras de número de carreras por año

Puede resultar útil a nivel informativo ya que una progresión ascendente en numero de carreras por año se traduce en que el mercado de la fórmula 1 admite expansión a nuevos países y regiones con todo lo que conlleva.

¿Cuáles han sido los mejores pilotos de la historia (con más podios o más veces primero)?

En la Figura 15 y Figura 14 se pueden ver los pilotos con más victorias (finalizar primero una carrera de Grand Prix) y podios (finalizar en un intervalo de 1º a 3º una carrera de Grand Prix) de la historia de la fórmula 1.

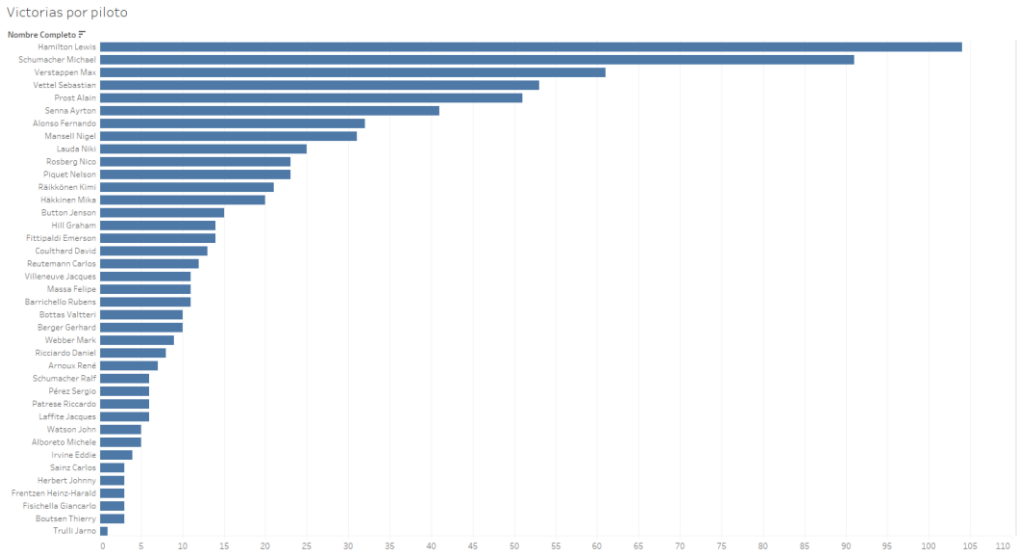

Figura 14: Los 50 pilotos con más victorias de la historia de la F1

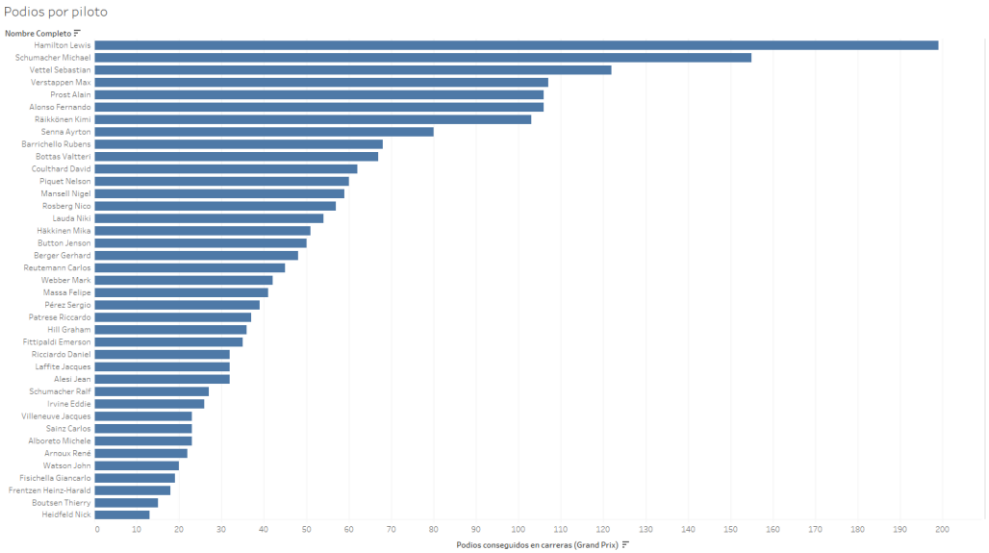

Figura 15: Los 50 pilotos con más podios a lo largo de la historia de la F1

Hay muchas formas de interpretar estas gráficas por que estar en una posición alta es más complicado para un piloto retirado que corriese en los orígenes de la formula 1 ya que el numero de carreras por año entonces era menos de la mitad de las que hay hoy en día. Sin embargo, puede resultar revelador para alguien ajeno al panorama competitivo que pilotos retirados hace años sigan apareciendo en los rankings. Mi conclusión es que fueron pilotos que ganaron casi todo lo ganable en su tiempo y los pilotos en activo que se ponen por encima de ellos siguen la misma dinámica como *Hamilton* o *Verstappen* que siguiendo la tendencia de que a lo largo de una temporada hay un claro favorito ya sea por su coche o por la habilidad del piloto.

¿Qué resultados tuvo cada gran premio?

Durante la realización de este proyecto me resultó muy atractiva la idea de realizar una pantalla interactiva que mostrase los resultados de cada Gran Premio. En la Figura 16 puede verse el visualizador que se realizó. Esto permite una exploración interactiva de resultados por año y gran premio. Además, aporta información extra sobre la meteorología del día de la carrera con progresión en horas (Figura 17). Así podemos saber si llovió en algún momento de la carrera, temperatura, viento o incluso condiciones de presión ambiental (fundamental en el deporte) o visibilidad reducida (hoy en día hay normas de seguridad más estrictas respecto a esta variable).

