CIENCIA DE DATOS II

Docente: M.Sc. Alcides RAMOS CALCINA.

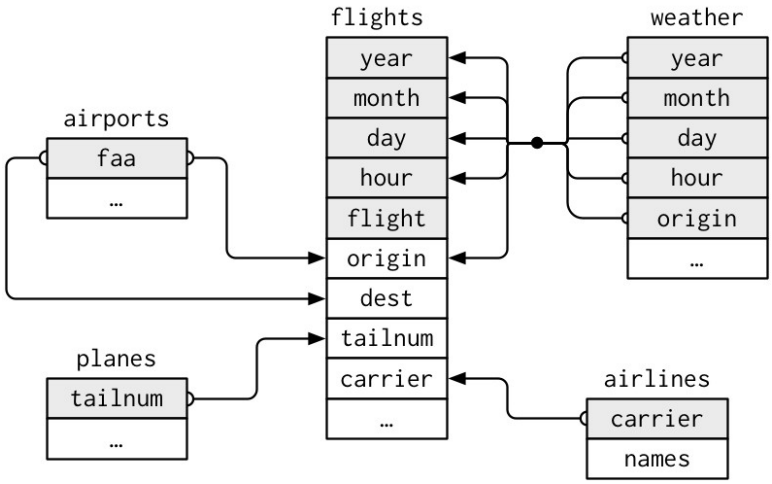
ANÁLISIS DESCRIPTIVO – *NYC Flights 13*

El paquete **nycflights13** contiene información sore todos los vuelos que partieron desde Nueva York (EWR, JFL, LGA) en destinos a los Estados Unidos en 2013. Fueron 336,776 vuelos en total. Para ayudar a comprender las causas de los retrasos, también incluye otros conjuntos de datos útiles.

Este paquete incluye las siguientes tablas:

- flights : todos los vuelos que salieron de Nueva York en el 2013
- weather : datos meteorológicos por hora de cada aeropuerto
- planes : información de construcción de cada avión
- airports : nombres y ubicaciones de aeropuertos
- airlines : relación entre nombres y códigos de las aerolíneas.

Las tablas de este paquete y sus relaciones son las siguientes:



Integración y limpieza

1) Integrar las tablas “flights” y “airlines” por “carrier”
2) A la tabla anterior integrar “planes” por “tailnum”
3) Consulta la estructura de la base de datos “flights”, así mismo; clase, dimensiones y resultados descriptivos.
4) Verificar la frecuencia y datos faltantes
5) Buscar filas o registros duplicados

Situación del problema 1

Te acabas de incorporar a una empresa consultora en Inteligencia de Negocios, actualmente están brindando servicios de análisis para la industria de la aviación y les interesa tener a la aerolínea American Airlines como cliente ya que es una de las aerolíneas líderes en los aeropuertos de Nueva York, motivo por el cuál te han contratado.

Te han pedido que identifiques, ¡cómo puede dicha aerolínea mejorar su posición competitiva!

Para identificar oportunidades de mejorar la posición competitiva de la aerolínea American Airlines, necesitas realizar algunos análisis, para determinar si hay variaciones en la posición de liderazgo de dicha aerolínea.

Manejo de datos

1. Consulta y explora el data frame planes y weather para que conozcas su contenido.
2. Se necesita saber de cada vuelo, la aerolínea, el aeropuerto de origen y el aeropuerto destino.
3. En la consulta anterior se necesita conocer el nombre de la aerolínea.
4. Se necesita saber la cantidad de vuelos por cada destino para identificar cuáles son los destinos más buscados. Agregar el nombre de la aerolínea al data frame anterior.
5. Se necesita conocer las aerolíneas (clave y nombre) y destinos que vuelan por la Mañana: de 6 a 12, Tarde: de 12 a 19, Noche: de 19 a 24 y Madrugada de 24 a 6. Agrega un nuevo campo a la tabla con el nombre de clas_horario y agrega, mañana, tarde, noche y madrugada según sea el caso.
6. Se necesita saber la cantidad de vuelos por aerolínea y destino que hay por la mañana, tarde, noche y madrugada. (group_by() y count()).
7. Se necesita saber a qué destinos vuela la aerolínea American Airlines Inc.-AA, durante la madrugada. (select(), filter(), group_by()).
8. ¿Qué aviones utiliza la aerolínea AA? Aerolínea, tipo, motor y número de asientos. ¿Cuántos vuelos se han realizado con cada uno? Elimina los NA'S' (LEFT_JOIN(), SELECT(), FILTER(), GROUP_BY(), COUNT).