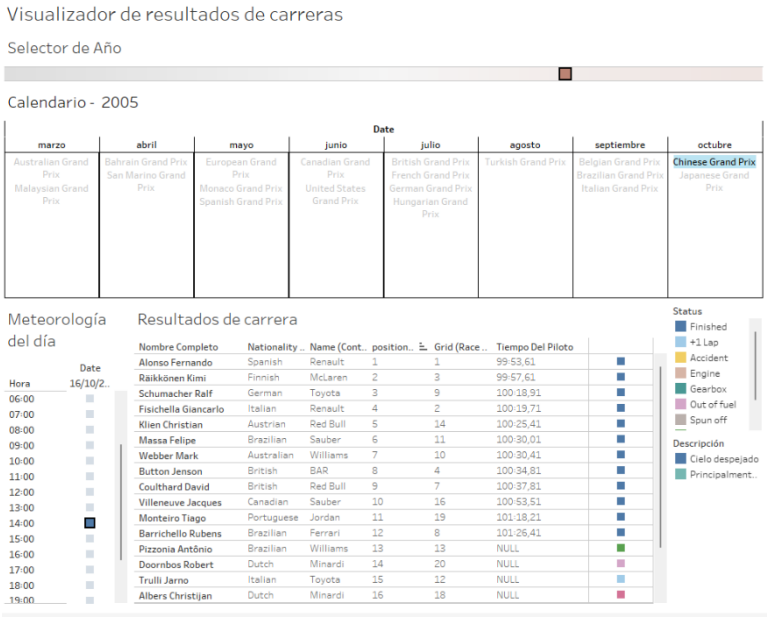

Figura 16: Visualizador de resultados de carreras

Cobertura Porcentaje:	0
Date:	16/10/2005
Dirección Viento Grados:	219
Hora:	14:00
Humedad Relativa:	67
Lluvia Mm:	0
Presión Hpa:	1017,9
Radiacion Directa Wm2:	0
Rafagas Viento Kmh:	7,6
Temperatura C:	15,9
Velocidad Viento Kmh:	2,3
Weather Description:	Cielo despejado

Figura 17: Informe meteorológico

Haciendo uso de este tablero veo que los tiempos de carrera parecen ir a menos según avanzan los años. Esto puede deberse a la mejor tecnología de los coches, la reducción del número de vueltas y el perfeccionamiento de técnica de los pilotos.

¿Cuáles son las comunicaciones entre piloto y equipo entre carrera?

Se generó el *dashboard* de la Figura 18 que permite escuchar las comunicaciones entre piloto y equipo que podrán buscarse por gran premio. Sobre este panel es importante que no hay datos recopilados de antes de 2023.

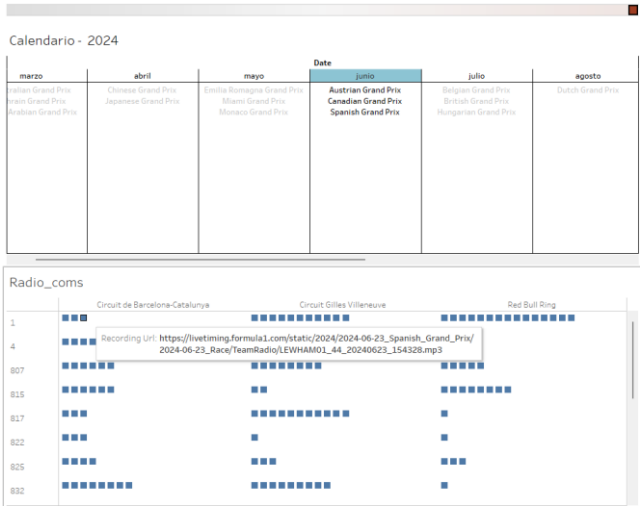

Figura 18: Dashboard comunicaciones radio

El análisis de los tableros muestra una expansión significativa de la Fórmula 1 a nivel global, evidenciada por el incremento en el número de circuitos y carreras anuales, lo que refleja un mercado en constante crecimiento. Además, los datos destacan la excelencia de pilotos históricos y actuales, demostrando que el rendimiento está influenciado tanto por la habilidad individual como por las ventajas tecnológicas. Asimismo, se identificó una tendencia a la reducción de tiempos de carrera gracias a avances en la ingeniería automotriz y la profesionalización de los pilotos. Por último, los *dashboards* interactivos, como los que presentan resultados de carreras y comunicaciones por radio, subrayan la importancia de la información en tiempo real y la meteorología en la toma de decisiones estratégicas, ofreciendo una nueva perspectiva sobre la evolución del deporte.

6. Bibliografía

[1] M. Ehsan, «Formula 1 World Championship History (1950-2024),» kaggle, [En línea]. Available: https://www.kaggle.com/datasets/muhammadehsan02/formula-1-world-championship-history-1950-2024.

[2] openf1.org, «openf1.org,» openf1.org, [En línea]. Available: https://openf1.org/#team-radio.

[3] open-meteo, «open-meteo,» open-meteo, [En línea]. Available: https://open-meteo.com/en/docs/historical-weather-api.

[4] W. M. Organization, «World Meteorological Organization,» World Meteorological Organization, [En línea]. Available: https://wmo.int/.