Visualización

9. Se solicita analizar para la aerolínea American Airlines si los vuelos que tienen retraso en la partida también tienen retraso en la hora de llegada. Utilice un diagrama de dispersión para el Retraso en la Partida y Retraso en la Llegada.
10. Visualiza la tendencia de la temperatura durante los primeros 15 días del mes de enero en los vuelos que parten de aeropuerto "Newark, EWR", utiliza una gráfica de línea.
11. Visualiza la temperatura más frecuente en los primeros 15 días del mes de enero, utiliza un histograma.
12. Crear un histograma con Facets para mostrar la temperatura en cada mes.

13. Número de vuelos que salieron de Nueva York en 2013 por aerolínea (mostrar solamente las 10 aerolíneas con más vuelos), utilizando una gráfica de barras.
14. Visualiza el punto anterior en una gráfica de pie.
15. Relaciona el data frame fligths con el data frame airports a través del campo destino ¿cómo lograr esta relación?
16. Crea un nuevo data frame con el punto anterior únicamente con los 5 carriers con más vuelos por destino.
17. Gráficos
    a. Gráfico de barras apiladas o agrupadas para aerolíneas según destino.
    b. Gráficos de barras apilados verticalmente, uno por color.

Situación del problema 2

¿Por qué se retrasan los vuelos?

18. Gráfico - Atraso promedio por origen
    a. Calcular el número total de vuelos por origen
    b. Calcular el número de vuelos retrasados por origen
    c. Unir los datos y calcular el porcentaje de vuelos retrasados
    d. Gráfico de barras para el porcentaje de vuelos retrasados por origen
19. Gráfico - Atraso promedio por Carrier
    a. Grafico de barras del atraso promedio por **Carrier**
    b. Calcular el número total de vuelos por transportista
    c. Calcular el número de vuelos retrasados por transportista
    d. Unir los datos y calcular el porcentaje de vuelos retrasados
    e. Gráfico de barras para el porcentaje de vuelos retrasados por transportista
20. Gráfico - Atraso promedio por mes
    a. Gráfico de línea del atraso promedio por mes
    b. Calcular el número total de vuelos por mes
    c. Calcular el número de vuelos retrasados por mes
    d. Unir los datos y calcular el porcentaje de vuelos retrasados
    e. Gráfico de línea para el porcentaje de vuelos retrasados por mes
21. Retraso por tipo de avión
    a. Retraso por tipo de avión
    b. Gráfico de barras para retraso por tipo de avión
    c. Calcular el número total de vuelos por tipo de avión
    d. Calcular el número de vuelos retrasados por tipo de avión
    e. Unir los datos y calcular el porcentaje de vuelos retrasados
    f. Gráfico de barras para porcentaje de Vuelos Retrasados por Tipo de Avión

22. Retraso por día de la semana
    a. Retraso por día de la semana
    b. Ordenar días de la semana
    c. Gráfico de barras para retraso promedio por día de la semana
23. Gráfico de barras para el retraso por hora del día
24. Conclusiones del retraso de los vuelos
25. Realice un análisis del clima para determinar si es un factor que influye en los retrasos.
    a. Gráfico para el retraso promedio por humedad relativa
    b. Gráfico de dispersión con suavizado para el retraso promedio por temperatura
    c. Gráfico de suavizado para el retraso promedio por velocidad del viento
    d. Gráfico de barras para retraso promedio por dirección del viento
    e. Gráfico de líneas para retraso promedio por visibilidad en el aire
26. Conclusiones de análisis del clima